Data Science in Education Using R

Emily A. Bovee, Ryan A. Estrellado, Jesse Mostipak, Joshua M. Rosenberg, and Isabella C. Velásquez

Table of Contents

# Welcome

Welcome to Data Science in Education Using R! Inspired by {bookdown}, this book is open source. Its contents are reproducible and publicly accessible for people worldwide. The online version of the book is hosted at [datascienceineducation.com](https://datascienceineducation.com/).

## Prologue

There’s this story going around the internet about an eagle egg that hatches in a chicken farm. The eagle egg hatches near the chicken eggs. The local hens are so busy doing their thing that they don’t notice the baby eagle egg is not their own. The eagle chick is born into the world and, having no knowledge of its own eagleness, joins its new family on a nervous and exciting first day of life. Over the next few years the baby eagle lives as chickens live. It eats chicken feed, learns to fly in short choppy hops a few feet at a time, and masters the rapid head jabs of the chicken strut.

One day, while strutting around the chicken farm, the young eagle sees something soaring through the sky. The flying creature has long wings, which it stretches wide before tucking them back in and angling itself downward for a dive towards the earth. The sight of this other-worldly bird stirs something in the young eagle.

Over the next few weeks the eagle finds it can’t shake the vision of the soaring eagle from its mind. It tests the conversational waters during feeding time. It wonders out loud, “What if we tried to fly more than two feet off the ground?” The other chickens stare back. The young eagle, uncertain if these stares are ambivalence or the default chicken eye position, begins to ponder the only way forward. It must learn to fly high while living with the chicken family he loves.

This is a book about learning to program in R while working in education. It’s for folks who feel at home in the education community but are looking out into the world and wondering how to use data better. It’s about being a great educator and wondering if it’s too late to learn to code. It’s about being an educator who’s learning to code and wondering if there are others you can learn with.

We were on Twitter a lot in November of 2017. We talked about things like debugging R code, interpreting model coefficients, and working on spreadsheets with three header rows. We kept coming back to these topics over and over again. It was like having an obscure hobby with online friends because it’s hard to find local knitters who only knit Friends characters, or vinyl collectors who only collect Swedish disco albums. When you work as a data science consultant in education or as an educator learning data science, it’s hard to find that professional community that just gets you. Going to education conferences is great, but the eyes glaze over when you start talking about regression models. The data science conferences are super, but the group at the cocktail table gets smaller when you vent about the state of aggregate test score data.

We started talking about data science in education online because we wanted to be around folks who do data science in education. We wrote this book for you, so you can learn data science with datasets you can find in education work. We don’t claim to be experts at education or data science, but we’re pretty good at talking about what it’s like to do both in a time where doing both is just starting to take off.

So give your chicken family a big hug, open up your laptop, and let’s start learning together. Turns out, there are a lot more hatchlings wanting to be eagles and chickens at the same time.



Figure 1: The Tweet That Started It All

## Acknowledgements

This work was supported by many individuals from the [DataEdu Slack channel](https://dataedu.slack.com/) (<https://dataedu.slack.com/>). Thank you to everyone who contributed code, suggested changes, asked questions, filed issues, and even designed a logo for us: Abi Aryan, Jason Becker, William Bork, Erin Grand, Ludmila Janda, Jake Kaupp, Nathan Kenner, David Ranzolin, Kris Stevens, Gustavo Velásquez, and Bret Staudt Willet.

Thank you to the data scientists in education that took time share their stories with us: Isabella Fante, LaCole Foots, Tobie Irvine, Arpi Karapetyan, John LaPlante, and Andrew Morozov.

## Citation

If you would like to cite this book, please use the below:

Bovee, E. A., Estrellado, R. A., Motsipak, J., Rosenberg, J. M., & Velásquez, I. C. (under contract). Data science in education using R. London, England: Routledge. Nb. All authors contributed equally

# Introduction: Data Science in Education - You’re Invited to the Party!

Dear Data Scientists, Educators, and Data Scientists who are Educators:

This book is a warm welcome and an invitation. If you’re a data scientist in education or an educator in data science, your role isn’t exactly straightforward. This book is our contribution to a growing movement to merge the paths of data analysis and education. We wrote this book to make your first step on that path a little clearer and a little less scary.

Whether you’re a data scientist using your skills in an education job or an educator who wants to learn data science skills, we invite you to read this book and put these techniques to work in the real world. We think that your work in the eduation community will help decide how education and data science come together going forward.

## Learning Data Science in Education

Over the coming chapters we’ll be learning together about what data science in education can look like. But to understand why we were compelled to write about the topic, we need to talk about why data science in education is not such a straightforward thing.

Learning data science in education is challenging because there isn’t a universal vision for that role yet. Data science in education isn’t straightforward because the role itself is not straightforward. If education were a building, it would be multi-storied with many rooms. There are privately and publicly funded schools. There are more than eighteen possible grade levels. Students can learn alone or with others in a classroom.

This imaginary building we call education also has rooms most residents never see: rooms where business and finance staff plan the most efficient use of limited funds. The transportation department plans bus routes across vast spaces. University administrators search for the best way to measure career readiness. Education consultants study how students perform on course work and even how they feel about class materials.

There are a lot of ways one *could* do data science in education, but building consensus on ways one *should* do data science in education is just getting started. The data science in education community is still working out how it all fits together. And for someone just getting started, it can all seem very overwhelming.

Even if we did have perfect clarity on the topic, there’s still the issue of helping education systems learn to leverage these new analytic tools. In many education settings, school administrators and their staff may have never had someone around who understands education, knows how to code, and uses statistical techniques (Conway [2010](#ref-conway2010data)) all at once.

## Making the Path a Little Clearer

As data science in education grows, the way we talk about and conceptualize it (Rosenberg et al. [2020](#ref-rosenberg2020mdsc)) also needs to grow. We begin this book by offering a primer for data science in education, including a discussion of unique challenges and foundational skills in the programming language R.

Next, you’ll take what you’ve learned and apply it in our data analysis in education walkthroughs. The walkthroughs in this book are our contribution towards a more example-driven approach to learning. They’re meant to make the ambiguous path of learning data science in education a little clearer by way of recognizable and actionable demonstrations. These examples fall into three different education data themes, with walkthroughs for each theme:

Student perceptions of learning

* [Walkthrough 1: The Education Dataset Science Pipeline](#c07)
* [Walkthrough 5: Text Analysis With Social Media Data](#c16)
* [Walkthrough 7: The Role (and Usefulness) of Multi-Level Models](#c13)

Analyze student performance data

* [Walkthrough 2: Approaching Gradebook Data From a Data Science Perspective](#c08)
* [Walkthrough 8: Predicting Students’ Final Grades Using Machine Learning Methods](#c14)

Get value from publicly available data

* [Walkthrough 3: Introduction to Aggregate Data](#c09)
* [Walkthrough 4: Longitudinal Analysis With Federal Students With Disabilities Data](#c10)

We’ll end the book by discussing how to bring data science skills into your education job.

We hope after reading this book you’ll feel like you’re not alone in learning to do data science in education. We hope your experience with this book is the right balance of challenging and fun. Finally, we hope you’ll take what you learned and share it with others who are looking to start this journey.

# How to Use this Book

We’ve heard it from fellow data scientists and experienced it ourselves - learning a programming language is hard. Like learning a foreign language, it is not just about mastering vocabulary. It’s also about learning the language’s norms, its underlying structure, and the metaphors that hold the whole thing together.

The beginning of the learning journey is particularly challenging because it feels slow. If you have experience as an educator or consultant, you already have efficient solutions you use in your day-to-day work. Introducing code to your workflow slows you down at first because you won’t be as fast as you are with your favorite spreadsheet software. However, you’re probably reading this book because you realize that learning how to analyze data using R is like investing in your own personal infrastructure–it takes time while you’re building the initial skills, but the investment pays off when you start solving complex problems faster and at scale. One person we spoke with shared this story about their learning journey:

“The first six months were hard. I knew how quickly I could do a pivot table in Excel. It took longer in R because I had to go through the syntax and take the book out. I forced myself to do it, though. In the long-term, I’d be a better data scientist. I’m so glad I thought that way, but it was hard the first few months.”

Our message is this: learning R for your education job is doable, challenging, and rewarding all at once. We wrote this book for you because we do this work every day. We’re not writing as education data science masters. We’re writing as people who learned R and data science *after* we chose education. And like you, improving the lives of students is our daily practice. Learning to use R and data science helped us do that. Join us in enjoying all that comes with R and data science – both the challenge of learning and the joy of solving problems in creative and efficient ways.

## Different Strokes for Different Data Scientists in Education

As we learned in the introduction, it’s tough to define data science in education because people are educated in all kinds of settings and in all kinds of age groups. Education organizations require different roles to make it work, which creates different kinds of data science uses. A teacher’s approach to data analysis is different from an administrator’s or an operations manager’s perspective.

We also know that learning data science and R is not in the typical job description. Most readers of this book are educators working with data and looking to expand their tools. You might even be an educator who **doesn’t** work with data, but you’ve discovered a love for learning about the lives of students through data. Either way, learning data science and R is probably not in your job description.

Like most professionals in education, you’ve got a full work schedule and challenging demands in the name of improving the student experience. Your busy workday doesn’t include regular professional development time or self-driven learning. You also have a life outside of work, including family, hobbies, and relaxation. We struggle with this ourselves, so we’ve designed this book to be used in lots of different ways. The important part in learning this material is to establish a routine that allows you to engage and practice the content every day, even if for just a few minutes at a time. That will make the content ever-present in your mind and will help you shift your mindset so you start seeing even more opportunities for practice.

We want all readers to have a rewarding experience, and so we believe there should be different ways to use this book. Here are some of those ways:

### Read the Book Cover to Cover (and How to Keep Going)

We wrote this book assuming you’re at the start of your journey learning R and using data science in your education job. The book takes you from installing R to practicing more advanced data science skills like text analysis.

If you’ve never written a line of R code, we welcome you to the community! We wrote this book for you. Consider reading the book cover to cover and doing all the analysis walkthroughs. Remember that you’ll get more from a few minutes of practice every day than you will from long hours of practice every once in awhile. Typing code every day, even if it doesn’t always run, is a daily practice that invites learning and a-ha moments. We know how easy it is to avoid coding when it doesn’t feel successful (we’ve been there), so we’ve designed the book to deliver frequent small wins to keep the momentum going. But even then, we all eventually hit a wall in our learning. When that happens, take a break and then come back and keep coding. When daily coding becomes a habit, so does the learning.

If you get stuck in an advanced chapter and you need a break, try reviewing an earlier chapter. You’ll be surprised at how much you learn from reviewing old material with the benefit of new experience. Sometimes that kind of back-to-basics attitude is what we need to get a fresh perspective on new challenges.

### Pick a Chapter of Interest and Start There

We interviewed R users in education as research for this book. We chose people with different levels of experience in R, in the education field, and in statistics. We asked each interviewee to rate their level of experience on a scale from 1 to 5, with 1 being “no experience” and 5 being “very experienced”. You can try this now–take a moment to rate your level of experience in:

* Using R
* Education as a field
* Statistics

If you rated yourself as a 1 in Using R, we recommend reading the book from beginning to end as part of a daily practice. If you rated yourself higher than a 1, consider reviewing the table of contents and skimming all the chapters first. If a particular chapter calls to you, feel free to start your daily practice there. Eventually, we do hope you choose to experience the whole book, even if you start somewhere in the middle.

For example, you might be working through a specific use case in your education job–perhaps you are analyzing student quiz scores, evaluating a school program, introducing a data science technique to your teammates, or designing data dashboards. If this describes your situation, feel free to find a section in the book that inspires you or shows you techniques that apply to your project.

This book is primarily about learning to use R as a tool for data science in education. Your experience level with R should be the main factor when you decide how to enjoy the book. But do consider how you rated your level of experience with education and statistics. If these are areas you want to focus on, take your time understanding the education scenarios and statistics techniques we describe. All three disciplines are important parts of being a data scientist in education.

### Read Through the Walkthroughs and Run the Code

If you’re experienced in data analysis using R, you may be interested in starting with the walkthroughs. Each walkthrough is designed to demonstrate basic analytic routines using datasets that look familiar to people working in the education field.

In this approach, we suggest readers be intentional about what they want to learn from the walkthroughs. For example, readers may seek out examples of aggregated datasets, exploratory data analysis, the {ggplot2} package, or the pivot\_longer() function. Read the walkthrough and run the code in your R console as you go. After you successfully run the code, experiment with the functions and techniques you learned by changing the code and seeing new results (or new error messages!). After running the code in the walkthroughs, reflect on how what you learned can be applied to the datasets, problems, and analytic routines in your education work.

One last note on this approach to the book: we believe that doing data science in education using R is, at its heart, an endeavor aimed at improving the student experience. The skills taught in the walkthroughs are only one part of doing data science in education using R. As an experienced R user, you know that this endeavor involves complex problems and collaboration. Since part of your task may be to convince others around you of the merits of your analytic tools and approaches, we’ve written this book with that context in mind. [Chapter 15](#c15) in particular explores ways to introduce these skills to your education job and invite others into analytic activities. We believe you’ll glean useful perspectives from chapters on concepts you’re already familiar with, too.

## A Note on Statistics

Data science is the intersection between content expertise, programming, and statistics. You’ll want to grow all three of these as you learn more about using data science in your education job. Your education knowledge will lead you to the right problems, your statistics skills will bring rigor to your analysis, and your programming skills will scale your analysis to reach more people.

What happens when we remove one of these pieces? Consider a data scientist working in education who is an expert programmer and statistician but has not learned about the real life conditions that generate education data. She might make analysis decisions that overlook the nuances in the data. As another example, consider a data scientist who is an expert statistician and an education veteran, but who has not learned to code. He will find it difficult to scale his analysis up, thereby foregoing the chance to make the largest possible improvement to the student experience. Finally, consider a data scientist who is an expert programmer and an education veteran. She can only scale surface level analysis and might miss chances to understand causal relationships or predict student outcomes.

In this book we will spend a lot of time learning R by way of recognizable education data examples. But doing a deep dive into statistics and how to use statistical techniques responsibly is better covered by books dedicated solely to the topic. It’s hard to understate how important this part of the learning is on the lives of students and educators. One education data scientist we spoke to said this about the difference between building a model for an online retailer and building a model in education:

“It’s not a big deal if an online shopper gets mistakenly shown 1000 brooms but if I got my model wrong and we close a school, that will change a child’s entire life.”

We want this book to be your go-to R reference as you start integrating data science tools into your education job. Our aim is to help you learn R by teaching data science techniques using education datasets. We’ll demonstrate statistics techniques like hypothesis testing and model building and how to run these operations in R. However, the explanations in our chapters will not provide a complete background about the statistical techniques.

We wrote within these boundaries because we believe that the technical and ethical use of statistics techniques deserves its own space. If you already have a foundation in statistics, you will learn how to implement some familiar processes in R. If you have no foundation in statistics, you will be able to take a satisfying leap forward in your learning by successfully using R to run the models and experiencing the model interpretations in our walkthroughs. We provide enough background for you to understand the purpose of the analysis and its results. We encourage you to explore other excellent books like [Learning Statistics With R](https://learningstatisticswithr.com/) (<https://learningstatisticswithr.com/>) (**???**), as you learn the required nuances of applying statistical techniques to scenarios outside our walkthroughs.

## What This Book Is Not About

While we wrote *Data Science in Education Using R* to be a wide-ranging introduction to the topic, there is a great deal that this book is not about. Some of these topics are those that we would have liked to have been able to include, but we did not because they did not fit our intention of providing a solid foundation in doing data science in education. We chose to not include other topics because, frankly, excellent resources already exist. We detail some of what we had to not include in the book here.

* git/GitHub: Git and GitHub are version control software programs, which means that they help keep track of different versions of coding files and which specific changes were made for each version. Git and GitHub are parts of many data scientists’ workflows for solo or collaborative work. However, there is a steep learning curve and these tools are not necessary to get started with coding in R. An outstanding introduction to git and Github is Bryan ([2020](#ref-bryan2020)) ’s freely-available book [Happy git with R](https://happygitwithr.com/) (<https://happygitwithr.com/>).
* Building R packages: If you are carrying out the same analyses many times, it may be helpful to create your own package. Packages are collections of code and sometimes data, such as the {roomba} (for tidying complex, nested lists) and {tidyLPA} (for carrying out Latent Profile Analysis) packages that authors of this book created. However, building an R package is not the focus of this book. Hadley Wickham wrote a very helpful - and freely-available - book on the topic called [R Packages](http://r-pkgs.had.co.nz/) (<http://r-pkgs.had.co.nz/>) (Wickham [2015](#ref-rpackages)).
* Advanced statistical methodologies: As noted above, there are other excellent books for learning statistics. While we do discuss basic and advanced statistical methods, this is not a statistical methods book. One advanced statistical book that we think is excellent from a machine learning perspective is James et al. ([2013](#ref-james2013)) *An Introduction to Statistical Learning with Applications in R*.
* Creating a website (or book): As you might already suspect, R is versatile and can be used for more than just performing data analyses. In fact, R can be used to write books (like this one, which we wrote using the {bookdown} package) and create websites (which some of the authors have done using the {blogdown} package). This book does not describe how to create books or websites; there are excellent, freely available books on these topics as well (see Xie, Thomas, and Hill ([2019](#ref-xie2019blogdown))’s *blogdown: Creating Websites with R Markdown* (<https://bookdown.org/yihui/blogdown/>) and Xie ([2019](#ref-xie2019bookdown))’s *bookdown: Authoring Books and Technical Documents with R Markdown* (<https://bookdown.org/yihui/bookdown/>).

## Supporting the Book

If you find this book useful, please support it by:

* Letting people about it!
* Communicating about the book on social media e.g. via the #dataedu hashtag on Twitter
* Citing or linking to it
* Starring the GitHub repository for either the book (<https://github.com/data-edu/data-science-in-education>) or the package (<https://github.com/data-edu/dataedu>)
* Reviewing it, e.g. on Amazon or Goodreads
* Buying a copy

## Contributing to the Book

We designed this book to be useful and practical for our readers in education. We wrote it as a guide to getting up and running in R, but we know this book does not comprehensively cover every topic related to R. We did this to create a reference that is not intimidating to new users and that creates frequent, small wins while learning to use R.

One question we asked ourselves was: How do we expand this work as data science in education expands as a field? We want readers of this book to be equipped with an agile skillset, and we want this book to continue to provide that even as new R packages are developed and new methods arise. We wrote this book in the open on GitHub so that community members can help us evolve the work, even after it is formally published.

We want this to be the book new data scientists in education have with them as they grow their craft. To achieve that goal, it’s important to us that the stories and examples in the book are based on **your** stories and examples. Therefore, we’ve built ways for you to share with us.

If you have some experience with git and want to contribute that way, here’s how you can can contribute:

* Submit an “issue” to our [GitHub repository](https://github.com/data-edu/data-science-in-education) (<https://github.com/data-edu/data-science-in-education/issues>) that describes a data science problem that is unique to the education setting
* Submit a pull request to share a solution for the problems discussed in the book to the education setting
* Share an anonymized dataset

If you are new to data science in education, welcome!! We would love to have your feedback by [email](mailto:authors@datascienceineducation.com) ([authors@datascienceineducation.com](mailto:authors@datascienceineducation.com)).

We hope that as the book evolves, it grows to reflect the changing needs of data scientists in education.

# What Does Data Science in Education Look Like?

You can think of a data scientist as someone who combines three skills to do data analysis: programming, statistics, and content knowledge (Conway [2010](#ref-conway2010data)). However, if you Google “what is a data scientist,” you won’t find a simple answer. In reality, “data scientist” is not a clear description of a single job functionality: it’s sort of like saying you are a “business person.” Data science as a field describes a wide array of job functionalities: some data scientists work on database architecture, while others focus on data analysis and interpretation. Despite this heterogeneity, in this chapter we’ll provide a working definition of data science in education by sharing some of the roles that professionals occupy in this line of work. We’ll also share some common day-to-day tasks for a data scientist in education.

## Data Roles in Education

We learned from talking with data scientists in the education field that their roles and specializations can be very different from each other. People working in education have specialized skills and passions that allow them to add value to their organizations’ data culture. Here are some of the roles and specializations data scientists in education might take on.

### Building Systems That Get Data to the Right People

School staff and leadership can’t make data-informed decisions unless they have good data. Data scientists in education who specialize in data engineering and data warehousing build systems that organize data in one place. They also keep the data secure to protect the information of students and staff, and they distribute datasets to the people who need it. In this area of data science, you might also find people who specialize in data governance: the creation and maintenance of policies used to keep data collection, documentation, security, and communication to a high standard.

### Measuring the Impact of Our Work on the Student Experience

When it is a goal to continuously improve the systems that serve students, scientific evaluation can help explore the impact of student-facing policies. Measuring the impact of instructional interventions is important because it informs the allocation of time, money, and attention to future improvements to education systems. Data scientists who specialize in measuring impact know how to use statistical techniques to isolate the effect of an intervention and estimate its value. For example, an education system may choose to work with their data analysts to quantify gains in student attendance that result from a new intervention aimed at chronic absenteeism.

### Looking for Patterns in Student Data

Now more than ever, students and school staff are generating data as they go about their day learning and teaching. Online quizzes generate quiz data. Student systems collect data about attendance, discipline, behavior, and native language. Online individualized education program (IEP) systems house information about students with disabilities. Statewide testing assessments are scored, stored in a database, and reported back to families. Much of this data gets reported to the state education agency (SEA) for processing and publishing online as a part of an accountability system.

School systems that learn to explore this data as it is generated get a lot of value from it. Data analysts are experts at systematically exploring these data and finding useful ways to compare data across different categories. This technique, called exploratory data analysis, is used to generate plausible hypotheses about relationships between variables in the data. These hypotheses can be tested and then can become the material educational organizations use to learn about what’s happening with their students. For example, one way for school systems to strive for equity in student outcomes is to frequently examine any differences in outcomes among student subgroups.

### Improving How We Use Statistical Models in Education

When we spoke with data scientists in education during the research phase of this book, we learned about the ways they use tried and true methods for analysis in schools. Even so, we also learned that they are innovators. They take techniques that are commonly found in other industries, like business, and explore how these techniques can improve the state of analytics in education. In particular, the data scientists that we spoke with talked about going beyond exploratory data analysis and introducing more advanced techniques like inferential statistics and predictive modeling to the data culture of the schools where they work. This work is not only about improving how well schools implement their current practices, but it is also about exploring our curiosities about how we might apply new techniques to the task of improving the learning experience of our students.

## Common Activities of Data Scientists

Now let’s explore what a data scientist in education can expect to do on a typical day at the office. We’ll preview some common activities data scientists do daily. Later in the book, we’ll learn and practice these techniques.

### Processing Data

Processing data, or cleaning data, is the act of taking data in its raw form and preparing it for analysis. When you start a data analysis, the data you have is in the same state it was in when it was generated and stored. It very often isn’t designed to support the specific analysis that that you’re tasked with doing.

Here are some examples of common things you’ll need to do to prepare your data:

It’s common for raw datasets to have generic variable names that don’t describe the values in that dataset’s column. The variable names have to be reworked so they’re convenient to reference in your code. For example, a dataset indicating students’ grades at various points in the semester might have variable names that are just the date of the measurement. In this case, the variable name doesn’t fully describe the data captured in the column: it just captures the date of the measurement of that data. These variable names should be changed into something that intuitively represents the values in that column. There are also format-related problems with variables. Things like spaces between words, lengthy variable names, or symbols in the variables names can cause inconvenience in your code or make it hard to keep track of the steps in a complicated analysis.

Datasets also have to be filtered to the subset that you’re interested in analyzing. It’s possible that the dataset you’re given contains a larger group of students than you need for your project. For example, a principal at a school site may give you a dataset of every student and the number of days they’ve missed this school year. Now imagine she asks you to do an analysis of attendance patterns in first, second, and third graders. One step you’ll need to take before you start the analysis is filtering the dataset so that it only contains first, second, and third graders.

Sometimes you’ll need to do computations to get the summary figures your stakeholders are asking you to generate. Imagine that the director of curriculum and instruction asks you to report the percentage of students that have scored in the “proficient” range on a statewide assessment. Now imagine that the datasets you’re given are (1) a list of students, (2) a list of the schools they attend, and (3) a list of their test scores. To produce the requested report, you’ll need to merge these lists together so that all the data for each student is in one place: (student, school, and test score). Next, you’ll need to identify the number of students who scored above the “proficient” threshold on the test at each school. Finally, you’ll be able to calculate the percentage of students who met that threshold at each school.

### Doing Analysis

This is the part of our workflow that most people associate with data science. Analysis is the application of statistics techniques to identify the underlying structure of the dataset. This means that you are making educated guesses about the real life conditions that generated the dataset.

We realize this may be the first time you’ve heard data analysis described this way. We choose to describe it this way because in the end, data analysis in education is about understanding what the data tells us about the student experience. If we can understand the underlying structure of a dataset, we can improve our understanding of the students whose academic behaviors generated the numbers.

Let’s look at a concrete examples of this. Imagine that you are an education consultant and your client is a school district superintendent. The superintendent has asked you to evaluate the impact of a new teacher coaching initiative the school district has been using for a year. After processing a dataset that contains teachers, the number of hours they’ve spent in coaching sessions, and the change in quiz scores, you set out to explore the data and fit a statistical model. Your initial visualization of the dataset - a line graph of the relationship between hours the teachers spent in coaching and the test scores of those teachers’ students - suggests there might be a linear relationship: the more hours a teacher has spent in coaching, the higher that teacher’s students score on quizzes. While it might seem intuitive to say that teachers spending more time in coaching leads to an increase in students’ test scores, you can’t definitively draw that conclusion just from the visualization. One thing that the visualization *doesn’t* tell you is whether the relationship between those two variables is meaningful.

Using a statistical model to analyze this dataset is one way of understanding the underlying structure of the dataset. Specifically, a statistical model can help estimate how much of the change in test scores can be explained by the hours a teacher spent in coaching sessions, and how much can be explained by some other factor (even random chance!). In this example, one possibility is that more conscientious and passionate teachers seek out additional hours of coaching. The correlation that was suggested by the data visualization between coaching and quiz scores might actually be reflecting a relationship between effective teaching style and quiz scores: it’s just that more effective teachers are more likely to attend more hours of coaching.

As you can see, when we try to describe human behavior, things tend to get complicated quickly. Data scientists in education are fundamentally interested in the people who generated the numbers. In this case, that’s the students who took the quizzes and the teachers who participated in the coaching sessions. If you can analyze the dataset and understand its underlying structure, you might learn more about what’s happening with the student and teacher experience.

### Sharing Results

So far, we’ve discussed processing data and analyzing data. At these stages, the audiences for your output are usually you, other data scientists, or stakeholders who are in a position to give feedback about the process so far. But when you’ve sorted through your findings and have selected conclusions you want to share, the audience shifts to a wider group of people. Now you’re tasked with communicating with leadership, staff, parents, the community, or some combination of those audiences.

Your thought process and techniques for sharing with a wider audience are different from the ones you use when processing and analyzing data. The sharing of your results includes developing visualizations that clearly communicate a finding, writing narratives that give context and story to your analysis, and developing presentations that spark conversations about the student experience.

## Who We Are and What We Do

We’ve talked about the roles of data scientists in education and the daily tasks they can be expected to do. Now, let’s talk about how who data scientists in education are.

In some fields the relationship between what you do today and how you got there is prescribed. If you want to help people by performing cardiac surgery, you have to go to medical school first. If you want to hear trials in court, you have to go to law school first. To prepare for this book, we talked to lots of folks who do data analysis in the education field. We found that there’s quite a bit of variety in how people work with data in education. There’s also variety in the journeys that people took to arrive at their education data science roles.

We see this as good news for people who want to start working with data in education in a more formalized way. You don’t need a Ph.D. to do this kind of work, though some people we talked to had pursued graduate education. Some people we interviewed had a statistics background, but you don’t need to be an expert in statistical modeling. We talked to people who are consultants that came to the education field. We also talked to teachers and administrators who became consultants. We talked to people who are the lone data scientist in their education organizations and we talked to people who are part of an analytics team.

You might not think of yourself as a data scientist because your job title does not have those words in it. However, we believe data science is more about the things that you do than the title on your business card. Our own paths toward doing data science in education are very different. Here’s a little about us and how we practice data science:

*Leading Office Culture Toward a Data-Driven Approach*

Jesse, a director at an education non-profit in Texas, is setting up a database to house student achievement data. This project requires a number of data science skills we’ll discuss in chapter five, including processing data into a consistent format. Once the data is prepared, Jesse builds dashboards to help her teammates explore the data.

However, not all of Jesse’s work can be found in a how-to manual for data scientists. She manages a team and serves as the de facto project manager for IT initiatives. Given her expertise and experience in data science, she’s also leading the charge towards a more data-driven approach within the organization.

*Helping School Districts Plan to Meet Their Goals*

Ryan, a special education administrator in California, uses data science to reproduce the state department of education’s special education compliance metrics. Then, he uses the results to build an early warning system for compliance based on local datasets. In this case, Ryan uses foundational data science skills like data processing, visualization, and modeling to help school districts monitor and meet their compliance requirements.

*Doing and Empowering Research On Data Scientists in Education*

Joshua, Assistant Professor of STEM Education at University of Tennessee in Knoxville, researches how students do data science and helps teachers teach the next generation of data-informed citizens. He makes this work possible by building R packages - self-contained groups of data tools - that he and other researchers use to analyze datasets efficiently.

*Supporting Student Success with Data*

Emily, a dental education administrator in Wisconsin, guides faculty members on best practices in assessing student learning. Like Jesse, Emily works on merging multiple data sources together to get a better understanding of the educational experience. For example, she merges practice national board exam scores with actual national board performance data. Later, Emily conducts statistical analyses to help identify a practice national board score threshold at which students are ready to sign up for the real national board exam. All this is possible because of R!

*Placing Schools and Districts in Context*

Isabella, a data analyst at a large philanthropy, uses publicly available aggregated data to understand what schools and districts look like, how they’ve changed over time, and other contextual information needed to better understand the field of education. These datasets are often in messy formats (or even PDFs!), and sometimes data from the same agency are organized in a slightly different way every year. Using R allows the downloading and cleaning process to be reproducible when new data comes in. The code clearly shows the decision rules made to make the aggregated data useful in models or visualizations. Packages and projects allow the entire process to be shared and reused across the analytics team.

## Next Steps for Data Science in Education

As you saw above, there are a wide variety of ways to apply statistics and programming techniques to support educators and students and to create new knowledge in the education field. We hope this book is part of a movement to develop common norms and expectations for the field as the relationship between data science and education grows.

Because the relationship between data science and education is still young, it is important that the people growing the field of data science in education understand the culture and unique challenges in their education job. After all, the defining feature that will differentiate data science in education from data science in general will be the ability to meet the unique needs of students, staff, and administration in education.

As you progress through this book, we hope you begin to understand where your particular data interests and passions are. There is more variety in educational backgrounds and in the daily work of education data analysis than one might think. You can bring your own unique background and interests to this field. We hope this book will help you combine your unique experiences with some new learning so you can create a practice that improves the education experience of students while using the talents you already have.

# Special Considerations

Data science in education is a new domain. It presents opportunities, like those discussed in the [previous chapter](#c03), but also some challenges. These challenges vary a lot: we consider doing data science in education to include not only accessing, processing, and modeling data, but also social and cultural factors, like the training and support that these data scientists have available to them. Some of these challenges are common to *all* domains in which data science is carried out, while others are very particular to the field of education.

## Things to Consider when Doing Data Science in any Domain

### Learning to Code

Data scientists everywhere are combining content knowledge, programming, and statistics to solve problems. Many people are not experts in all three areas when they begin their data science journeys. In education, if the piece of the data science trifecta that you are missing is programming, you are not alone. Learning to code can seem like a daunting task, but we don’t want you to feel paralyzed. Data scientists everywhere are combining content knowledge, programming, and statistics to solve problems. Many people are not experts in all three areas when they begin their data science journeys. If you work in education and the piece of the data science trifecta that you are missing is the programming, you are not alone. Programming is powerful but challenging, and many of us in education do not have prior experience with it. To help those who are writing their first few lines of code, we wrote this book for learners without a computer science background or even any informal training related to coding. The great thing about entering a field as flexible as data science is that you are joining a vast crowd of individuals who are self-taught. You will find that there is a very supportive online community to help as you learn through this book.

### Addressing Ambiguity: A Reproducible Approach

One way to address concerns of of those feeling wary of data collection and reporting is to build analytic processes that are transparent. Specifically, it is helpful if the data scientist is open about what data is collected, how it is sourced, how it is analyzed, and how it is considered alongside other data when used in decision-making conversations. This transparency can be achieved through a number of activities, including having regular conversations about analytic methods, providing written reports describing data collection, and receiving input about analytic goals from staff members.

One such process for achieving openness in data collection and analysis is called reproducible research. The concept of reproducible work (Wikipedia [2020](#ref-wikipedia2020reproducible)) is the idea that a completed analysis should come with all the necessary materials, including a description of methodology and programming code, needed for someone else to run the analysis and achieve the same results. If school staff are apprehensive about how school data is collected and used, a more transparent method showing how the data was transformed into insights could help put their minds at ease.

A reproducible approach can be especially beneficial in transition periods. If a data-science-in-school advocate leaves their original position, they would leave behind not just descriptions of the analyses that they did, but also the specific files needed to run the analyses again. The new individual who takes their place will be able to seamlessly transition into the new role. If asked to run “the same report I always got from your predecessor,” the new person will understand immediately what files were needed to create that original report and will be able to request all necessary data to generate a new version of the report.

To implement a reproducible approach in your organization, you can start by keeping all files related to each project you do in their own folders. As you create reports, keeping notes in the files will help you to easily generate similar reports in the future. Many data scientists find that even though changing administration might mean changing requests, having careful documentation of past processes allows for more efficiency in the way they use data to answer those requests.

## Things to Consider when Doing Data Science in Education

### Addressing Organizational Resistance: A Self-Driven Analytic Approach

One consideration when adopting data science in education is that often, there is a lack of precedent for applying data science techniques. For example, it is not common for a teacher to be conducting regression analyses on student data. However, it’s not necessary to wait for a district-wide or state-wide initiative to begin to implement the techniques you will learn in this book.

An organization should encourage their staff to do their own data analyses primarily for the purpose of testing their own hypotheses. In a school, for example, a teacher might wonder about student learning in their classroom and might want to utilize data to directly guide decisions about how they deliver instruction. There are at least two benefits to this approach. First, staff begin to realize the value of doing data analysis as an ongoing inquiry into their outcomes, instead of a special event once a year ahead of school board presentations. Second - and more important for the idea of reducing apprehension around data analysis in schools - school staff begin to demystify data analysis as a process. When school staff collect and analyze their own data, they know exactly how it is collected and exactly how it is analyzed. The long-term effect of this self-driven analytic approach might be more openness to analysis, whether it is self-driven or conducted by the school district.

Building and establishing a data governance system that advocates for an open and transparent analytic process is difficult and long-term work, but the likely result will be less apprehension about how data is used and more channels for school staff to participate in the analysis. Here are more practical steps a school district can take towards building a more open approach to analysis:

* Make technical write-ups of data analyses available so interested parties can learn more about how data was collected and analyzed
* Make datasets available to staff within the organization, to the extent that privacy laws and policies allow
* Establish an expectation that analysts present their work in a way that is accessible to many levels of data experience
* Hold regular forums to discuss how the organization collects and uses data

By adopting a self-driven analytic approach, individuals can help their education organization to embrace the potential of utilizing data to anticipate and possibly forestall problems in the future.

### Lack of Processes and Guidelines

One challenge is the ambiguity around the process and practice of doing data science in the education context specifically. While there is a body of past research on *students’* work with data (see (**???**) for a review), there is limited information from case- or design-based research on how others in education –teachers, administrators, and data scientists– use data in their work. In other words, we do not have a good idea of what best practices in our field are. This challenge is reflected in part in the variability in the job titles of those who work with data: some are data analysts, some are research associates, and the list continues. However, as data science in education emerges as a field, some school districts are now hiring for data scientist positions. Even so, there is a lack of an organizing body that brings all these people together. There are a multitude of discipline-specific (e.g., science teaching) or department-specific (e.g., institutional research) conferences, but no overarching norms universal to those who work with data in education.

Education is a field that is rich with data: survey, assessment, written, and policy and evaluation data, and more. Nevertheless, there often is a lack of common consensus on processes and procedures for educators and data scientists to share data and the results of data analysis with each other. Academic and research settings sometimes can lead to silos of information. A group of researchers at one university could do a survey, and another group doing similar work could not hear about the results until the study is published years later. Sometimes, the second group of researchers may never become aware that the survey had even happened. The good news about this is that many education organizations are curious about and passionate about supporting student success. It is likely that even if many separate data collection efforts are being implemented (rather than one unified strategy), you will not be dealing with the problem of “I don’t have enough data to analyze.” As a pioneer for data science in your organization, you can help to clarify these redundant processes and can offer your skills to help make sense of the wealth of information already being gathered.

### Limited Training and Educational Opportunities

Data science in education is new. At the present time, there are limited opportunities for those working in education to build their capabilities in data science (though this is changing to an extent; see Anderson and colleagues’ work to create a data science in education certificate program at the University of Oregon and Baker’s educational data mining Massive Open Online Course offered through Coursera). Many data scientists in education have been trained in fields other than statistics, business analytics, or research. Moreover, the training in terms of particular tools and approaches that data scientists in education utilize are highly varied. However, we believe this diversity of training creates unique opportunities for the development of data science in education as a field. Indeed, educators’ diverse training and backgrounds positions them to tackle challenges in education creatively. One challenge is the ambiguity around the process and practice of doing data science in the educational context specifically. While there is a body of past research on *students’* work with data (see (**???**) for a review), there is limited information from case- or design-based research on how others in education –teachers, administrators, and data scientists– use data in their work. In other words, we do not have a good idea of what best practices in our field are. This challenge is reflected in part in the variability in the job titles of those who work with data: some are data analysts, some are research associates, and the list continues. However, as data science in education emerges as a field, some school districts are now hiring for data scientist positions. Even so, there is a lack of an organizing body that brings all these people together. There are a multitude of discipline-specific (e.g., science teaching) or department-specific (e.g., institutional research) conferences, but no overarching norms universal to those who work with data in education.

Education is a field that is rich with data: survey, assessment, written, and policy and evaluation data, and more. Nevertheless, the lack of a common consensus on processes and procedures for educators and data professionals to share data and the results of data analysis with each other. Academic and research settings sometimes can lead to silos of information. A group of researchers at one university could do a survey, and another group doing similar work could not hear about the results until years later, when the study is published. Sometimes, the second group of researchers may never become aware that the survey had even happened. The good news about this is that many educational organizations are curious about and passionate about supporting student success. It is likely that even if many separate data collection efforts are being implemented (rather than one unified strategy), you will not be dealing with the problem of “I don’t have enough data to analyze.” As a pioneer for data science in your organization, you can help to clarify these redundant processes and can offer your skills to help make sense of the wealth of information already being gathered.

### Limited Training and Educational Opportunities

Data science in education is new. At the present time, there are limited opportunities for those working in education to build their capabilities in data science (though this is changing to an extent; see Anderson and colleagues’ work to create a data science in education certificate program at the University of Oregon and Baker’s data mining in education Massive Open Online Course offered through Coursera). Many data scientists have been trained in fields other than statistics, business analytics, or research. Moreover, the training in terms of particular tools and approaches that data scientists utilize are highly varied. However, we believe this diversity of training creates unique opportunities for the development of data science in education as a field. Indeed, educators’ diverse training and backgrounds positions them to tackle challenges in education creatively.

### Advancing Equity

Data science can be used to inform decisions that reduce inequities in the education system. However, data science can also be a tool that exacerbates the marginalization of students we want to serve. An example is an algorithm that is not transparent and implemented poorly, and prompts people to make decisions that have adverse effects.

For a data scientist in education, it is crucial that before beginning an analysis, we fully understand how we/our organization defines equity. Additionally, we should formulate clear equity goals and consider the ways we will continuously check against our biases. After defining equity and our equity goals, we can then work to ensure that our data science life-cycle reflects what we are trying to learn (and that we incorporate these learnings).

Thoughtful decisions during the project design and data collection, analysis, and the presentation of the results, can increase the data’s ability to move an organization towards its equity goals. For example, if an organization hopes to decrease the opportunity gap between students affected by poverty and students not affected by poverty, then it is important that they (1) define what ‘affected by poverty’ means, (2) identify the type of project design that will help them understand if they are moving towards their goals, and (3) determine whether their data collection allows them to disaggregate these demographics (see [Walkthrough 03](#c09)). We can then make sure the analyses take these disaggregations into account. The final report should be conscientious of any potential blind spots we may have about the results, as all data is biased and can only ever tell a partial story.

Thoughtful and deliberate data science can help us understand what to do so our students reach their highest potential. Data science can make us more efficient in our tasks. It can increase transparency about what we are doing to help our students. It can also help monitor how we are progressing. However, we must continuously inspect our processes and work to make sure we do not do unintentional harm.

R and RStudio, both freely-available and open, also serve to increase equity in data science. As opposed to proprietary tools, they are accessible to anybody with a computer and internet. The code behind the packages is available online, opening up the “black box” of research. If code is submitted alongside analyses and reports, we can see what decisions were taken to produce the analysis and rerun the analysis ourselves. Using R can enable more audiences to learn, understand, and reuse analyses.

### The Complex Nature of Education Data

Education data are complex to collect and to analyze. Education data are often hierarchical, in that data at multiple “levels” are collected. These levels include classrooms, schools, districts, states, and countries - quite the hierarchy! Additionally, an education dataset often require linking with other datasets. For example, when data is collected on students at the school level, it might be important to also know about the training of the teachers in the school. Contextual data about the funding provided by the community in terms of per-pupil spending would be helpful to merge with data about the academic outcomes of students in that school district. The complexity does not end when the data are collected and merged with other relevant information: education data are not simple.

Education data includes characteristics and behaviors of students, teachers, and other individuals. These variables are not always numeric. A categorical variable is a descriptive type of variable with multiple levels for which the levels do not signify quantity but instead signify groups, such as sex or grade level. Education data can also involve open-ended responses that are stored as string variables (a type of variable used to store text), or even recordings that consist of audio and video data. As with the diversity of training for data scientists, the complexity of education data also presents opportunities for educators to creatively approach their tasks. There are specific techniques to efficiently handle each type of data listed above and we will explore some of those techniques in this book.

Often, the variables gathered in education are numeric, but just as often, they are not. Education data involves characteristics of students, teachers, and other individuals that are categorical. A categorical variable is a descriptive type of variable with multiple levels for which the levels do not signify quantity but instead signify groups, such as sex or grade level. It is not quite right to interpret these data as numeric. Additionally, education data can involve open-ended responses that are stored as string variables (a type of variable used to store text), or even recordings that consist of audio and video data. All of these types of data present challenges to the data scientist in education. As with the diversity of training for data scientists, the complexity of data in education also presents opportunities for practitioners to creatively approach their tasks. There are specific techniques to efficiently handle each type of data listed above, and we will explore some of those techniques in this book.

If you are faced with a large and complicated dataset, you might begin by asking yourself what you are curious about and carving out just a couple variables that you can use to answer your question. Your colleague might be interested in an entirely different question, and might consider different variables from the same dataset in their analysis. The complexity of education data need not discourage educators from pursuing their interests.

### Ethical and Legal Concerns

There are a number of ethical and legal concerns in working responsibly with education data. At the K-12 level, most datasets require safeguards because minors are a protected population. There might be physical limitations to the places that a data scientist in education could access confidential data, and there might be limitations about the ways that results of a data analysis can be shared with others within the organization. A closely related issue concerns the aims of education within predetermined constraints. Those working in education often seek to improve it and often work to do so with a scarcity of school and community resources. These ethical, legal, and even values-related concerns may become amplified as the role of data in education increases. They should be carefully considered and emphasized from the outset by those involved in data science in education. If you feel resistance in your organization as you begin to adopt the principles you learn in this book, you might begin by offering to analyze “de-identified” or “anonymous” data. In this way, you could show your administration what is possible and foster additional buy-in further down the road.

### Analytic Considerations

Due to the particular nature of education data, analysis can be difficult, too. The data is often not ready to be used: it may be in a format that is difficult to open without specialized software, or it may need to be “cleaned” before it is usable. In data science, “cleaning” or “processing” data refers to reorganizing or restructuring the dataset to make it easier to analyze. This process would be analogous to the steps you would take if you received an Excel spreadsheet but found that the columns were in an order that didn’t make sense to you and that there were some duplicate columns. The process you’d go through to reorganize the data to make it logical is data cleaning. Because of the different *types* of data, data scientists must often use a variety of analytic approaches, such as multi-level models, models for longitudinal data, or even models and analytic approaches for text data. In later chapters of this book, you will learn more specifics about building models.

## Conclusion

While there are many challenges to working with education data, there are many opportunities as well. Once they unlock the power of data science to reveal insights about their organizational context (their students, their policies, etc.), many practitioners in education will become more interested in gathering more data and continuing on this path. Data science becomes a useful tool to help connect with the purpose of your job. Once you begin to rely on data science, it can be hard to stop! As a data scientist in education, remember that you are more closely acquainted with your context than any outside analyst could ever be. This affords you the unique opportunity to become the data and analysis guru in your area.

In summary, educators that want to evolve their data analysis processes into something practical and meaningful to student progress will need to address some unique challenges in order to help all stakeholders to understand the benefits of the questions being answered with data. That hard work will pay off.

# Getting Started with R and RStudio

## Chapter Overview

This chapter is designed to take you from installing R and RStudio all the way through the very basics of data loading and manipulation using the {tidyverse} (Wickham et al. [2019](#ref-wickham2019)).

We will be covering the following topics in this chapter:

* Installing R and RStudio
* RStudio environment, layout, and customization
* How to run code in R
* Installing the {dataedu} package
* Help documentation
* Steps for working through new and unfamiliar content
* Getting started with {swirl}

## Downloading R and RStudio

First, you will need to download the latest versions of R (R Core Team [2019](#ref-rcoreteam)) and RStudio (RStudio Team [2015](#ref-rstudio)). R is a free environment for statistical computing and graphics using the programming language R. RStudio is a set of integrated tools that allows for a more user-friendly experience for using R.

Although you will likely use RStudio as your main console and editor, *you must first install R*, as RStudio uses R behind-the-scenes.  
Both R and RStudio are freely-available, cross-platform, and open-source.

### To Download R:

* Visit [CRAN](https://cran.r-project.org/) (https://cran.r-project.org/) to download R
* Find your operating system (Mac, Windows, or Linux)
* Download the ‘latest release’ on the page for your operating system and download and install the application

Don’t worry; you will not mess anything up if you download (or even install!) the wrong file. Once you’ve installed R, you can get started.

### To Download RStudio:

* Visit [RStudio’s website](https://www.rstudio.com/products/rstudio/download/) (https://www.rstudio.com/products/rstudio/download/) to download RStudio s
* Under the column called “RStudio Desktop FREE”, click Download
* Find your operating system (Mac, Windows, or Linux)
* Download the ‘latest release’ on the page for your operating system and download and install the application

If you do have issues, consider [the Data Carpentry page](https://datacarpentry.org/R-ecology-lesson/) (https://datacarpentry.org/R-ecology-lesson/) and then reach out for help. Another excellent place to get help is the [RStudio Community foums](https://community.rstudio.com/) (https://community.rstudio.com/).

## RStudio Layout and Customization: Getting to Know R through RStudio

Now that we’ve installed both R and RStudio, we will be accessing R *through* RStudio. One of the most reliable ways to tell if you’re opening R or RStudio is to look at the icons:

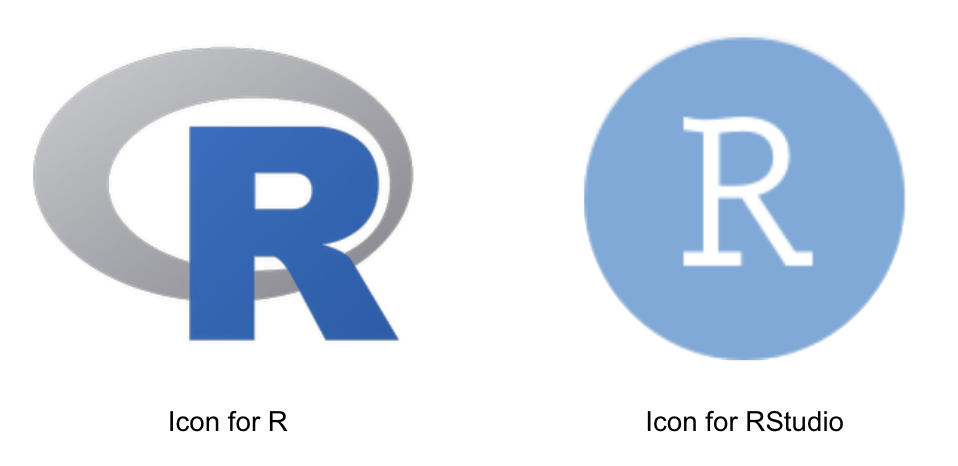


Figure 2: Icons

Whenever we want to work with R, we’ll open RStudio. RStudio interfaces directly with R, and is an **I**ntegrated **D**evelopment **E**nvironment (IDE). This means that RStudio comes with built-in features that make using R a little easier. If you’d like more information on the difference between R and RStudio, we recommend the **Getting Started** section of the [Modern Dive](https://moderndive.com/1-getting-started.html#) (https://moderndive.com/1-getting-started.html#) Ismay and Kim ([2019](#ref-statisticalinf)) textbook.

You do not *have* to use RStudio to access R, and many people don’t!

Other IDEs that work with R include:

* [Jupyter notebook](https://jupyter.org/) (https://jupyter.org/)
* [VisualStudio](https://visualstudio.microsoft.com/services/visual-studio-online/) (https://visualstudio.microsoft.com/services/visual-studio-online/)
* [VIM](https://github.com/jalvesaq/Nvim-R) (https://github.com/jalvesaq/Nvim-R)
* [IntelliJ IDEA](https://plugins.jetbrains.com/plugin/6632-r-language-for-intellij) (https://plugins.jetbrains.com/plugin/6632-r-language-for-intellij)
* [EMACS Speaks Statistics (ESS)](https://ess.r-project.org/) (https://ess.r-project.org/)

This is a non-exhaustive list, and most of these options require a good deal of familiarity with a given IDE. However we bring up alternative IDEs – particularly ESS – because RStudio, as of this writing, is not fully accessible for learners who utilize screen readers. We have chosen to use RStudio in this text in order to standardize the experience, but encourage you to choose the IDE that best suits your needs!

### RStudio layout

When we open RStudio for the first time, we’re should see something similar to this:

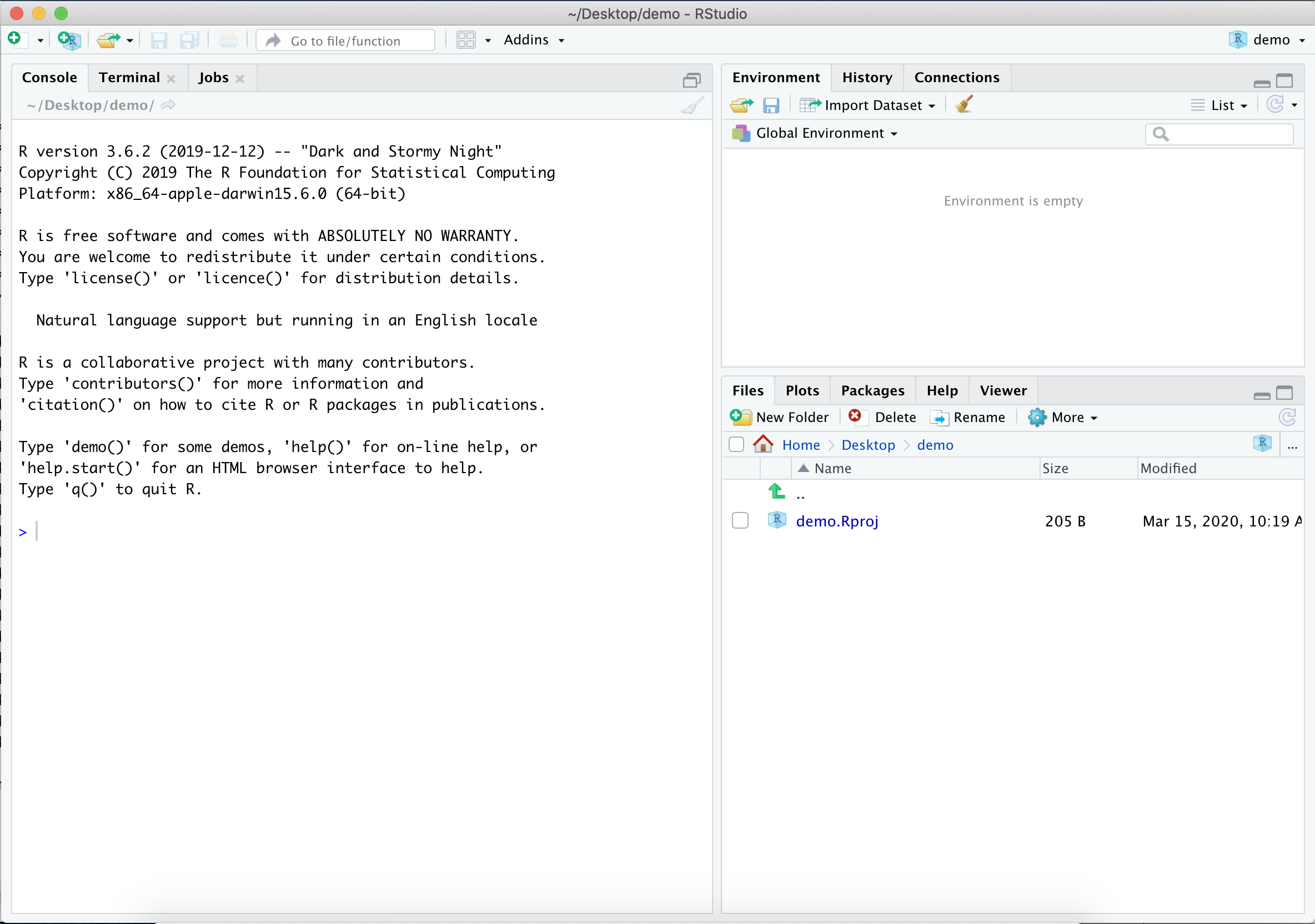


Figure 3: RStudio Layout

We’ll refer to these three “panes” as the **Console** pane, the **Environment** pane, and the **Files** pane. The large square on the left is the **Console** pane, the square in the top right is the **Environment** pane, and the square in the bottom right is the **Files** pane.

As you work with R more, you’ll find yourself using the tabs within each of the panes.

When we create a new file, such as an R script, an R Markdown file, or a Shiny app, RStudio will open a fourth pane, known as the **source** pane.  
The source pane should show up as a square in the top left. We can open up an .R script in the source pane by going to File, selecting New File, and then selecting R Script:

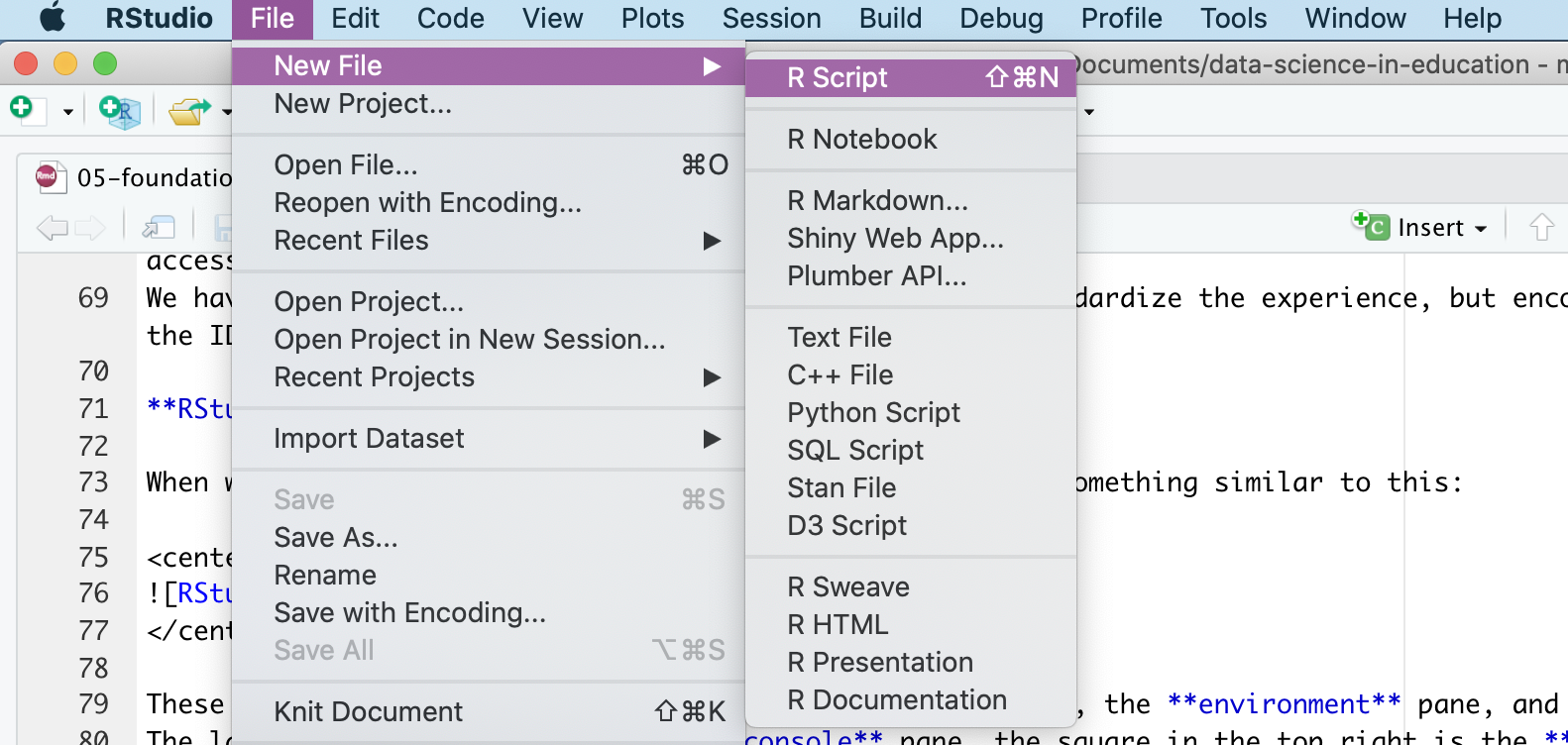


Figure 4: Creating a new R Script in RStudio

You do not need to do anything specific with this file, but we encourage you to experiment with it if you would like!

### Customizing RStudio

One of the balances we’ve tried to strike in this text is a balance between best practices in your *workflow* (how you’ll use R in your projects) and your *R code*. A best practice for your *workflow* is to ensure that you’re starting with a blank slate every time you open R (through RStudio). To accomplish this, go to Tools and select Global Options from the dropdown menu.

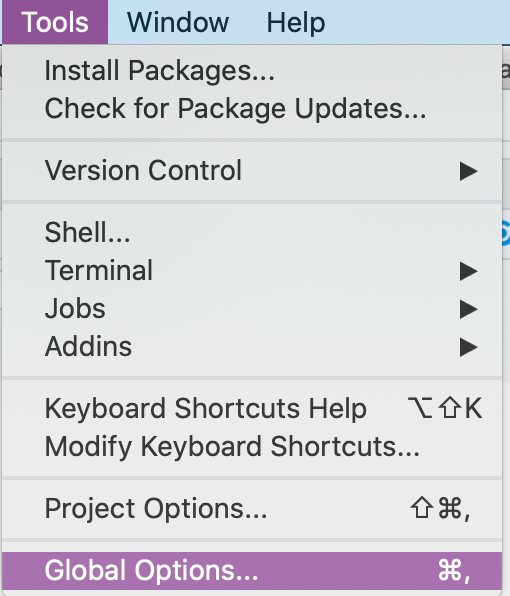


Figure 5: Selecting Global Options from the Tool Dropdown Menu

The General tab will open, with several checkboxes selected and unselected. The most important thing you can do is select “Never” next to the **Save workspace to .RData on exit:** prompt. After selecting “Never”, go through and check and uncheck boxes so that your General tab looks like this:

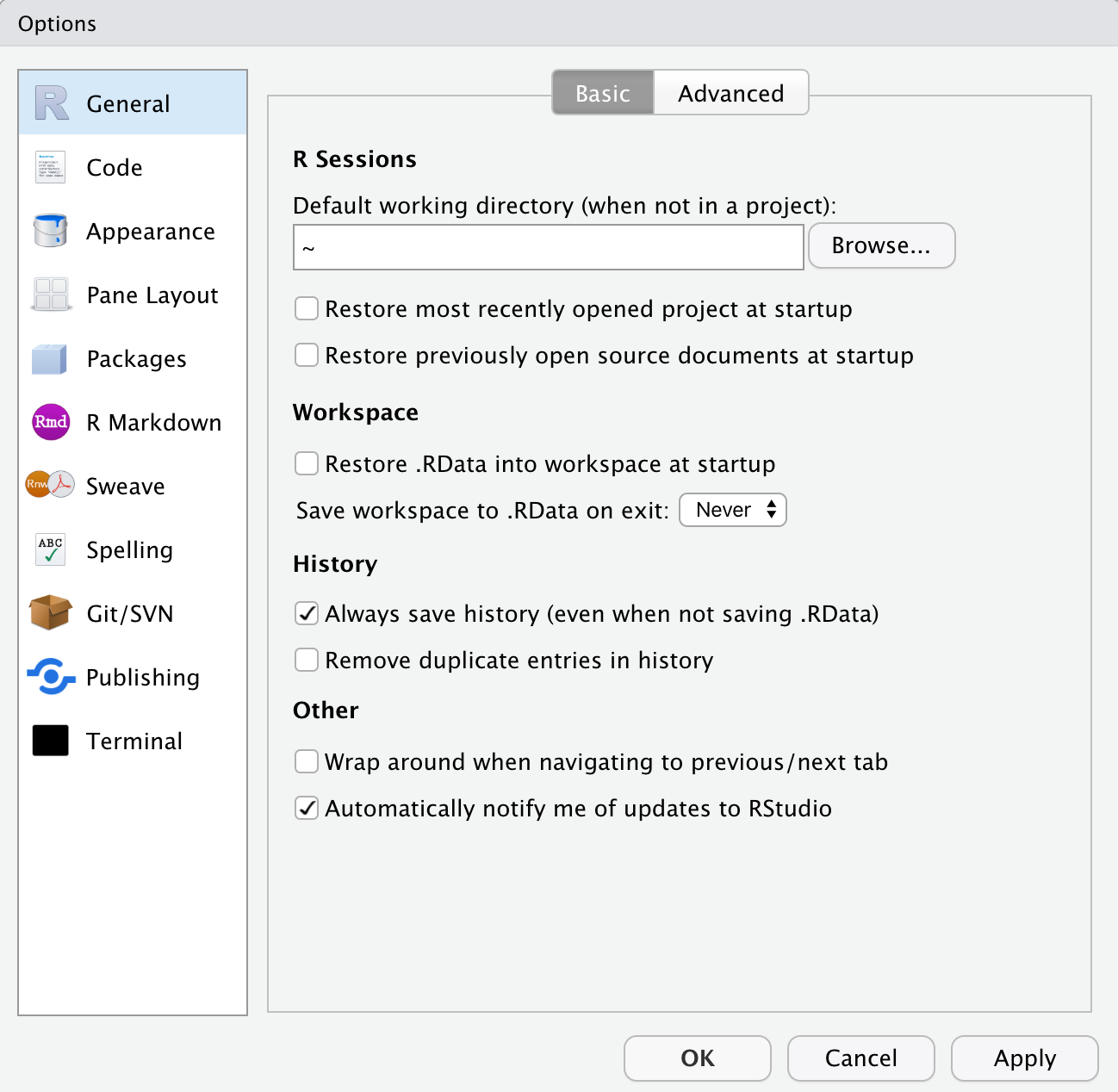


Figure 6: General tab from Global Options

Last, but certainly not least, click on the “Appearance” tab from within the Global Options. From here you can select your RStudio Font, Font Size, and Theme. Go through the options and select an appearance that works best for you, and know that you can *always* come back and change it!

### Minimized and Missing Panes

If at any point you find that one of your panes seems to have “disappeared,” one of two things has likely happened:

* A pane has been minimized
* A pane has been closed

Let’s look at the Environment pane as an example. If the Environment pane has been minimized, we’ll see something like this:

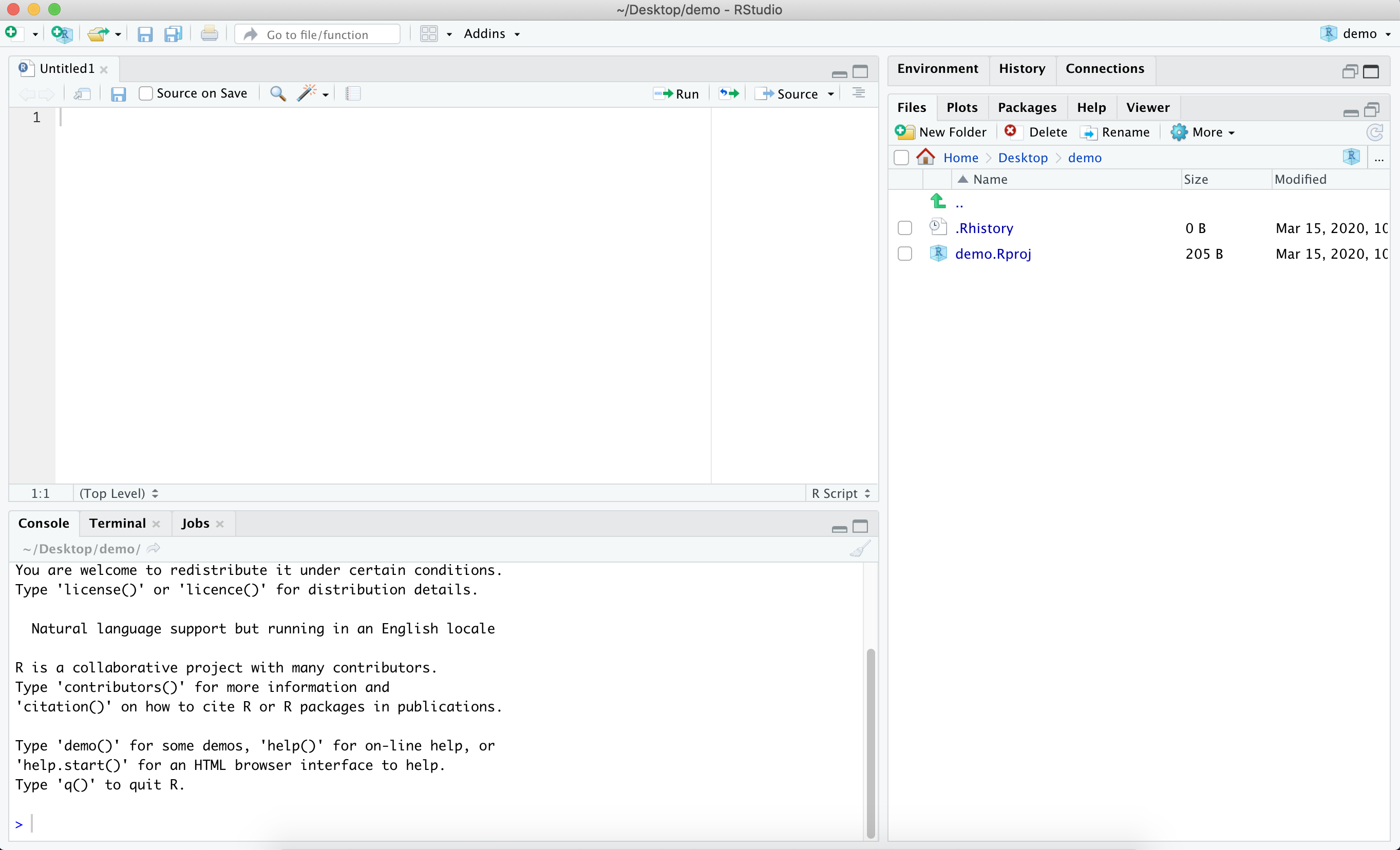


Figure 7: RStudio layout with the Environment Pane Minimized

We know that the Environment pane has been minimized, because we can see the pane headers in the top right, we just can’t see the information *within* the Environment pane. To fix this we can click on the icon of two squares in the top right of the Environment pane. If you click on the icon of the large square in the top right of the Environment pane you’ll maximize the Environment pane and minimize the Files pane.

If the Environment pane has somehow been closed, you can recover it by going to the View menu, selecting Panes, and then selecting Pane Layout, like so:

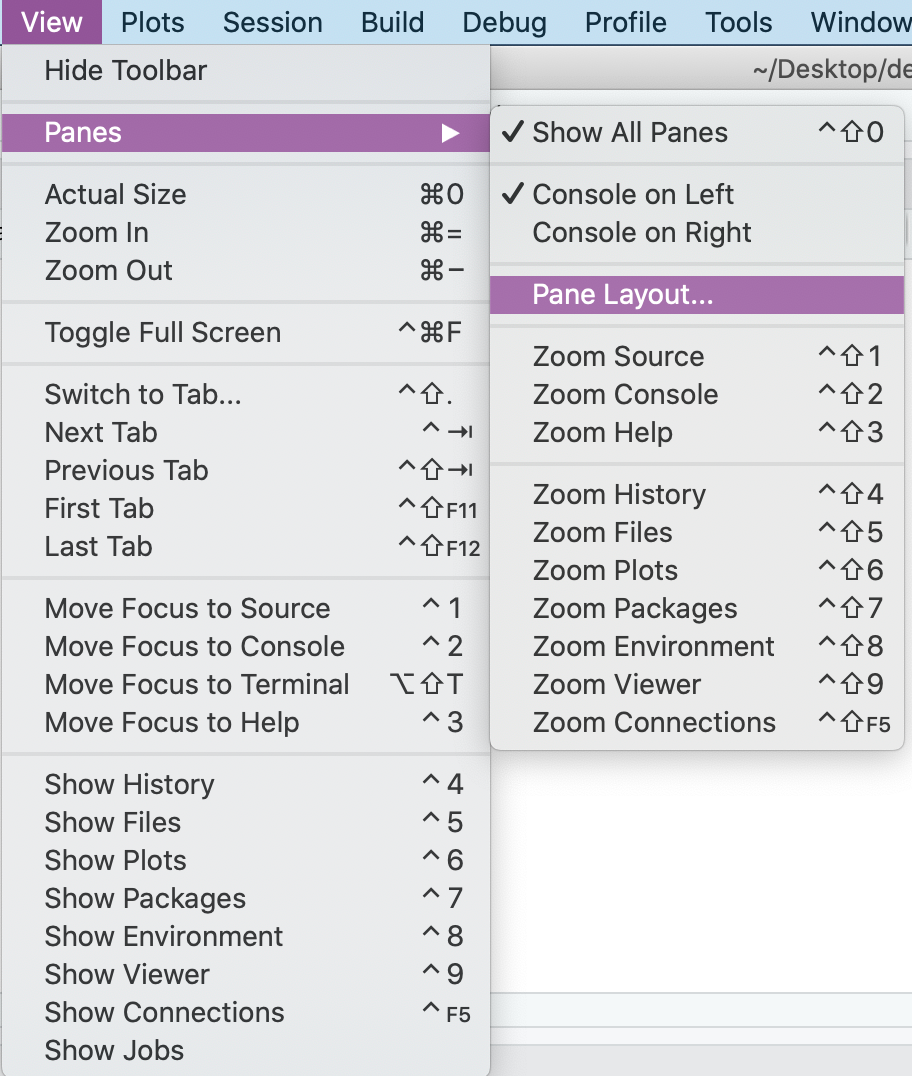


Figure 8: Accessing the Pane Layout from the View Dropdown Menu

When we select Pane Layout, we’ll see this:

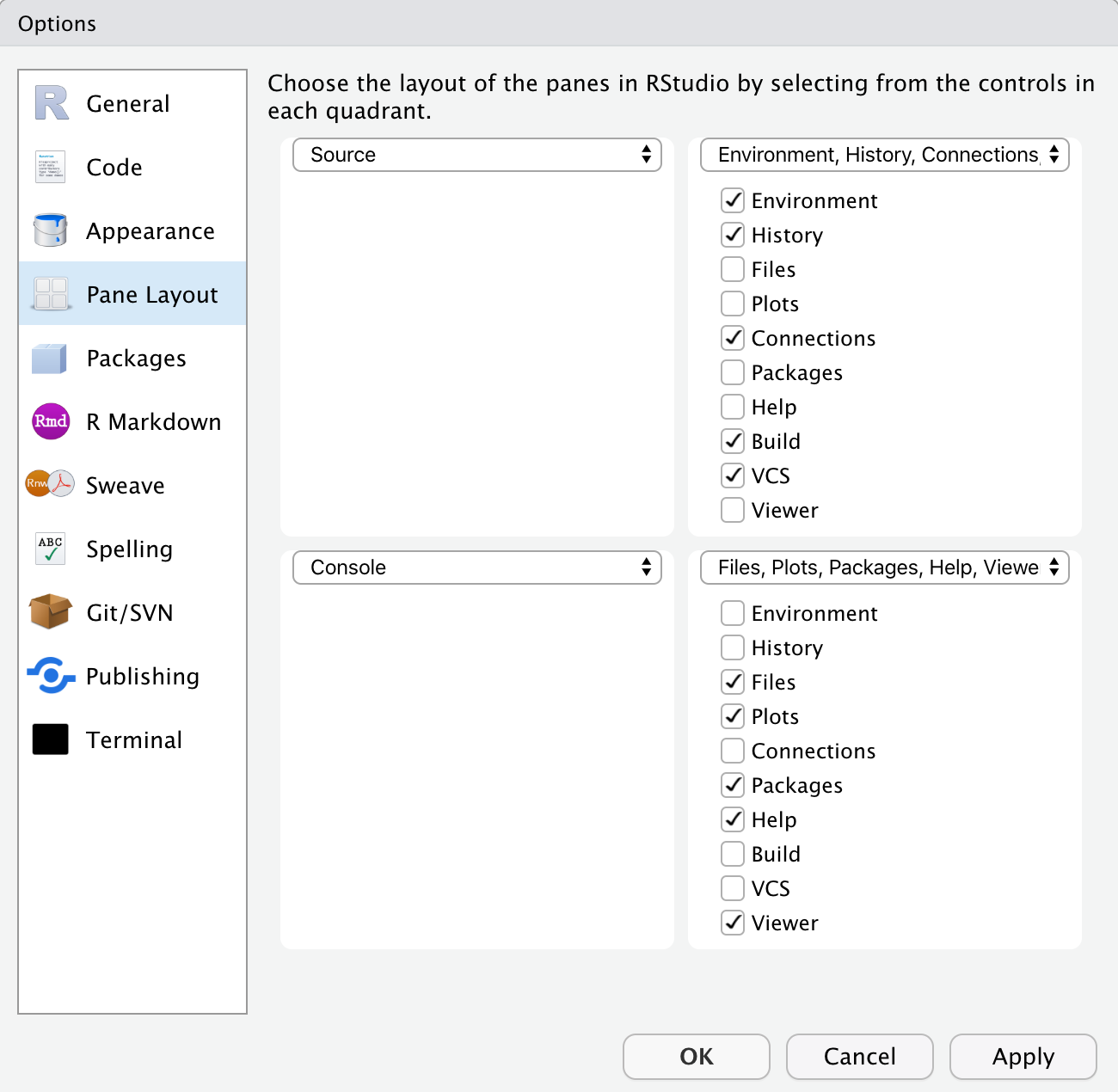


Figure 9: Pane Layout options within RStudio

From here you can select which tabs you’d like to appear within each pane, and even change where each pane appears within RStudio. So if our Environment Pane had been closed, we would select it from the Pane Layout in order to re-open it within RStudio.

## Writing and Running Code in RStudio

Up to this point we’ve been exploring the RStudio interface and setting up our preferences. Now we’ll shift to some basic coding practices. In order to run code in R you need to type your code either in the Console or within an .R script.

We generally recommend creating an .R script as you’re learning, as it allows you to type all of your code, add comments, and then save your .R script for reference. If instead you work entirely in the Console, anything that you type in the Console will disappear as soon as you restart or close R and you will not be able to reference it in the future.

### Writing Code in the Console

To run code in the Console, you type your code next to the > and hit ‘Enter’.  
We’ll spend a little time practicing running code in the Console by exploring some basic properties of coding in R.

In the Console, type 3 + 4 and hit Enter. You should see the following:



Figure 10: Using the Console as a Calculator

We’ve just used R to add the numbers 3 and 4. R has returned the sum of 3 + 4 on a new line, next to [1]. The [1] tells us that there is one row of data.

We can also use R to print out text. Type the following in the Console and hit Enter:

print("I am learning R")

We should see this in the Console:

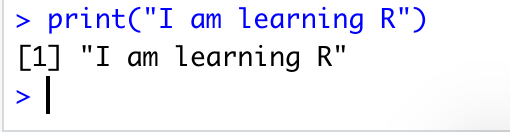


Figure 11: Printing Text to the Console

There’s one error that you’re likely going to come across, both when running code in the Console as well as in an R script. Let’s explore that error now, by running the following code in the Console and hitting Enter:

print("This is going to cause a problem"

Make sure that you left off the closing parentheses! What you’ll see in the Console is:

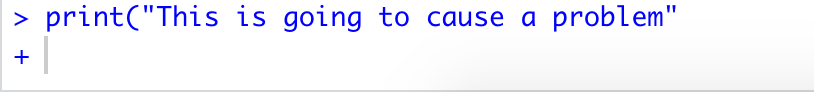


Figure 12: Incomplete Parentheses Change what R Expects Next

When we’re missing a closing parentheses, R is expecting us to provide more code. We know this because instead of seeing a carat > in our Console, we see a +, and R has not returned the print statement that we were expecting! There are two ways to fix this problem:

* Type the closing ) in the Console and hit Enter
* Hit the Esc key

Go ahead and run this intentional error and try each of the options above. Compare the output of each, and think about how they’re different. Can you think of when you might want to use one option instead of the other?

### Writing Code in an R Script

There are three main ways to run code in an .R script:  
- Highlight the line(s) of code you’d like to run and press **Ctrl + Enter**  
- Highlight the line(s) of code you’d like to run and click the **Run** button in the R script pane  
- To run *every* line of code in your file you can press **Ctrl + Shift + Enter**

Create a new .R script, or using the one you created earlier in this chapter, type in the following code and run it using each of the options listed above.

print("We're going to use R as a calculator.")  
print("First up, addition!")  
  
12 + 8  
632 + 41  
  
print("Next, subtraction!")  
  
48 - 6  
0.65 - 1.42

Feel free to spend some more time writing and running code within your .R script, or move on to the next section, where we’ll add comments to our code.

### Commenting Your Code in R

It is considered good practice to comment your code when working in an .R script. Even if you are the only person to ever work on your code, it can be helpful to write yourself notes about what you were trying to do with a specific piece of code. Moreover, writing comments in your code as you work through the examples in this book is a great way to help reinforce what you’re learning. Comments are ignored by R when running a script, so they will not affect your code or analysis.

To comment out a line of code, you can place a pound sign (also called an octothorpe!) # in front of the line of code that you want to exclude when you’re running your script. Be careful when doing this, especially in longer files, as it can be easy to forget where you’ve commented out code. It is often better to simply start a new section of code to tinker with until you get it working as expected, rather than commenting out lines of code.

We can also write comments in-line with our code, like this:

#' this will be a short code example.  
#' you are not expected to know what this does,  
#' nor do you need to try running it on your computer.  
  
library(readr) # load the readr package  
library(here) # load the here package  
  
data <- read\_csv(here("file\_path", "file\_name.csv")) # save file\_name.csv as data

If you think you’ll be writing more than one line of code, you can do a pound sign followed by a single quotation mark (#'). This will continue to comment out lines of text or code each time you hit “Enter.” You can delete the #' on a new line where you want to write code for R to run. This method is useful when you’re writing a description of what you’re doing in R.

*Note: when we refer to “commenting” we’re referring to adding in actual text comments, whereas “commenting out” refers to using the pound sign (octothorpe) in front of a line of code so that R ignores it.* *We will also use the phrase “uncomment code,” which means you should delete (or omit when typing out) the # or #' in an example.*

## Installing the {dataedu} package

This next section will briefly go over installing the {dataedu} package that’s used throughout this book. We created the {dataedu} package to provide our readers an opportunity to jump into R however they see fit.

The package serves three main functions:

1. Mass installation of all the packages used in the book
2. Reproducible code for the walkthroughs
3. Access to the data used in each of the walkthroughs

If you feel that you need more information before you’re ready to install the package, you can skip this section and rest assured that we’ll cover packages, their installation, and how to load them into R in more depth in Chapter 6. However, if you’re feeling a bit adventurous, go ahead and give it a shot by running the code below:

# install devtools  
install.packages("devtools", repos = "http://cran.us.r-project.org")  
  
# install the dataedu package  
devtools::install\_github("data-edu/dataedu")

## Exploring R with the {swirl} package

If you were able to install the {dataedu} package without any issues or concerns, and you’re eager to get started exploring everything that R can do, you can supplement your learning through [{swirl}](https://swirlstats.com/students.html) (https://swirlstats.com/students.html).

You can install {swirl} by running the following code:

install.packages("swirl")

{swirl} is set of packages (more on packages in Chapter 6!) that you can download, providing an interactive method for learning R by using R in the RStudio Console.  
You can follow the instructions on the {swirl} webpage (you’ve already installed R, RStudio, and the {swirl} package), or run the following code *in your console pane* to get started with a beginner-level course in {swirl}:

library(swirl)  
install\_course("R\_Programming\_E")  
swirl()

There are multiple courses available on {swirl}, and you can access them by installing them and then running the swirl() command in your console. We are not affiliated with {swirl} in any way, nor is it required to use {swirl} in order to progress through this text, but it’s a great resource that we want to make sure that you’re aware of!

## Conclusion

Congratulations! At this point in the book you’ve installed R and RStudio, explored the RStudio IDE, and even written some basic code. At this point you’re set up to either move on to Chapter 6, where we’ll go in-depth on Projects, packages, and functions, and how those relate to the data you’ll work with, as well as information on Help documentation and some skills for when you’re working with new or unfamiliar information. If that information is something you’re already comfortable with, you can jump ahead to a walkthrough of your choosing!

# Foundational Skills

## Chapter Overview

This chapter is designed to give you the skills and knowledge necessary to *get started* in any of the walk through chapters. Our goal is to get you working with R using the RStudio *I*ntegrated *D*evelopment *E*nvironment (IDE) through a series of applied examples. If you have not yet installed R and/or RStudio, please go through the steps outlined in Chapter 05 before beginning this chapter.

Please note that this chapter is not intended to be a full and complete introduction to programming with R, nor for using R for data science. There are many excellent resources available which provide this kind of instruction, and we’ve listed them for you in Chapter 17, [Additional Resources](#c18).

## Foundational Skills Framework

No two data science projects are the same, and rather than be overly prescriptive, this chapter errs on the side of creating a general framework for you to use as a home base as you work through this text. The four basic concepts we will use to build our framework are:

* **Projects**
* **Packages**
* **Functions**
* **Data**

### Projects

Within RStudio we set up a **Project**. This is the home for all of the files, images, reports, and code that we’ll create for a single project.

We use Projects because they create a self-contained folder for a given analysis in R. This means that if you want to share your Project with a colleague they will not have to reset file paths (or even know anything about file paths!) in order to re-run your analysis.

And even if the only person you ever collaborate with is a future version of yourself, using a Project for each of your analyses will mean that you can move the Project folder around on your computer, or even move it to a new computer, and remain confident that the analysis will run in the future (at least in terms of file path structures). #### Setting up your Project

**How do I create a Project?**  
Creating a Project is one of the first steps in working on an R-based data science project in RStudio. To create a Project you will need to be in RStudio.

From within RStudio, follow these steps:

1. Click on File
2. Select New Project
3. Choose New Directory
4. Click on New Project
5. Enter your Project’s name in the box that says “Directory name.” We recommend choosing a Project name that helps you to remember that this is a project that involves data science in education. Avoid using spaces in your Project name, and instead separate words with hyphens or underscore characters.
6. Choose where to save your Project by clicking on “Browse” next to the box labeled “Create project as a subdirectory of:” If you are just using this to learn and to test out creating a Project, consider placing it in your downloads or another temporary directory so that you remember to remove it later.
7. Click “Create Project”

We should point out that it is not *necessary* to create a Project for your work, although we *strongly* recommend it, as when it is coupled with the {here} package, will set you up with an easy to use workflow. However if you do not create a Project, you can always check where your working directory is by running getwd(). You can then change your working directory manually, by running setwd() and providing your file path name as an argument.

### The {here} package

*We don’t use the {here} package, do we?*

We’ve talked about what Projects are and why you should use them, but what really makes Projects in RStudio shine is the use of the here() function from the {here} package. What here() does is eliminate the need for you to do what’s called “hard-coding” your file path. For example, if we had loaded the data using the file path on one of our individual computers, it would be difficult for someone to open our folders on *their* computer and re-run the analysis because they do not have the same file structure. What here() does is tell R that the file structure starts at the Project-level, and so every subsequent call starts at the Project-level, and allows you to navigate throughout each of the folders and files within a given Project.

### Packages

“Packages” are shareable collections of R code that provide functions (i.e., a command to perform a specific task), data, and documentation. Packages increase the functionality of R by providing access to additional functions to suit a variety of needs, thereby expanding on base R.

If an R user feels that their package would be of benefit to a broad audience, they may choose to submit their package to [CRAN](https://cran.r-project.org/) (https://cran.r-project.org/), the **C**omprehensive **R** **A**rchive **N**etwork. Most of the packages we’ll be working with in this book are available on CRAN, which means that we can install them using the install.packages() function, as we did above.

The process of submitting a package and having it published through CRAN is beyond the scope of this book, and it’s important to point out that you - yes, you! - can create a package that you use all for yourself, share with colleagues, or submit to CRAN. Packages do not need to be submit to CRAN to be used by the public, and many packages are available directly from their respective developers via GitHub.

**The tidyverse, a package of packages** The {tidyverse} is a single package that contains additional packages, and within each of those individual packages are a set of functions. The {tidyverse} is an excellent tool that has a cohesive syntax across all functions, and the packages allow you do to the bulk of an analytical workflow. As of this writing, the {tidyverse} is the only known package of packages.

## Installing and Loading a package

*Installing a package*

If you read Chapter 05, you might have noticed at the very end that we both installed and loaded packages, but didn’t talk too much about what we were doing. We’ll get into more detail on installing and loading packages now!

While it is entirely possible to do all of your work in R without ever using a package, we do not recommend that approach due to the wealth of packages available that help reduce both the learning curve associated with R, as well as the amount of time spent on any given analytical project.

In order to access the functions within a package, you must first install the package on your computer. If the package is on CRAN we can install it by running the following code:

# template for installing a package  
install.packages("package\_name")  
  
# example of installing a package  
install.packages("dplyr")

You can also navigate to the Packages pane, click “Install”, and then search for and Install one or more packages.

“Image of the Packages pane, which is found in the bottom right corder of the RStudio IDE, along with the Files, Plots, Help, and Viewer panes”

“Image of the Packages pane, which is found in the bottom right corder of the RStudio IDE, along with the Files, Plots, Help, and Viewer panes”

*Loading a package*

Once a package is installed on your computer you do not have to re-install it in order to use the functions in the package, however you need to give R instructions telling it where to look for the functions in order to access the functions in the package.

If you know that you’ll be using a package’s functions multiple times in an R script (a file ending in .R), you can load the package into your R environment using library(). Loading a package into our R enviroment signals to R that we would like to use any of the functions available to us in that package.

We load a package, like the {dplyr} package (Wickham et al. [2020](#ref-R-dplyr)) below, using the following code:

# template for loading a package  
library(package\_name)   
  
# example of loading a package  
library(dplyr)

We only have to install a package once, but to use it, we have to load it each time we start a new R session.

A package is a like a book, a library is like a library; you use library() to check a package out of the library. - Hadley Wickham, Chief Scientist, RStudio

Sometimes you’ll see require() instead of library(). We strongly advocate for the use of library(), as it forces R to load the package, and if the package is not installed or there are issues with the package, will give you an error message. require() on the other hand will not give an error if the package is not available or if there are issues with the package. Using library() will help to eliminate sources of confusion later on.

### How to Find a Package

As you begin your R learning journey, the bulk of the packages that you will need to use are either already included when you install R, or available on CRAN. [CRAN TaskViews](https://cran.r-project.org/web/views/) (https://cran.r-project.org/web/views/) is one of the best resources for seeing what packages are available and might be relevant to your work. Other great resources to learn about various R packages are through Twitter (following the #rstats hashtag) as well as through Google searches. As R has grown in popularity, Google has gotten significantly better at returning R-related results.

### How to Learn More About a Package

Sometimes when you look up a package, you’re able to pull the function that you need and continue on your way. Other times you may need (or want!) to learn more about a specific package. Packages on CRAN all come with something called a “vignette,” which is a worked example using various functions from within the package. You can access a package’s vignette(s) on CRAN TaskViews:

Package authors may also publish vignettes or blog posts about their package, and other R users may *also* publish tutorials about a specific package. You may also find yourself on GitHub looking at information for a package - more often than not the README file will have good information for getting started with a package.

### Functions

**Recognizing a Function**

Functions in R can be spotted by a word followed by a set of parentheses, like so: word()

The word (or set of words) represent the name of the function, and the parentheses are where we can provide arguments to a function, if arguments are needed for the function to run. Many functions in R packages do not *require* arguments, and will use a set of default arguments unless you provide something different from the default.

### The Relationship Between Packages and Functions

**From the outside in**  
Packages are a collection of functions, and most packages are designed for: a specific data set, a specific field, and/or a specific set of tasks. Functions are individual components within a package, and functions are what we use to interact with our data.

**From the inside out**  
To put it another way, an R user might write a series of functions that they find themselves needing to use repeatedly in a variety of projects. Instead of re-writing (or copying and pasting) the functions each time they need to use them, an R user can collect all of these individual functions inside a package. The R user can then load the package any time that they want to use the functions, using a single line of code instead of tens to tens of thousands of lines of code for each function.

### Verb-Based Functions

Another pattern you may have picked up on is that every function is a verb. That is, the name of the function is telling us what the function is going to do! This is common both within the Tidyverse as well as in other packages.

### When Functions Need Arguments and When They Don’t

You also may have noticed that some of our functions needed arguments, that is, some kind of additional information provided inside the parentheses. There are not any hard and fast rules about when a function needs an argument (or series of arguments), however if you are having trouble running your code, first check for typos, then check the help documentation to see if you can provide arguments to more clearly direct R as to what to do.

### Functions within Functions

One of the (many!) cool things about R is that we can nest our functions. When we nest our functions, R will read the functions from the innermost function to the outermost. We saw this both in the section of code above that used arrange(desc(n)), as well as when we imported data using read\_csv(here()).

### Writing Your Own Functions

As you work in R more and more, you may find yourself copying and pasting the same lines of code, and then making small modifications. This is perfectly fine while you’re learning, but eventually you’re going to come across a large enough dataset that’s going to take you a couple hours to type out by hand (and increase the opportunity to introduce errors!). This is when you *know* you need to write a function. *(We could argue that you need functions* ***much*** *sooner! For example, a general premise in programming is DRY, or* ***D****on’t* ***R****epeat* ***Y****ourself.* *What this translates to is the idea that once you find yourself copying and pasting code for the third time, it’s time to write a function!)*

Functions are reusable pieces of code that you can write and use over and over again. You might write a function for a specific script, or you might find yourself using a function across multiple scripts (at which point you might want to consider creating a package!)

We’ll cover the very basics of writing a function, but would strongly suggest you check out [R For Data Science](https://r4ds.had.co.nz) (https://r4ds.had.co.nz).

The template for writing a function is:

name\_of\_function <- function(argument\_1, argument\_2, argument\_n){  
 code\_that\_does\_something  
 code\_that\_does\_something\_else  
}

Functions can be as simple or as complex as you would like. For example, if we wanted to create a function that adds two numbers together, we would write:

# writing our function  
# we've named the function "addition"  
# and asked for two arguments, "number\_1" and "number\_2"  
addition <- function(number\_1, number\_2) {  
 number\_1 + number\_2  
}  
  
# using our function  
# note that we provide each argument separated by commas  
addition(number\_1 = 3, number\_2 = 1)

## [1] 4

addition(0.921, 12.01)

## [1] 12.9

addition(62, 34)

## [1] 96

What happens if we only provide one argument? What happens if we provide more than two arguments?

You’ll encounter several (more advanced) functions that we’ll write together in various walkthroughs in this book!

### Data

We have **data** that we bring into a Project within RStudio. RStudio is the interface that we use to access and manipulate R. Sometimes you’ll hear RStudio referred to as an “eye-dee-eee”, “IDE”, or “Interactive Development Environment.” We use an IDE - in this case RStudio, although you could use others - because it adds features that make our analytical lives a little (and sometimes a lot!) easier. You’re likely using this text because you have some data that you’d like to do something with, and you’d like to try doing said thing using R. The framework we’ll use pulls data directly from the {dataedu} package, however there are a multitude of file types that data can be stored in. We’ve provided additional resources for loading data from Excel, SAV, and Google Sheets in the [Technical Appendix A](#technical-appendix-a) at the end of this chapter.

While it is possible to connect directly to a database from within R, we do not cover those skills in this text. For those curious as to how to accomplish this, we recommend starting with the [Databases using R](https://db.rstudio.com/) (https://db.rstudio.com/) resource from RStudio.

### Data Types

There are different types of data within R. When you loaded the Massachusetts Public Schools data from 2017, you likely saw a series of messages like this:

[IMG of import message]

This message is telling you how R decided to “code” each column, as you can only have one data type per column in R. We can see that the default is to code everything as a double (a number) where it says .default = col\_double(). Where columns were coded as something \_other\_ than double, we seecol\_character()(text that has at least one non-numeric character),col\_number()(for our purposes a number, and analogous tocol\_double()), andcol\_logical()` (TRUE or FALSE).

When read\_csv() reads through the first 1,000 rows of data, it uses what it encounters to make a best guess as to what type of data is in the column. If R sees 990 rows of numbers, but 10 rows of numbers mixed with letters, R will *coerce* the data in the column into character data. This means that 193 and 192A are both seen as *text*. One of the (many) implications of this is that you can no longer do mathematical operations on 193 when it’s been coerced into a character string.

Not all is lost! There are methods for safely coercing your data into the correct format, and we’ll cover a handful throughout this text. If you’re looking for more in-depth information on data types and coercion rules, we recommend both [Hands-On Programming with R](https://rstudio-education.github.io/hopr/) (https://rstudio-education.github.io/hopr/) (**???**) and [Advanced R](https://adv-r.hadley.nz/) (https://adv-r.hadley.nz/) Wickham ([2019](#ref-wickham2019advr)).

## Introduction to Help Documentation

Very few - if any - people in the world know everything there is to know about R. This means that we all need to look things up, sometimes every few minutes! Thankfully there are some excellent built-in resources that we can leverage as we use R.

From within RStudio we can access the Help documentation by using ? or ?? in the console. For example, if I wanted to look up information on the data() function, I can type ?data or ?data() next to the carat > in the Console and hit Enter. Try this now, and you should see the Help panel on the bottom right side of your RStudio environment populate with documentation on the data() function.

This works because the data() function is part of something called **base R** - that is, all of the functions included with R when you first install it. R also comes with packages pre-installed, however as you use R throughout this book, we’ll be asking you to install additional packages. These additional packages extend the functionality of base R and its pre-installed packages by providing us with access to new functions. This means that instead of writing a function to do a common data analysis task, such as creating a new variable out of existing variables, someone has written that function and made it available for you to use (almost always at no charge! Don’t worry - all of the packages we’ll be using in this text are considered **O**pen **S**ource **S**oftware, and you will not have to purchase anything to complete any of the exercises or walkthroughs in this text.)

One of the functions that can accomplish the task of creating a new variable out of existing variables is called mutate(). What happens when you type ?mutate (or ?mutate()) into the Console and hit Enter?

We’ve gotten one of our first error messages!

“Error message when running ?mutate reads: No documentation for ‘mutate’ in specified packages and libraries: you could try ‘??mutate’”

“Error message when running ?mutate reads: No documentation for ‘mutate’ in specified packages and libraries: you could try ‘??mutate’”

This is a fantastic error message because not only has it told us that something is wrong (there is no documentation for mutate), it tells us what we should try to do to solve the error. Let’s see what happens when we follow the error message instructions by typing ??mutate (or ??mutate()) into the Console and hitting Enter. What do you see?

## Steps for working through new and unfamiliar content

A sure sign of a great educator is the ability to ask questions. Asking the right questions at the right time of the learners in your classroom can facilitate understanding, uncover misconceptions, and indicate whether or not learners have mastered the material.

However, when you’re learning on your own you have to simultaneously fill the roles of both learner and educator, and not only know both how and when to ask yourself questions, but also answer your questions, evaluate your answers, and redirect your learning path as you progress.

This section is intended to give you a series of steps you can use as you encounter new and unfamiliar content both in reading this book as well as in your broader data science learning endeavors. For this section we’ll use the example of encountering a function for the first time, but you can use these steps with any new piece of information that you encounter! We’ll cover functions in greater depth in subsequent chapters, but for now all you need to know about a function is that it’s a reusable piece of code that allows us to consistently repeat a programming task.

**Activate prior knowledge** You’ve been reading through a tutorial and have come across the coalesce() function in the vignette for the [{janitor} function](https://github.com/sfirke/janitor) (https://github.com/sfirke/janitor):

library(tidyverse)  
library(janitor)  
  
roster <- roster\_raw %>%   
 clean\_names() %>%   
 remove\_empty(c("rows", "cols")) %>%   
 mutate(hire\_date = excel\_numeric\_to\_date(hire\_date),  
 cert = coalesce(certification, certification\_1)) %>%   
 select(-certification, -certification\_1)

*Note: you aren’t expected to know what the chunk of code that you’ve just read does, but by the time you’ve finished this book you’ll be able to run and understand everything in the code example!*

Take a moment to think through the following questions:

* What does the word “coalesce” mean?
* Have you ever seen the coalesce() function before? If so, in what context?

**Look for context clues**

* Read a couple of lines of code both above and below where the coalesce() function appears - are there any clues as to what this function might do?

**Check the docs**

* What information is available in the Help documentation?
* Are there any examples from the Help documentation that seem similar to what you’re trying to accomplish? For example, this seems somewhat related to what we’re trying to do:

Example from the coalesce() Help documentation

Example from the coalesce() Help documentation

**Find the limits** Work through examples in the Help documentation (or examples that you’ve found online) and test the limits. Testing the limits is a way of understanding the code by seeing how it handles different situations. Ultimately what you’re doing is recognizing a pattern, developing a hypothesis, and testing whether or not that hypothesis is true. Some methods for testing the limits include:

* Substituting obviously larger (or smaller values)
* Substituting different data types
* What happens if you introduce NA values?
* Is the order of values important?

**Test (and refine) your understanding** Take a moment to think through whether or not you could explain what you’ve just learned to someone else. Imagine the questions that they might ask of you, and try to answer them. If you can’t answer them, dig deeper into the documentation, online forums, or even in testing your own knowledge, until you feel like you can!

You won’t necessarily have the time (or interest!) in doing this for each new or unfamiliar piece of content that you come across, but we hope that this provides you with a starting framework for furthering your understanding when you *do* come across content that you want to explore a bit more in-depth!

## CODE WALKTHROUGH

Pay attention to when asking you to run code in your .R script vs. the console. Won’t break anything if make a mistake, but if you get lost or confused, check this first. - Set up a Project in RStudio - Create a new .R script and save it as “chapter\_06\_walkthrough”

You can see what the available arguments are for a given function by using the help documentation.

What are the arguments for library()?

### Run This Code

Take a few minutes to type out and run each of the following lines of code in an .R script, one by one, and notice what you see happening in the Console after you run each line.

# Installing packages  
install.packages("tidyverse")  
install.packages("janitor")  
install.packages("skimr")

# Setting up your environment  
library(tidyverse)  
library(janitor)  
library(skimr)

Based on what we’ve covered so far in this chapter, what did running the above lines of code accomplish? How do you know?

### Function Conflicts between Packages

In your Console you may have noticed the following message:

“List of attached packages and associated conflicts when loading the Tidyverse”

“List of attached packages and associated conflicts when loading the Tidyverse”

This isn’t error, but rather some important information that we need to consider! When we first open R (via RStudio) we are working with base R – that is, everything that comes with R along with a (relative) handful of pre-installed packages. These are packages and functions that exist in R that we have access to without needing to first load the package using library().

If you would like to see what functions are available to you in base R, you can run library(help = "base") in the Console.

If you would like to see the packages that came pre-installed with R, you can run installed.packages()[ installed.packages()[,"Priority"] %in% "base", c("Package", "Priority")] in the Console.

Additionally, if you would like to see a list of *all* of the packages that have been installed (both pre-installed with base R as well as those that you have installed), you can run rownames(installed.packages()) in the Console.

However because of the broad array of packages that have been created for use in R, it’s not uncommon for two (or more!) packages to have functions with the same name. What this message is telling us then is that if we use the filter() function, R will use the filter() function from the {dplyr} package (a package within the {tidyverse}) rather than the filter() function from within the {stats} package (one of the packages that accompanies base R).

How can we use the Help documentation to explore how the two functions might differ?

R will give precedence the most recently loaded package.

What happens if we want to use the filter() function from the {stats} package *and* the filter() function from the {dplyr} package in the same R session?

One solution would be to reload the library you want to use each time you want to change the package you’re using the filter() function from. But this can be tricky for several reasons: - It’s best practice to keep your library() calls at the very top of your R script, so re-loading a package using library() throughout your script can clutter things and cause you headaches down the road - If you scroll to the top of your script and reload the packages as you need them, it can be difficult to keep track of which one you recently loaded.

Instead there’s an easier way to handle this kind of problem. When we have conflicting function names from different packages we can tell R which package we’d like to pull a function from by using ::. Using the example of the filter() function above, coupled with the examples in the Help documentation, we can specify which package to pull the filter() function using ::, as outlined below:

*Note: we haven’t covered what any of this code does yet, but see what you can learn from running the code and using the Help documentation!*

# using the filter() function from the stats package  
x <- 1:100  
stats::filter(x, rep(1, 3))  
  
# using the filter() function from the dplyr package  
starwars %>%   
 dplyr::filter(mass > 85)

### Installing the {dataedu} package using {devtools}

Here, we will show how to install the {dataedu} package that accompanies this book.

As of this writing, the {dataedu} package is not available on CRAN. This means we’ll have to install the package using {devtools}. To do this you first have to install the {devtools} package, and then the {dataedu} package by running the following code either in your RStudio console or in a script (an .R file):

# install devtools  
install.packages("devtools", repos = "http://cran.us.r-project.org")  
  
# install the dataedu package  
devtools::install\_github("data-edu/dataedu")

From here we can load the {dataedu} package using the library() call, as follows:

# load the dataedu package into our R environment  
library(dataedu)

### Mass Installation of Packages

We strived to use packages that we use in our daily work when creating the walkthroughs in the book. Because we covered a variety of subjects, that means we used a lot of packages! As described in the Foundational Skills chapter, you can install the packages individually as they suit your needs.

However, if you want to quickly get started and download all the packages at once, please use install\_dataedu().

dataedu::install\_dataedu()

To see the packages used in the book, run:

dataedu::dataedu\_packages

## [1] "apaTables" "caret" "dummies" "ggraph" "here"   
## [6] "janitor" "lme4" "lubridate" "performance" "readxl"   
## [11] "rtweet" "randomNames" "sjPlot" "textdata" "tidygraph"   
## [16] "tidylog" "tidyverse" "tidytext"

**A special note on {tabulizer}:** One of the walkthroughs uses [tabulizer](https://github.com/ropensci/tabulizer) (https://github.com/ropensci/tabulizer), created by ROpenSci to read PDFs. {tabulizer} requires the installation of [RJava](https://cran.r-project.org/web/packages/rJava/index.html) (https://cran.r-project.org/web/packages/rJava/index.html), which can be a tricky process on Mac computers. {tabulizer} is not included in mass\_install() and we recommend reading through the notes on its Github repo if installing.

### Downloading and Accessing the Datasets Used in this Book

Throughout this book you’ll see data accessed in a multitude of ways. Sometimes we’ve pulled the data directly from a website, while other times we ask you to load the data from a .csv or .xls file. We’ve also provided each of the datasets used in this book as .rda files that are accessible via the {dataedu} package (Bovee et al. [2020](#ref-R-dataedu)). More details about the {dataedu} package on be found [on GitHub](https://github.com/data-edu/dataedu) (https://github.com/data-edu/dataedu), however we’ll walk through the basic steps here as well.

Once you’ve installed the {dataedu} package (described above), you can use the install\_dataedu() function to download all of the packages we’ll be using in this text. This step is optional, and you’re welcome to install packages as you use them.

To install all of the packages used in this text, run the following code in your RStudio console:

# see which packages are used in the Data Science in Education Using R text  
dataedu::dataedu\_packages  
  
# install all of the packages used in the Data Science in Education Using R text  
dataedu::install\_dataedu()

## Assigning Data to a Variable//Everything’s An Object!

### Run this Code

Take a few minutes to type out and run each of the following lines of code, one by one, and notice what you see happening in the Console after you run each line.

*Note: although we provide all of the data sets used in this book in the dataedu package, we would strongly suggest* [*downloading the dataset from Kaggle*](https://www.kaggle.com) *(https://www.kaggle.com) so that we can show you the true power of Projects in R.*

dataedu::ma\_data\_init  
  
dataedu::ma\_data\_init -> my\_data  
  
my\_data <- dataedu::ma\_data\_init

Each of the three code examples above look slightly different, but two of them do almost the exact same thing. The first example provided loads the data into our R environment, but not in a format that’s immediately useful to us. The second and third examples read in the data and assign it to a variable,my\_data and ma\_data\_init respectively. In our Environment pane we can see how each of the data types has been brought into R, and even click on the table icon to get an interactive table (The data set is rather large, so RStudio may lag slightly as you open the table and manipulate it.)

### The Assignment Operator

The second and third examples in the code chunk above are how you’ll most commonly see things in R being saved to a variable. When we save something to a variable, we do so using what’s called an “assignment operator,” which in R is either a left- or right-facing arrow (<- or ->). Writing the name of your variable followed by a left-facing arrow is currently the most common convention used in R, but it is also perfectly acceptable to use the right-facing arrow.

Intuitively the right-facing arrow may make more sense for those of us who work predemoninantly in languages that read left to right, as what we’re essentially saying is “take this entire chunk of code and save it to this variable name.” Regardless of which option you choose, both are different means to the same end.

## Reading Code and Writing Functions - THIS IS A BAD TITLE

### Run This Code

Take a few minutes to type out and run each of the following lines of code, one by one, and notice what you see happening in the Console after you run each line. If you’d like, practice commenting your code by noting what you see happening with each line of code that you run.

*Note: we have intentionally included errors in this chapter to help highlight concepts as well as introduce you to error messages early on!*

# Exploring and manipulating your data  
names(ma\_data\_init)  
  
glimpse(ma\_dat\_init)   
glimpse(ma\_data\_init)  
summary(ma\_data\_init)  
  
glimpse(ma\_data\_init$Town)  
summary(ma\_data\_init$Town)  
  
glimpse(ma\_data\_init$AP\_Test Takers)  
glimpse(ma\_data\_init$`AP\_Test Takers`)  
summary(ma\_data\_init$`AP\_Test Takers`)

What differences do you see between each line of code? What changes in the output to the Console with each line of code that you run?

### Common Errors: Typos, Spaces, and Parentheses

There were two lines of code that resulted in errors, and both were due to one of the most common sources of error in programming - typos! The first was glimpse(ma\_dat\_init). This might be a tricky error to spot, because at first glance it might seem like nothing is wrong! However we’ve left off the “a” in “data,” which has caused problems with R.

R will do exactly as you tell it to do. This means if you want to run a function on a data set, R will only run the function on the data sets that are available in its environment. Looking at our Environment pane we can see that there is no data set called ma\_dat\_init, which is what R is trying to tell us with its error message of Error in glimpse(ma\_dat\_init) : object 'ma\_dat\_init' not found.

The second error was with glimpse(ma\_data\_init$AP\_Test Takers). What do you think the error is here?

R is unhappy with the space in the file name, and doesn’t know how to read the code. To get around this there are a couple of things we can do. First, we could make sure that data column names never have spaces in them. This is unlikely to be within our control, so a second option would be to use R to convert the column names upon import, before we start doing any exploration. Another method for dealing with it is to leave the column names as they are, but to use single backticks to surround the column header with spaces in it.

*Note: the single backtick key is usually in the top-left of your keyboard. It’s common to try and use a set of single quotation marks ' ' instead of the actual backticks !*

### Closing your parentheses

Run the following code. What happens? How do we fix it? Seeing the + sign in the Console Fixing it by pressing Esc

### The $ Operator

There are many ways to isolate and explore a single variable within your data set. In this set of examples we used the $ symbol. The pattern for using the $ symbol is name\_of\_dataset$variable\_in\_dataset. It’s important that the spelling, punctuation, and capitalization that you use in your code match what’s in your data set, otherwise R will tell you that it can’t find anything!

## The Pipe Operator

The pipe operator %>% tends to throw R learners for a loop, until all of a sudden something clicks for them and they decide that they either love it or hate it. We use the pipe operator throughout this text because we also heavily rely on use of the tidyverse. The {magrittr} package {dplyr} and package dependencies “Correct” pipe length (careful - will need to stick to this throughout text)

*Note: As you progress in your R learning journey you will likely find that you need to move well beyond the Tidyverse for accomplishing your analytical goals - and that’s OK!* *We like the Tidyverse for teaching and learning because it relies on the same syntax across packages, so as you learn how to use functions within one package, you’re learning the syntax for functions in other Tidyverse packages.*

It’s worth taking a few moments to talk about the pipe operator and its package. The pipe operator first appeared in the {magrittr} package, and is a play on a famous painting by the artist Magritte, who painted The Treachery of Images. In these images he would paint an object, such as a pipe, and accompany it with the text “Ceci n’est pas une pipe,” which is French for “This is not a pipe.”

At the risk of spoiling a joke by over-explaining it, it’s common in the R programming world to name a package by choosing a word that represents what the package does (or what the package is for) and either capitalizing the letter R if it appears in the package name, or adding an R to the end of the package ({dplyr}, {tidyr}, {stringr}, and even {purrr}).

### Run This Code

Take a few minutes to type out and run each of the following lines of code, one by one, and notice what you see happening in the Console after you run each line.

ma\_data\_init %>%   
 group\_by(District Name) %>%   
 count()  
  
ma\_data\_init %>%   
 group\_by(`District Name`) %>%   
 count()   
   
ma\_data\_init %>%   
 group\_by(`District Name`) %>%   
 count() %>%   
 filter(n > 10)  
  
# ma\_data\_init %>%   
# group\_by(`District Name`) %>%   
# count() %>%   
# filter(n > 10) %>%   
# arrange(desc(n)

### “Reading” Code

When you encounter new-to-you code, it’s helpful to pause and read through the code to see if you can come up with a hypothesis as to what the code is trying to accomplish. Doing this will help you not only understand code a bit better, but also help you spot errors more quickly when the code doesn’t do what you thought it was going to do.

The way that we would read the first chunk of code is:

Take the ma\_data\_init data set *and then* **group** it **by** District Name *and then* **count** (the number of schools in a district) *and then* **filter** for Districts with more than 10 schools *and then* **arrange** the list of Districts and the number of schools in each District in descending order, based on the number of schools.

That’s a mouthful! But there are a couple of consistent points to make regarding this paragraph. Every time we see the pipe, we say “and then.” This is because we’re starting with our data set, ma\_data\_init, *and then* doing one thing after another to it. Because we’re using the pipe operator between each function, R knows that all of our functions are being applied to the ma\_data\_init data set. We do not need to call or refer to the ma\_data\_init data set each time.

When we link together functions using the pipe operator in this manner, we often refer to it as “chaining together functions.”

## Assignment vs. Equality

### Run This Code

Take a few minutes to read through the code before typing or running anything in R. Try to guess what is happening in each code chunk by writing out a sentence for each line of code so that you have a small paragraph for each chunk of code. Once you’ve done that, type out and run each of the following lines of code, one by one, and notice what you see happening in the Console after you run each line.

ma\_data\_init %>%  
 group\_by(`District Name`) %>%  
 count() %>%  
 filter(n > 10) %>%  
 arrange(desc(n))  
  
ma\_data\_init %>%  
 group\_by(`District Name`) %>%  
 count() %>%  
 filter(n = 10)   
  
ma\_data\_init %>%  
 group\_by(`District Name`) %>%  
 count() %>%  
 filter(n == 10)   
  
ma\_data\_init %>%   
 rename(district\_name = `District Name`,  
 grade = Grade) %>%   
 select(district\_name, grade)

**Wait I thought you said that <- and -> were the assignment operators?**

### The Difference Between = and ==

We talked earlier about using a left- or right-facing arrow to assign values or code to a variable, but we could also use an equals sign (=) to accomplish the same thing. When R encounters an equal sign (=) it is almost always looking to create an object by assigning a value to a variable. So when we saw filter(n = 10), R didn’t understand why we were trying to filter something we were naming, and told us so with an error message.

When we are looking to determine whether or not values are equal, we use a double equals sign (==), as we did in filter(n == 10). When R sees a double equals sign (==) it is evaluating whether or not the value on the left is equivalent to the value on the right.

## Basics of Names

Naming things is important! The more you use R the more you’ll develop your own sense of how you prefer to name things, either as an organization or an individual programmer. But there are some hard and fast rules that R has about naming things, and we’ll cover them in this section.

### Run This Code

Take a few minutes to type out and run each of the following lines of code, one by one, and notice what you see happening in the Console after you run each line.

ma\_data %>%   
 clean\_names()  
  
01\_ma\_data <- ma\_data\_init %>%   
 clean\_names()  
  
$\_ma\_data <- ma\_data\_init %>%   
 clean\_names()  
  
ma\_data\_01 <- ma\_data\_init %>%   
 clean\_names()

As you saw in the above examples, R doesn’t like it when you create a name that starts with a number or symbol. In addition, R is going to squawk when you give it a name with a space in it.

## Conclusion

It would be impossible for us to cover *everything* you can do with R in a single chapter of a book, but it is our hope that this chapter gives you a strong foundation from which to explore both subsequent chapters as well as additional R resources. In this chapter we’ve covered the concepts of Projects, packages, functions, and data, and also walked through foundational ideas, concepts, and skills related to doing data science in R. It is our hope that…

## Technical Appendix A

Here, we show how to use data from a few other common sources - your data likely exists somewhere in the world. That is to say, somewhere besides a package.

## Using Functions to Import Data

You might be thinking that an Excel file is the first type of data that we would load, but there happens to be a format which you can open and edit in Excel that is even easier to use between Excel and R, as well as SPSS and other statistical software (like MPlus) and even other programming languages, like Python. That format is .csv, or a comma-separated-values file.

The .csv file is useful because you can open it with Excel and save Excel files as .csv files. Additionally, and as its name indicates, a .csv file is rows of a spreadsheet with the columns separated by commas, so you can also view it in a text editor, like TextEdit for Macintosh. Not surprisingly, Google Sheets easily converts .csv files into a Sheet, and also easily saves Sheets as .csv files. However we would be remiss if we didn’t point out that there is a package, {googlesheets4}, which can be used to read a Google Sheet directly into R.

For these reasons, we start with - and emphasize - reading .csv files.

### Saving a File from the Web

You’ll need to copy this URL:

https://goo.gl/bUeMhV

Here’s what it resolves to (it’s a .csv file):

https://raw.githubusercontent.com/data-edu/data-science-in-education/master/data/pisaUSA15/stu-quest.csv

This next chunk of code downloads the file to your working directory. Run this to download it so in the next step you can read it into R. As a note: There are ways to read the file directory (from the web) into R. Also, of course, you could do what the next (two) lines of code do manually: Feel free to open the file in your browser and to save it to your computer (you should be able to ‘right’ or ‘control’ click the page to save it as a text file with a .csv extension).

student\_responses\_url <-  
 "https://goo.gl/bUeMhV"  
  
student\_responses\_file\_name <-  
 paste0(getwd(), "/data/student-responses-data.csv")  
  
download.file(  
 url = student\_responses\_url,  
 destfile = student\_responses\_file\_name)

It may take a few seconds to download as it’s around 20 MB.

The process above involves many core data science ideas and ideas from programming/coding. We will walk through them step-by-step.

1. The *character string* "https://goo.gl/wPmujv" is being saved to an *object* called student\_responses\_url.

student\_responses\_url <-  
 "https://goo.gl/bUeMhV"

1. We concatenate your working directory file path to the desired file name for the .csv using a *function* called paste0. This is stored in another *object* called student\_reponses\_file\_name. This creates a file name with a *file path* in your working directory and it saves the file in the folder that you are working in.

student\_responses\_file\_name <-  
 paste0(getwd(), "/data/student-responses-data.csv")

1. The student\_responses\_url *object* is passed to the url argument of the *function* called download.file() along with student\_responses\_file\_name, which is passed to the destfile argument.

In short, the download.file() function needs to know - where the file is coming from (which you tell it through the url) argument and - where the file will be saved (which you tell it through the destfile argument).

download.file(  
 url = student\_responses\_url,  
 destfile = student\_responses\_file\_name)

Understanding how R is working in these terms can be helpful for troubleshooting and reaching out for help. It also helps you to use functions that you have never used before because you are familiar with how some functions work.

Now, in RStudio, you should see the downloaded file in the Files tab. This should be the case if you created a project with RStudio; if not, it should be whatever your working directory is set to. If the file is there, great. If things are *not* working, consider downloading the file in the manual way and then move it into the directory that the R Project you created it.

### Loading a .csv File

Okay, we’re ready to go. The easiest way to read a .csv file is with the function read\_csv() from the package readr, which is contained within the Tidyverse.

Let’s load the tidyverse library:

library(tidyverse) # so tidyverse packages can be used for analysis

You may have noticed the hash symbol after the code that says library(tidyverse). It reads# so tidyverse packages can be used for analysis`. That is a comment and the code after it (but not before it) is not run (the code before it runs just like normal). Comments are useful for showing *why* a line of code does what it does.

After loading the tidyverse packages, we can now load a file. We are going to call the data student\_responses:

# readr::write\_csv(pisaUSA15::stu\_quest, here::here("data", "pisaUSA15", "stu\_quest.csv"))  
student\_responses <-  
 read\_csv("./data/student-responses-data.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## CNT = col\_character(),  
## CYC = col\_character(),  
## NatCen = col\_character(),  
## STRATUM = col\_character(),  
## Option\_Read = col\_character(),  
## Option\_Math = col\_character(),  
## ST011D17TA = col\_character(),  
## ST011D18TA = col\_character(),  
## ST011D19TA = col\_character(),  
## ST124Q01TA = col\_logical(),  
## IC001Q01TA = col\_logical(),  
## IC001Q02TA = col\_logical(),  
## IC001Q03TA = col\_logical(),  
## IC001Q04TA = col\_logical(),  
## IC001Q05TA = col\_logical(),  
## IC001Q06TA = col\_logical(),  
## IC001Q07TA = col\_logical(),  
## IC001Q08TA = col\_logical(),  
## IC001Q09TA = col\_logical(),  
## IC001Q10TA = col\_logical()  
## # ... with 420 more columns  
## )

## See spec(...) for full column specifications.

Since we loaded the data, we now want to look at it. We can type its name in the function glimpse() to print some information on the dataset (this code is not run here).

glimpse(student\_responses)

Woah, that’s a big data frame (with a lot of variables with confusing names, to boot)!

Great job loading a file and printing it! We are now well on our way to carrying out analysis of our data.

### Saving Files

Using our data frame student\_responses, we can save it as a .csv (for example) with the following function. The first argument, student\_reponses, is the name of the object that you want to save. The second argument, student-responses.csv, what you want to call the saved dataset.

write\_csv(student\_responses, "student-responses.csv")

That will save a .csv file entitled student-responses.csv in the working directory. If you want to save it to another directory, simply add the file path to the file, i.e. path/to/student-responses.csv. To save a file for SPSS, load the haven package and use write\_sav(). There is not a function to save an Excel file, but you can save as a .csv and directly load it in Excel.

In Technical Appendix A (at the end of this chapter), we show how to access directly data from a few other sources: Excel, SPSS (via .SAV files), and Google Sheets.

### Loading Excel Files

You might be thinking that you can open the file in Excel and then save it as a .csv. This is generally a good idea. At the same time, sometimes you may need to directly read a file from Excel. Note that, when possible, we recommend the use of .csv files. They work well across platforms and software (i.e., even if you need to load the file with some other software, such as Python).

The package for loading Excel files, {readxl}, is not a part of the tidyverse, so we will have to install it first (remember, we only need to do this once), and then load it using library(readxl). Note that the command to install {readxl} is grayed-out below: The # symbol before install.packages("readxl") indicates that this line should be treated as a comment and not actually run, like the lines of code that are not grayed-out. It is here just as a reminder that the package needs to be installed if it is not already.

Once we have installed readxl, we have to load it (just like tidyverse):

install.packages("readxl")

library(readxl)

We can then use the function read\_excel() in the same way as read\_csv(), where “path/to/file.xlsx” is where an Excel file you want to load is located (note that this code is not run here):

my\_data <-  
 read\_excel("path/to/file.xlsx")

Of course, were this run, you can replace my\_data with a name you like. Generally, it’s best to use short and easy-to-type names for data as you will be typing and using it a lot.

Note that one easy way to find the path to a file is to use the “Import Dataset” menu. It is in the Environment window of RStudio. Click on that menu bar option, select the option corresponding to the type of file you are trying to load (e.g., “From Excel”), and then click The “Browse” button beside the File/URL field. Once you click on the, RStudio will automatically generate the file path - and the code to read the file, too - for you. You can copy this code or click Import to load the data.

### Loading SAV Files

The same factors that apply to reading Excel files apply to reading SAV files (from SPSS). NOte that you can also read .csv file directly into SPSS and so because of this and the benefits of using CSVs (they are simple files that work across platforms and software), we recommend using CSVs when possible. First, install the package haven, load it, and the use the function read\_sav():

```r  
install.packages("haven")  
```

library(haven)  
my\_data <-  
 read\_sav("path/to/file.xlsx")

### Google Sheets

Finally, it can sometimes be useful to load a file directly from Google Sheets, and this can be done using the Google Sheets package.

install.packages("googlesheets")

library(googlesheets)

When you run the command below, a link to authenticate with your Google account will open in your browser.

my\_sheets <- gs\_ls()

You can then simply use the gs\_title() function in conjunction with the gs\_read() function:

```r  
df <- gs\_title('title')  
df <- gs\_read(df)  
```

# Walkthrough 1: The Education Data Science Pipeline With Online Science Class Data

## Introduction to the walkthroughs

This chapter is the first of eight walkthroughs included in the book. In it, we present *one approach* to analyzing a specific dataset; in this case, the approach is what we call the education data science pipeline using data from a number of online science classes. While the walkthroughs are very different, the structure (and section headings) will be consistent throughout the walkthroughs. For example, every walkthrough will begin with a vocabulary section, followed by an introduction to the dataset (and the question or problem) explored in the walkthrough.

## Vocabulary

* item
* joins
* keys
* log-trace data
* pass
* reverse scale
* regression
* survey
* tibble
* vectorize

## Chapter Overview

In this walkthrough, we explore some of the key steps that are a part of many data science in education projects. In particular, we explore how to process and prepare data: what is sometimes referred to as data wrangling. To do so, we rely heavily on a set of tools that we use throughout *all* of the walkthroughs, those associated with the tidyverse, the set of packages for data manipulation, exploration, and visualization using the design philosophy of ‘tidy’ data (Wickham et al. [2019](#ref-wickham2019)). For more information, see the Foundational Skills chapters or <https://www.tidyverse.org/>.

The tidyverse is predicated on the concept of tidy data (Wickham [2014](#ref-wickham2014)). Tidy data has a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. We’ll discuss both the tidyverse and tidy data much more throughout the book.

### Background

The online science classes from which the data used in this walkthrough was obtained were designed and taught by instructors through a statewide online course provider designed to supplement (and not replace) students’ enrollment in their local school. For example, students may choose to enroll in an online physics class because one was not offered at their school. The data were originally collected for a research study, which involved a number of different data sources which were explored to understand students’ motivation, or their reasons for taking the course. The datasets included:

1. A self-report survey for three distinct but related aspects of students’ motivation
2. Log-trace data, such as data output from the learning management system
3. Discussion board data (not used in this walkthrough)
4. Achievement-related (i.e., final grade) data

Our *purpose* for this walkthrough is to begin to understand what explains students’ performance in these online courses. The *problem* we are facing is a very common one when it comes to data science in education: the data are complex and in need of further processing before we can get to answering questions (or running analyses).

To understand students’ performance, we will focus on a variable that was available through the learning management system used for the courses on the amount of time students’ spent on the course. We will also explore how different (science) subjects as well as being in a particular class may help to explain student performance.

First, these different data sources will be described in terms of how they were provided by the school.

### Data Sources

#### Data Source #1: Self-Report Survey about Students’ Motivation

The first data source is a self-report survey. This was data collected before the start of the course via self-report survey. The survey included 10 items, each corresponding to one of three *measures*: interest, utility value, and perceived competence:

1. I think this course is an interesting subject. (Interest)
2. What I am learning in this class is relevant to my life. (Utility value)
3. I consider this topic to be one of my best subjects. (Perceived competence)
4. I am not interested in this course. (Interest - reverse coded)
5. I think I will like learning about this topic. (Interest)
6. I think what we are studying in this course is useful for me to know. (Utility value)
7. I don’t feel comfortable when it comes to answering questions in this area. (Perceived competence)
8. I think this subject is interesting. (Interest)
9. I find the content of this course to be personally meaningful. (Utility value)
10. I’ve always wanted to learn more about this subject. (Interest)

### Data Source #2: Log-Trace Data

*Log-trace data* is data generated from our interactions with digital technologies, such as archived data from social media postings (see [Chapter 11](#c11) and [Chapter 12](#c12)). In education, an increasingly common source of log-trace data is that generated from interactions with learning management systems and other digital tools (Siemens and d Baker [2012](#ref-siemens2012)). The data for this walk-through is a *summary of* log-trace data, namely, the number of minutes students spent on the course. While this data is rich, you can imagine even more complex sources of log-trace data (e.g. time stamps associated with when students started and stopped accessing the course!).

### Data Source #3: Achievement-Related and Gradebook Data

This is a common source of data: one associated with graded assignments students completed. In this walkthrough, we just examine students’ final grade.

### Data Source #4: Discussion Board Data

Discussion board data is both rich and unstructured, in that it is primarily in the form of written text. We also collected discussion board data for this project.

### Methods

In this walkthrough, we will concentrate on the different joins available in the {dplyr} package. We will also start exploring how to run linear models in R.

## Load Packages

This analysis uses R packages, which are collections of R code that help users code more efficiently, as you will recall from [Chapter 1](#c1). We load these packages with the function library(). In particular, the packages we’ll use will help us organize the structure of the data, work with dates in the data using {lubridate} (Spinu, Grolemund, and Wickham [2018](#ref-R-lubridate)), create formatted tables using {apaTables} (Stanley [2018](#ref-R-apaTables)) and {sjPlot} (Lüdecke [2020](#ref-R-sjPlot)), and navigate file directories using {here} (Müller [2017](#ref-R-here)).

library(readxl)  
library(tidyverse)  
library(lubridate)  
library(here)  
library(dataedu)  
library(apaTables)  
library(sjPlot)

## Import Data

This code chunk loads the log-trace data from the {dataedu} package. Note that we assign a dataset to an object three times, once for each of the three log-trace datasets. We assign each of the datasets a name using <-.

# Gradebook and log-trace data for F15 and S16 semesters  
course\_data <- dataedu::course\_data  
  
# Pre-survey for the F15 and S16 semesters  
  
pre\_survey <- dataedu::pre\_survey  
  
# Log-trace data for F15 and S16 semesters - this is for time spent  
  
course\_minutes <- dataedu::course\_minutes

## View Data

Now that we’ve successfully loaded all three log-trace datasets, we can visually inspect the data by typing the names that we assigned to each dataset.

pre\_survey

## # A tibble: 1,102 x 12  
## opdata\_username opdata\_CourseID Q1Maincellgroup… Q1Maincellgroup…  
## <chr> <chr> <dbl> <dbl>  
## 1 \_80624\_1 FrScA-S116-01 4 4  
## 2 \_80623\_1 BioA-S116-01 4 4  
## 3 \_82588\_1 OcnA-S116-03 NA NA  
## 4 \_80623\_1 AnPhA-S116-01 4 3  
## 5 \_80624\_1 AnPhA-S116-01 NA NA  
## 6 \_80624\_1 AnPhA-S116-02 4 2  
## 7 \_80624\_1 AnPhA-T116-01 NA NA  
## 8 \_80624\_1 BioA-S116-01 5 3  
## 9 \_80624\_1 BioA-T116-01 NA NA  
## 10 \_80624\_1 PhysA-S116-01 4 4  
## # … with 1,092 more rows, and 8 more variables: Q1MaincellgroupRow3 <dbl>,  
## # Q1MaincellgroupRow4 <dbl>, Q1MaincellgroupRow5 <dbl>,  
## # Q1MaincellgroupRow6 <dbl>, Q1MaincellgroupRow7 <dbl>,  
## # Q1MaincellgroupRow8 <dbl>, Q1MaincellgroupRow9 <dbl>,  
## # Q1MaincellgroupRow10 <dbl>

course\_data

## # A tibble: 29,711 x 8  
## CourseSectionOr… Bb\_UserPK Gradebook\_Item Grade\_Category FinalGradeCEMS  
## <chr> <dbl> <chr> <chr> <dbl>  
## 1 AnPhA-S116-01 60186 POINTS EARNED… <NA> 86.3  
## 2 AnPhA-S116-01 60186 WORK ATTEMPTED <NA> 86.3  
## 3 AnPhA-S116-01 60186 0.1: Message … <NA> 86.3  
## 4 AnPhA-S116-01 60186 0.2: Intro As… Hw 86.3  
## 5 AnPhA-S116-01 60186 0.3: Intro As… Hw 86.3  
## 6 AnPhA-S116-01 60186 1.1: Quiz Qz 86.3  
## 7 AnPhA-S116-01 60186 1.2: Quiz Qz 86.3  
## 8 AnPhA-S116-01 60186 1.3: Create a… Hw 86.3  
## 9 AnPhA-S116-01 60186 1.3: Create a… Hw 86.3  
## 10 AnPhA-S116-01 60186 1.4: Negative… Hw 86.3  
## # … with 29,701 more rows, and 3 more variables: Points\_Possible <dbl>,  
## # Points\_Earned <dbl>, Gender <chr>

course\_minutes

## # A tibble: 598 x 3  
## Bb\_UserPK CourseSectionOrigID TimeSpent  
## <dbl> <chr> <dbl>  
## 1 44638 OcnA-S116-01 1383.  
## 2 54346 OcnA-S116-01 1191.  
## 3 57981 OcnA-S116-01 3343.  
## 4 66740 OcnA-S116-01 965.  
## 5 67920 OcnA-S116-01 4095.  
## 6 85355 OcnA-S116-01 595.  
## 7 85644 OcnA-S116-01 1632.  
## 8 86349 OcnA-S116-01 1601.  
## 9 86460 OcnA-S116-01 1891.  
## 10 87970 OcnA-S116-01 3123.  
## # … with 588 more rows

## Process Data

Often, survey data needs to be processed in order to be (most) useful. Here, we process the self-report items into three scales for 1) interest, 2) self-efficacy, and 3) utility value. We do this by:

* Renaming the question variables to something more manageable
* Reversing the response scales on questions 4 and 7
* Categorizing each question into a measure
* Computing the mean of each measure

Let’s take these steps in order:

1. Rename the question columns to something much simpler:

pre\_survey <-  
 pre\_survey %>%  
 # Rename the qustions something easier to work with because R is case sensitive  
 # and working with variable names in mix case is prone to error  
 rename(  
 q1 = Q1MaincellgroupRow1,  
 q2 = Q1MaincellgroupRow2,  
 q3 = Q1MaincellgroupRow3,  
 q4 = Q1MaincellgroupRow4,  
 q5 = Q1MaincellgroupRow5,  
 q6 = Q1MaincellgroupRow6,  
 q7 = Q1MaincellgroupRow7,  
 q8 = Q1MaincellgroupRow8,  
 q9 = Q1MaincellgroupRow9,  
 q10 = Q1MaincellgroupRow10  
 ) %>%  
 # Convert all question responses to numeric  
 mutate\_at(vars(q1:q10), list( ~ as.numeric(.)))

Let’s take a moment to discuss the {dplyr} function mutate\_at(). mutate\_at() is a version of mutate(), which changes the values in an existing column or creates new columns. It’s useful in education datasets because you’ll often need to transform your data before analyzing it. Try this example, where we create a new total\_students column by adding the number of male students and female students:

# Dataset of students  
df <- tibble(  
 male = 5,   
 female = 5  
)  
  
df %>% mutate(total\_students = male + female)

## # A tibble: 1 x 3  
## male female total\_students  
## <dbl> <dbl> <dbl>  
## 1 5 5 10

mutate\_at() is a special version of mutate(), which conveniently changes the values of multiple columns. In our dataset pre\_survey, we let mutate() know we want to change the variables q1 through q10. We do this with the argument vars(q1:q10)

1. Next we’ll reverse the scale of the survey responses on questions 4 and 7 so the responses for all questions can be interpreted in the same way. Rather than write a lot of code once to reverse the scales for question 4 then writing it again to reverse the scales on question 7, we’ll build a function that does that job for us. Then we’ll use the same function for question 4 and question 7. This will result in much less code, plus it will make it easier for us to change in the future.

We’ll use case\_when() in our function to reverse the scale of the item responses. case\_when() is useful when you need to replace the values in a column with other values based on some criteria. Education datasets use a lot of codes to describe demographics, like numerical codes for disability categories, race groups, or proficiency in a test. When you work with codes like this, you’ll often want to change the codes to values that are easier to understand. For a example, a consultant analyzing how students did on state testing might use case\_when() to replace proficiency codes like 1, 2, or 3 to more descriptive words like “below proficiency”, “proficient”, or “advanced”.

case\_when() lets you vectorize the rules you want to use to change values in a column. When a sequence of criteria is vectorized, R will evaluate a value in a column against each criteria in your case\_when() sequence. case\_when() is helpful because it does this without complicated loops. By using code that is compact and readable once you understand how all the arguments work.

The left hand side of each case\_when() argument will be a formula that returns either a TRUE or a FALSE. In the function below we’ll use logical operators in the left hand side of the formula like this: question == 1 ~ 5. Here are some other logical operators you can use in the future:

* >: greater than
* <: lesser than
* >=: greater than or equal to
* <=: lesser than or equal to
* ==: equal to
* !=: not equal to
* !: not
* &: and
* |: or

Let’s make this all concrete and use it here in our function that reverses the scale of the survey responses:

# This part of the code is where we write the function:  
# Function for reversing scales   
reverse\_scale <- function(question) {  
 # Reverses the response scales for consistency  
 # Args:  
 # question: survey question  
 # Returns: a numeric converted response  
 # Note: even though 3 is not transformed, case\_when expects a match for all  
 # possible conditions, so it's best practice to label each possible input  
 # and use TRUE ~ as the final statement returning NA for unexpected inputs  
 x <- case\_when(  
 question == 1 ~ 5,  
 question == 2 ~ 4,  
 question == 4 ~ 2,  
 question == 5 ~ 1,  
 question == 3 ~ 3,  
 TRUE ~ NA\_real\_  
 )  
 x  
}  
  
# And here's where we use that function to reverse the scales  
# Reverse scale for questions 4 and 7  
pre\_survey <-  
 pre\_survey %>%  
 mutate(q4 = reverse\_scale(q4),  
 q7 = reverse\_scale(q7))

1. We’ll accomplish the last two steps in one chunk of code. First we’ll create a column called measure and we’ll fill that column with one of three question categories:

* int: interest
* uv: utility value
* pc: self efficacy

After that we’ll find the mean response of each category using mean() function.

# Add measure variable   
measure\_mean <-  
 pre\_survey %>%  
 # Gather questions and responses  
 pivot\_longer(cols = q1:q10,  
 names\_to = "question",  
 values\_to = "response") %>%  
 # Here's where we make the column of question categories  
 mutate(  
 measure = case\_when(  
 question %in% c("q1", "q4", "q5", "q8", "q10") ~ "int",  
 question %in% c("q2", "q6", "q9") ~ "uv",  
 question %in% c("q3", "q7") ~ "pc",  
 TRUE ~ NA\_character\_  
 )  
 ) %>%  
 group\_by(measure) %>%  
 # Here's where we compute the mean of the responses  
 summarise(  
 # Mean response for each measure  
 mean\_response = mean(response, na.rm = TRUE),  
 # Percent of each measure that had NAs in the response field  
 percent\_NA = mean(is.na(response))  
 )  
  
measure\_mean

## # A tibble: 3 x 3  
## measure mean\_response percent\_NA  
## <chr> <dbl> <dbl>  
## 1 int 4.25 0.178  
## 2 pc 3.65 0.178  
## 3 uv 3.74 0.178

### Processing the Course Data

We also can process the course data in order to create new variables which we can use in analyses. Information about the course subject, semester, and section are stored in a single column, CourseSectionOrigID. If we give each of these their own columns, we’ll have more opportunities to analyze them as their own variables. We’ll use a function called separate() to do this. This pulls out the subject, semester, and section from the course ID so we can use them later on.

# split course section into components  
course\_data <-   
 course\_data %>%  
 # Give course subject, semester, and section their own columns  
 separate(  
 col = CourseSectionOrigID,  
 into = c("subject", "semester", "section"),  
 sep = "-",  
 remove = FALSE  
 )

### Joining the Data

To join the course data and pre-survey data, we need to create similar *keys*. In other words, our goal here is to have one variable that matches across both datasets, so that we can merge the datasets on the basis of that variable.

For these data, both have variables for the course and the student, though they have different names in each. Our first goal will be to rename two variables in each of our datasets so that they will match. One variable will correspond to the course, and the other will correspond to the student. We are not changing anything in the data itself at this step - instead, we are just cleaning it up so that we can look at the data all in one place.

Let’s start with the pre-survey data. We will rename RespondentID and opdata\_CourseID to be student\_id and course\_id, respectively.

pre\_survey <-  
 pre\_survey %>%  
 rename(student\_id = opdata\_username,  
 course\_id = opdata\_CourseID)  
  
pre\_survey

## # A tibble: 1,102 x 12  
## student\_id course\_id q1 q2 q3 q4 q5 q6 q7 q8 q9  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 \_80624\_1 FrScA-S1… 4 4 4 5 5 4 5 5 5  
## 2 \_80623\_1 BioA-S11… 4 4 3 4 4 4 4 3 4  
## 3 \_82588\_1 OcnA-S11… NA NA NA NA NA NA NA NA NA  
## 4 \_80623\_1 AnPhA-S1… 4 3 3 4 3 3 3 4 2  
## 5 \_80624\_1 AnPhA-S1… NA NA NA NA NA NA NA NA NA  
## 6 \_80624\_1 AnPhA-S1… 4 2 2 4 4 4 5 4 4  
## 7 \_80624\_1 AnPhA-T1… NA NA NA NA NA NA NA NA NA  
## 8 \_80624\_1 BioA-S11… 5 3 3 5 5 4 5 5 3  
## 9 \_80624\_1 BioA-T11… NA NA NA NA NA NA NA NA NA  
## 10 \_80624\_1 PhysA-S1… 4 4 3 4 4 4 4 4 3  
## # … with 1,092 more rows, and 1 more variable: q10 <dbl>

Looks better now!

In addition to needing to be renamed, the student\_id variable had an issue - the variable has some additional characters before and after *the actual ID* that we will need to be able to join this data with the other data sources we have. Why does this variable have these additional characters? We are not sure! Sometimes, educational data from different systems (used for different purposes) may have additional “meta”-data added on. In any event, here is what the variables look like before and processing:

head(pre\_survey$student\_id)

## [1] "\_80624\_1" "\_80623\_1" "\_82588\_1" "\_80623\_1" "\_80624\_1" "\_80624\_1"

What we need is the five characters in between the underscore symbols - these: \_.

One way to do this is to use the str\_sub() function from the {stringr} package. You can specify the indices of the variables you want the string to *start* and *end* with. Here, for example, is how we can start with the second character, skipping the first underscore in the process.

str\_sub("\_80624\_1", start = 2)

## [1] "80624\_1"

We can do the same with the last few characters:

str\_sub("\_80624\_1", end = -3)

## [1] "\_80624"

Putting the pieces together, the following should return what we want:

str\_sub("\_80624\_1", start = 2, end = -3)

## [1] "80624"

We can apply this to our data using mutate(). We convert the string into a number using as.numeric() at the same time, so the data can be joined to the other, numeric student\_id variables (in the other data sets):

pre\_survey <- pre\_survey %>%   
 mutate(student\_id = as.numeric(str\_sub(student\_id, start = 2, end = -3)))

Let’s proceed to the course data. Our goal is to rename two variables that correspond to the course and the student so that we can match with the other variables we just created for the pre-survey data.

course\_data <-  
 course\_data %>%  
 rename(student\_id = Bb\_UserPK,  
 course\_id = CourseSectionOrigID)

Now that we have two variables that are consistent across both datasets - we have called them course\_id and student\_id - we can join them using the {dplyr} function, left\_join().

Let’s save our joined data as a new object called dat.

dat <-  
 left\_join(course\_data, pre\_survey,  
 by = c("student\_id", "course\_id"))  
dat

## # A tibble: 40,348 x 21  
## course\_id subject semester section student\_id Gradebook\_Item Grade\_Category  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 AnPhA-S1… AnPhA S116 01 60186 POINTS EARNED… <NA>   
## 2 AnPhA-S1… AnPhA S116 01 60186 WORK ATTEMPTED <NA>   
## 3 AnPhA-S1… AnPhA S116 01 60186 0.1: Message … <NA>   
## 4 AnPhA-S1… AnPhA S116 01 60186 0.2: Intro As… Hw   
## 5 AnPhA-S1… AnPhA S116 01 60186 0.3: Intro As… Hw   
## 6 AnPhA-S1… AnPhA S116 01 60186 1.1: Quiz Qz   
## 7 AnPhA-S1… AnPhA S116 01 60186 1.2: Quiz Qz   
## 8 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 9 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 10 AnPhA-S1… AnPhA S116 01 60186 1.4: Negative… Hw   
## # … with 40,338 more rows, and 14 more variables: FinalGradeCEMS <dbl>,  
## # Points\_Possible <dbl>, Points\_Earned <dbl>, Gender <chr>, q1 <dbl>,  
## # q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>, q7 <dbl>, q8 <dbl>,  
## # q9 <dbl>, q10 <dbl>

left\_join() is named based on the ‘direction’ that the data is being joined. Note the order of the data frames passed to our “left” join. Left joins retain all of the rows in the data frame on the “left”, and joins every matching row in the right data frame to it.

Let’s hone in on how this code is structured. After left\_join(), we see course\_data and then pre\_survey. In this case, course\_data is the “left” data frame (passed as the *first* argument), while pre\_survey is the “right” data frame (passed as the *second* argument). So, in the above, what happens? You can run the code yourself to check.

What our aim - and what should happen - is that all of the rows in course\_data are retained in our new data frame, dat, with matching rows of pre\_survey joined to it. An important note is that there are not multiple matching rows of pre\_survey; otherwise, you would end up with more rows in dat than expected. There is a lot packed into this one function. Joins are, however, extremely powerful - and common - in many data analysis processing pipelines, both in education and in any field. Think of all of the times you have data in more than one data frame, and want them to be in a single data frame! As a result, we think that joins are well worth investing the time to be able to use.

With education (and other) data, left\_join() is helpful for carrying out most tasks related to joining datasets. However, there are functions for other types of joins. They may be less important than left\_join() but are still worth mentioning (note that for all of these, the “left” data frame is always the first argument, and the “right” data frame is always the second):

#### semi\_join()

semi\_join(): joins and retains all of the *matching* rows in the “left” and “right” data frame; it is useful when you are only interested in keeping the rows (or cases/observations) that are able to be joined. semi\_join() will not create duplicate rows of the left data frame, even when it finds multiple matches on the right data frame. It will also keep only the columns from the left data frame.

For example, the following returns only the rows that are present in both course\_data and pre\_survey:

dat\_semi <-   
 semi\_join(course\_data,  
 pre\_survey,  
 by = c("student\_id", "course\_id"))  
  
dat\_semi

## # A tibble: 28,655 x 11  
## course\_id subject semester section student\_id Gradebook\_Item Grade\_Category  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 AnPhA-S1… AnPhA S116 01 60186 POINTS EARNED… <NA>   
## 2 AnPhA-S1… AnPhA S116 01 60186 WORK ATTEMPTED <NA>   
## 3 AnPhA-S1… AnPhA S116 01 60186 0.1: Message … <NA>   
## 4 AnPhA-S1… AnPhA S116 01 60186 0.2: Intro As… Hw   
## 5 AnPhA-S1… AnPhA S116 01 60186 0.3: Intro As… Hw   
## 6 AnPhA-S1… AnPhA S116 01 60186 1.1: Quiz Qz   
## 7 AnPhA-S1… AnPhA S116 01 60186 1.2: Quiz Qz   
## 8 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 9 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 10 AnPhA-S1… AnPhA S116 01 60186 1.4: Negative… Hw   
## # … with 28,645 more rows, and 4 more variables: FinalGradeCEMS <dbl>,  
## # Points\_Possible <dbl>, Points\_Earned <dbl>, Gender <chr>

#### anti\_join()

anti\_join(): *removes* all of the rows in the “left” data frame that can be joined with those in the “right” data frame.

dat\_anti <-  
 anti\_join(course\_data,  
 pre\_survey,  
 by = c("student\_id", "course\_id"))  
  
dat\_anti

## # A tibble: 1,056 x 11  
## course\_id subject semester section student\_id Gradebook\_Item Grade\_Category  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 AnPhA-S1… AnPhA S116 01 85865 POINTS EARNED… <NA>   
## 2 AnPhA-S1… AnPhA S116 01 85865 WORK ATTEMPTED <NA>   
## 3 AnPhA-S1… AnPhA S116 01 85865 0.1: Message … <NA>   
## 4 AnPhA-S1… AnPhA S116 01 85865 0.2: Intro As… Hw   
## 5 AnPhA-S1… AnPhA S116 01 85865 0.3: Intro As… Hw   
## 6 AnPhA-S1… AnPhA S116 01 85865 1.1: Quiz Qz   
## 7 AnPhA-S1… AnPhA S116 01 85865 1.2: Quiz Qz   
## 8 AnPhA-S1… AnPhA S116 01 85865 1.3: Create a… Hw   
## 9 AnPhA-S1… AnPhA S116 01 85865 1.3: Create a… Hw   
## 10 AnPhA-S1… AnPhA S116 01 85865 1.4: Negative… Hw   
## # … with 1,046 more rows, and 4 more variables: FinalGradeCEMS <dbl>,  
## # Points\_Possible <dbl>, Points\_Earned <dbl>, Gender <chr>

#### right\_join()

right\_join(): perhaps the least helpful of the three, right\_join() works the same as left\_join(), but by retaining all of the rows in the “right” data frame, and joining matching rows in the “left” data frame (so, the opposite of left\_join()).

dat\_right <-  
 right\_join(course\_data,  
 pre\_survey,  
 by = c("student\_id", "course\_id"))  
  
dat\_right

## # A tibble: 39,593 x 21  
## course\_id subject semester section student\_id Gradebook\_Item Grade\_Category  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 FrScA-S1… <NA> <NA> <NA> 80624 <NA> <NA>   
## 2 BioA-S11… <NA> <NA> <NA> 80623 <NA> <NA>   
## 3 OcnA-S11… <NA> <NA> <NA> 82588 <NA> <NA>   
## 4 AnPhA-S1… <NA> <NA> <NA> 80623 <NA> <NA>   
## 5 AnPhA-S1… <NA> <NA> <NA> 80624 <NA> <NA>   
## 6 AnPhA-S1… <NA> <NA> <NA> 80624 <NA> <NA>   
## 7 AnPhA-T1… <NA> <NA> <NA> 80624 <NA> <NA>   
## 8 BioA-S11… <NA> <NA> <NA> 80624 <NA> <NA>   
## 9 BioA-T11… <NA> <NA> <NA> 80624 <NA> <NA>   
## 10 PhysA-S1… <NA> <NA> <NA> 80624 <NA> <NA>   
## # … with 39,583 more rows, and 14 more variables: FinalGradeCEMS <dbl>,  
## # Points\_Possible <dbl>, Points\_Earned <dbl>, Gender <chr>, q1 <dbl>,  
## # q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>, q7 <dbl>, q8 <dbl>,  
## # q9 <dbl>, q10 <dbl>

If we wanted this to return exactly the same output as left\_join() (and so to create a data frame that is identical to the dat data frame above), we could simply switch the order of the two data frames to be the opposite of those used for the left\_join() above:

dat\_right <-  
 semi\_join(pre\_survey,  
 course\_data,  
 by = c("student\_id", "course\_id"))  
  
dat\_right

## # A tibble: 801 x 12  
## student\_id course\_id q1 q2 q3 q4 q5 q6 q7 q8 q9  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 85791 FrScA-S1… 3 3 3 3 4 3 3 3 2  
## 2 87010 FrScA-S1… 5 3 3 5 4 4 3 5 2  
## 3 87027 FrScA-S1… 5 5 4 4 4 5 4 5 4  
## 4 85649 FrScA-S1… NA NA NA NA NA NA NA NA NA  
## 5 86216 BioA-S11… 5 3 4 4 5 4 3 5 4  
## 6 68476 OcnA-S11… 4 4 2 2 3 4 4 4 4  
## 7 68476 OcnA-S11… 4 4 4 4 4 4 4 4 4  
## 8 87866 FrScA-T1… 5 4 3 5 4 4 3 3 3  
## 9 64930 FrScA-T1… 4 3 3 4 4 4 3 4 4  
## 10 86280 FrScA-S1… 4 3 2 4 4 4 4 4 3  
## # … with 791 more rows, and 1 more variable: q10 <dbl>

Just one more data frame to merge:

course\_minutes <-  
 course\_minutes %>%  
 rename(student\_id = Bb\_UserPK,  
 course\_id = CourseSectionOrigID)  
  
course\_minutes <-  
 course\_minutes %>%  
 # Change the data type for student\_id in course\_minutes so we can match to   
 # student\_id in dat  
 mutate(student\_id = as.integer(student\_id))  
  
dat <-   
 dat %>%   
 left\_join(course\_minutes,   
 by = c("student\_id", "course\_id"))

Note that they’re now combined, even though the course data has many more rows: The pre\_survey data has been joined for each student by course combination.

We have a pretty large data frame! Let’s take a quick look.

dat

## # A tibble: 40,348 x 22  
## course\_id subject semester section student\_id Gradebook\_Item Grade\_Category  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 AnPhA-S1… AnPhA S116 01 60186 POINTS EARNED… <NA>   
## 2 AnPhA-S1… AnPhA S116 01 60186 WORK ATTEMPTED <NA>   
## 3 AnPhA-S1… AnPhA S116 01 60186 0.1: Message … <NA>   
## 4 AnPhA-S1… AnPhA S116 01 60186 0.2: Intro As… Hw   
## 5 AnPhA-S1… AnPhA S116 01 60186 0.3: Intro As… Hw   
## 6 AnPhA-S1… AnPhA S116 01 60186 1.1: Quiz Qz   
## 7 AnPhA-S1… AnPhA S116 01 60186 1.2: Quiz Qz   
## 8 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 9 AnPhA-S1… AnPhA S116 01 60186 1.3: Create a… Hw   
## 10 AnPhA-S1… AnPhA S116 01 60186 1.4: Negative… Hw   
## # … with 40,338 more rows, and 15 more variables: FinalGradeCEMS <dbl>,  
## # Points\_Possible <dbl>, Points\_Earned <dbl>, Gender <chr>, q1 <dbl>,  
## # q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>, q7 <dbl>, q8 <dbl>,  
## # q9 <dbl>, q10 <dbl>, TimeSpent <dbl>

It looks like we have 40348 observations from 30 variables.

There is one last step to take. If we were interested in a fine-grained analysis of how students performed (according to the teacher) on different assignments (see the Gradebook\_Item column), we would keep all rows of the data. But, our goal (for now) is more modest: to calculate the percentage of points students earned as a measure of their final grade (noting that the teacher may have assigned a different grade - or weighted their grades in ways not reflected through the points).

dat <-  
 dat %>%  
 group\_by(student\_id, course\_id) %>%  
 mutate(Points\_Earned = as.integer(Points\_Earned)) %>%  
 summarize(  
 total\_points\_possible = sum(Points\_Possible, na.rm = TRUE),  
 total\_points\_earned = sum(Points\_Earned, na.rm = TRUE)  
 ) %>%  
 mutate(percentage\_earned = total\_points\_earned / total\_points\_possible) %>%  
 ungroup() %>%  
 # note that we join this back to the original data frame to retain all of the variables  
 left\_join(dat)

## Joining, by = c("student\_id", "course\_id")

### Finding Distinct Cases at the Student-Level

This last step calculated a new column for the percentage of points each student earned. That value is the same for the same student (an easy way we would potentially use to check this is View(), i.e., View(dat)). But, because we are not carrying out a finer-grained analysis using the Gradebook\_Item, the duplicate rows are not necessary. We only want variables at the student-level (and not at the level of different gradebook items). We can do this using the distinct() function. This function takes the name of the data frame and the name of the variables used to determine what counts as a unique case.

Imagine having a bucket of Halloween candy that has 100 pieces of candy. You know that these 100 pieces are really just a bunch of duplicate pieces from a relatively short list of candy brands. distinct() takes that bucket of 100 pieces and returns a bucket containing only one of each distinct piece. Another thing to note about distinct() is that it will only return the variable(s) (we note that you can pass more than one variable to distinct()) you used to determine uniqueness, *unless* you include the argument .keep\_all = TRUE. For the sake of making it very easy to view the output, we omit this argument (only for now).

Were we to run distinct(dat, Gradebook\_Item), what do you think would be returned?

distinct(dat, Gradebook\_Item)

## # A tibble: 222 x 1  
## Gradebook\_Item   
## <chr>   
## 1 POINTS EARNED & TOTAL COURSE POINTS   
## 2 WORK ATTEMPTED   
## 3 0-1.1: Intro Assignment - Send a Message to Your Instructor  
## 4 0-1.2: Intro Assignment - DB #1   
## 5 0-1.3: Intro Assignment - Submitting Files   
## 6 1-1.1: Lesson 1-1 Graphic Organizer   
## 7 1-2.1: Explore a Career Assignment   
## 8 1-2.2: Explore a Career DB #2   
## 9 PROGRESS CHECK 1 @ 02-18-16   
## 10 1-2.3: Lesson 1-2 Graphic Organizer   
## # … with 212 more rows

What is every distinct gradebook item is what is returned. You might be wondering (as we were) whether some gradebook items have the same values across courses; we can return the unique *combination* of courses and gradebook items by simply adding another variable to distinct():

distinct(dat, course\_id, Gradebook\_Item)

## # A tibble: 1,269 x 2  
## course\_id Gradebook\_Item   
## <chr> <chr>   
## 1 FrScA-S216-02 POINTS EARNED & TOTAL COURSE POINTS   
## 2 FrScA-S216-02 WORK ATTEMPTED   
## 3 FrScA-S216-02 0-1.1: Intro Assignment - Send a Message to Your Instructor  
## 4 FrScA-S216-02 0-1.2: Intro Assignment - DB #1   
## 5 FrScA-S216-02 0-1.3: Intro Assignment - Submitting Files   
## 6 FrScA-S216-02 1-1.1: Lesson 1-1 Graphic Organizer   
## 7 FrScA-S216-02 1-2.1: Explore a Career Assignment   
## 8 FrScA-S216-02 1-2.2: Explore a Career DB #2   
## 9 FrScA-S216-02 PROGRESS CHECK 1 @ 02-18-16   
## 10 FrScA-S216-02 1-2.3: Lesson 1-2 Graphic Organizer   
## # … with 1,259 more rows

It looks like *a lot* of gradebook items were repeated - likely across the different sections of the same course (we would be curious to hear what you find if you investigate this!).

Let’s use what we just did, but to find the unique values at the student-level. Thus, instead of exploring unique gradebook items, we will explore unique students (still accounting for the course, as students could enroll in more than one course.) This time, we will add the keep\_all = TRUE argument.

dat <-  
 distinct(dat, course\_id, student\_id, .keep\_all = TRUE)

This is a much smaller data frame - with one row for each student in the course (instead of the 29,701 rows which we would be interested in were we analyzing this data at the level of specific students’ grades for specific gradebook items). Now that our data are ready to go, we can start to ask some questions of the data,

## Analysis

In this section, we focus on some initial analyses in the form of visualizations and some models. We expand on these in [Chapter 13](#c13). Before we start visualizing relationships between variables in our survey dataset, let’s introduce {ggplot2}, a visualization package we’ll be using in our walkthroughs.

### About {ggplot2}

{ggplot2} is a package we’ll be using a lot for graphing our education datasets. {ggplot2} is designed to build graphs layer by layer, where each layer is a building block for your graph. Making graphs in layers is useful because we can think of building up our graphs in separate parts–the data comes first, then the x- and y-axis, then other components like text labels and graph shapes. When something goes wrong and your ggplot2 code returns an error, you can learn about what’s happening by removing one layer at a time and running it again until the code works properly. Once you know which line is causing the problem, you can focus on fixing it.

The first two lines of most {ggplot2} code look similar in most graphs. The first line tells R which dataset to graph and which columns the x-axis and y-axis will represent. The second line tells R which shape to use when drawing the graph. You can tell R which shape to use in your graphs with a family of {ggplot2} functions that start with geom\_. {ggplot2} has many graph shapes you can use, including points, bars, lines, and boxplots. Here’s a {ggplot2} example using a dataset of school mean test scores to graph a bar chart:

# make dataset  
students <-   
 tibble(  
 school\_id = c("a", "b", "c"),   
 mean\_score = c(10, 20, 30)  
 )  
  
# tell R which dataset to plot and which columns the x-axis and y-axis will represent  
ggplot(data = students, aes(x = school\_id, y = mean\_score)) +   
 # draw the plot  
 geom\_bar(stat = "identity",  
 fill = dataedu\_colors("darkblue")) +  
 theme\_dataedu()

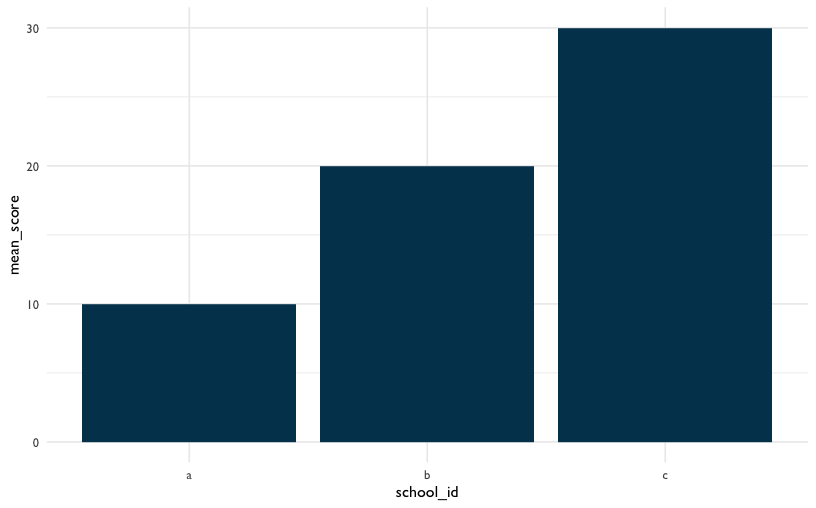


Figure 13: Example Plot

The data argument in the first line tells R we’ll be using the dataset called students. The aes argument tells R we’ll be using values from the school\_id column for the x-axis and values from the mean\_score column for the y-axis. In the second line, the geom\_bar function tells R we’ll drawing the graph using the bar chart format. Each line of ggplot code is connected by a + at the end to tell R the next line of code is an additional ggplot layer to add.

### The Relationship between Time Spent on Course and Percentage of Points Earned

One thing we might be wondering is how time spent on course is related to students’ final grade.

dat %>%  
 # aes() tells ggplot2 what variables to map to what feature of a plot  
 # Here we map variables to the x- and y-axis  
 ggplot(aes(x = TimeSpent, y = percentage\_earned)) +   
 # Creates a point with x- and y-axis coordinates specified above  
 geom\_point(color = dataedu\_colors("green")) +   
 theme\_dataedu() +  
 xlab("Time Spent") +  
 ylab("Percentage Earned")



Figure 14: Percentage Earned vs. Time Spent

There appears to be *some* relationship. What if we added a line of best fit - a linear model?

dat %>%  
 ggplot(aes(x = TimeSpent, y = percentage\_earned)) +  
 geom\_point(color = dataedu\_colors("green")) + # same as above  
 # this adds a line of best fit  
 # method = "lm" tells ggplot2 to fit the line using linear regression  
 geom\_smooth(method = "lm") +  
 theme\_dataedu() +  
 xlab("Time Spent") +  
 ylab("Percentage Earned")

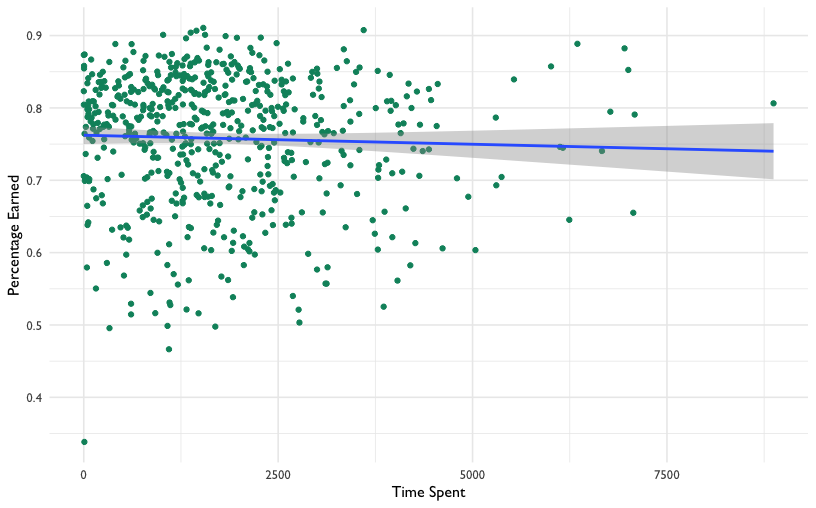


Figure 15: Adding a Line of Best Fit

So, it appears that the more time students spent on the course, the more points they earned.

### Linear Model (Regression)

We can find out exactly what the relationship is using a linear model. We also discuss linear models in [Chapter 10](#c10).

Let’s use this technique to model the relationship between the time spent on the course and the percentage of points earned. Here, we predict percentage\_earned, or the percentage of the total points that are possible for a student to earn. Percentage earned is the dependent, or *y*-variable, and so we enter it first, after the lm() command and before the tilde (~) symbol. To the right of the tilde is one independent variable, TimeSpent, or the time that students spent on the course. We also pass, or provide, the data frame, dat. At this point, we’re ready to run the model. Let’s run this line of code and save the results to an object - we chose m\_linear, but any name will work, as well as the summary() function on the output.

m\_linear <-  
 lm(percentage\_earned ~ TimeSpent, data = dat)

Another way that we can generate table output is with a function from the {sjPlot} package, tab\_model().

tab\_model(m\_linear,  
 title = "Table 7.1")

Table 7.1

percentage earned

Predictors

Estimates

CI

p

(Intercept)

0.76

0.75 – 0.77

<0.001

TimeSpent

-0.00

-0.00 – 0.00

0.364

Observations

598

R2 / R2 adjusted

0.001 / -0.000

This will work well for R Markdown documents (or simply to interpret the model in R). If you want to save the model for use in a Word document, the [{apaTables}](https://cran.r-project.org/web/packages/apaTables/vignettes/apaTables.html) (https://cran.r-project.org/web/packages/apaTables/vignettes/apaTables.html) package may be helpful. To do so, just pass the name of the regression model, like we did with tab\_model(). Then, you can save the output to a Word document, simply by adding a filename argument:

apa.reg.table(m\_linear, filename = "regression-table-output.doc")

We encourage you to read more about the package at <https://cran.r-project.org/web/packages/apaTables/index.html> to see what else the {apaTables} package does. The vignette is especially helpful.

Before we proceed to the next code chunk, let’s talk about some functions we’ll be using a lot in this book. filter(), group\_by(), and summarize() are functions in the {dplyr} package that you will see a lot in upcoming chapters.

* filter() removes rows from the dataset that don’t match a criteria. Use it for tasks like only keeping records for students in the fifth grade
* group\_by() groups records together so you can perform operations on those groups instead of on the entire dataset. Use it for tasks like getting the mean test score of each school instead of a whole school district
* summarize() and summarise() reduce your dataset down to a summary statistic. Use it for tasks like turning a dataset of student test scores into a dataset of grade levels and their mean test score

So let’s use these {dplyr} functions on our survey analysis. We will create the same measures (based on the survey items) that we used earlier to understand how they relate to one another:

survey\_responses <-  
 pre\_survey %>%  
 # Gather questions and responses  
 pivot\_longer(cols = q1:q10,  
 names\_to = "question",  
 values\_to = "response") %>%  
 mutate(  
 # Here's where we make the column of question categories  
 measure = case\_when(  
 question %in% c("q1", "q4", "q5", "q8", "q10") ~ "int",  
 question %in% c("q2", "q6", "q9") ~ "uv",  
 question %in% c("q3", "q7") ~ "pc",  
 TRUE ~ NA\_character\_  
 )  
 ) %>%  
 group\_by(student\_id, measure) %>%  
 # Here's where we compute the mean of the responses  
 summarize(  
 # Mean response for each measure  
 mean\_response = mean(response, na.rm = TRUE)  
 ) %>%  
 filter(!is.na(mean\_response)) %>%  
 pivot\_wider(names\_from = measure,   
 values\_from = mean\_response)  
  
survey\_responses

## # A tibble: 515 x 4  
## # Groups: student\_id [515]  
## student\_id int pc uv  
## <dbl> <dbl> <dbl> <dbl>  
## 1 43146 5 4.5 4.33  
## 2 44638 4.2 3.5 4   
## 3 47448 5 4 3.67  
## 4 47979 5 3.5 5   
## 5 48797 3.8 3.5 3.5   
## 6 49147 4.25 3.73 3.71  
## 7 51943 4.6 4 4   
## 8 52326 5 3.5 5   
## 9 52446 3 3 3.33  
## 10 53248 4 3 3.33  
## # … with 505 more rows

Now that we’ve prepared the survey responses, we can use the apa.cor.table() function:

survey\_responses %>%   
 apa.cor.table()

##   
##   
## Means, standard deviations, and correlations with confidence intervals  
##   
##   
## Variable M SD 1 2 3   
## 1. student\_id 85966.07 10809.12   
##   
## 2. int 4.22 0.59 .00   
## [-.08, .09]   
##   
## 3. pc 3.60 0.64 .04 .59\*\*   
## [-.05, .13] [.53, .64]   
##   
## 4. uv 3.71 0.71 .02 .57\*\* .50\*\*   
## [-.06, .11] [.51, .62] [.43, .56]  
##   
##   
## Note. M and SD are used to represent mean and standard deviation, respectively.  
## Values in square brackets indicate the 95% confidence interval.  
## The confidence interval is a plausible range of population correlations   
## that could have caused the sample correlation (Cumming, 2014).  
## \* indicates p < .05. \*\* indicates p < .01.  
##

The time spent variable is on a very large scale (minutes); what if we transform it to represent the number of hours that students spent on the course? Let’s use the mutate() function we used earlier. We’ll end the variable name in \_hours, to represent what this variable means.

# creating a new variable for the amount of time spent in hours  
dat <-   
 dat %>%   
 mutate(TimeSpent\_hours = TimeSpent / 60)  
  
# the same linear model as above, but with the TimeSpent variable in hours  
m\_linear\_1 <-   
 lm(percentage\_earned ~ TimeSpent\_hours, data = dat)  
  
# viewing the output of the linear model  
tab\_model(m\_linear\_1,  
 title = "Table 7.2")

Table 7.2

percentage earned

Predictors

Estimates

CI

p

(Intercept)

0.76

0.75 – 0.77

<0.001

TimeSpent\_hours

-0.00

-0.00 – 0.00

0.364

Observations

598

R2 / R2 adjusted

0.001 / -0.000

The scale still does not seem quite right. What if we standardized the variable to have a mean of zero and a standard deviation of one?

# this is to standardize the TimeSpent variable to have a mean of zero and a standard deviation of 1  
dat <-   
 dat %>%   
 mutate(TimeSpent\_std = scale(TimeSpent))  
  
# the same linear model as above, but with the TimeSpent variable standardized  
m\_linear\_2 <-   
 lm(percentage\_earned ~ TimeSpent\_std, data = dat)  
  
# viewing the output of the linear model  
tab\_model(m\_linear\_2,  
 title = "Table 7.3")

Table 7.3

percentage earned

Predictors

Estimates

CI

p

(Intercept)

0.76

0.75 – 0.77

<0.001

TimeSpent\_std

-0.00

-0.01 – 0.00

0.364

Observations

598

R2 / R2 adjusted

0.001 / -0.000

That seems to make more sense. However, there is a different interpretation now for the time spent variable: for every one standard deviation increase in the amount of time spent on the course, the percentage of points a student earns increases by .11, or 11 percentage points.

## Results

Let’s extend our regression model and consider the following to be the final model in this sequence of models: What other variables may matter? Perhaps there are differences based on the subject of the course. We can add subject as a variable easily, as follows:

# a linear model with the subject added   
# independent variables, such as TimeSpent\_std and subject, can simply be separated with a plus symbol:  
m\_linear\_3 <-   
 lm(percentage\_earned ~ TimeSpent\_std + subject, data = dat)

We can use tab\_model() once again to view the results:

tab\_model(m\_linear\_3,  
 title = "Table 7.4")

Table 7.4

percentage earned

Predictors

Estimates

CI

p

(Intercept)

0.75

0.74 – 0.77

<0.001

TimeSpent\_std

-0.00

-0.01 – 0.00

0.346

subject [BioA]

-0.00

-0.03 – 0.03

0.872

subject [FrScA]

0.00

-0.02 – 0.02

0.766

subject [OcnA]

0.02

-0.00 – 0.04

0.094

subject [PhysA]

0.00

-0.03 – 0.03

0.900

Observations

598

R2 / R2 adjusted

0.008 / -0.001

It looks like subject FrSc - forensic science - and subject Ocn - oceanography - are associated with a higher percentage of points earned, overall. This indicates that students in those two classes earned higher grades than students in other science classes in this dataset.

## Conclusion

In this walkthrough, we focused on taking unprocessed, or raw data, and loading, viewing, and then processing it through a series of steps. The result was a data set which we could use to create visualizations and a simple (but powerful!) model, a linear model, also known as a regression model. We found that the time that students spent on the course was positively (and statistically significantly) related to students’ final grades, and that there appeared to be differences by subject. While we focused on using this model in a traditional, explanatory sense, it could also (potentially) be used for predictive analytics, in that knowing how long students spent on the course and what subject their course is could be used to estimate what that students’ final grade might be. We focus on uses of predictive models further in [Chapter 14](#c14).

In the follow-up to this walkthrough (see [Chapter 13](#c13)), we will focus on visualizing and then modeling the data using an advanced methodological technique, multi-level models, using the data we prepared as a part of this data processing pipeline used in this chapter.

# Walkthrough 2: Approaching Gradebook Data From a Data Science Perspective

## Vocabulary

* correlation
* directory
* environment
* linear model
* linearity
* missing values/ NA
* outliers
* pivot
* string

## Chapter Overview

Whereas [Walkthrough 1/Chapter 7](#c07) focused on the the education data science pipeline in the context of online science class data, this walkthrough explores a different dataset - ubiquitous but not-often-analyzed using data science tools and techniques classroom gradebook dataset - and focuses more on analyses, including correlations and linear models.

There are a variety of data sources to explore in the education field. Student assessment scores can be examined for progress towards goals. The text from a teacher’s written classroom observation notes about a particular learner’s in-class behavior or emotional status can be analyzed for trends. We can tap into the exportable data available from common learning software or platforms popular in the K-12 education space.

### Background

This walkthrough goes through a series of analyses using the data science framework. The first analysis centers around a common K-12 classroom tool: the gradebook. While gradebook data is nearly ubiquitous in education, it is sometimes ignored in favor of data collected by evaluators and researchers or data from state-wide tests. Nevertheless, it represents an important, occasionally untapped data source, and one for which a data science approach can reveal the potential of the data science framework for analyzing a range of educational data sources.

### Data Sources

We use an Excel gradebook template, [*Assessment Types - Points*](https://web.mit.edu/jabbott/www/excelgradetracker.html) (https://web.mit.edu/jabbott/www/excelgradetracker.html), coupled with simulated student data. On your first read through this section try using our simulated dataset found in this book’s data/ folder.

### Methods

This analysis uses a linear model, which relates one or more X, or independent variables, to a Y, or dependent variable, and a correlation analysis.

## Load Packages

As mentioned in the Foundational Skills chapter, begin by loading the libraries that will be used. We will load the {tidyverse} package used in [Walkthrough 1](#c07). This chapter has an example of using the {readxl} package to read and import Excel spreadsheets, file types are very common in the education field. We will also use the {janitor} package, which provides a number of functions related to cleaning and preparing data, for the first time (Firke [2020](#ref-R-janitor)).

Make sure you have installed the packages in R on your computer before starting (see Foundational Skills chapter). Load the libraries, including as they must be loaded each time we start a new project.

# Load libraries  
library(tidyverse)  
library(here)  
library(readxl)  
library(janitor)  
library(dataedu)

## Import Data

Recall how the Foundational Skills chapter recommended favoring .csv files, or comma-separated values files, when working with datasets in R? This is because .csv files, with the .csv file extension, are common in the digital world. However, data won’t always come in the preferred file format. Fortunately, R can import a variety of data file types. This walkthrough imports an Excel file because these file types, with the .xlsx or .xls extensions, are very likely to be encountered in the K-12 education world. We’ll show you two ways to import the gradebook dataset: The first is uses a file path. The second uses the here() function from the {here} package. We recommend using here(), but it’s worthwhile to review both methods.

### Import Using a File Path

First let’s look at importing the dataset using a file path. This code uses the read\_excel() function of the {readxl} package to find and read the data of the desired file. Note the file path that read\_excel() takes to find the simulated dataset file named ExcelGradeBook.xlsx which sits in a folder on your computer if you have downloaded it. The function getwd() will help locate your current working directory. This tells where on the computer R is currently working with files.

# See the current working directory  
getwd()

For example, an R user on Linux or Mac might see their working directory as: /home/username/Desktop. A Windows user might see their working directory as: C:\Users\Username\Desktop.

From this location, go deeper into files to find the desired file. For example, if you downloaded the [book repository](https://github.com/data-edu/data-science-in-education) (https://github.com/data-edu/data-science-in-education) from Github to your Desktop the path to the Excel file might look like one of these below:

* /home/username/Desktop/data-science-in-education/data/gradebooks/ExcelGradeBook.xlsx (on Linux & Mac)
* C:\Users\Username\Desktop\data-science-in-education\data\gradebooks\ExcelGradeBook.xlsx (on Windows)

After locating the sample Excel file, use the code below to run the function read\_excel() which reads and saves the data from ExcelGradeBook.xlsx to an object also called ExcelGradeBook. Note the two arguments specified in this code sheet = 1 and skip = 10. This Excel file is similar to one you might encounter in real life with superfluous features that we are not interested in. This file has 3 different sheets and the first 10 rows contain things we won’t need. Thus, sheet = 1 tells read\_excel() to just read the first sheet in the file and disregard the rest. While skip = 10 tells read\_excel() to skip reading the first 10 rows of the sheet and start reading from row 11 which is where the column headers and data actually start inside the Excel file. Remember to replace path/to/file.xlsx your own path to the file you want to import.

ExcelGradeBook <- read\_excel("path/to/file.xlsx", sheet = 1, skip = 10)

### Import Using here()

Whenever possible, we prefer to use here() from the {here} package because it conveniently guesses at the correct file path based on the working directory. In your working directory, place the ExcelGradeBook.xlsx file in a folder called gradebooks. Then place the gradebooks folder in a folder called data. The last step is to make sure your new data folder and all its contents are in your working directory. Following those steps, use this code to read the data in:

# Use readxl package to read and import file and assign it a name  
ExcelGradeBook <-  
 read\_excel(  
 here("data", "gradebooks", "ExcelGradeBook.xlsx"),  
 sheet = 1,  
 skip = 10  
 )

The ExcelGradeBook file has been imported into RStudio. Next, assign the data frame to a new name using the code below. Renaming cumbersome filenames can improve the readability of the code and make is easier for the user to call on the dataset later on in the code.

# Rename data frame  
gradebook <- ExcelGradeBook

Your environment will now have two versions of the dataset. There is ExcelGradeBook which was the original dataset imported. There is also gradebook which is currently a copy of ExcelGradeBook. As you progress through this section, we will work primarily with the gradebook dataset. Additionally, while working onward in this section of the book, if you make a mistake and mess up the gradebook data frame and are not able to fix it, you can reset the data frame to return to the same state as the original ExcelGradeBook data frame by running gradebook <- ExcelGradeBook again. This will overwrite any errors in the gradebook data frame with the originally imported ExcelGradeBook data frame. Afterwards, just continue running code from this point in the text.

## Process Data

### Tidy Data

This walkthrough uses an Excel data file because it is one that we are likely to encounter. Moreover, the messy state of this file mirrors what might be encountered in real life. The Excel file contains more than one sheet, has rows we don’t need, and uses column names that have spaces between words. All these things make the data tough to work with. The data is **not** tidy. We can begin to overcome these challenges before importing the file into RStudio by deleting the unnecessary parts of the Excel file then saving it as a .csv file. However, if you clean the file outside of R, this means if you ever have to clean it up (say, if the dataset is accidentally deleted and you need to re-download it from the original source) you would have to do everything from the beginning, and may not recall exactly what you did in Excel prior to importing the data to R. We recommend cleaning the original data in R so that you can recreate all the steps necessary for your analysis. Also, the untidy Excel file provides realistic practice for tidying up the data programmatically (using a computer program) with R itself (instead of doing these steps manually).

First, modify the column names of the gradebook data frame to remove any spaces and replace them with an underscore. Using spaces as column names in R can present difficulties later on when working with the data.

Second, we want the column names of our data to be easy to use and understand. The original dataset has column names with uppercase letters and spaces. We can use the {janitor} package to quickly change them to a more usable format.

### About {janitor}

The {janitor} package is a great resource for anybody who works with data, and particularly fantastic for data scientists in education. Created by Sam Firke, the Analytics Director of The New Teacher Project, it is a package created by a practitioner in education with education data in mind.

{janitor} has many handy functions to clean and tabulate data. Some examples include:

* clean\_names(), which takes messy column names that have periods, capitalized letters, spaces, etc. into R-friendly column names
* get\_dupes(), which identifies and examines duplicate records
* tabyl(), which tabulates data in a data.frame format, and can be ‘adorned’ with the adorn\_ functions to add total rows, percentages, and other dressings

Let’s use {janitor} with these data!

First, let’s have a look at the original column names. The output will be long, so let’s just look at the first ten by using head().

# look at original column names  
head(colnames(gradebook))

## [1] "Class" "Name" "Race" "Gender"   
## [5] "Age" "Repeated Grades"

You can look at the full output by removing the call to head().

# look at original column names  
colnames(gradebook)

Now let’s look at the cleaned names:

gradebook <-   
 gradebook %>%   
 clean\_names()  
  
# look at cleaned column names  
head(colnames(gradebook))

## [1] "class" "name" "race" "gender"   
## [5] "age" "repeated\_grades"

Review what the gradebook data frame looks like now. It shows 25 students and their individual values in various columns like projects or formative\_assessments.

view(gradebook)

The data frame looks cleaner now but there still are some things we can remove. For example, there are rows without any names in them. Also, there are entire columns that are unused and contain no data (such as gender). These are called missing values and are denoted by *NA*. Since our simulated classroom only has 25 learners and doesn’t use all the columns for demographic information, we can safely remove these to tidy our dataset up even more.

We can remove the extra columns rows that have no data using the {janitor} package. The handy remove\_empty() removes columns, rows, or both that have no information in them.

# Removing rows with nothing but missing data  
gradebook <-   
 gradebook %>%   
 remove\_empty(c("rows", "cols"))

Now that the empty rows and columns have been removed, notice there are two columns, absent and late, where it seems someone started putting data into but then decided to stop. These two columns didn’t get removed by the last chunk of code because they technically contained some data in those columns. Since the simulated enterer of this simulated class data decided to abandon using the absent and late columns in this gradebook, we can remove it from our data frame as well.

In the Foundational Skills chapter, we introduced the select() function, which tells R which columns we want to keep. Let’s do that again here. This time we’ll use negative signs to say we want the dataset without absent and late.

# Remove a targeted column because we don't use absent and late at this school.  
gradebook <-   
 gradebook %>%   
 select(-absent, -late)

At last, the formerly untidy Excel sheet has been turned into a useful data frame. Inspect it once more to see the difference.

view(gradebook)

### Create new variables and further process the data

R users transform data to facilitate working with the data during later phases of visualization and analysis. A few examples of data transformation include creating new variables, grouping data, and more. This code chunk first creates a new data frame named classwork\_df, then selects particular variables from our gradebook dataset using select(), and finally gathers all the homework data under new variables into new columns.

As mentioned previously, select() is very powerful. In addition to explicitly writing out the columns you want to keep, you can also use functions from the package {stringr} within select(). The {stringr} package is contained within the {tidyverse} meta-package. Here, we’ll use the function contains() to tell R to select columns that contain a certain string (that is, text). Here, it searches for any column with the string classwork\_. The underscore makes sure the variables from classwork\_1 all the way to classwork\_17 are included in classwork\_df.

pivot\_longer() transforms the dataset into tidy data.

Note that scores are in character format. We use mutate() to transform them to numeric.

# Creates new data frame, selects desired variables from gradebook, and gathers all classwork scores into key/value pairs  
classwork\_df <-  
 gradebook %>%  
 select(  
 name,  
 running\_average,  
 letter\_grade,  
 homeworks,  
 classworks,  
 formative\_assessments,  
 projects,  
 summative\_assessments,  
 contains("classwork\_")) %>%  
 mutate\_at(vars(contains("classwork\_")), list(~ as.numeric(.))) %>%  
 pivot\_longer(  
 cols = contains("classwork\_"),  
 names\_to = "classwork\_number",  
 values\_to = "score"  
 )

View the new data frame and note which columns were selected for this new data frame. Also, note how all the classwork scores were gathered under new columns classwork\_number and score. The contains() function. We will use this classwork\_df data frame later.

view(classwork\_df)

## Analysis

### Visualize Data

Visual representations of data are more human friendly than just looking at numbers alone. This next line of code shows a summary of the data by each column similar to what we did in Walkthrough 1.

# Summary of the data by columns  
summary(gradebook)

But R can do more than just print numbers to a screen. We’ll use the {ggplot2} package from within {tidyverse} to graph some of the data to help get a better grasp of what the data looks like. This code uses {ggplot2} to graph categorical variables into a bar graph. Here we can see the variable Letter\_grade is plotted on the x-axis showing the counts of each letter grade on the y-axis.

# Bar graph for categorical variable  
gradebook %>%  
 ggplot(aes(x = letter\_grade,  
 fill = running\_average > 90)) +  
 geom\_bar() +  
 labs(title = "Bar Graph of Student Grades",  
 x = "Letter Grades",  
 y = "Count",  
 fill = "A or Better") +  
 scale\_fill\_dataedu() +  
 theme\_dataedu()

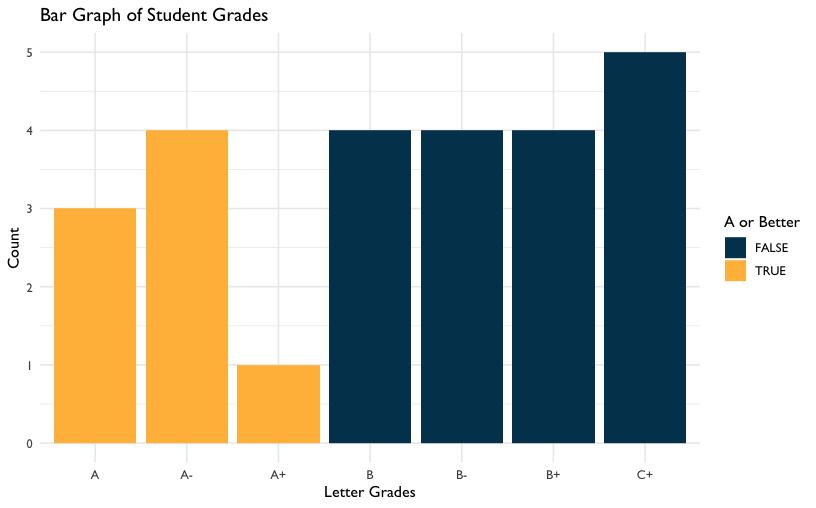


Figure 16: Bar Graph of Student Grades

Using {ggplot2} we can create many types of graphs. Using our classwork\_df from earlier, we can see the distribution of scores and how they differ from classwork to classwork using boxplots. We are able to do this because we have made the classworks and scores columns into tidy formats.

# Scatterplot of continuous variable  
classwork\_df %>%  
 ggplot(aes(x = classwork\_number,  
 y = score,  
 fill = classwork\_number)) +  
 geom\_boxplot() +  
 labs(  
 title = "Distribution of Classwork Scores",  
 x = "Classwork",  
 y = "Scores"  
 ) +  
 scale\_fill\_dataedu() +  
 theme\_dataedu() +  
 theme(  
 # removes legend  
 legend.position = "none",  
 # angles the x axis labels  
 axis.text.x = element\_text(angle = 45, hjust = 1)  
 )

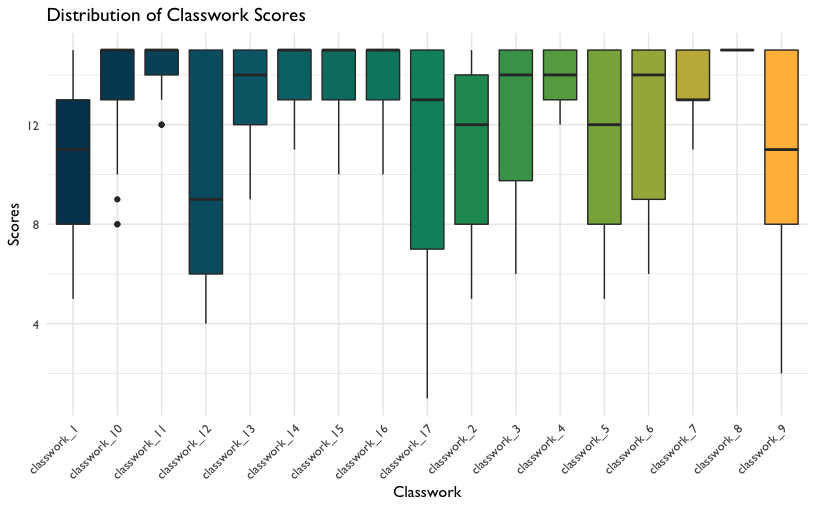


Figure 17: Distribution of Classwork Scores

### Model Data

#### Deciding on an Analysis

Using this spreadsheet, we can start to form hypotheses about the data. For example, we can ask ourselves, “Can we predict overall grade using formative assessment scores?” For this, we will try to predict a response variable Y (overall grade) as a function of a predictor variable Y (formative assessment scores). The goal is to create a mathematical equation for overall grade as a function of formative assessment scores when only formative assessment scores are known.

#### Visualize Data to Check Assumptions

It’s important to visualize data to see any distributions, trends, or patterns before building a model. We use {ggplot2} to understand these variables graphically.

##### Linearity

First, we plot X and Y to determine if we can see a linear relationship between the predictor and response. The x-axis shows the formative assessment scores while the y-axis shows the overall grades. The graph suggests a correlation between overall class grade and formative assessment scores. As the formative scores goes up, the overall grade goes up too.

# Scatterplot between formative assessment and grades by percent  
# To determine linear relationship  
gradebook %>%  
 ggplot(aes(x = formative\_assessments,  
 y = running\_average)) +  
 geom\_point(color = dataedu\_colors("green")) +  
 labs(title = "Relationship Between Overall Grade and Formative Assessments",  
 x = "Formative Assessment Score",  
 y = "Overall Grade in Percentage") +  
 theme\_dataedu()

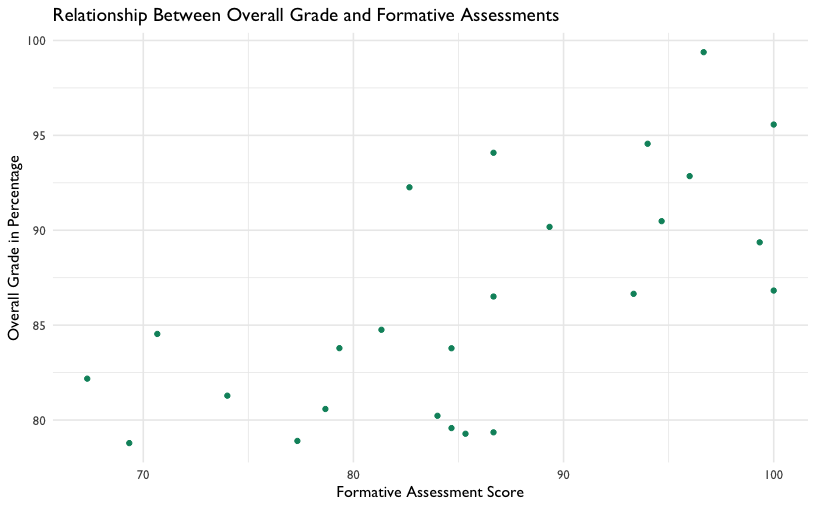


Figure 18: Relationship Between Overall Grade and Formative Assessments

We can layer different types of plots on top of each other in {ggplot2}. Here the scatterplot is layered with a line of best fit, suggesting a positive linear relationship.

# Scatterplot between formative assessment and grades by percent  
# To determine linear relationship  
# With line of best fit  
gradebook %>%  
 ggplot(aes(x = formative\_assessments,  
 y = running\_average)) +  
 geom\_point(color = dataedu\_colors("green")) +  
 geom\_smooth(method = "lm",  
 se = TRUE) +  
 labs(title = "Relationship Between Overall Grade and Formative Assessments",  
 x = "Formative Assessment Score",  
 y = "Overall Grade in Percentage") +  
 theme\_dataedu()

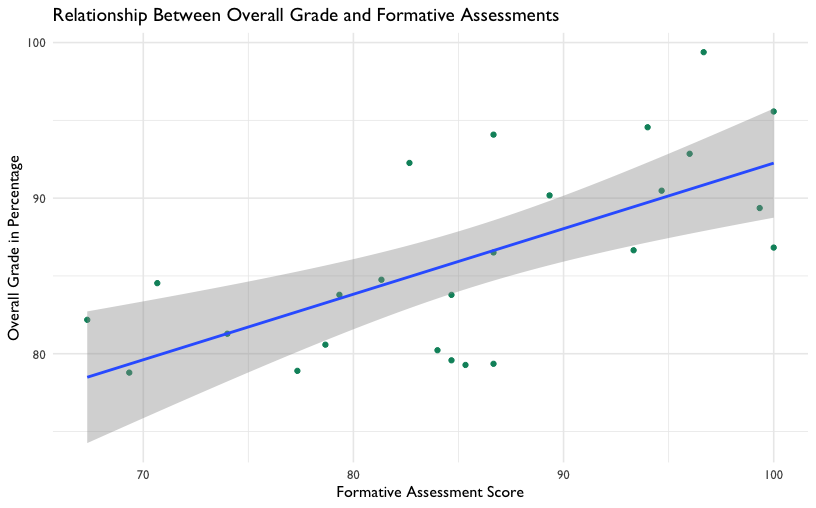


Figure 19: Relationship Between Overall Grade and Formative Assessments (with Line of Best Fit)

##### Outliers

Now we use boxplots to determine if there are any outliers in formative assessment scores or overall grades. As we would like to conduct a linear regression, we’re hoping to see no outliers in the data. We don’t see any for these two variables, so we can proceed with the model.

# Boxplot of formative assessment scores  
# To determine if there are any outliers  
gradebook %>%  
 ggplot(aes(x = "",  
 y = formative\_assessments)) +  
 geom\_boxplot(fill = dataedu\_colors("yellow")) +  
 labs(title = "Distribution of Formative Assessment Scores",  
 x = "Formative Assessment",  
 y = "Score") +  
 theme\_dataedu()

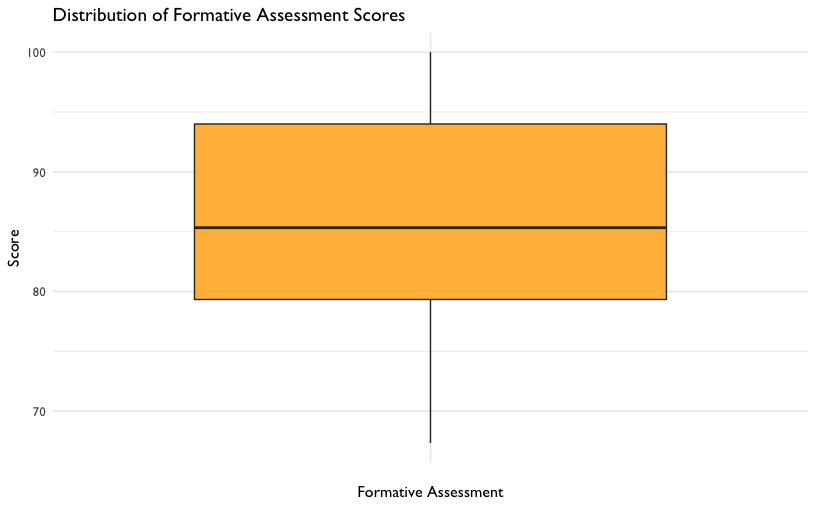


Figure 20: Distribution of Formative Assessment Scores

# Boxplot of overall grade scores in percentage  
# To determine if there are any outliers  
gradebook %>%  
 ggplot(aes(x = "",  
 y = running\_average)) +  
 geom\_boxplot(fill = dataedu\_colors("yellow")) +  
 labs(title = "Distribution of Overall Grade Scores",  
 x = "Overall Grade",  
 y = "Score in Percentage") +  
 theme\_dataedu()

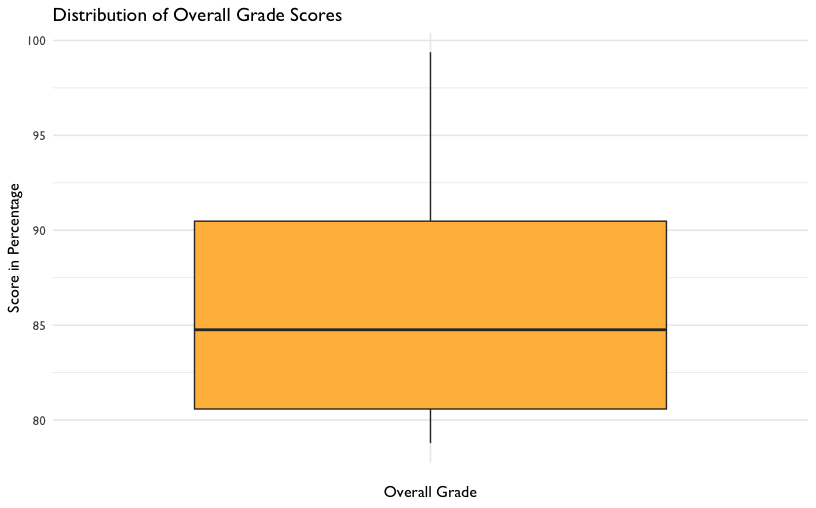


Figure 21: Distribution of Overall Grade Scores

### Correlation Analysis

We want to know the strength of the relationship between the two variables, formative assessment scores and overall grade percentage. The strength is denoted by the “correlation coefficient.” The correlation coefficient goes from -1 to 1. If one variable consistently increases with the increasing value of the other, then they have a positive correlation (towards 1). If one variable consistently decreases with the increasing value of the other, then they have a negative correlation (towards -1). If the correlation coefficient is 0, then there is no relationship between the two variables.

Correlation is good for finding relationships but it does *not* imply that one variable causes the other (correlation does not mean causation).

cor(gradebook$formative\_assessments, gradebook$running\_average)

## [1] 0.663

## Results

In [Chapter 7](#c07), we introduced the concept of linear models. Let’s use that same technique here. Now that you’ve checked your assumptions and seen a linear relationship, we can build a linear model, that is, a mathematical formula that calculates your running average as a function of your formative assessment score. This is done using the lm() function, where the arguments are:

* Your predictor (formative\_assessments)
* Your response (running\_average)
* The data (gradebook)

lm() is available in “base R” - that is, no additional packages beyond what is loaded with R automatically are necessary.

linear\_mod <-   
 lm(running\_average ~ formative\_assessments, data = gradebook)  
  
summary(linear\_mod)

##   
## Call:  
## lm(formula = running\_average ~ formative\_assessments, data = gradebook)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.281 -2.793 -0.013 3.318 8.535   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.1151 8.5477 5.86 5.6e-06 \*\*\*  
## formative\_assessments 0.4214 0.0991 4.25 3e-04 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.66 on 23 degrees of freedom  
## Multiple R-squared: 0.44, Adjusted R-squared: 0.416   
## F-statistic: 18.1 on 1 and 23 DF, p-value: 0.000302

When you fit a model to two variables, you create an equation that describes the relationship between those two variables on average. This equation uses the (Intercept), which is 50.11511, and the coefficient for formative\_assessments, which is .42136. The equation reads like this:

running\_average = 50.11511 + 0.42136\*formative\_assessments

We interpret these results by saying “For every one unit increase in formative assessment scores, we can expect a .421 unit increase in running average scores.” This equation estimates the relationship between formative assessment scores and running average scores in the student population. Think of it as an educated guess about any one particular student’s running average, if all you had was their formative assessment scores.

*More on Interpreting Models*

Challenge yourself to apply your education knowledge to the way you communicate a model’s output to your audience. Consider the difference between describing the relationship between formative assessment scores and running averages for a large group of students and for an individual student.

If you were describing the formative assessment system to stakeholders, you might say something like, “We can generally expect our students to show a .421 increase in their running average score for every one point increase in their formative assessment scores.” That makes sense, because your goal is to explain what happens **in general**.

But we can rarely expect every prediction about individual students to be correct, even with a reliable model. So when using this equation to inform how you support an individual student, it’s important to consider all the real-life factors, visible and invisible, that influence an individual student outcome.

To illustrate this concept, consider predicting how long it takes for you to run around the block right outside your office. Imagine you ran around the block five times and after each lap you jotted your time down on a post-it. After the fifth lap you do a calculation on your cell phone and see that your average lap time is five minutes. If you were to guess how long your sixth lap would take, you’d be smart to guess five minutes. But intuitively you know there’s no guarantee the sixth lap time will land right on your average. Maybe you’ll trip on a crack in the sidewalk and lose a few seconds, or maybe your favorite song pops into your head and gives you a thirty second advantage. Statisticians would call the difference between your predicted lap time and your actual lap time a “residual” value. Residuals are the differences between predicted values and actual values that aren’t explained by your linear model equation.

It takes practice to interpret and communicate these concepts well. A good start is exploring model outputs in two contexts: first as a general description of a population and second as a practical tool for helping individual student performance.

## Conclusion

This walkthrough chapter followed the basic steps of a data analysis project. We first *imported* our data, then *cleaned and transformed* it. Once we had the data in a tidy format, we were able to *explore* the data using data visualization before *modeling* the data using linear regression. Imagine that you ran this analysis for someone else: a teacher or an administrator in a school. In such cases, you might be interested in sharing the results in the form of a report or document. Thus, the only remaining step in this analysis would be to communicate our findings using a tool such as [RMarkdown](https://rmarkdown.rstudio.com/) (https://rmarkdown.rstudio.com/). While we do not discuss RMarkdown in this book, we note that it provides the functionality to easily generate reports that include both text (like the words you just read) as well as code and the output from code that are displayed together in a single document (PDF, Word, HTML, and other formats format). We note that while we began to explore models in this walkthrough, in later chapters (i.e., [Chapter 9 on aggregate data](#c9), [Chapter 10 on longitudinal analyses](#c10), [Chapter 13 on multi-level models](#c13), and [Chapter 14 on random forest machine learning models](#c14)), we will discuss such analyses and statistical modeling in more detail.

# Walkthrough 3: Using School-Level Aggregate Data to Illuminate Educational Inequities

## Vocabulary

* aggregate data
* disaggregated data
* data frame
* Free/Reduced Price Lunch (FRPL)
* histogram
* lists
* subgroup
* trim
* weighted average

## Chapter Overview

### Background

A common situation encountered when searching for education data, particularly by analysts who are not directly working with schools or districts, is the prevalence of publicly available, aggregate data. Aggregate data refers to numerical or non-numerical information that is (1) collected from multiple sources and/or on multiple measures, variables, or individuals and (2) compiled into data summaries or summary reports, typically for the purposes of public reporting or statistical analysis (Schools, [n.d.](#ref-greatschools2014)). Example of publicly available aggregate data include school-level graduation rates, state test proficiency scores by grade and subject, or mean survey responses.

Aggregated data are essential both for accountability purposes and for providing useful information about schools and districts to those who are monitoring them. For example, district administrators might aggregate row-level (also known as individual-level or student-level) enrollment reports over time. This allows them to see how many students enroll in each school, in the district overall, and any grade-level variation. Depending on their state, the district administrator might submit these aggregate data to their state education agency for reporting purposes. These datasets might be posted on the state’s department of education website for anybody to download and use.

Federal and international education datasets provide additional context for evaluating education systems. In the US, some federal datasets aim to consolidate important metrics from all states. This can be quite useful because each state has its own repository of data and to go through each state to download a particular metric is a significant effort. The federal government also funds assessments and surveys which are disseminated to the public. However, the federal datasets often have more stringent data requirements than the states, so the datasets may be less usable.

For education data practitioners, these reports and datasets can be analyzed to answer questions related to their field of interest. However, publicly available, aggregate datasets are large and often suppressed to protect privacy. Sometimes they are already a couple of years old by the time they’re released. Because of their coarseness, they can be difficult to interpret and use. Generally, aggregated data are generally used to surface of broader trends and patterns in education as opposed to diagnosing underlying issues or making causal statements. It is very important that we consider the limitations of aggregate data before analyzing them.

Analysis of aggregate data can help us identify patterns that may not have previously been known. When we have gained new insight, we can create research questions, craft hypotheses around our findings, and make recommendations on how to make improvements for the future. This walkthrough, then, uses and explores aggregate data, with a focus on educational equity, from a single school district.

*Disaggregating Aggregated Data*

Aggregated data can tell us many things, but in order for us to examine subgroups and their information, we must have data *disaggregated* by the subgroups we hope to analyze. This data is still aggregated from row-level data but provides information on smaller components than the grand total (**???**). Common disaggregations for students include gender, race/ethnicity, socioeconomic status, English learner designation, and whether they are served under the *Individuals with Disabilities Education Act (IDEA)*.

*Disaggregating Data and Equity*

Disaggregated data is essential to monitor equity in educational resources and outcomes. If only aggregate data is provided, we are unable to distinguish how different groups of students are doing and what support they need. With disaggregated data, we can identify where solutions are needed to solve disparities in opportunity, resources, and treatment.

It is important to define what equity means to your team so you know whether you are meeting your equity goals.

## Data Sources

There are many education-related, publicly available aggregate datasets. On the international level, perhaps the most well known is:

* [Programme for International Student Assessment (PISA)](http://www.oecd.org/pisa/) (http://www.oecd.org/pisa/), which measures 15-year-old school pupils’ scholastic performance on mathematics, science, and reading.

On the federal level, examples include:

* [Civil Rights Data Collection (CRDC)](https://www2.ed.gov/about/offices/list/ocr/data.html) (https://www2.ed.gov/about/offices/list/ocr/data.html), which reports many different variables on educational program and services disaggregated by race/ethnicity, sex, limited English proficiency, and disability. These data are school-level.
* [Common Core of Data (CCD)](https://nces.ed.gov/ccd/) (https://www2.ed.gov/about/offices/list/ocr/data.html), which is the U.S. Department of Education’s primary database on public elementary and secondary education.
* [EdFacts](https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html) (https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html), which includes state assessments and adjusted cohort graduation rates. These data are school- and district-level.
* [Integrated Postsecondary Education Data System (IPEDS)](https://nces.ed.gov/ipeds/) (https://nces.ed.gov/ipeds/), which is the U.S. Department of Education’s primary database on postsecondary education.
* [National Assessment for Educational Progress (NAEP) Data](https://nces.ed.gov/nationsreportcard/researchcenter/datatools.aspx) (https://nces.ed.gov/nationsreportcard/researchcenter/datatools.aspx), an assessment of educational progress in the United States. Often called the “nation’s report card.” The NAEP reading and mathematics assessments are administered to a representative sample of fourth- and eighth-grade students in each state every two years.

On the state and district level, examples include:

* [California Department of Education](https://www.cde.ca.gov/ds/) (https://www.cde.ca.gov/ds/), which is the state department of education website. It includes both downloadable CSV files and “Data Quest”, which lets you query the data online.
* [Minneapolis Public Schools](https://mpls.k12.mn.us/reports_and_data) (https://mpls.k12.mn.us/reports\_and\_data), which is a district-level website with datasets beyond those listed in the state website.

*Selecting Data*

For the purposes of this walkthrough, we will be looking at a particular school district’s data. This district reports their student demographics in a robust, complete way. Not only do they report the percentage of students in a subgroup, but they also include the number of students in each subgroup. This allows a deep look into their individual school demographics. Their reporting of the composition of their schools provides an excellent opportunity to explore inequities in a system.

### Methods

In this chapter, we will walk through how running analyses on a publicly available dataset can help education data practitioners understand the landscape of needs and opportunities in the field of education. As opposed to causal analysis, which aims to assess the root cause of an phenomenon or the effects of an intervention, we use analysis on an aggregate dataset to find out whether there *is* a phenomenon, *what* it is, and *what* we’d be trying to solve through interventions (**???**).

## Load Packages

As usual, we begin our code by calling the libraries we will use.

library(tidyverse)  
library(here)  
library(janitor)  
library(dataedu)

ROpenSci created the [{tabulizer}](https://github.com/ropensci/tabulizer) (https://github.com/ropensci/tabulizer) package which provides R bindings to the Tabula java library, which can be used to computationally extract tables from PDF documents. {rJava} is a required package to load {tabulizer}. Unfortunately, installing {rJava} on Macs can be very tedious. If you find yourself unable to install {tabulizer} (Leeper [2018](#ref-R-tabulizer)), or would like to skip to the data processing, the data pulled from the PDFs is available in the {dataedu} package. You can decide whether to skip the steps requiring {rJava} (Urbanek [2019](#ref-R-rJava)).

library(tabulizer)

## Import Data

{tabulizer} pulls the PDF data into lists using extract\_tables().

# if using tabulizer  
race\_pdf <-  
 extract\_tables("https://studentaccounting.mpls.k12.mn.us/uploads/mps\_fall2018\_racial\_ethnic\_by\_school\_by\_grade.pdf")  
  
# if reading in the tabulizer output  
race\_pdf <-  
 readRDS(here("data", "wt05\_agg\_data", "race\_pdf.Rds"))

It is important to consistently check what we’re doing with the actual PDF’s to ensure we’re getting the data that we need.

* We then transform the list to a data frame with map(as\_tibble).
* The slice() in map\_df() removes unnecessary rows from the PDF.
* Now we create readable column names using set\_names().

race\_df <-  
 race\_pdf %>%  
 # turn each page into a tibble  
 map(as\_tibble) %>%   
 # remove unnecessary rows  
 map\_df( ~ slice(.,-1:-2)) %>%  
 # use descriptive column names  
 set\_names(  
 c(  
 "school\_group",  
 "school\_name",  
 "grade",  
 "na\_num",  
 # native american number of students  
 "na\_pct",  
 # native american percentage of students  
 "aa\_num",  
 # african american number of students  
 "aa\_pct",  
 # african american percentage  
 "as\_num",  
 # asian number of students  
 "as\_pct",  
 # asian percentage  
 "hi\_num",  
 # hispanic number of students  
 "hi\_pct",  
 # hispanic percentage  
 "wh\_num",  
 # white number of students  
 "wh\_pct",  
 # white percentage  
 "pi\_pct",  
 # pacific islander percentage  
 "blank\_col",  
 "tot" # total number of students  
 )  
 )

For the Race/Ethnicity table, we want the totals for each district school as we won’t be looking at grade-level variation. When analyzing the PDF, we see the totals have “Total” in the School Name.

We clean up this dataset by:

1. Removing unnecessary or blank columns using select(). Negative selections means those columns will be removed.
2. Removing all Grand Total rows (otherwise they’ll show up in our data, when we just want district-level data) using filter(). We keep schools that have “Total” in the name but remove any rows that are Grand Total.
3. Then we trim white space from strings.
4. The data in the “percentage” columns are provided with a percentage sign. This means we will have to remove all of them to be able to do math on these columns (for example, adding them together). Also, we want to divide the numbers by 100 so they are true percentages.

race\_clean <-  
 race\_df %>%  
 # remove unnecessary columns  
 select(-school\_group,-grade,-pi\_pct,-blank\_col) %>%  
 # filter to get grade-level numbers  
 filter(str\_detect(school\_name, "Total"),  
 school\_name != "Grand Total") %>%  
 # clean up school names  
 mutate(school\_name = str\_replace(school\_name, "Total", "")) %>%  
 # remove white space  
 mutate\_if(is.character, trimws) %>%  
 # turn numbers into percentages  
 mutate\_at(vars(contains("pct")), list( ~ as.numeric(str\_replace(., "%", "")) / 100))

We will import the Free Reduced Price Lunch (FRPL) PDF’s now.

FRPL stands for Free/Reduced Price Lunch, often used as a proxy for poverty (**???**). Students from a household with an income up to 185 percent of the poverty threshold are eligible for free or reduced price lunch. (Sidenote: Definitions are very important in disaggregated data. FRPL is used because it’s ubiquitous and reporting is mandated but there is debate as to whether it actually reflects the level of poverty among students.)

# if using tabulizer  
frpl\_pdf <-  
 extract\_tables("https://studentaccounting.mpls.k12.mn.us/uploads/fall\_2018\_meal\_eligiblity\_official.pdf")  
  
# if reading in the tabulizer output  
race\_pdf <-  
 readRDS(here("data", "wt05\_agg\_data", "race\_pdf.Rds"))

Similar to the Race/Ethnicity PDF, there are rows that we don’t need from each page, which we remove using slice().

frpl\_df <-  
 frpl\_pdf %>%  
 map(as\_tibble) %>%  
 map\_df( ~ slice(.,-1)) %>%  
 set\_names(  
 c(  
 "school\_name",  
 "not\_eligible\_num",  
 # number of non-eligible students,  
 "reduce\_num",  
 # number of students receiving reduced price lunch  
 "free\_num",  
 # number of students receiving free lunch  
 "frpl\_num",  
 # total number of students  
 "frpl\_pct" # free/reduced price lunch percentage  
 )  
 )

To clean it up, we remove the rows that are blank. When looking at the PDF, we notice that there are aggregations inserted into the table that are not district-level. For example, they have included ELM K\_08, presumably to aggregate FRPL numbers up to the K-8 level. Although this is useful data, we don’t need it for this district-level analysis. There are different ways we can remove these rows but we will just filter them out.

frpl\_clean <-  
 frpl\_df %>%  
 filter(  
 # remove blanks  
 school\_name != "",  
 # filter out the rows in this list  
 school\_name %in% c(  
 "ELM K\_08",  
 "Mid Schl",  
 "High Schl",  
 "Alt HS",  
 "Spec Ed Total",  
 "Cont Alt Total",  
 "Hospital Sites Total",  
 "Dist Total"  
 )  
 ) %>%  
 # turn numbers into percentages  
 mutate(frpl\_pct = as.numeric(str\_replace(frpl\_pct, "%", "")) / 100)

Because we want to look at race/ethnicity data in conjunction with free/reduced price lunch percentage, we join the two datasets by the name of the school.

* Use mutate\_at() to specify columns to which to apply a function (in this case, as.numeric.)

# create full dataset, joined by school name  
joined\_df <-  
 left\_join(race\_clean, frpl\_clean, by = c("school\_name")) %>%  
 mutate\_at(2:17, as.numeric)

**Did you notice?** The total number of students from the Race/Ethnicity table does **not** match the total number of students from the FRPL table, even though they’re referring to the same districts in the same year. Why? Perhaps the two datasets were created by different people, who used different rules when aggregating the dataset. Perhaps the counts were taken at different times of the year, when students may have moved around in the meantime. We don’t know but it does require us to make strategic decisions about which data we consider the ‘truth’ for our analysis.

Now we move on to the fun part of creating new columns based on the merged dataset.

1. We want to calculate, for each race, the number of students in ‘high poverty’ schools. This is defined by NCES as schools that are over 75% FRPL Education Statistics U.S. Department of Education ([2019](#ref-ncesfrpl)). When a school has over 75% FRPL, we count the number of students for that particular race under the variable [racename]\_povnum.
2. The {janitor} package has a handy adorn\_totals() function that sums the columns for you. This is important because we want a weighted average of students in each category, so we need the total number of students in each group.
3. We create the weighted average of the percentage of each race by dividing the number of students by race by the total number of students.
4. To get FRPL percentage for all schools, we have to recalculate frpl\_pct.
5. To calculate the percentage of students by race who are in high poverty schools, we must divide the number of students in high poverty schools by the total number of students in that race.

merged\_df <-  
 joined\_df %>%  
 # calculate high poverty numbers  
 mutate(  
 hi\_povnum = case\_when(frpl\_pct > .75 ~ hi\_num),  
 aa\_povnum = case\_when(frpl\_pct > .75 ~ aa\_num),  
 wh\_povnum = case\_when(frpl\_pct > .75 ~ wh\_num),  
 as\_povnum = case\_when(frpl\_pct > .75 ~ as\_num),  
 na\_povnum = case\_when(frpl\_pct > .75 ~ na\_num)  
 ) %>%  
 adorn\_totals() %>%  
 # create percentage by demographic  
 mutate(  
 na\_pct = na\_num / tot,  
 aa\_pct = aa\_num / tot,  
 as\_pct = as\_num / tot,  
 hi\_pct = hi\_num / tot,  
 wh\_pct = wh\_num / tot,  
 frpl\_pct = (free\_num + reduce\_num) / frpl\_num,  
 # create percentage by demographic and poverty  
 hi\_povsch = hi\_povnum / hi\_num[which(school\_name == "Total")],  
 aa\_povsch = aa\_povnum / aa\_num[which(school\_name == "Total")],  
 as\_povsch = as\_povnum / as\_num[which(school\_name == "Total")],  
 wh\_povsch = wh\_povnum / wh\_num[which(school\_name == "Total")],  
 na\_povsch = na\_povnum / na\_num[which(school\_name == "Total")]  
 )

To facilitate the creation of plots later on, we put this data in tidy format.

tidy\_df <-  
 merged\_df %>%  
 pivot\_longer(cols = -matches("school\_name"),  
 names\_to = "category",  
 values\_to = "value")

Running the above code, particularly the download of the PDFs, takes a lot of time. We’ve saved copies of the merged and tidy data in the data folder and can be accessed by running the code below.

tidy\_df <-  
 read\_csv(here("data", "wt05\_agg\_data", "wt05\_aggdat\_tidy\_dat.csv"))  
  
merged\_df <-  
 read\_csv(here("data", "wt05\_agg\_data", "wt05\_aggdat\_merged\_dat.csv"))

## View Data

### Discovering Distributions

What do the racial demographics in this district look like? For this, a barplot can quickly visualize the different proportion of subgroups.

tidy\_df %>%  
 # filter for Total rows, since we want district-level information  
 filter(school\_name == "Total",  
 str\_detect(category, "pct"),  
 category != "frpl\_pct") %>%  
 # reordering x axis so bars appear by descending value  
 ggplot(aes(x = reorder(category, -value), y = value)) +  
 geom\_bar(stat = "identity", aes(fill = category)) +  
 ggtitle("Percentage of Population by Subgroup") +   
 xlab("Subgroup") +  
 ylab("Percentage of Population") +  
 scale\_x\_discrete(  
 labels = c(  
 "aa\_pct" = "Black",  
 "wh\_pct" = "White",  
 "hi\_pct" = "Hispanic",  
 "as\_pct" = "Asian",  
 "na\_pct" = "Native Am."  
 )  
 ) +  
 scale\_y\_continuous(labels = scales::percent) + # makes labels present as percentages  
 scale\_fill\_dataedu() +  
 theme(legend.position = "none") +  
 theme\_dataedu()

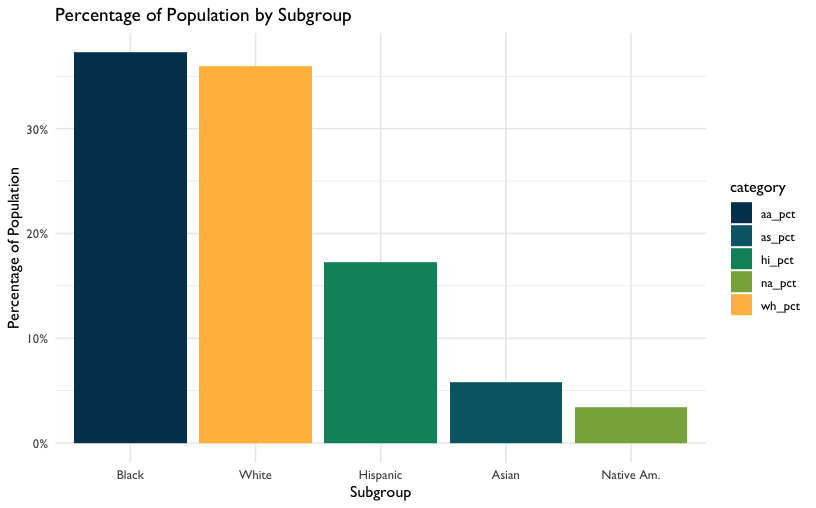


Figure 22: Percentage of Population by Subgroup

When we look at these data, the district looks very diverse. Almost **40% of students are Black** and around **36% are White.**

Note: this matches the percentages provided in the original PDF’s. This shows our calculations above were accurate. Hooray!

In terms of free/reduced price lunch, we have that calculated under frpl\_pct:

tidy\_df %>%  
 filter(category == "frpl\_pct",  
 school\_name == "Total")

## # A tibble: 1 x 3  
## school\_name category value  
## <chr> <chr> <dbl>  
## 1 Total frpl\_pct 0.569

**56.9% of the students are eligible for FRPL**, compared to [the U.S. average of 52.1%.](https://nces.ed.gov/programs/digest/d17/tables/dt17_204.10.asp?current=yes) (https://nces.ed.gov/programs/digest/d17/tables/dt17\_204.10.asp?current=yes)

This also matches the PDF’s. Great!

Now, we dig deeper to see if there is more to the story.

### Analyzing Spread

Another view of the data can be visualizing the distribution of students with different demographics across schools. Here is a histogram for the percentage of White students within the schools for which we have data.

merged\_df %>%  
 # remove district totals  
 filter(school\_name != "Total") %>%  
 # x-axis will be the percentage of white students within schools  
 ggplot(aes(x = wh\_pct)) +  
 geom\_histogram(breaks = seq(0, 1, by = .1),  
 fill = dataedu\_colors("darkblue")) +  
 ggtitle("Count of Schools by White Population") +  
 xlab("White Percentage") +  
 ylab("Count") +  
 scale\_x\_continuous(labels = scales::percent) +   
 theme(legend.position = "none") +  
 theme\_dataedu()

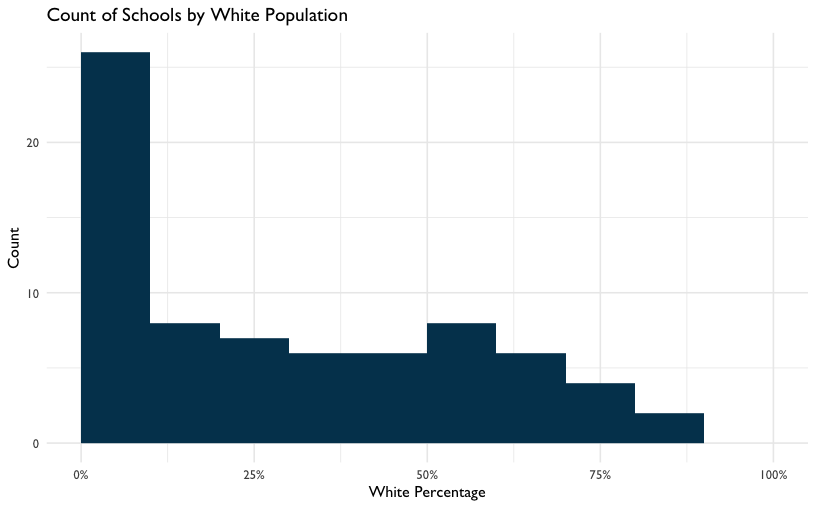


Figure 23: Count of Schools by White Population

**26 of the 74 (35%) of schools have between 0-10% White students.** This implies that even though the school district may be diverse, the demographics are not evenly distributed across the schools. More than half of schools enroll fewer than 30% of White students even though White students make up 35% of the district student population.

The school race demographics are not representative of the district populations but does that hold for socioeconomic status as well?

## Analysis

### Creating Categories

High-poverty schools are defined as public schools where more than 75% of the students are eligible for FRPL. According to NCES, 24% of public school students attended high-poverty schools (Education Statistics U.S. Department of Education ([2019](#ref-ncesfrpl))). However, different subgroups were overrepresented and underrepresented within the high poverty schools. Is this the case for this district?

tidy\_df %>%  
 filter(school\_name == "Total",  
 str\_detect(category, "povsch")) %>%  
 ggplot(aes(x = reorder(category,-value), y = value)) +  
 geom\_bar(stat = "identity", aes(fill = factor(category))) +  
 ggtitle("Distribution of Subgroups in High Poverty Schools") +  
 xlab("Subgroup") +  
 ylab("Percentage in High Poverty Schools") +  
 scale\_x\_discrete(  
 labels = c(  
 "aa\_povsch" = "Black",  
 "wh\_povsch" = "White",  
 "hi\_povsch" = "Hispanic",  
 "as\_povsch" = "Asian",  
 "na\_povsch" = "Native Am."  
 )  
 ) +  
 scale\_y\_continuous(labels = scales::percent) +  
 scale\_fill\_dataedu() +  
 theme(legend.position = "none") +  
 theme\_dataedu()

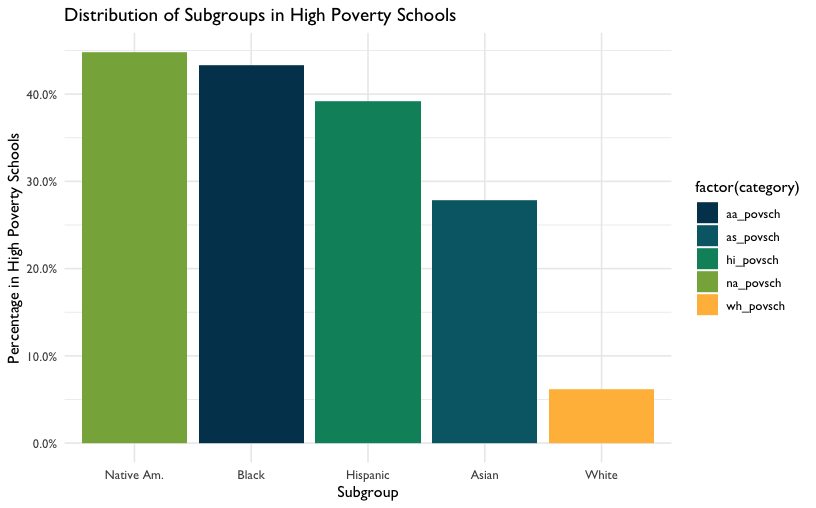


Figure 24: Distribution of Subgroups in High Poverty Schools

**8% of White students** attend high poverty schools, compared to **43% of Black students, 39% of Hispanic students, 28% of Asian students, and 45% of Native American students.** We can conclude these students are disproportionally attending high poverty schools.

### Reveal Relationships

Let’s explore what happens when we correlate race and FRPL percentage by school.

merged\_df %>%  
 filter(school\_name != "Total") %>%  
 ggplot(aes(x = wh\_pct, y = frpl\_pct)) +  
 geom\_point(color = dataedu\_colors("green")) +  
 ggtitle("FRPL Percentage vs. White Percentage") +  
 xlab("White Percentage") +  
 ylab("FRPL Percentage") +  
 scale\_y\_continuous(labels = scales::percent) +  
 scale\_x\_continuous(labels = scales::percent) +  
 theme(legend.position = "none") +  
 theme\_dataedu()

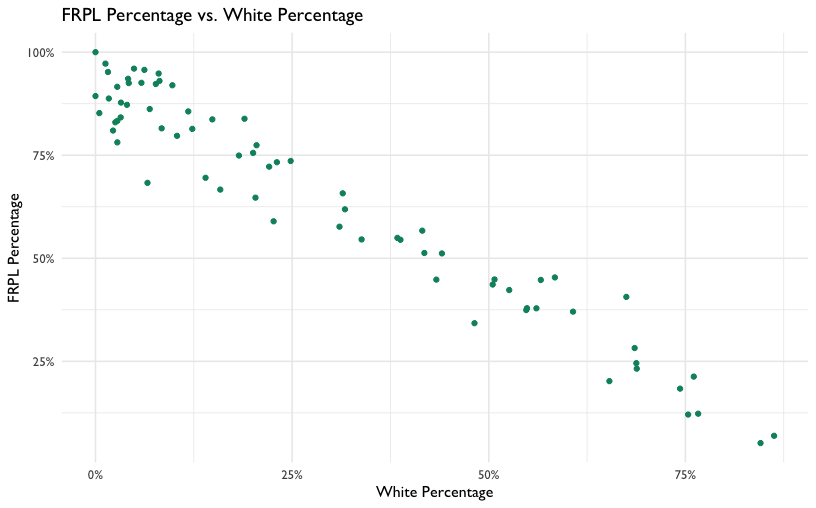


Figure 25: FRPL Percentage vs. White Percentage

Related to the result above, there is a strong negative correlation between FRPL percentage and the percentage of White students in a school. That is, high poverty schools have a lower percentage of White students and low poverty schools have a higher percentage of White students.

## Results

Because of the disaggregated data this district provides, we can go deeper than the average of demographics across the district and see what it looks like on the school level. These views display that (1) there exists a distribution of race/ethnicity within schools that are not representative of the district, (2) that students of color are overrepresented in high poverty schools, and (3) there is a negative relationship between the percentage of White students in a school and the percentage of students eligible for FRPL.

## Conclusion

According to the Urban Institute, the disproportionate percentage of students of color attending high poverty schools “is a defining feature of almost all Midwestern and northeastern metropolitan school systems” (**???**). Among other issues, “high poverty schools tend to lack the educational resources—like highly qualified and experienced teachers, low student-teacher ratios, college prerequisite and advanced placement courses, and extracurricular activities—available in low-poverty schools.” This has a huge impact on these students and their futures.

In addition, research shows that racial and socioeconomic diversity in schools can provide students with a range of cognitive and social benefits. Therefore, this deep segregation that exists in the district can have adverse effects on students.

As a data scientist in education, we can use these data to suggest interventions for what we can do to improve equity in the district. In addition, we can advocate for more datasets such as these, which allow us to dig deep.

# Walkthrough 4: Longitudinal Analysis With Federal Students With Disabilities Data

## Vocabulary

* aggregate data
* file path
* list
* read in
* tidy format
* statistical model
* student-level data
* subset
* vector

## Chapter Overview

### Background

Data scientists working in education don’t always have access to student level data, so knowing how to model publicly available datasets is a useful skill. This walkthrough has two goals. First, we’ll be learning some ways to explore data over time. Second, we’ll be learning how to explore a publicly available dataset. Like most public datasets, this one contains aggregate data. This means that someone totaled up the student counts so it doesn’t reveal any private information.

You can download the datasets for this walkthrough on the United States Department of Education website (see of Education ([2020](#ref-usdoe2019)))[[1]](#footnote-270). This walkthrough uses datasets of student with disabilities counts in each state.

### Methods

In this walkthrough, we’ll learn how to read multiple datasets in using the map() function. Next, we’ll prepare our data for analysis by cleaning the variable names. Finally, we’ll explore this dataset by visualizing student counts and comparing male to female ratios over time.

## Load Packages

For this walkthrough, we’ll be using four packages: {tidyverse}, {here}, {dataedu}, and {lubridate} (Spinu, Grolemund, and Wickham [2018](#ref-R-lubridate)); of these, {lubridate} is the sole package we have not used previously. You can load those packages by running this code:

library(tidyverse)  
library(here)  
library(dataedu)  
library(lubridate)

## Import Data

In this analysis we’ll be importing and combining six datasets that describe the number of students with disabilities in a given year. Let’s spend some time carefully reviewing how to get the .csv files we’ll need downloaded and stored on your computer. If you want to run the code exactly as written here, you’ll need to store the same datasets in the right location. It’s possible to use this walkthrough on different datasets or to store them in different locations on your computer, but you’ll need to make adjustments to your code based on the datasets you used and where you stored them. We suggest only doing this if you already have some experience using R.

### What to Download

In this walkthrough, we’ll be using six separate datasets of child counts, one for each year between 2012 and 2017. If you’re copying and pasting the code in this walkthrough, we recommend downloading the datasets from our GitHub repository for the most reliable results. Here’s a link to each file:

* [2012 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2012.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2012.csv)
* [2013 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2013.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2013.csv)
* [2014 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2014.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2014.csv)
* [2015 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2015.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2015.csv)
* [2016 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2016.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2016.csv)
* [2017 data](https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal_data/bchildcountandedenvironments2017-18.csv) (https://github.com/data-edu/data-science-in-education/raw/master/data/longitudinal\_data/bchildcountandedenvironments2017-18.csv)

You can also find these files on the [United States Department of Education website](https://www2.ed.gov/programs/osepidea/618-data/state-level-data-files/index.html) (https://www2.ed.gov/programs/osepidea/618-data/state-level-data-files/index.html)

### A Note on File Paths

When you download these files, be sure to store them in a folder in your working directory. To get to the data in this walkthrough, we’ll be using this file path in our working directory: “data/longitudinal\_data”. We’ll be using the {here} package, which convenieniently fills in all the folders in the file path of your working directory all the way up to the folders you specify in the arguments. So when referencing the file path “data/longitudinal\_data”, we’ll use code like this:

here::here("data",  
 "longitudinal\_data",  
 "bchildcountandedenvironments2012.csv")

You can use a different file path if you like, just take note of where your downloaded files are so you can use the correct file path when writing your code to import the data.

### Reading in One Dataset

We’ll be learning how to read in more than one dataset using the map() function. Let’s try it first with one dataset, then we’ll scale our solution up to multiple datasets. When you are analyzing multiple datasets that all have the same structure, you can read in each dataset using one code chunk. This code chunk will store each dataset as an element of a list.

Before doing that, you should explore one of the datasets to see what you can learn about its structure. Clues from this exploration inform how you read in all the datasets at once later on. For example, we can see that the first dataset has some lines at the top that contain no data:

## # A tibble: 16,234 x 31  
## `Extraction Dat… `6/12/2013` X3 X4 X5 X6 X7 X8 X9 X10   
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  
## 1 Updated: 2/12/2014 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>   
## 2 Revised: <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>   
## 3 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>   
## 4 Year State Name SEA … SEA … Amer… Asia… Blac… Hisp… Nati… Two …  
## 5 2012 ALABAMA Corr… All … - - - - - -   
## 6 2012 ALABAMA Home All … 1 1 57 12 0 2   
## 7 2012 ALABAMA Home… All … - - - - - -   
## 8 2012 ALABAMA Insi… All … - - - - - -   
## 9 2012 ALABAMA Insi… All … - - - - - -   
## 10 2012 ALABAMA Insi… All … - - - - - -   
## # … with 16,224 more rows, and 21 more variables: X11 <chr>, X12 <chr>,  
## # X13 <chr>, X14 <chr>, X15 <chr>, X16 <chr>, X17 <chr>, X18 <chr>,  
## # X19 <chr>, X20 <chr>, X21 <chr>, X22 <chr>, X23 <chr>, X24 <chr>,  
## # X25 <chr>, X26 <chr>, X27 <chr>, X28 <chr>, X29 <chr>, X30 <chr>, X31 <chr>

The rows containing “Extraction Date:”, “Updated:” and “Revised:” aren’t actually rows. They’re notes the authors left at the top of the dataset to show when the dataset was changed.

read\_csv() uses the first row as the variable names unless told otherwise, so we need to tell read\_csv() to skip those lines using the skip argument. If we don’t, read\_csv() assumes the very first line–the one that says “Extraction Date:”–is the correct row of variable names. That’s why calling read\_csv() without the skip argument results in column names like X4. When there’s no obvious column name to read in, read\_csv() names them X[...] and let’s you know in a warning message.

Try using skip = 4 in your call to read\_csv():

read\_csv(here::here(  
 "data",  
 "longitudinal\_data",  
 "bchildcountandedenvironments2012.csv"  
),  
skip = 4)

## # A tibble: 16,230 x 31  
## Year `State Name` `SEA Education … `SEA Disability… `American India…  
## <dbl> <chr> <chr> <chr> <chr>   
## 1 2012 ALABAMA Correctional Fa… All Disabilities -   
## 2 2012 ALABAMA Home All Disabilities 1   
## 3 2012 ALABAMA Homebound/Hospi… All Disabilities -   
## 4 2012 ALABAMA Inside regular … All Disabilities -   
## 5 2012 ALABAMA Inside regular … All Disabilities -   
## 6 2012 ALABAMA Inside regular … All Disabilities -   
## 7 2012 ALABAMA Other Location … All Disabilities 7   
## 8 2012 ALABAMA Other Location … All Disabilities 1   
## 9 2012 ALABAMA Parentally Plac… All Disabilities -   
## 10 2012 ALABAMA Residential Fac… All Disabilities 0   
## # … with 16,220 more rows, and 26 more variables: `Asian Age 3-5` <chr>, `Black  
## # or African American Age 3-5` <chr>, `Hispanic/Latino Age 3-5` <chr>,  
## # `Native Hawaiian or Other Pacific Islander Age 3-5` <chr>, `Two or More  
## # Races Age 3-5` <chr>, `White Age 3-5` <chr>, `Female Age 3 to 5` <chr>,  
## # `Male Age 3 to 5` <chr>, `LEP Yes Age 3 to 5` <chr>, `LEP No Age 3 to  
## # 5` <chr>, `Age 3 to 5` <chr>, `Age 6-11` <chr>, `Age 12-17` <chr>, `Age  
## # 18-21` <chr>, `Ages 6-21` <chr>, `LEP Yes Age 6 to 21` <chr>, `LEP No Age 6  
## # to 21` <chr>, `Female Age 6 to 21` <chr>, `Male Age 6 to 21` <chr>,  
## # `American Indian or Alaska Native Age 6 to21` <chr>, `Asian Age 6  
## # to21` <chr>, `Black or African American Age 6 to21` <chr>, `Hispanic/Latino  
## # Age 6 to21` <chr>, `Native Hawaiian or Other Pacific Islander Age 6  
## # to21` <chr>, `Two or more races Age 6 to21` <chr>, `White Age 6 to21` <chr>

The skip argument told read\_csv() to make the line containing “Year”, “State Name”, and so on as the first line. The result is a dataset that has “Year”, “State Name”, and so on as variable names.

### Reading in Many Datasets

Will the read\_csv() and skip = 4 combination work on all our datasets? To find out, we’ll use this strategy:

* Store a vector of filenames and paths in a list. These paths point to our datasets
* Pass the list of filenames as arguments to read\_csv() using purrr::map(), including skip = 4, in our read\_csv() call
* Examine the new list of datasets to see if the variable names are correct

Imagine a widget-making machine that works by acting on raw materials it receives on a conveyer belt. This machine executes one set of instructions on each of the raw materials it receives. You are the operator of the machine and you design instructions to get a widget out of the raw materials. Your plan might look something like this:

* *Raw materials*: a list of filenames and their paths
* *Widget-making machine*: purrr:map()
* *Widget-making instructions*: `read\_csv(path, skip = 4)
* *Expected widgets*: a list of datasets

Let’s create the raw materials first. Our raw materials will be file paths to each of the CSVs we want to read. Use list.files to make a vector of filename paths and name that vector filenames. list.files returns a vector of file names in the folder specified in the path argument. When we set the full.names argument to “TRUE”, we get a full path of these filenames. This will be useful later when we need the file names and their paths to read our data in.

# Get filenames from the data folder   
filenames <-  
 list.files(path = here::here("data", "longitudinal\_data"),  
 full.names = TRUE)  
  
# A list of filenames and paths  
filenames

That made a vector of six filenames, one for each year of child count data stored in the data folder. Now pass our raw materials, the vector called filenames, to our widget-making machine called map() and give the machine the instructions read\_csv(., skip = 4). Name the list of widgets it cranks out all\_files:

# Pass filenames to map and read\_csv  
all\_files <-  
 filenames %>%  
 # Apply the function read\_csv to each element of filenames  
 map(., ~ read\_csv(., skip = 4))

It is important to think ahead here. The goal is to combine the datasets in all\_files into one dataset using bind\_rows(). But that will only work if all the datasets in our list have the same number of columns and the same column names. We can check our column names by using map() and names():

We can use identical() to see if the variables from two datasets match. We see that the variable names of the first and second datasets don’t match, but the variables from the second and third do.

# Variables of first and second dataset don't match  
identical(names(all\_files[[1]]), names(all\_files[[2]]))

## [1] FALSE

# Variables of third and second files match  
identical(names(all\_files[[2]]), names(all\_files[[3]]))

## [1] TRUE

And we can check the number of columns by using map() and ncol():

all\_files %>%   
 # apply the function ncol to each element of all\_files  
 map(ncol)

## [[1]]  
## [1] 31  
##   
## [[2]]  
## [1] 50  
##   
## [[3]]  
## [1] 50  
##   
## [[4]]  
## [1] 50  
##   
## [[5]]  
## [1] 50  
##   
## [[6]]  
## [1] 50

Congratulations on finding an extremely common problem in education data! You’ve discovered that neither the number of columns nor the column names match. This is a problem because we won’t be able to combine the datasets into one. When we try, bind\_rows() returns a dataset with 100 columns instead of the expected 50.

# combining the datasets at this stage results in the incorrect   
# number of columns  
bind\_rows(all\_files) %>%   
 # check the number of columns  
 ncol()

## [1] 100

We’ll correct this in the next section by selecting and renaming our variables, but it’s good to notice this problem early in the process so you know to work on it later.

## Process Data

Transforming your dataset before visualizing it and fitting models is critical. It’s easier to write code when variable names are concise and informative. Many functions in R, especially those in the {ggplot2} package, work best when datasets are in a “tidy” format. It’s easier to do an analysis when you have just the variables you need. Any unused variables can confuse your thought process.

Let’s preview the steps we’ll be taking:

1. Fix the variable names in the 2016 data
2. Combine the datasets
3. Pick variables
4. Filter for the desired categories
5. Rename the variables
6. Standardize the state names
7. Transform the column formats from wide to long using pivot\_longer
8. Change the data types of variables
9. Explore NAs

In real life, data scientists don’t always know the cleaning steps until they dive into the work. Learning what cleaning steps are needed requires exploration, trial and error, and clarity on the analytic questions you want to answer.

After a lot of exploring, we settled on these steps for this analysis. When you do your own, you will find different things to transform. As you do more and more data analysis, your instincts for what to transform will improve.

### Fix the Variable Names in the 2016 Data

When we print the 2016 dataset, we notice that the variable names are incorrect. Let’s verify that by looking at the first ten variable names of the 2016 dataset, which is the fifth element of all\_files:

# Look at the first 10 column names of 2016  
names(all\_files[[5]])[1:10]

## [1] "2016" "Alabama"   
## [3] "Correctional Facilities" "All Disabilities"   
## [5] "-" "-\_1"   
## [7] "-\_2" "-\_3"   
## [9] "-\_4" "-\_5"

We want the variable names to be Year and State Name, not 2016 and Alabama. But first, let’s go back and review how to get at the 2016 dataset from all\_files. We need to identify which element the 2016 dataset was in the list. The order of the list elements was set all the way back when we fed map() our list of filenames. If we look at filenames again, we see that its fifth element is the 2016 dataset. Try looking at the first and fifth elements of filenames:

filenames[[1]]  
filenames[[5]]

Once we know the 2016 dataset is the fifth element of our list, we can pluck it out by using double brackets:

all\_files[[5]]

## # A tibble: 16,230 x 50  
## `2016` Alabama `Correctional F… `All Disabiliti… `-` `-\_1` `-\_2` `-\_3`  
## <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  
## 1 2016 Alabama Home All Disabilities 43 30 35 0   
## 2 2016 Alabama Homebound/Hospi… All Disabilities - - - -   
## 3 2016 Alabama Inside regular … All Disabilities - - - -   
## 4 2016 Alabama Inside regular … All Disabilities - - - -   
## 5 2016 Alabama Inside regular … All Disabilities - - - -   
## 6 2016 Alabama Parentally Plac… All Disabilities - - - -   
## 7 2016 Alabama Residential Fac… All Disabilities 5 3 4 0   
## 8 2016 Alabama Residential Fac… All Disabilities - - - -   
## 9 2016 Alabama Separate Class All Disabilities 58 58 98 0   
## 10 2016 Alabama Separate School… All Disabilities 11 20 19 0   
## # … with 16,220 more rows, and 42 more variables: `-\_4` <chr>, `-\_5` <chr>,  
## # `-\_6` <chr>, `-\_7` <chr>, `-\_8` <chr>, `-\_9` <chr>, `-\_10` <chr>,  
## # `-\_11` <chr>, `-\_12` <chr>, `-\_13` <chr>, `-\_14` <chr>, `0` <chr>,  
## # `0\_1` <chr>, `0\_2` <chr>, `0\_3` <chr>, `0\_4` <chr>, `0\_5` <chr>,  
## # `0\_6` <chr>, `0\_7` <chr>, `0\_8` <chr>, `1` <chr>, `2` <chr>, `4` <chr>,  
## # `14` <chr>, `22` <chr>, `30` <chr>, `4\_1` <chr>, `0\_9` <chr>, `7` <chr>,  
## # `70` <chr>, `77` <chr>, `0\_10` <chr>, `77\_1` <chr>, `1\_1` <chr>,  
## # `76` <chr>, `0\_11` <chr>, `0\_12` <chr>, `68` <chr>, `0\_13` <chr>,  
## # `0\_14` <chr>, `0\_15` <chr>, `9` <chr>

We used skip = 4 when we read in the datasets in the list. That worked for all datasets except the fifth one. In that one, skipping four lines left out the variable name row. To fix it, we’ll read the 2016 dataset again using read\_csv() and the fifth element of filenames but this time will use the argument skip = 3. We’ll assign the newly read dataset to the fifth element of the all\_files list:

Try printing all\_files now. You can confirm we fixed the problem by checking that the variable names are correct.

### Pick Variables

Now that we know all our datasets have the correct variable names, we simplify our datasets by picking the variables we need. This is a good place to think carefully about which variables to pick. This usually requires a fair amount of trial and error, but here is what we found we needed:

* Our analytic questions are about gender, so let’s pick the gender variable
* Later, we’ll need to filter our dataset by disability category and program location so we’ll want SEA Education Environment and SEA Disability Category
* We want to make comparisons by state and reporting year, so we’ll also pick State Name and Year

Combining select() and contains() is a convenient way to pick these variables without writing a lot of code. Knowing that we want variables that contain the acronym “SEA” and variables that contain “male” in their names, we can pass those characters to contains():

all\_files[[1]] %>%  
 select(  
 Year,  
 contains("State", ignore.case = FALSE),  
 contains("SEA", ignore.case = FALSE),  
 contains("male")  
 )

## # A tibble: 16,230 x 8  
## Year `State Name` `SEA Education … `SEA Disability… `Female Age 3 t…  
## <dbl> <chr> <chr> <chr> <chr>   
## 1 2012 ALABAMA Correctional Fa… All Disabilities -   
## 2 2012 ALABAMA Home All Disabilities 63   
## 3 2012 ALABAMA Homebound/Hospi… All Disabilities -   
## 4 2012 ALABAMA Inside regular … All Disabilities -   
## 5 2012 ALABAMA Inside regular … All Disabilities -   
## 6 2012 ALABAMA Inside regular … All Disabilities -   
## 7 2012 ALABAMA Other Location … All Disabilities 573   
## 8 2012 ALABAMA Other Location … All Disabilities 81   
## 9 2012 ALABAMA Parentally Plac… All Disabilities -   
## 10 2012 ALABAMA Residential Fac… All Disabilities 6   
## # … with 16,220 more rows, and 3 more variables: `Male Age 3 to 5` <chr>,  
## # `Female Age 6 to 21` <chr>, `Male Age 6 to 21` <chr>

That code chunk verifies that we got the variables we want, so now we will turn the code chunk into a function called pick\_vars(). We will then use map() to apply pick\_vars() to each dataset of our list, all\_files, to the function. In this function, we’ll use a special version of select() called select\_at(), which conveniently picks variables based on criteria we give it. The argument vars(Year, contains("State", ignore.case = FALSE), contains("SEA", ignore.case = FALSE), contains("male")) tells R we want to keep any column whose name has “State” in upper or lower case letters, has “SEA” in the title, and has “male” in the title. This will result in a newly transformed all\_files list that contains six datasets, all with the desired variables.

# build the function  
pick\_vars <-  
 function(df) {  
 df %>%  
 select\_at(vars(  
 Year,  
 contains("State", ignore.case = FALSE),  
 contains("SEA", ignore.case = FALSE),  
 contains("male")  
 ))  
 }  
  
# use the function with `all\_files`  
all\_files <-  
 all\_files %>%  
 map(pick\_vars)

### Combine Six Datasets into One

Now we’ll turn our attention to combining the datasets in our list all\_files into one. We’ll use bind\_rows(), which combines datasets by adding each one to the bottom of the one before it. The first step is to check and see if our datasets have the same number of variables and the same variable names. When we use names() on our list of newly changed datasets, we see that each dataset’s variable names are the same:

# check variable names  
all\_files %>%   
 map(names)

## [[1]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"   
##   
## [[2]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"   
##   
## [[3]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"   
##   
## [[4]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"   
##   
## [[5]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"   
##   
## [[6]]  
## [1] "Year" "State Name"   
## [3] "SEA Education Environment" "SEA Disability Category"   
## [5] "Female Age 3 to 5" "Male Age 3 to 5"   
## [7] "Female Age 6 to 21" "Male Age 6 to 21"

That means that we can combine all six datasets into one using bind\_rows(). We’ll call this newly combined dataset child\_counts:

child\_counts <-  
 all\_files %>%  
 # combine all datasets in `all\_files`  
 bind\_rows()

Since we know that a) each of our six datasets had eight variables and b) our combined dataset also has eight variables, we can conclude that all our rows combined together correctly. But let’s use str() to verify:

str(child\_counts)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 97387 obs. of 8 variables:  
## $ Year : num 2012 2012 2012 2012 2012 ...  
## $ State Name : chr "ALABAMA" "ALABAMA" "ALABAMA" "ALABAMA" ...  
## $ SEA Education Environment: chr "Correctional Facilities" "Home" "Homebound/Hospital" "Inside regular class 40% through 79% of day" ...  
## $ SEA Disability Category : chr "All Disabilities" "All Disabilities" "All Disabilities" "All Disabilities" ...  
## $ Female Age 3 to 5 : chr "-" "63" "-" "-" ...  
## $ Male Age 3 to 5 : chr "-" "174" "-" "-" ...  
## $ Female Age 6 to 21 : chr "4" "-" "104" "1590" ...  
## $ Male Age 6 to 21 : chr "121" "-" "130" "3076" ...

### Filter for the Desired Disabilities and Age Groups

We want to explore gender related variables, but our dataset has additional aggregate data for other subgroups. For example, we can use count() to explore all the different disability groups in the dataset. Here’s the number of times an SEA Disability Category appears in the dataset:

child\_counts %>%  
 # count number of times the category appears in the dataset  
 count(`SEA Disability Category`)

## # A tibble: 16 x 2  
## `SEA Disability Category` n  
## <chr> <int>  
## 1 All Disabilities 6954  
## 2 Autism 6954  
## 3 Deaf-blindness 6954  
## 4 Developmental delay 4636  
## 5 Developmental delay (valid only for children ages 3-9 when defined by … 2318  
## 6 Emotional disturbance 6954  
## 7 Hearing impairment 6954  
## 8 Intellectual disability 6954  
## 9 Multiple disabilities 6954  
## 10 Orthopedic impairment 6954  
## 11 Other health impairment 6954  
## 12 Specific learning disability 6954  
## 13 Speech or language impairment 6954  
## 14 Traumatic brain injury 6954  
## 15 Visual impairment 6954  
## 16 <NA> 31

Since we will be visualizing and modeling gender variables for all students in the dataset, we’ll filter out all subgroups except “All Disabilities” and the age totals:

child\_counts <-  
 child\_counts %>%  
 filter(  
 # filter all but the All Disabilities category  
 `SEA Disability Category` == "All Disabilities",  
 # filter all but the age totals  
 `SEA Education Environment` %in% c("Total, Age 3-5", "Total, Age 6-21")  
 )

### Rename the Variables

In the next section we’ll prepare the dataset for visualization and modeling by “tidying” it. When we write code to transform datasets, we’ll be typing the column names a lot, so it’s useful to change them to ones with more convenient names.

child\_counts <-  
 child\_counts %>%  
 rename(  
 # change these columns to more convenient names  
 year = Year,  
 state = "State Name",  
 age = "SEA Education Environment",  
 disability = "SEA Disability Category",  
 f\_3\_5 = "Female Age 3 to 5",  
 m\_3\_5 = "Male Age 3 to 5",  
 f\_6\_21 = "Female Age 6 to 21",  
 m\_6\_21 = "Male Age 6 to 21"  
 )

### Clean State Names

You might have noticed that some state names in our dataset are in upper case letters and some are in lower case letters:

child\_counts %>%  
 count(state) %>%  
 head()

## # A tibble: 6 x 2  
## state n  
## <chr> <int>  
## 1 Alabama 8  
## 2 ALABAMA 4  
## 3 Alaska 8  
## 4 ALASKA 4  
## 5 American Samoa 8  
## 6 AMERICAN SAMOA 4

If we leave it like this, R will treat state values like “CALIFORNIA” and “California” as two different states. We can use mutate and tolower to transform all the state names to lowercase letters.

child\_counts <-  
 child\_counts %>%  
 mutate(state = tolower(state))

### Tidy the Dataset

Visualizing and modeling our data will be much easier if our dataset is in a “tidy” format. In his paper *Tidy Data* (Wickham [2014](#ref-wickham2014)), Hadley Wickham defines tidy datasets where:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

*A note on the gender variable in this dataset*

This dataset uses a binary approach to data collection about gender. Students are described as either male or female. The need for an inclusive approach to documenting gender identity is discussed in a paper by (“Reachable: Data Collection Methods for Sexual Orientation and Gender Identity” [2016](#ref-park2016)) of The Williams Institute at UCLA.

The gender variables in our dataset are spread across four columns, with each one representing a combination of gender and age range. We can use pivot\_longer() to bring the gender variable into one column. In this transformation, we create two new columns: a gender column and a total column. The total column will contain the number of students in each row’s gender and age category.

child\_counts <-  
 child\_counts %>%  
 pivot\_longer(cols = f\_3\_5:m\_6\_21,   
 names\_to = "gender",   
 values\_to = "total")

To make the values of the gender column more intuitive, we’ll use case\_when() to transform the values to either “f” or “m”:

child\_counts <-   
 child\_counts %>%  
 mutate(  
 gender = case\_when(  
 gender == "f\_3\_5" ~ "f",  
 gender == "m\_3\_5" ~ "m",  
 gender == "f\_6\_21" ~ "f",  
 gender == "m\_6\_21" ~ "m",  
 TRUE ~ as.character(gender)  
 )  
 )

### Convert Data Types

The values in the total column represent the number of students from a specific year, state, gender, and age group. We know from the chr under their variable names that R is treating these values like characters instead of numbers. While R does a decent job of treating numbers like numbers when needed, it’s much safer to prepare the dataset by changing these character columns to numeric columns. We’ll use mutate() to change the count columns.

child\_counts <-  
 child\_counts %>%  
 mutate(total = as.numeric(total))  
  
child\_counts

## # A tibble: 2,928 x 6  
## year state age disability gender total  
## <dbl> <chr> <chr> <chr> <chr> <dbl>  
## 1 2012 alabama Total, Age 3-5 All Disabilities f 2228  
## 2 2012 alabama Total, Age 3-5 All Disabilities m 5116  
## 3 2012 alabama Total, Age 3-5 All Disabilities f NA  
## 4 2012 alabama Total, Age 3-5 All Disabilities m NA  
## 5 2012 alabama Total, Age 6-21 All Disabilities f NA  
## 6 2012 alabama Total, Age 6-21 All Disabilities m NA  
## 7 2012 alabama Total, Age 6-21 All Disabilities f 23649  
## 8 2012 alabama Total, Age 6-21 All Disabilities m 48712  
## 9 2012 alaska Total, Age 3-5 All Disabilities f 676  
## 10 2012 alaska Total, Age 3-5 All Disabilities m 1440  
## # … with 2,918 more rows

Converting these count columns from character classes to number classes resulted in two changes. First, the chr under these variable names has now changed to dbl, short for “double-precision”. This lets us know that R recognizes these values as numbers with decimal points. Second, the blank values changed to NA. When R sees a character class value like "4", it knows to change it to numeric class 4. But there is no obvious number represented by a value like "" or -, so it changes it to NA:

# Convert a character to a number   
as.numeric("4")

## [1] 4

# Convert a blank character or a symbol to a number  
as.numeric("")

## [1] NA

as.numeric("-")

## [1] NA

Similarly, the variable year needs to be changed from the character format to the date format. Doing so will make sure R treats this variable like a point in time when we plot our dataset. The package {lubridate} has a handy function called ymd that can help us. We just have to use the truncated argument to let R know we don’t have a month and date to convert.

child\_counts <-  
 child\_counts %>%  
 mutate(year = ymd(year, truncated = 2))

### Explore and Address NAs

You’ll notice that some rows in the total column contain an NA. When we used pivot\_longer() to create a gender column, R created unique rows for every year, state, age, disability, and gender combination. Since the original dataset had both gender and age range stored in a column like Female Age 3 to 5, R made rows where the total value is NA . For example, there is no student count for the age value “Total, Age 3-5” that also has the gender value for female students who were age 6-21. You can see that more clearly by sorting the dataset by year, state, and gender.

In our Foundational Skills chapter, we introduced a {dplyr} function called arrange() to sort the rows of a dataset by the values in a column. Let’s use arrange() here to sort the dataset by the year, state and gender columns. When you pass arrange() a variable, it will sort by the order of the values in that variable. If you pass it multiple variables, arrange() will sort by the first variable, then by the second, and so on. Let’s see what it does on child\_counts when we pass it the year, state, and gender variables:

child\_counts %>%  
 arrange(year, state, gender)

## # A tibble: 2,928 x 6  
## year state age disability gender total  
## <date> <chr> <chr> <chr> <chr> <dbl>  
## 1 2012-01-01 alabama Total, Age 3-5 All Disabilities f 2228  
## 2 2012-01-01 alabama Total, Age 3-5 All Disabilities f NA  
## 3 2012-01-01 alabama Total, Age 6-21 All Disabilities f NA  
## 4 2012-01-01 alabama Total, Age 6-21 All Disabilities f 23649  
## 5 2012-01-01 alabama Total, Age 3-5 All Disabilities m 5116  
## 6 2012-01-01 alabama Total, Age 3-5 All Disabilities m NA  
## 7 2012-01-01 alabama Total, Age 6-21 All Disabilities m NA  
## 8 2012-01-01 alabama Total, Age 6-21 All Disabilities m 48712  
## 9 2012-01-01 alaska Total, Age 3-5 All Disabilities f 676  
## 10 2012-01-01 alaska Total, Age 3-5 All Disabilities f NA  
## # … with 2,918 more rows

We can simplify our dataset by removing the rows with NAs, leaving us with one row for each category:

* females age 3-5
* females age 6-21
* males age 3-5
* males age 6-21

Each of these categories will be associated with a state and reporting year:

child\_counts <-  
 child\_counts %>%  
 filter(!is.na(total))

We can verify we have the categories we want by sorting again:

child\_counts %>%  
 arrange(year, state, gender)

## # A tibble: 1,390 x 6  
## year state age disability gender total  
## <date> <chr> <chr> <chr> <chr> <dbl>  
## 1 2012-01-01 alabama Total, Age 3-5 All Disabilities f 2228  
## 2 2012-01-01 alabama Total, Age 6-21 All Disabilities f 23649  
## 3 2012-01-01 alabama Total, Age 3-5 All Disabilities m 5116  
## 4 2012-01-01 alabama Total, Age 6-21 All Disabilities m 48712  
## 5 2012-01-01 alaska Total, Age 3-5 All Disabilities f 676  
## 6 2012-01-01 alaska Total, Age 6-21 All Disabilities f 5307  
## 7 2012-01-01 alaska Total, Age 3-5 All Disabilities m 1440  
## 8 2012-01-01 alaska Total, Age 6-21 All Disabilities m 10536  
## 9 2012-01-01 american samoa Total, Age 3-5 All Disabilities f 45  
## 10 2012-01-01 american samoa Total, Age 6-21 All Disabilities f 208  
## # … with 1,380 more rows

## Analysis

In the last section we focused on importing our dataset. In this section, we will ask: how have child counts changed over time? First, we’ll use visualization to explore the number of students in special education over time. In particular, we’ll compare the count of male and female students. Next, we’ll use what we learn from our visualizations to quantify any differences that we see.

### Visualize the Dataset

Showing this many states in a plot can be overwhelming, so to start we’ll make a subset of the dataset. We can use a function in the {dplyr} package called top\_n() to help us learn which states have the highest mean count of students with disabilities:

child\_counts %>%  
 group\_by(state) %>%  
 summarize(mean\_count = mean(total)) %>%  
 # which six states have the highest mean count of students with disabilities  
 top\_n(6, mean\_count)

## # A tibble: 6 x 2  
## state mean\_count  
## <chr> <dbl>  
## 1 california 180879.  
## 2 florida 92447.  
## 3 new york 121751.  
## 4 pennsylvania 76080.  
## 5 texas 115593.  
## 6 us, outlying areas, and freely associated states 1671931.

These six states have the highest mean count of students in special education over the six years we are examining. For reasons we will see in a later visualization, we are going to exclude outlying areas and freely associated states. That leaves us us with five states: California, Florida, New York, Pennsylvania, and Texas. We can remove all other states but these by using filter(). We’ll call this new dataset high\_count:

high\_count <-  
 child\_counts %>%  
 filter(state %in% c("california", "florida", "new york", "pennsylvania", "texas"))

Now we can use high\_count to do some initial exploration. Our analysis is about comparing counts of male and female students in special education, but visualization is also a great way to explore related curiosities. You may surprise yourself with what you find when visualizing your datasets. You might come up with more interesting hypotheses, find that your initial hypothesis requires more data transformation, or find interesting subsets of the data–we saw a little of that in the surprisingly high mean\_count of freely associated states in the state column. Let your curiosity and intuition drive this part of the analysis. It’s one of the activities that makes data analysis a creative process.

In that spirit, we’ll start by visualizing specific genders and age groups. Feel free to try these, but also try the other student groups for practice and more exploration.

Start by copying and running this code in your console to see what it does:

high\_count %>%  
 filter(gender == "f", age == "Total, Age 6-21") %>%  
 ggplot(aes(x = year, y = total, color = state)) +  
 geom\_freqpoly(stat = "identity", size = 1) +  
 labs(title = "Count of Female Students in Special Education Over Time",  
 subtitle = "Ages 6-21") +  
 scale\_color\_dataedu() +  
 theme\_dataedu()

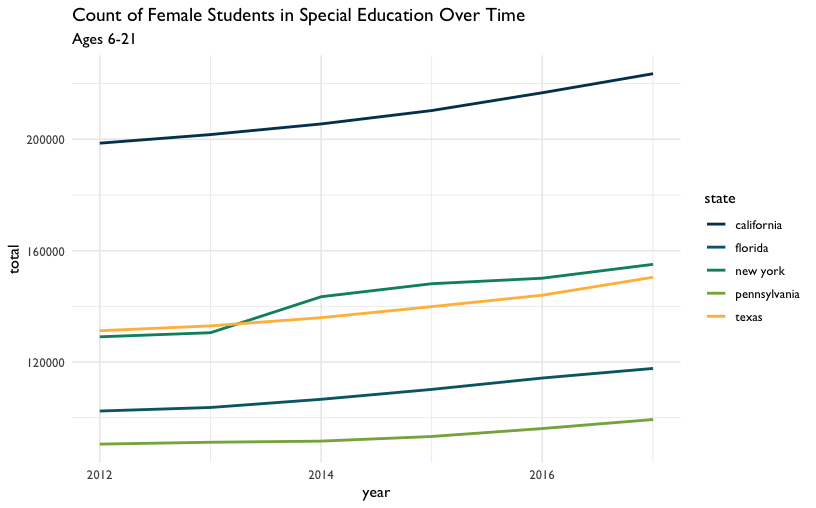


Figure 26: Count of Female Students in Special Education Over Time

That gives us a plot that has the years in the x-axis and a count of female students in the y-axis. Each line takes a different color based on the state it represents.

Let’s look at that closer: We used filter() to subset our dataset for students who are female and ages 6 to 21. We used aes to connect visual elements of our plot to our data. We connected the x-axis to year, the y-axis to total, and the color of the line to state.

It’s worth calling out one more thing, since it’s a technique we’ll be using as we explore further. Note here that, instead of storing our new dataset in a new variable, we filter the dataset then use the pipe operator %>% to feed it to {ggplot}. Since we’re exploring freely, we don’t need to create a lot of new variables we probably won’t need later.

We can also try the same plot, but subsetting for male students instead. We can use the same code we used for the last plot, but filter for the value “m” in the gender field:

high\_count %>%  
 filter(gender == "m", age == "Total, Age 6-21") %>%  
 ggplot(aes(x = year, y = total, color = state)) +  
 geom\_freqpoly(stat = "identity", size = 1) +  
 labs(title = "Count of Male Students in Special Education Over Time",  
 subtitle = "Ages 6-21") +  
 scale\_color\_dataedu() +  
 theme\_dataedu()

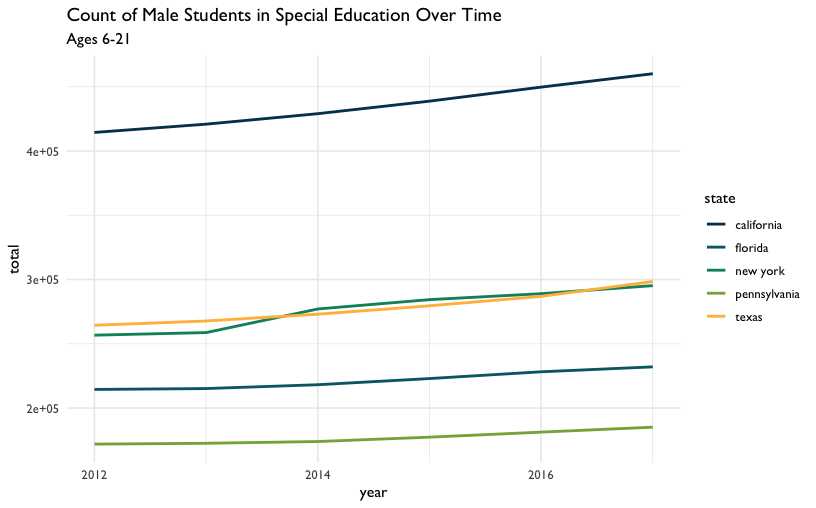


Figure 27: Count of Male Students in Special Education Over Time

We’ve looked at each gender separately. What do these lines look like if we visualized the total amount of students each year per state? To do that, we’ll need to add both gender values together and both age group values together. We’ll do this using a very common combination of functions: group\_by() and summarize().

high\_count %>%  
 group\_by(year, state) %>%  
 summarize(n = sum(total)) %>%  
 ggplot(aes(x = year, y = n, color = state)) +  
 geom\_freqpoly(stat = "identity", size = 1) +  
 labs(title = "Total Count of Students in Special Education Over Time",  
 subtitle = "Ages 3-21") +  
 scale\_color\_dataedu() +  
 theme\_dataedu()

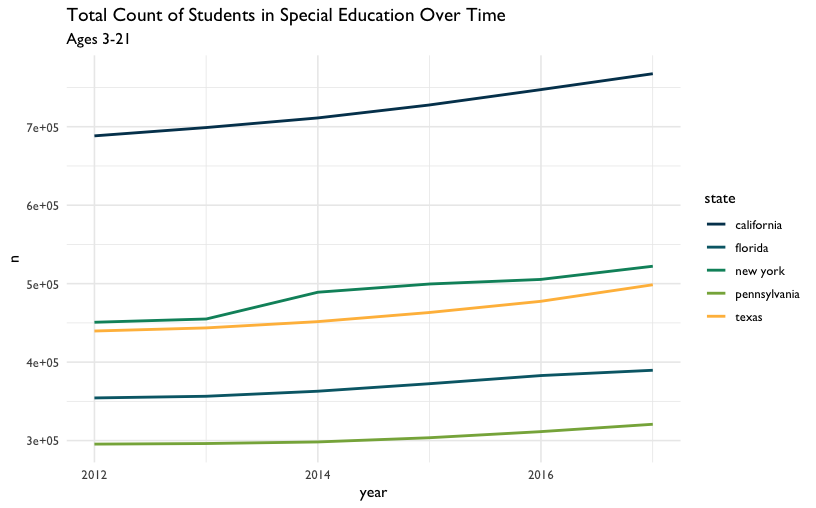


Figure 28: Total Count of Students in Special Education Over Time

So far we’ve looked at a few ways to count students over time. In each plot, we see that while counts have grown overall for all states, each state has different sized populations. Let’s see if we can summarize that difference by looking at the median student count for each state over the years:

high\_count %>%  
 group\_by(year, state) %>%  
 summarize(n = sum(total)) %>%  
 ggplot(aes(x = state, y = n)) +  
 geom\_boxplot(fill = dataedu\_colors("yellow")) +  
 labs(title = "Median Students with Disabilities Count",  
 subtitle = "All ages and genders, 2012-2017") +  
 theme\_dataedu()

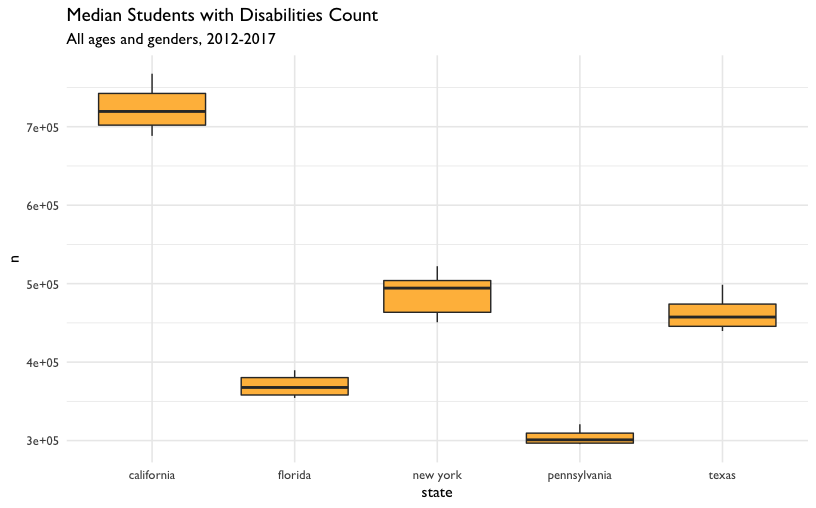


Figure 29: Median Students with Disabilities Count

The boxplots show us what we might have expected from our freqpoly plots before it. The highest median student count over time is California and the lowest is Pennsylvania.

What have we learned about our data so far? The five states in the US with the highest total student counts (not including outlying areas and freely associated states) do not have similar counts to each other. The student counts for each state also appear to have grown over time.

But how can we start comparing the male student count to the female student count? One way is to use a ratio:

In mathematics, a ratio is a relationship between two numbers indicating how many times the first number contains the second. Wikipedia

We can use the count of male students in each state and divide it by the count of each female student. The result is the number of times male students are in special education more or less than the female students in the same state and year. Our coding strategy will be to

* Use pivot\_wider() to create separate columns for male and female students.
* Use mutate() to create a new variable called ratio. The values in this column will be the result of dividing the count of male students by the count of female students

Note here that we can also accomplish this comparison by dividing the number of female students by the number of male students. In this case, the result would be the number of times female students are in special education more or less than male students.

high\_count %>%  
 group\_by(year, state, gender) %>%  
 summarize(total = sum(total)) %>%  
 # Create new columns for male and female student counts  
 pivot\_wider(names\_from = gender,   
 values\_from = total) %>%   
 # Create a new ratio column  
 mutate(ratio = m / f) %>%  
 ggplot(aes(x = year, y = ratio, color = state)) +  
 geom\_freqpoly(stat = "identity", size = 1) +  
 scale\_y\_continuous(limits = c(1.5, 2.5)) +  
 labs(title = "Male Student to Female Student Ratio Over Time",  
 subtitle = "Ages 6-21") +  
 scale\_color\_dataedu() +  
 theme\_dataedu()

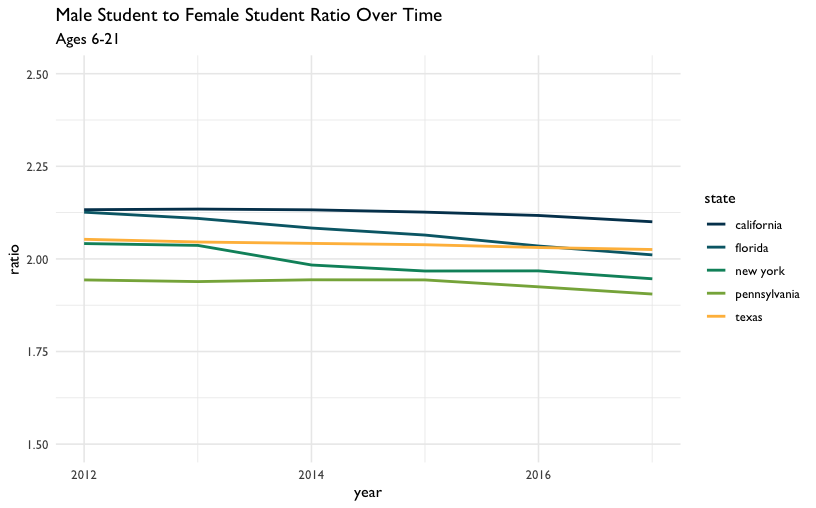


Figure 30: Male Student to Female Student Ratio Over Time

By visually inspecting, we can hypothesize that there was no significant change in the male to female ratio between the years 2012 and 2017. But very often we want to understand the underlying properties of our education dataset. We can do this by quantifying the relationship between two variables. In the next section, we’ll explore ways to quantify the relationship between male student counts and female student counts.

### Model the Dataset

When you visualize your datasets, you are exploring possible relationships between variables. But sometimes visualizations can be misleading because of the way we perceive graphics. In his book *Data Visualization: A Practical Introduction*, Healy ([2019](#ref-healy2019)) teaches us that

Visualizations encode numbers in lines, shapes, and colors. That means that our interpretation of these encodings is partly conditional on how we perceive geometric shapes and relationships generally.

What are some ways we can combat these errors of perception and at the same time draw substantive conclusions about our education dataset? When you spot a possible relationship between variables, the relationship between female and male counts for example, you’ll want to quantify that relationship by fitting a statistical model. Practically speaking, this means you are selecting a distribution that represents your dataset reasonably well. This distribution will help you quantify and predict relationships between variables. This is an important step in the analytic process, because it acts as a check on what you saw in your exploratory visualizations.

In this example, we’ll follow our intuition about the relationship between male and female student counts in our special education dataset. In particular, we’ll test the hypothesis that this ratio has decreased over the years. Fitting a linear regression model that estimates the year as a predictor of the male to female ratio will help us do just that.

*Do we have enough information for our model?*

At the start of this section, we chose to exclude outlying areas and freely associated states. This visualization suggests that there are some states that have a child count so high it leaves a gap in the x-axis values. This can be problematic when we try to interpret our model later. Here’s a plot of female students compared to male students. Note that the relationship appears linear, but there is a large gap in the distribution of female student counts somewhere between the values of 250,000 and 1,750,000:

child\_counts %>%  
 filter(age == "Total, Age 6-21") %>%  
 pivot\_wider(names\_from = gender,   
 values\_from = total) %>%  
 ggplot(aes(x = f, y = m)) +  
 geom\_point(size = 3, alpha = .5, color = dataedu\_colors("green")) +  
 geom\_smooth() +  
 labs(  
 title = "Comparison of Female Students to Male Students in Special Education",  
 subtitle = "Counts of students in each state, ages 6-21",  
 x = "Female students",  
 y = "Male students",  
 caption = "Data: US Dept of Education"  
 ) +  
 theme\_dataedu()

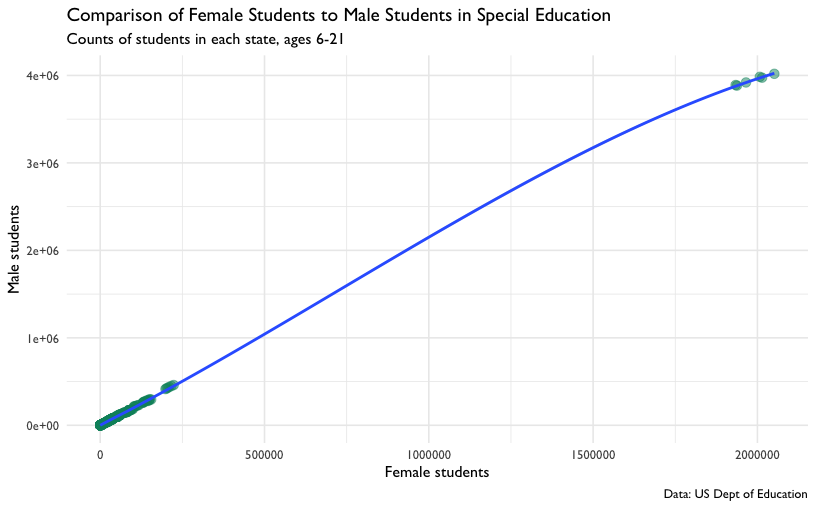


Figure 31: Comparison of Female Students to Male Students in Special Education

If you think of each potential point on the linear regression line as a ratio of male to female students, you’ll notice that we don’t know a whole lot about what happens in states where there are between 250,000 and 1,750,000 female students in any given year.

To learn more about what’s happening in our dataset, we can filter it for only states that have more than 500,000 female students in any year:

child\_counts %>%  
 filter(age == "Total, Age 6-21") %>%  
 pivot\_wider(names\_from = gender,   
 values\_from = total) %>%   
 filter(f > 500000) %>%  
 select(year, state, age, f, m)

## # A tibble: 6 x 5  
## year state age f m  
## <date> <chr> <chr> <dbl> <dbl>  
## 1 2012-01-01 us, outlying areas, and freely associ… Total, Age 6… 1933619 3.89e6  
## 2 2013-01-01 us, outlying areas, and freely associ… Total, Age 6… 1937726 3.88e6  
## 3 2014-01-01 us, outlying areas, and freely associ… Total, Age 6… 1965204 3.92e6  
## 4 2015-01-01 us, outlying areas, and freely associ… Total, Age 6… 2007174 3.98e6  
## 5 2016-01-01 us, outlying areas, and freely associ… Total, Age 6… 2014120 3.97e6  
## 6 2017-01-01 us, outlying areas, and freely associ… Total, Age 6… 2051438 4.02e6

This is where we discover that each of the data points in the upper right hand corner of the plot are from the state value “us, us, outlying areas, and freely associated states”. If we remove these outliers, we have a distribution of female students that looks more complete.

child\_counts %>%  
 filter(age == "Total, Age 6-21") %>%  
 pivot\_wider(names\_from = gender,   
 values\_from = total) %>%  
 # Filter for female student counts less than 500,000  
 filter(f <= 500000) %>%  
 ggplot(aes(x = f, y = m)) +  
 geom\_point(size = 3, alpha = .5, color = dataedu\_colors("green")) +  
 labs(  
 title = "Comparison of Female Students to Male Students with Disabilities",  
 subtitle = "Counts of students in each state, ages 6-21.\nDoes not include outlying areas and freely associated states",  
 x = "Female students",  
 y = "Male students",  
 caption = "Data: US Dept of Education"  
 ) +  
 theme\_dataedu()

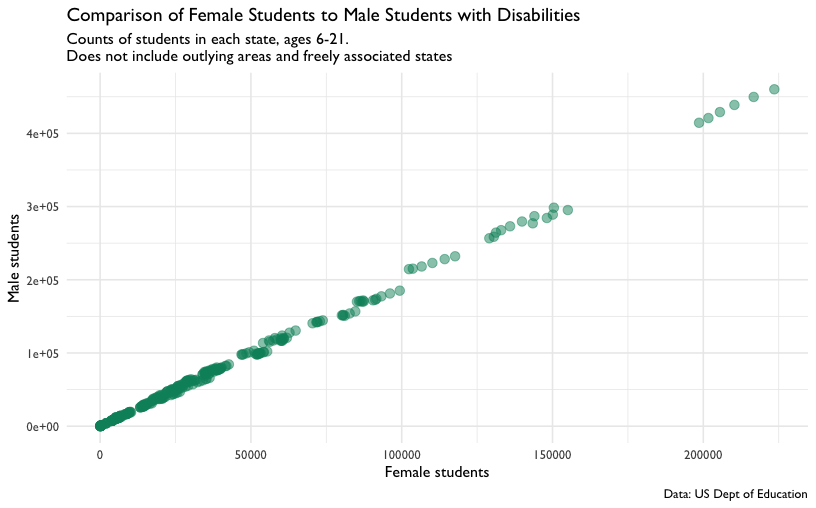


Figure 32: Comparison of Female Students to Male Students with Disabilities

This should allow us to fit a better model for the relationship between male and female student counts, albeit only the ones where the count of female students takes a value between 0 and 500,000.

*Male to Female Ratio Over Time*

Earlier we asked the question Do we have enough data points for the count of female students to learn about the ratio of female to male students? Similarly, we should ask the question Do we have enough data points across our year variable to learn about how this ratio has changed over time?

To answer that question, let’s start by making a new dataset that excludes any rows where the f variables has a value that is less than or equal to 500000. We’ll convert the year variable to a factor data type–we’ll see how this helps in a bit. We’ll also add a column called ratio that contains the male to female count ratio.

model\_data <- child\_counts %>%  
 filter(age == "Total, Age 6-21") %>%  
 mutate(year = as.factor(year(year))) %>%  
 pivot\_wider(names\_from = gender,   
 values\_from = total) %>%   
 # Exclude outliers  
 filter(f <= 500000) %>%  
 # Compute male student to female student ratio  
 mutate(ratio = m / f) %>%  
 select(-c(age, disability))

We can see how much data we have per year by using count():

model\_data %>%  
 count(year)

## # A tibble: 6 x 2  
## year n  
## <fct> <int>  
## 1 2012 59  
## 2 2013 56  
## 3 2014 56  
## 4 2015 58  
## 5 2016 57  
## 6 2017 55

Let’s visualize the ratio values across all years as an additional check. Note the use of geom\_jitter() to spread the points horizontally so we can estimate the quantities better:

ggplot(data = model\_data, aes(x = year, y = ratio)) +  
 geom\_jitter(alpha = .5, color = dataedu\_colors("green")) +  
 labs(title = "Male to Female Ratio Across Years (Jittered)") +  
 theme\_dataedu()

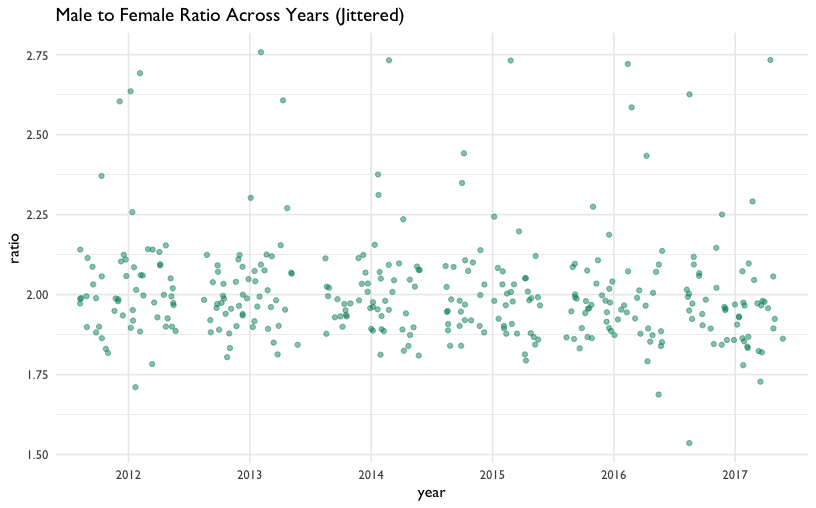


Figure 33: Male to Female Ratio Across Years (Jittered)

Each year seems to have data points that can be considered when we fit the model. This means that there are enough data points to help us learn how the year variable predicts the ratio variable.

We fit the linear regression model by passing the argument ratio ~ year to the function lm(). In R, the ~ usually indicates a formula. In this case, the formula is the variable year as a predictor of the variable ratio. The final argument we pass to lm is data = model\_data, which tells R to look for the variables ratio and year in the dataset model\_data. The results of the model are called a “model object.” We’ll store the model object in ratio\_year:

ratio\_year <-   
 lm(ratio ~ year, data = model\_data)

Each model object is filled with all sorts of model information. We can look at this information using the function summary():

summary(ratio\_year)

##   
## Call:  
## lm(formula = ratio ~ year, data = model\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4402 -0.1014 -0.0281 0.0534 0.7574   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.0336 0.0220 92.42 <2e-16 \*\*\*  
## year2013 -0.0120 0.0315 -0.38 0.70   
## year2014 -0.0237 0.0315 -0.75 0.45   
## year2015 -0.0310 0.0313 -0.99 0.32   
## year2016 -0.0396 0.0314 -1.26 0.21   
## year2017 -0.0576 0.0317 -1.82 0.07 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.169 on 335 degrees of freedom  
## Multiple R-squared: 0.0122, Adjusted R-squared: -0.00259   
## F-statistic: 0.824 on 5 and 335 DF, p-value: 0.533

Here’s how we can interpret the Estimate column: The estimate of the (Intercept) is 2.03356, which is the estimated value of the ratio variable when the year variable is “2012”. Note that the value year2012 isn’t present in the in the list of rownames. That’s because the (Intercept) row represents year2012. In linear regression models that use factor variables as predictors, the first level of the factor is the intercept. Sometimes this level is called a “dummy variable”. The remaining rows of the model output show how much each year differs from the intercept, 2012. For example, year2013 has an estimate of -0.01205, which suggests that on average the value of ratio is .01205 less than 2.03356. On average, the ratio of year2014 is .02372 less than 2.03356. The t value column tells us the size of difference between the estimated value of the ratio for each year and the estimated value of the ratio of the intercept. Generally speaking, the larger the t value, the larger the chance that any difference between the coefficient of a factor level and the intercept are significant.

Though the relationship between year as a predictor of ratio is not linear (recall our previous plot), the linear regression model still gives us useful information. We fit a linear regression model to a factor variable, like year, as a predictor of a continuous variable, likeratio. In doing so, we got the average ratio at every value of year. We can verify this by taking the mean ratio of ever year:

model\_data %>%  
 group\_by(year) %>%  
 summarize(mean\_ratio = mean(ratio))

## # A tibble: 6 x 2  
## year mean\_ratio  
## <fct> <dbl>  
## 1 2012 2.03  
## 2 2013 2.02  
## 3 2014 2.01  
## 4 2015 2.00  
## 5 2016 1.99  
## 6 2017 1.98

This verifies that our intercept, the value of ratio during the year 2012, is 2.033563 and the value of ratio for 2013 is .01205 less than that of 2012 on average. Fitting the model gives us more details about these mean ratio scores– namely the coefficient, t value, and p value. These values help us apply judgement when deciding if differences in ratio values suggest an underlying difference between years or simply differences you can expect from randomness. In this case, the absence of "\*" in all rows except the Intercept row suggest that any differences occurring between years are within the range you’d expect by chance.

If we use summary() on our model\_data dataset, we can verify the intercept again:

model\_data %>%  
 filter(year == "2012") %>%  
 summary()

## year state f m ratio   
## 2012:59 Length:59 Min. : 208 Min. : 443 Min. :1.71   
## 2013: 0 Class :character 1st Qu.: 5606 1st Qu.: 11467 1st Qu.:1.93   
## 2014: 0 Mode :character Median : 22350 Median : 44110 Median :1.99   
## 2015: 0 Mean : 32773 Mean : 65934 Mean :2.03   
## 2016: 0 3rd Qu.: 38552 3rd Qu.: 77950 3rd Qu.:2.09   
## 2017: 0 Max. :198595 Max. :414466 Max. :2.69

The mean value of the ratio column when the year column is 2012 is 2.034, just like in the model output’s intercept row.

Lastly, we may want to communicate to a larger audience that there were roughly twice amount of male students in this dataset than female students and that this did not change significantly between the years 2012 and 2017. When you are not communicating to an audience of other data scientists, it’s helpful to illustrate your point without the technical details of the model output. Think of yourself as an interpreter: since you can speak the language of model outputs and the language of data visualization, your challenge is to take what you learned from the model output and tell that story in a way that is meaningful to your non-data scientist audience.

There are many ways to do this, but we’ll choose boxplots to show our audience that there was roughly twice as many male students in special education than female students between 2012 and 2017. For our purposes, let’s verify this by looking at the median male to female ratio for each year:

model\_data %>%  
 group\_by(year) %>%  
 summarize(median\_ratio = median(ratio))

## # A tibble: 6 x 2  
## year median\_ratio  
## <fct> <dbl>  
## 1 2012 1.99  
## 2 2013 1.99  
## 3 2014 1.98  
## 4 2015 1.98  
## 5 2016 1.97  
## 6 2017 1.96

Now let’s visualize this using our boxplots:

model\_data %>%  
 pivot\_longer(cols = c(f, m),   
 names\_to = "gender",   
 values\_to = "students") %>%   
 ggplot(aes(x = year, y = students, color = gender)) +  
 geom\_boxplot() +  
 scale\_y\_continuous(labels = scales::comma) +  
 labs(  
 title = "Median Male and Female Student Counts in Special Education",  
 subtitle = "Ages 6-21. Does not include outlying areas and freely associated states",  
 x = "",  
 y = "",  
 caption = "Data: US Dept of Education"  
 ) +  
 scale\_color\_dataedu() +  
 theme\_dataedu()

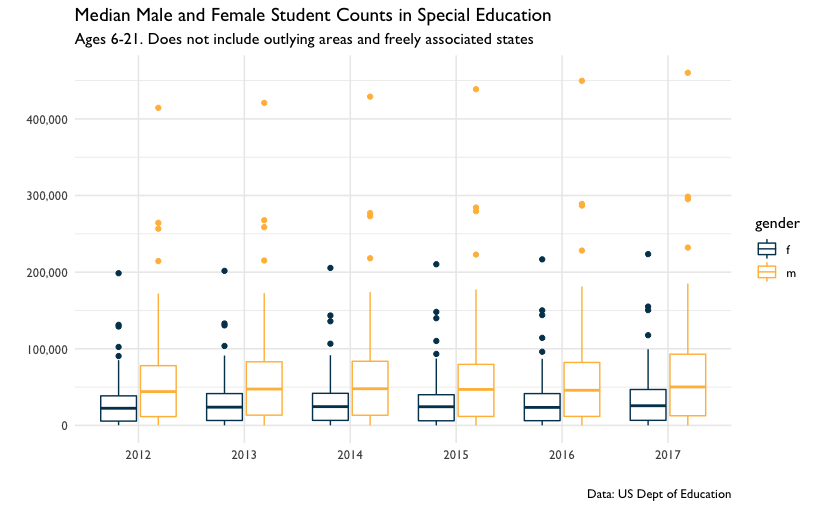


Figure 34: Median Male and Female Student Counts in Special Education

Once we learned from our model that male to female ratios did not change in any meaningful way from 2012 to 2017 and that the median ratio across states was about 2 male students to every female student, we can present these two ideas using this plot. When discussing the plot, it helps to have your model output in your notes so you can reference specific coefficient estimates when needed.

## Results

We learned that each state has a different count of students with disabilities–so different that we need to use statistics like ratios or visualizations to compare across states. Even when we narrow our focus to the five states with the highest counts of students with disabilities, we see that there are differences in these counts.

When we look at these five states over time, we see that despite the differences in total count each year, all five states have increased their student counts. We also learned that though the male to female ratios for students with disabilities appears to have gone down slightly over time, our model suggests that these decreases do not represent a big difference.

The comparison of student counts across each state is tricky because there is a lot of variation in total enrollment across all fifty states. While we explored student counts across each state and verified that there is variation in the counts, a good next step would be to combine these data with total enrollment data. This would allow us to compare counts of students with disabilities as a percentage of total enrollment. Comparing proportions like this is a common way to compare subgroups of a population across states when each state’s population varies in size.

## Conclusion

Education data science is about using data science tools to learn about and improve the lives of our students. So why choose a publicly available aggregate dataset instead of a student-level dataset? We chose to use an aggregate dataset because it reflects an analysis that a data scientist in education would typically do.

Using student-level data requires that the data scientist is either an employee of the school agency or that works under a memorandum of understanding (MOU) that allows her to access this data. Without either of these conditions, the education data scientist learns about the student experience by working on publicly available datasets, almost all of which are aggregated student-level datasets.

Before we discuss the benefits of aggregate data, let’s take some time to understand the differences between aggregate and student-level data. Publicly available data–like the US Federal Government special education student count we used in this walkthrough–is a summary of student-level data. That means that student-level data is totaled up to protect the identities of students before making them publicly available. We can use R to demonstrate this concept.

Here are rows in a student-level dataset:

# student-level data   
tibble(  
 student = letters[1:10],  
 school = rep(letters[11:15], 2),  
 test\_score = sample(0:100, 10, replace = TRUE)  
)

## # A tibble: 10 x 3  
## student school test\_score  
## <chr> <chr> <int>  
## 1 a k 70  
## 2 b l 70  
## 3 c m 22  
## 4 d n 87  
## 5 e o 76  
## 6 f k 31  
## 7 g l 99  
## 8 h m 72  
## 9 i n 92  
## 10 j o 86

Aggregate data totals up a variable–the variable test\_score in this case–to “hide” the student-level information. The rows of the resulting dataset represent a group. The group in our example is the school variable:

tibble(  
 student = letters[1:10],  
 school = rep(letters[11:15], 2),  
 test\_score = sample(0:100, 10, replace = TRUE)  
) %>%  
 # Aggregate by school  
 group\_by(school) %>%  
 summarize(mean\_score = mean(test\_score))

## # A tibble: 5 x 2  
## school mean\_score  
## <chr> <dbl>  
## 1 k 65   
## 2 l 50   
## 3 m 55.5  
## 4 n 37.5  
## 5 o 75.5

Notice here that this dataset no longer identifies individual students.

*Student-level data for analysis of local populations. Aggregate data for base rate and context.*

Longitudinal analysis is typically done with student-level data because educators are interested in what happens to students over time. So if you cannot access student-level data, how do we use aggregate data to offer value to the analytic conversation?

Aggregate data is valuable because it allows us to learn from populations that are larger or different from the local student-level population. Think of it as an opportunity to learn from totaled up student data from other states or the whole country.

In the book *Thinking Fast and Slow*, Kahneman ([2011](#ref-kahneman2011)) discusses the importance of learning from larger populations, a context he refers to as the base rate. The base rate fallacy is the tendency to only focus on conclusions we can draw from immediately available information. It’s the difference between computing how often a student at one school is identified for special education services (student-level data) and how often students are identified for special educations services nationally (base rate data). We can use aggregate data to combat the baserate fallacy by putting what we learn from local student data in the context of surrounding populations.

For example, consider an analysis of student-level data in a school district over time. Student-level data allows us to ask questions about our local population: One such question is: Are the rates of special education identification for male students different from other gender identities *in our district*? This style of question looks *inward* at your own educational system.

Taking a cue from Daniel Kahneman, we should also ask what this pattern looks like in other states or in the country. Aggregate data allows us to ask questions about a larger population: One such question is Are the rates of special education identification for male students different from other gender identities *in the United States*? This style of question looks for answers *outside* your own educational system. The combination of the two lines of inquiry are powerful way to generate new knowledge about the student experience.

So education data scientists should not despair in situations where they cannot access student-level data. Aggregate data is a powerful way to learn from state level or national level data when a data sharing agreement for student-level data is not possible. In situations where student-level data *is* available, including aggregate data is an excellent way to combat the base rate fallacy.

# Walkthrough 5: Text Analysis With Social Media Data

## Vocabulary

* anti\_join
* RDS files
* sample\_n()
* set.seed
* stop words
* slice
* token
* tokenize

## Chapter Overview

The ability to work with many kinds of datasets is one of the great features of doing data science with programming. So far we’ve analyzed data in .csv files, but that’s not the only way data is stored. If we can learn some basic techniques for analyzing text, we increase the number of places we can find information to learn about the student experience.

In this chapter, we focus on analyzing data from a data visualization-focused learning community on Twitter, #tidytuesday. That community, #tidytuesday is one sparked by the work of one of the Data Science in Education Using R co-authors, Jesse Mostipak, who created the #r4ds community from which #tidytuesday was created. #tidytuesday is a weekly data visualization challenge. A great place to see examples from past #tidytuesday challenges is an interactive Shiny application (<https://github.com/nsgrantham/tidytuesdayrocks>).

We note that while we focused on #tidytuesday because we think it exemplifies the new kinds of learning-related data that a data science toolkit allows an analyst to try to understand, we also chose this because it is straightforward to access data from Twitter, and - due to the presence of an interactive Shiny application - because it is particularly easy to access data on #tidytuesday. Relatedly, this chapter includes a technical appendix which is not necessary to read to understand text analysis, but which describes a number of techniques for accessing data from Twitter, such as the data from the #tidytuesday explored in this and the subsequent chapter.

### Background

When we think about data science in education, our minds tends to go data in spreadsheets. But what can we learn about the student experience from text data? Take a moment to mentally review all the moments in your work day that you generated or consumed text data. In education, we’re surrounded by it. We do our lessons in word processor documents, our students submit assignments online, and the school community expresses themselves on public social media platforms. The text we generate can be an authentic reflection of reality in schools, so how might we learn from it?

Even the most basic text analysis techniques will expand your data science toolkit. For example, you can use text analysis to count the number of key words that appear in open ended survey responses. You can analyze word patterns in student responses or message board posts.

Analyzing a collection of text is different from analyzing large numerical datasets because words don’t have agreed upon values the way numbers do. The number 2 will always be more than 1 and less than 3. The word “fantastic,” on the other hand, has multiple ambiguous levels of degree depending on interpretation and context.

Using text analysis can help to broadly estimate what is happening in the text. When paired with observations, interviews, and close review of the text, this approach can help education staff learn from text data. In this chapter, we’ll learn how to count the frequency of words in a dataset and associate those words with common feelings like positivity or joy.

We’ll show these techniques using a dataset of tweets. We encourage you to complete the walkthrough, then reflect on how the skills learned can be applied to other texts, like word processing documents or websites.

### Data Source

It’s useful to learn these techniques from text datasets that are available for download. Take a moment to do an online search for “download tweet dataset” and note the abundance of tweet datasets available. Since there’s so much, it’s useful to narrow the tweets to a only those that help you answer your analytic questions. Hashtags are text within a tweet that act as a way to categorize content. Here’s an example:

RT @CKVanPay: I’m trying to recreate some Stata code in R, anyone have a good resource for what certain functions in Stata are doing? #RStats #Stata

Twitter recognizes any words that start with a “#” as a hashtag. The hashtags “#RStats” and “#Stata” make this tweet conveniently searchable. If Twitter uses search for “#RStats”, Twitter returns all the Tweets containing that hashtag.

In this example, we’ll be analyzing a dataset of tweets that have the hashtag #tidytuesday (<https://twitter.com/hashtag/tidytuesday>). #tidytuesday is a community sparked by the work of one of the *Data Science in Education Using R* co-authors, Jesse Mostipak, who created the (related) #r4ds community from which #tidytuesday was created. #tidytuesday is a weekly data visualization challenge. A great place to see examples from past #tidytuesday challenges is an interactive Shiny application (<https://github.com/nsgrantham/tidytuesdayrocks>).

The #tidytuesday hashtag (search Twitter for the hashtag, or see the results here: <http://bit.ly/tidytuesday-search>) returns tweets about the weekly TidyTuesday ritual, where folks learning R create and tweet data visualizations they made while learning to use tidyverse R packages.

You can view the TidyTuesday tweets dataset here: <http://bit.ly/tidytuesday-dataset>

### Methods

In this walkthrough, we’ll be learning how to count words in a text dataset. We’ll also use a technique called sentiment analysis to count and visualize the appearance of words that have a positive association. Lastly, we’ll learn how to get more context by selecting random rows of tweets for closer reading.

## Load Packages

For this analysis, we’ll be using the {tidyverse}, {here}, and {dataedu} packages. We will also use the {tidytext} package for working with textual data (Robinson and Silge [2019](#ref-R-tidytext)). Just a reminder, if you haven’t already installed the {dataedu} package, you can do so by typing this code:

devtools::install\_github("data-edu/dataedu")

library(tidyverse)  
library(here)  
library(dataedu)  
library(tidytext)

## Import Data

Let’s start by getting the data into our programming environment so we can start analyzing it. We’ve conveniently included the raw dataset of TidyTuesday tweets in the {dataedu} package. You can see the dataset by typing tt\_tweets. Let’s start by assigning the name raw\_tweets to this dataset:

raw\_tweets <- dataedu::tt\_tweets

### Using an API

It’s worth taking a short detour to talk about another way you can get a dataset like this. A more advanced but common way to import data from a social media website is to use something called an application programming interface (API). A full discussion and walkthrough of using an API is outside the scope of this book, but it’s worth introducing the idea so you have a sense of the possibilities. Once you learn how to do it, using an API offers some advantages over downloading the dataset.

Think of an API as a special door a home builder made for a house that has a lot of cool stuff in it. The home builder doesn’t want everyone to be able to walk right in and use a bunch of stuff in the house. But they also don’t want to make it too hard because, after all, sharing is caring! So imagine the home builder made a door just for folks who know how to use doors. In order to get through this door, users need to know where to find it along the outside of the house. Once they’re there, they have to know the code to open. And once they’re through the door, they have to know how to use the stuff inside. An API for social media platforms like Twitter and Facebook are the same way. You can download datasets of social media information, like tweets, using some code and authentication credentials organized by the website.

There are some advantages to using an API to import data at the start of your education dataset analysis. Every time you run the code in your analysis, you’ll be using the API to contact the social media platform and download a fresh dataset. Now your analysis is not just a one-off product. By using an API to import new data every time you run your code, you create an analysis that can be run again and again on future datasets.

A key point - and limitation - for how Twitter allows access to their data for the seven most recent days. There are a number of ways to access older data, which we discuss at the end of this chapter, though we focus on one way here: having access to the URLs to (or the status IDs for) tweets. We used this technique, which we describe in this chapter’s *Technical Appendix*, along with other strategies for collecting historical data from Twitter. The data that we processed is available in the {dataedu} R package as the tt-tweets dataset. We describe how to process and model this data, and conclude with a description of two powerful social network analysis models, for selection (to explore who interacts with whom) and influence (to determine how interactions relate to changes in what an individual knows or believes).

## View Data

Let’s return to our raw\_tweets dataset. Run glimpse(raw\_tweets) and notice the number of variables in this dataset. It’s good practice to use functions like glimpse() or str() to look at the data type of each variable. For this walkthrough, we won’t need all 90 variables so let’s clean the dataset and keep only the ones we want.

## Process Data

In this section we’ll select the columns we need for our analysis and we’ll transform the dataset so each row represents a word. After that, our dataset will be ready for exploring.

First, let’s use select() to pick the two columns we’ll need: status\_id and text. status\_id will help us associate interesting words with a particular tweet and text will give us the text from that tweet. We’ll also change status\_id to the character data type since it’s meant to label tweets and doesn’t actually represent a numerical value.

tweets <-  
 raw\_tweets %>%  
 #filter for English tweets  
 filter(lang == "en") %>%  
 select(status\_id, text) %>%  
 # Convert the ID field to the character data type  
 mutate(status\_id = as.character(status\_id))

Now the dataset has a column to identify each tweet and a column that shows the text that users tweeted. But each row has the entire tweet in the text variable, which makes it hard to analyze. If we kept our dataset like this, we’d need to use functions on each row to do something like count the number of times the word “good” appears. We can count words more efficiently if each row represented a single word. Splitting sentences in a row into single words in a row is called “tokenizing.” In their book *Text Mining With R*, Silge and Robinson ([2017](#ref-silge2017text)) describe tokens this way:

A token is a meaningful unit of text, such as a word, that we are interested in using for analysis, and tokenization is the process of splitting text into tokens. This one-token-per-row structure is in contrast to the ways text is often stored in current analyses, perhaps as strings or in a document-term matrix.

The {tidytext} package has a convenient function called unnest\_tokens() that tokenizes vectors of words. Let’s install the {tidytext} package so we can use it:

install.packages("tidytext")

Let’s use unnest\_tokens() to take our dataset of tweets and transform it into a dataset of words.

tokens <-   
 tweets %>%  
 unnest\_tokens(output = word, input = text)  
  
tokens

## # A tibble: 131,233 x 2  
## status\_id word   
## <chr> <chr>   
## 1 1163154266065735680 first   
## 2 1163154266065735680 tidytuesday  
## 3 1163154266065735680 submission   
## 4 1163154266065735680 roman   
## 5 1163154266065735680 emperors   
## 6 1163154266065735680 and   
## 7 1163154266065735680 their   
## 8 1163154266065735680 rise   
## 9 1163154266065735680 to   
## 10 1163154266065735680 power   
## # … with 131,223 more rows

We use output = word to tell unnest\_tokens() that we want our column of tokens to be called word. We use input = text to tell unnest\_tokens() to tokenize the tweets in the text column of our tweets dataset. The result is a new dataset where each row has a single word in the word column and a unique ID in the status\_id column that tells us which tweet the word appears in.

Notice that our tokens dataset has many more rows than our tweets dataset. This tells us a lot about how unnest\_tokens() works. In the tweets dataset, each row has an entire tweet and its unique ID. Since that unique ID is assigned to the entire tweet, each unique ID only appears once in the dataset. When we used unnest\_tokens() put each word on its own row, we broke each tweet into many words. This created additional rows in the dataset. And since each word in a single tweet shares the same ID for that tweet, an ID now appears multiple times in our new dataset.

We’re almost ready to start analyzing the dataset! There’s one more step we’ll take–removing common words that don’t help us learn about what people are tweeting about. Words like “the” or “a” are in a category of words called “stop words”. Stop words serve a function in verbal communication, but don’t tell us much on their own. As a result, they clutter our dataset of useful words and make it harder to manage the volume of words we want to analyze. The {tidytext} package includes a dataset called stop\_words that we’ll use to remove rows containing stop words. We’ll use anti\_join() on our tokens dataset and the stop\_words dataset to keep only rows that have words *not* appearing in the stop\_words dataset.

data(stop\_words)  
  
tokens <-  
 tokens %>%  
 anti\_join(stop\_words, by = "word")

Why does this work? Let’s look closer. inner\_join() matches the observations in one dataset to another by a specified common variable. Any rows that don’t have a match get dropped from the resulting dataset. anti\_join() does the same thing as inner\_join() except it drops matching rows and keeps the rows that *don’t* match. This is convenient for our analysis because we want to remove rows from tokens that contain words in the stop\_words dataset. When we call anti\_join(), we’re left with rows that *don’t* match words in the stop\_words dataset. These remaining words are the ones we’ll be analyzing.

One final note before we start counting words: Remember when we first tokenized our dataset and we passed unnest\_tokens() the argument output = word? We conveniently chose word as our column name because it matches the column name word in the stop\_words dataset. This makes our call to anti\_join() simpler because anti\_join() knows to look for the column named word in each dataset.

## Analysis: Counting Words

Now it’s time to start exploring our newly cleaned dataset of tweets. Computing the frequency of each word and seeing which words showed up the most often is a good start. We can pipe tokens to the count function to do this:

tokens %>%   
 count(word, sort = TRUE)

## # A tibble: 15,335 x 2  
## word n  
## <chr> <int>  
## 1 t.co 5432  
## 2 https 5406  
## 3 tidytuesday 4316  
## 4 rstats 1748  
## 5 data 1105  
## 6 code 988  
## 7 week 868  
## 8 r4ds 675  
## 9 dataviz 607  
## 10 time 494  
## # … with 15,325 more rows

We pass count() the argument sort = TRUE to sort the n variable from the highest value to the lowest value. This makes it easy to see the most frequently occurring words at the top. Not surprisingly, “tidytuesday” was the third most frequent word in this dataset.

We may want to explore further by showing the frequency of words as a percent of the whole dataset. Calculating percentages like this is useful in a lot of education scenarios because it helps us make comparisons across different sized groups. For example, you may want to calculate what percentage of students in each classroom receive special education services.

In our tweets dataset, we’ll be calculating the count of words as a percentage of all tweets. We can do that by using mutate() to add a column called percent. percent will divide n by sum(n), which is the total number of words. Finally, will multiply the result by 100.

tokens %>%  
 count(word, sort = TRUE) %>%  
 # n as a percent of total words  
 mutate(percent = n / sum(n) \* 100)

## # A tibble: 15,335 x 3  
## word n percent  
## <chr> <int> <dbl>  
## 1 t.co 5432 7.39   
## 2 https 5406 7.36   
## 3 tidytuesday 4316 5.87   
## 4 rstats 1748 2.38   
## 5 data 1105 1.50   
## 6 code 988 1.34   
## 7 week 868 1.18   
## 8 r4ds 675 0.919  
## 9 dataviz 607 0.826  
## 10 time 494 0.672  
## # … with 15,325 more rows

Even at 4316 appearances in our dataset, “tidytuesday” represents only about 6 percent of the total words in our dataset. This makes sense when you consider our dataset contains 15335 unique words.

## Analysis: Sentiment analysis

Now that we have a sense of the most frequently appearing words, it’s time to explore some questions in our tweets dataset. Let’s imagine that we’re education consultants trying to learn about the community surrounding the TidyTuesday data visualization ritual. We know from the first part of our analysis that the token “dataviz” (a short name for data visualization) appeared frequently relative to other words, so maybe we can explore that further. A good start would be to see how the appearance of that token in a tweet is associated with other positive words.

We’ll need to use a technique called sentiment analysis to get at the “positivity” of words in these tweets. Sentiment analysis tries to evaluate words for their emotional association. If we analyze words by the emotions they convey, we can start to explore patterns in large text datasets like our tokens data.

The functions we’ll be using for our sentiment analysis are in a package called {textdata}. Let’s start by installing that.

install.packages("textdata")

Earlier we used anti\_join() to remove stop words in our dataset. We’re going to do something similar here to reduce our tokens dataset to only words that have a positive association. We’ll use a dataset called the NRC Word-Emotion Association Lexicon to help us identify words with a positive association. This dataset was published in a work called Crowdsourcing a Word-Emotion Association Lexicon (Mohammad and Turney [2013](#ref-mohammad2013))

To explore this dataset more, we’ll use a {tidytext} function called get\_sentiments() to view some words and their associated sentiment. If this is your first time using the NRC Word-Emotion Association Lexicon in the {tidytext} package, you’ll be prompted to download the NRC lexicon. Respond “yes” to the prompt and the NRC lexicon will download. Note that you’ll only have to do this the first time you use the NRC lexicon.

get\_sentiments("nrc")

## # A tibble: 13,901 x 2  
## word sentiment  
## <chr> <chr>   
## 1 abacus trust   
## 2 abandon fear   
## 3 abandon negative   
## 4 abandon sadness   
## 5 abandoned anger   
## 6 abandoned fear   
## 7 abandoned negative   
## 8 abandoned sadness   
## 9 abandonment anger   
## 10 abandonment fear   
## # … with 13,891 more rows

This returns a dataset with two columns. The first is word and contains a list of words. The second is the sentiment column, which contains an emotion associated with each word. This dataset is similar to the stop\_words dataset. Note that this dataset also uses the column name word, which will again make it easy for us to match this dataset to our tokens dataset.

### Count Positive Words

Let’s begin working on reducing our tokens dataset down to only words that the NRC dataset associates with positivity. We’ll start by creating a new dataset, nrc\_pos, which contains the NRC words that have the positive sentiment. Then we’ll match that new dataset to tokens using the word column that is common to both datasets. Finally, we’ll use count() to total up the appearances of each positive word.

# Only positive in the NRC dataset  
nrc\_pos <-  
 get\_sentiments("nrc") %>%  
 filter(sentiment == "positive")  
  
# Match to tokens  
pos\_tokens\_count <-  
 tokens %>%  
 inner\_join(nrc\_pos, by = "word") %>%  
 # Total appearance of positive words  
 count(word, sort = TRUE)   
  
pos\_tokens\_count

## # A tibble: 644 x 2  
## word n  
## <chr> <int>  
## 1 fun 173  
## 2 top 162  
## 3 learn 131  
## 4 found 128  
## 5 love 113  
## 6 community 110  
## 7 learning 97  
## 8 happy 95  
## 9 share 90  
## 10 inspired 85  
## # … with 634 more rows

We can visualize these words nicely by using {ggplot2} to show the positive words in a bar chart. There are 644 words total, which is hard to convey in a compact chart. We’ll solve that problem by filtering our dataset to only words that appear 75 times or more.

pos\_tokens\_count %>%  
 # only words that appear 75 times or more  
 filter(n >= 75) %>%  
 ggplot(., aes(x = reorder(word, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = dataedu\_colors("darkblue")) +  
 labs(  
 title = "Count of Words Associated with Positivity",  
 subtitle = "Tweets with the hashtag #tidytuesday",  
 caption = "Data: Twitter and NRC",  
 x = "",  
 y = "Count"  
 ) +  
 theme\_dataedu()

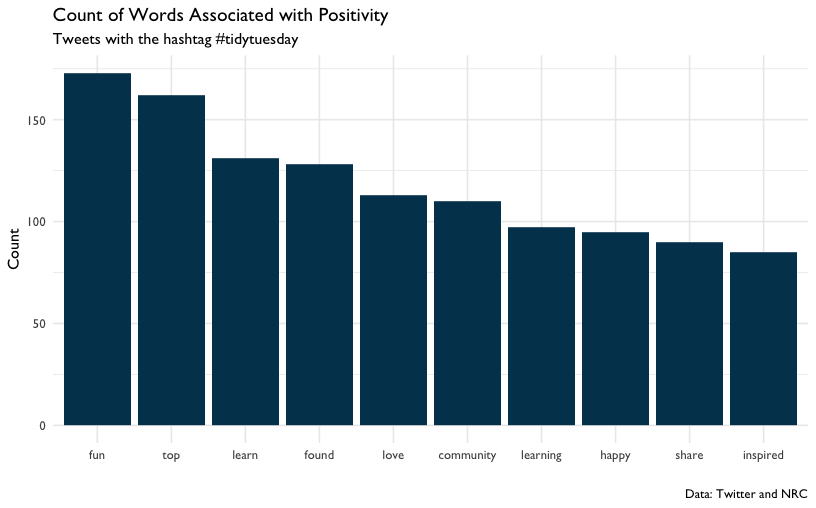


Figure 35: Count of Words Associated with Positivity

Note the use of reorder() when mapping the word variable to the x aesthetic. Using reorder() here sorts our x axis in descending order by the variable n. Sorting the bars from highest frequency to lowest makes it easier for the reader to identify and compare the most and least common words in the visualization.

### Dataviz and Other Positive Words

Earlier in the analysis we learned that “dataviz” was among the most frequently occurring words in this dataset. We can continue our exploration of TidyTuesday tweets by seeing how many tweets with “dataviz” also had at least one positive word from the NRC dataset. Looking at this might give us some clues about how people in the TidyTuesday learning community view dataviz as a tool.

There are a few steps to this part of the analysis, so let’s review our strategy. We’ll need to use the status\_id field in the tweets dataset to filter the tweets that have the word dataviz in them. Then we need to use the status\_id field in this new bunch of dataviz tweets to identify the tweets that include at least one positive word.

How do we know which status\_id values contain the word “dataviz” and which ones contain a positive word? Recall that our tokens dataset only has one word per row, which makes it easy to use functions like filter() and inner\_join() to make two new datasets: one of status\_id values that have “dataviz” in the word column and one of status\_id values that have a positive word in the word column.

We’ll explore the combinations of “dataviz” and any positive words in our tweets dataset using these three ingredients: our tweets dataset, a vector of status\_ids for tweets that have “dataviz” in them, and a vector of status\_ids for tweets that have positive words in them. Now that we have our strategy, let’s write some code and see how it works.

First, we’ll make a vector of status\_ids for tweets that have “dataviz” in them. This will be used later to identify tweets that contain “dataviz” in the text. We’ll use filter() on our tokens dataset to keep only the rows that have “dataviz” in the word column. Let’s name that new dataset dv\_tokens.

dv\_tokens <-  
 tokens %>%  
 filter(word == "dataviz")  
  
dv\_tokens

## # A tibble: 607 x 2  
## status\_id word   
## <chr> <chr>   
## 1 1116518351147360257 dataviz  
## 2 1098025772554612738 dataviz  
## 3 1161454327296339968 dataviz  
## 4 1110711892086001665 dataviz  
## 5 1151926405162291200 dataviz  
## 6 1095854400004853765 dataviz  
## 7 1157111441419395074 dataviz  
## 8 1154958378764046336 dataviz  
## 9 1105642831413239808 dataviz  
## 10 1108196618464047105 dataviz  
## # … with 597 more rows

The result is a dataset that has status\_ids in one column and the word “dataviz” in the other column. We can use $ to extract a vector of status\_ids for tweets that have “dataviz” in the text. This vector has hundreds of values, so we’ll use head to view just the first ten.

# Extract status\_id  
head(dv\_tokens$status\_id)

## [1] "1116518351147360257" "1098025772554612738" "1161454327296339968"  
## [4] "1110711892086001665" "1151926405162291200" "1095854400004853765"

Now let’s do this again, but this time we’ll we’ll make a vector of status\_ids for tweets that have positive words in them. This will be used later to identify tweets that contain a positive word in the text. We’ll use filter() on our tokens dataset to keep only the rows that have any of the positive words in the in the word column. If you’ve been running all the code up to this point in the walkthrough, you’ll notice that you already have a dataset of positive words called nrc\_pos, which can be turned into a vector of positive words by typing nrc\_pos$word. We can use the %in% operator in our call to filter() to find only words that are in this vector of positive words. Let’s name this new dataset pos\_tokens.

pos\_tokens <-   
 tokens %>%  
 filter(word %in% nrc\_pos$word)  
  
pos\_tokens

## # A tibble: 4,925 x 2  
## status\_id word   
## <chr> <chr>   
## 1 1163154266065735680 throne   
## 2 1001412196247666688 honey   
## 3 1001412196247666688 production  
## 4 1001412196247666688 increase   
## 5 1001412196247666688 production  
## 6 1161638973808287746 found   
## 7 991073965899644928 community   
## 8 991073965899644928 community   
## 9 991073965899644928 trend   
## 10 991073965899644928 white   
## # … with 4,915 more rows

The result is a dataset that has status\_ids in one column and a positive word from tokens in the other column. We’ll again use $ to extract a vector of status\_ids for these tweets.

# Extract status\_id  
head(pos\_tokens$status\_id)

## [1] "1163154266065735680" "1001412196247666688" "1001412196247666688"  
## [4] "1001412196247666688" "1001412196247666688" "1161638973808287746"

That’s a lot of status\_ids, many of which are duplicates. Let’s try and make the vector of status\_ids a little shorter. We can use distinct() to get a dataframe of status\_ids, where each status\_id only appears once:

pos\_tokens <-  
 pos\_tokens %>%   
 distinct(status\_id)

Note that distinct() drops all variables except for status\_id. For good measure, let’s use distinct() on our dv\_tokens dataframe too:

dv\_tokens <-  
 dv\_tokens %>%   
 distinct(status\_id)

Now we have a dataframe of status\_id for tweets containing “dataviz” and another for tweets containing a positive word. Let’s use these to transform our tweets dataset. First we’ll filter tweets for rows that have the “dataviz” status\_id. Then we’ll create a new column called positive that will tell us if the status\_id is from our vector of positive word status\_ids. We’ll name this filtered dataset dv\_pos.

dv\_pos <-  
 tweets %>%  
 # Only tweets that have the dataviz status\_id  
 filter(status\_id %in% dv\_tokens$status\_id) %>%  
 # Is the status\_id from our vector of positive word?  
 mutate(positive = if\_else(status\_id %in% pos\_tokens$status\_id, 1, 0))

Let’s take a moment to dissect how we use if\_else() to create our positive column. We gave if\_else() three arguments:

* status\_id %in% pos\_tokens$status\_id: a logical statement
* 1: the value of positive if the logical statement is true
* 0: the value of positive if the logical statement is false

So our new positive column will take the value 1 if the status\_id was in our pos\_tokens dataset and the value 0 if the status\_id was not in our pos\_tokens dataset. Practically speaking, positive is 1 if the tweet has a positive word and 0 if it does not have a positive word.

And finally, let’s see what percent of tweets that had “dataviz” in them also had at least one positive word:

dv\_pos %>%  
 count(positive) %>%  
 mutate(perc = n / sum(n))

## # A tibble: 2 x 3  
## positive n perc  
## <dbl> <int> <dbl>  
## 1 0 272 0.450  
## 2 1 333 0.550

About 55 percent of tweets that have “dataviz” in them also had at least one positive word and about 45 percent of them did not have at least one positive word. It’s worth noting here that this finding doesn’t necessarily mean users didn’t have anything good to say about 45 percent of the “dataviz” tweets. We can’t know precisely why some tweets had positive words and some didn’t, we just know that more dataviz tweets had positive words than not. To put this in perspective, we might have a different impression if 5 percent or 95 percent of the tweets had positive words.

Since the point of exploratory data analysis is to explore and develop questions, let’s continue to do that. In this last section we’ll review a random selection of tweets for context.

### Close Read of Randomly Selected Tweets

Let’s review where we are so far as we work to learn more about the TidyTuesday learning community through tweets. So far we’ve counted frequently used words and estimated the number of tweets with positive associations. This dataset is large, so we need to zoom out and find ways to summarize the data. But it’s also useful to explore by zooming in and reading some of the tweets. Reading tweets helps us to build intuition and context about how users talk about TidyTuesday in general. Even though this doesn’t lead to quantitative findings, it helps us to learn more about the content we’re studying and analyzing. Instead of reading all 4418 tweets, let’s write some code to randomly select tweets to review.

First, let’s make a dataset of tweets that had positive words from the NRC dataset. Remember earlier when we made a dataset of tweets that had “dataviz” and a column that had a value of 1 for containing positive words and 0 for not containing positive words? Let’s reuse that technique, but instead of applying to a dataset of tweets containing “dataviz”, let’s use it on our dataset of all tweets.

pos\_tweets <-  
 tweets %>%  
 mutate(positive = if\_else(status\_id %in% pos\_tokens$status\_id, 1, 0)) %>%  
 filter(positive == 1)

Again, we’re using if\_else to make a new column called positive that takes its value based on whether status\_id %in% pos\_tokens$status\_id is true or not.

We can use slice() to help us pick the rows. When we pass slice() a row number, it returns that row from the dataset. For example, we can select the 1st and 3rd row of our tweets dataset this way:

tweets %>%   
 slice(1, 3)

## # A tibble: 2 x 2  
## status\_id text   
## <chr> <chr>   
## 1 1163154266065735… "First #TidyTuesday submission! Roman emperors and their ri…  
## 2 1001412196247666… "My #tidytuesday submission for week 8. Honey production da…

Randomly selecting rows from a dataset is great technique to have in your toolkit. Random selection helps us avoid some of the biases we all have when we pick rows to review ourselves.

Here’s one way to do that using base R:

sample(x = 1:10, size = 5)

## [1] 7 6 8 1 10

Passing sample() a vector of numbers and the size of the sample you want returns a random selection from the vector. Try changing the value of x and size to see how this works.

{dplyr} has a version of this called sample\_n() that we can use to randomly select rows in our tweets dataset. Using sample\_n() looks like this:

set.seed(2020)  
  
pos\_tweets %>%   
 sample\_n(., size = 10)

## # A tibble: 10 x 3  
## status\_id text positive  
## <chr> <chr> <dbl>  
## 1 113347244173969… "Today is the day - excited to be leading this Bos… 1  
## 2 114436276456938… "UFO sightings with gganimate - need to tidy up so… 1  
## 3 996745975124430… "This week's #TidyTuesday dataset looks too fun to… 1  
## 4 108681181347857… "This week: 2019-01-15 #TidyTuesday #rstats my foc… 1  
## 5 103451450126419… "#TidyTuesday week 22.\nInteresting patterns in ho… 1  
## 6 113597645318988… ".@broadwym is kicking off the #TidyTuesday colear… 1  
## 7 114881375394674… "Night everyone i am heading to bed now i love you… 1  
## 8 107466544430342… "@sebastianhwells @jspairani Hi Sebastian - every … 1  
## 9 115468388823039… "Have you signed up for our next #rladies event? \… 1  
## 10 112789066770597… "Better late than never. My first #TidyTuesday plo… 1

That returned ten randomly selected tweets that we can now read through and discuss. Let’s look a little closer at how we did that. We used sample\_n(), which returns randomly selected rows from our tweets dataset. We also specified that size = 10, which means we want sample\_n() to give us 10 randomly selected rows. A few lines before that, we used set.seed(2020). This helps us ensure that, while sample\_n() theoretically plucks 10 random numbers, our readers can run this code and get the same result we did. Using set.seed(2020) at the top of your code makes sample\_n() pick the same ten rows every time. Try changing 2020 to another number and notice how sample\_n() picks a different set of ten numbers, but repeatedly picks those numbers until you change the argument in set.seed().

## Conclusion

The purpose of this walkthrough is to share code with you so you can practice some basic text analysis techniques. Now it’s time to make your learning more meaningful by adapting this code to text-based files you regularly see at work. Trying reading in some of these and doing a similar analysis:

* News articles
* Procedure manuals
* Open ended responses in surveys

There are also advanced text analysis techniques to explore. Consider trying topic modeling (<https://www.tidytextmining.com/topicmodeling.html>) or finding correlations between terms (<https://www.tidytextmining.com/ngrams.html>), both from Julia Silge’s and David Robinson’s book.

We use this data set further in the next chapter [on social network analysis](#c12).

## Technical Appendix: Accessing Twitter data

### Accessing Historical Twitter Data Using Already-Collected Status URLs

Because the creator of the interactive web application for exploring #tidytuesday content, #tidytuesday.rocks, searched for (and archived) #tidytuesday tweets on a regular basis, a large data set from more than one year of weekly #tidytuesday challenges is available through the [GitHub repository](https://github.com/nsgrantham/tidytuesdayrocks) (https://github.com/nsgrantham/tidytuesdayrocks) for the Shiny application. These Tweets (saved in the data directory as a .tsv (tab-separated-values) file) can be read with the following function:

raw\_tidytuesday\_tweets <-  
 read\_delim(  
 "https://raw.githubusercontent.com/nsgrantham/tidytuesdayrocks/master/data/tweets.tsv",  
 "\t",  
 escape\_double = FALSE,  
 trim\_ws = TRUE  
 )

## Parsed with column specification:  
## cols(  
## status\_url = col\_character(),  
## screen\_name = col\_character(),  
## created\_at = col\_datetime(format = ""),  
## favorite\_count = col\_double(),  
## retweet\_count = col\_double(),  
## dataset\_id = col\_character()  
## )

Then the URL for the tweet (the status\_url column) can be passed to a different rtweet function than the one we used, lookup\_statuses(). Before we do this, there is one additional step to take. Because most of the Tweets are from more than seven days ago, Twitter requires an additional authentication step. In short, you need to use keys and tokens for the Twitter API, or application programming interface. The rtweet vignette on accessing keys and tokens (<https://rtweet.info/articles/auth.html>) explains the process. The end result will be that you will create a token using rtweet that you will use along with your rtweet function (in this case, lookup\_statuses()):

token <-  
 create\_token(  
 consumer\_key = < add - your - key - here > ,  
 consumer\_secret = < add - your - secret - here >  
 )  
  
# here, we pass the status\_url variable from raw\_tidytuesday\_tweets as the statuses to lookup in the lookup\_statuses() function, as well as our token  
  
tidytuesday\_tweets <-  
 lookup\_statuses(raw\_tidytuesday\_tweets$status\_url,  
 token = token)

The end result will be a tibble, like that above for #rstats, for #tidytuesday tweets.

### Accessing Historical Data When You Do Not Have Access to Status URLs

In the above case, we had access to the URLs for tweets because they were saved for the #tidytuesday.rocks Shiny. But, in many cases, historical data will not be available. There are two strategies that may be helpful.

First is [TAGS](https://tags.hawksey.info/) (https://tags.hawksey.info/). TAGS is based in, believe it or not, Google Sheets, and it works great for collecting Twitter data over time - even a long period of time The only catch is that you need to setup and start to use a TAGS sheet *in advance of the period for which you want to collect data*. For example, you can start a TAGS archiver in August of one year, with the intention to collect data over the coming academic year; or, you can start a TAGS archiver before an academic conference for which you want to collect Tweets.

A second option is the Premium API through Twitter. This is an expensive option, but is one that can be done through rtweet, and can also access historical data, even if you haven not started a TAGS sheet and do not otherwise have access to the status URLs.

# Walkthrough 6: Exploring Relationships Using Social Network Analysis With Social Media Data

## Vocabulary

* Application Programming Interface (API)
* edgelist
* edge
* influence model
* regex
* selection model
* social network analysis
* sociogram
* vertex

## Chapter Overview

This chapter builds upon [Walkthrough 5/Chapter 11]](#c11) of #tidytuesday data. Whereas the previous chapter was focused on using text analysis to understand the *content* of tweets, this chapter focuses on the *interactions* between #tidytuesday participants. Thus, this chapter explores the use of social network analysis techniques. We note that this chapter, like the previous chapter, includes a technical appendix: This chapter’s technical appendix is focused on two social network processes, selection and influence, that are not the focus of this chapter, but are feasible to investigate using R, and so are, we think, worth introducing as a topic for further reading and study.

### Background

There are a few reasons to be interested in social media. For example, if you work in a school district, you may be interested in who is interacting with the content you share. If you are a researcher, you may wish to investigate what teachers, administrators, and others do through state-based hashtags (e.g., Rosenberg et al. ([2016](#ref-rosenberg2016))). Social media-based data can also be interesting because it provides new contexts for learning to take place, such as professional learning networks (Trust, Krutka, and Carpenter [2016](#ref-trust2016)).

In the past, if a teacher wanted advice about how to plan a unit or to design a lesson, they would likely turn to a trusted peer in their building or district (Spillane, Kim, and Frank [2012](#ref-spillane2012)). In the present, though, they are as likely to turn to someone in a social media-based network. Thus, while the kinds of interactions that take place through #tidytuesday do represent a new context for individuals to communicate and share, it is the kind of network that is becoming increasingly common in education, and, so, being able to use a data science approach to understand the nature of the activity that takes places through #tidytuesday is a potentially valuable capability.

### Packages, Data Sources and Import, and Methods

In this chapter, we access data using the {rtweet} package (Kearney [2016](#ref-kearney2016)). Through {rtweet}, it is easy to access data from Twitter as long as one has a Twitter account. We will load the {tidyverse} and {rtweet} packages to get started. Here is an example of searching the most recent 1,000 tweets which include the hashtag #rstats. When you run this code, you will be prompted to authenticate your access via Twitter. We will also load other packages that we will be using in this analysis, including two packages related to social network analysis (Pedersen [2019](#ref-R-tidygraph), @R-ggraph) as well as one that will help us to use not-anonymized names in a savvy way (Betebenner [2019](#ref-R-randomNames)).

library(tidyverse)  
library(rtweet)  
library(dataedu)  
library(randomNames)  
library(tidygraph)  
library(ggraph)

rstats\_tweets <-   
 search\_tweets("#rstats")

As described in [the previous chapter](#c11), the search term can be easily changed to other hashtags - or other terms. To search for #tidytuesday tweets, we can simply replace #rstats with #tidytuesday.

tidytuesday\_tweets <-   
 search\_tweets("#tidytuesday")

## View Data

We can see that there are *many* rows for the data:

tt\_tweets

## # A tibble: 4,418 x 90  
## user\_id status\_id created\_at screen\_name text source  
## <chr> <chr> <dttm> <chr> <chr> <chr>   
## 1 115921… 11631542… 2019-08-18 18:22:42 MKumarYYC "Fir… Twitt…  
## 2 107332… 11632475… 2019-08-19 00:33:11 cizzart "El … Twitt…  
## 3 107332… 11450435… 2019-06-29 18:57:17 cizzart "Pro… Twitt…  
## 4 107332… 11168648… 2019-04-13 00:45:15 cizzart "#Ar… Twitt…  
## 5 107332… 11228824… 2019-04-29 15:17:02 cizzart "Pes… Twitt…  
## 6 107332… 11176387… 2019-04-15 04:00:17 cizzart "Dat… Twitt…  
## 7 107332… 11245531… 2019-05-04 05:55:32 cizzart "El … Twitt…  
## 8 107332… 11407021… 2019-06-17 19:25:50 cizzart "#da… Twitt…  
## 9 107332… 11325299… 2019-05-26 06:12:46 cizzart "El … Twitt…  
## 10 107332… 11233585… 2019-04-30 22:48:43 cizzart "Vis… Twitt…  
## # … with 4,408 more rows, and 84 more variables: display\_text\_width <dbl>,  
## # reply\_to\_status\_id <chr>, reply\_to\_user\_id <chr>,  
## # reply\_to\_screen\_name <chr>, is\_quote <lgl>, is\_retweet <lgl>,  
## # favorite\_count <int>, retweet\_count <int>, quote\_count <int>,  
## # reply\_count <int>, hashtags <list>, symbols <list>, urls\_url <list>,  
## # urls\_t.co <list>, urls\_expanded\_url <list>, media\_url <list>,  
## # media\_t.co <list>, media\_expanded\_url <list>, media\_type <list>,  
## # ext\_media\_url <list>, ext\_media\_t.co <list>, ext\_media\_expanded\_url <list>,  
## # ext\_media\_type <chr>, mentions\_user\_id <list>, mentions\_screen\_name <list>,  
## # lang <chr>, quoted\_status\_id <chr>, quoted\_text <chr>,  
## # quoted\_created\_at <dttm>, quoted\_source <chr>, quoted\_favorite\_count <int>,  
## # quoted\_retweet\_count <int>, quoted\_user\_id <chr>, quoted\_screen\_name <chr>,  
## # quoted\_name <chr>, quoted\_followers\_count <int>,  
## # quoted\_friends\_count <int>, quoted\_statuses\_count <int>,  
## # quoted\_location <chr>, quoted\_description <chr>, quoted\_verified <lgl>,  
## # retweet\_status\_id <chr>, retweet\_text <chr>, retweet\_created\_at <dttm>,  
## # retweet\_source <chr>, retweet\_favorite\_count <int>,  
## # retweet\_retweet\_count <int>, retweet\_user\_id <chr>,  
## # retweet\_screen\_name <chr>, retweet\_name <chr>,  
## # retweet\_followers\_count <int>, retweet\_friends\_count <int>,  
## # retweet\_statuses\_count <int>, retweet\_location <chr>,  
## # retweet\_description <chr>, retweet\_verified <lgl>, place\_url <chr>,  
## # place\_name <chr>, place\_full\_name <chr>, place\_type <chr>, country <chr>,  
## # country\_code <chr>, geo\_coords <list>, coords\_coords <list>,  
## # bbox\_coords <list>, status\_url <chr>, name <chr>, location <chr>,  
## # description <chr>, url <chr>, protected <lgl>, followers\_count <int>,  
## # friends\_count <int>, listed\_count <int>, statuses\_count <int>,  
## # favourites\_count <int>, account\_created\_at <dttm>, verified <lgl>,  
## # profile\_url <chr>, profile\_expanded\_url <chr>, account\_lang <lgl>,  
## # profile\_banner\_url <chr>, profile\_background\_url <chr>,  
## # profile\_image\_url <chr>

## Process Data

Network data, in general, and network data from Twitter, particularly, requires some processing before it can be used in subsequent analyses. In particular, we are going to create an edgelist, a data structure that is especially helpful for understanding the nature of relationships.

An edgelist looks like the following, where the sender denotes who is initiating the interaction or relationship, and the receiver is who is the recipient of it:

## # A tibble: 12 x 2  
## sender receiver   
## <chr> <chr>   
## 1 al-Sawaya, Nabeeha Nuno Villanueva, Angie  
## 2 Tewolde, Jernayiah Cloud, Desiree   
## 3 Tewolde, Jernayiah Chase, Timothy   
## 4 Castillo-Halvorssen, Scott Cloud, Desiree   
## 5 Castillo-Halvorssen, Scott Nuno Villanueva, Angie  
## 6 Castillo-Halvorssen, Scott Mattie, Logan   
## 7 Vigil, Tiffany Chase, Timothy   
## 8 Vigil, Tiffany Martel, Alondra   
## 9 Vigil, Tiffany Mattie, Logan   
## 10 Proctor, Tina al-Imam, Samraa   
## 11 Ewald, Audrey Chase, Timothy   
## 12 Ewald, Audrey al-Imam, Samraa

In this edgelist, the sender could indicate, for example, someone who nominates someone else (the receiver) as someone they go to for help. The sender could also indicate someone who interacted with the receiver, such as by recognizing one of their tweets with a favorite (or a mention). In the following steps, we will work to create an edgelist from the data from #tidytuesday on Twitter.

### Extracting Mentions

Let’s extract the mentions. There is a lot going on in the code below; let’s break it down line-by-line, starting with the mutate():

* mutate(all\_mentions = str\_extract\_all(text, regex)): this line uses a regex, or regular expression, to identify all of the usernames in the tweet (*note*: the regex comes from from [this Stack Overflow page](https://stackoverflow.com/questions/18164839/get-twitter-username-with-regex-in-r) (https://stackoverflow.com/questions/18164839/get-twitter-username-with-regex-in-r))
* unnest(all\_mentions) this line uses a tidyr function, unnest() to move every mention to its own line, while keeping all of the other information the same (see more about unnest() here: <https://tidyr.tidyverse.org/reference/unnest.html>)).

regex <- "@([A-Za-z]+[A-Za-z0-9\_]+)(?![A-Za-z0-9\_]\*\\.)"  
  
tt\_tweets <-  
 tt\_tweets %>%  
 # Use regular expression to identify all the usernames in a tweet  
 mutate(all\_mentions = str\_extract\_all(text, regex)) %>%  
 unnest(all\_mentions)

Let’s put these into their own data frame, called mentions.

mentions <-  
 tt\_tweets %>%  
 mutate(all\_mentions = str\_trim(all\_mentions)) %>%  
 select(sender = screen\_name, all\_mentions)

### Putting the Edgelist Together

An edgelist is a common social network analysis data structure that has columns for the “sender” and “receiver” of interactions, or relations. For example, someone “sends” the mention to someone who is mentioned, who can be considered to “receive” it. This will require one last processing step. Let’s look at our data as it is now.

mentions

## # A tibble: 2,447 x 2  
## sender all\_mentions   
## <chr> <chr>   
## 1 cizzart @eldestapeweb   
## 2 cizzart @INDECArgentina   
## 3 cizzart @ENACOMArgentina  
## 4 cizzart @tribunalelecmns  
## 5 cizzart @CamaraElectoral  
## 6 cizzart @INDECArgentina   
## 7 cizzart @tribunalelecmns  
## 8 cizzart @CamaraElectoral  
## 9 cizzart @AgroMnes   
## 10 cizzart @AgroindustriaAR  
## # … with 2,437 more rows

What needs to happen to these to make them easier to work with in an edgelist? One step is to remove the “@” symbol from the columns we created and to save the results to a new tibble, edgelist.

edgelist <-   
 mentions %>%   
 # remove "@" from all\_mentions column  
 mutate(all\_mentions = str\_sub(all\_mentions, start = 2)) %>%   
 # rename all\_mentions to receiver  
 select(sender, receiver = all\_mentions)

## Analysis and Results

Now that we have our edgelist, it is straightforward to plot the network. We’ll use the {tidygraph} and {ggraph} packages to visualize the data.

### Plotting the Network

Because large networks (like this one) can present challenges, it is common to filter them to only include some individuals. Let’s explore how many interactions each individual in the network sent.

interactions\_sent <- edgelist %>%   
 # this counts how many times each sender appears in the data frame, effectively counting how many interactions each individual sent   
 count(sender) %>%   
 # arranges the data frame in descending order of the number of interactions sent  
 arrange(desc(n))  
  
interactions\_sent

## # A tibble: 618 x 2  
## sender n  
## <chr> <int>  
## 1 thomas\_mock 347  
## 2 R4DScommunity 78  
## 3 WireMonkey 52  
## 4 CedScherer 41  
## 5 allison\_horst 37  
## 6 mjhendrickson 34  
## 7 kigtembu 27  
## 8 WeAreRLadies 25  
## 9 PBecciu 23  
## 10 sil\_aarts 23  
## # … with 608 more rows

618 senders of interactions is a lot! What if we focused on only those who sent more than one interaction?

interactions\_sent <-   
 interactions\_sent %>%   
 filter(n > 1)

That leaves us with only 349, perhaps a more reasonable number.

We now need to filter the edgelist to only include these 349 individuals. The following code simply uses the filter() function combined with the %in% operator to do this:

edgelist <- edgelist %>%   
 # the first of the two lines below filters to include only senders in the interactions\_sent data frame  
 # the second line does the same, for receivers  
 filter(sender %in% interactions\_sent$sender,  
 receiver %in% interactions\_sent$sender)

We’ll use the as\_tbl\_graph() function, which (by default) identified the first column as the “sender” and the second as the “receiver.” Let’s look at the object it creates, too.

g <-   
 as\_tbl\_graph(edgelist)  
  
g

## # A tbl\_graph: 267 nodes and 975 edges  
## #  
## # A directed multigraph with 7 components  
## #  
## # Node Data: 267 x 1 (active)  
## name   
## <chr>   
## 1 dgwinfred   
## 2 datawookie   
## 3 jvaghela4   
## 4 FournierJohanie  
## 5 JonTheGeek   
## 6 jakekaupp   
## # … with 261 more rows  
## #  
## # Edge Data: 975 x 2  
## from to  
## <int> <int>  
## 1 1 32  
## 2 1 36  
## 3 2 120  
## # … with 972 more rows

We can see that the network now consists of 267 individuals - the 267 who sent more than one interaction.

Next, we’ll use the ggraph() function:

g %>%  
 # we chose the kk layout as it created a graph which was easy-to-interpret, but others are available; see ?ggraph  
 ggraph(layout = "kk") +  
 # this adds the points to the graph  
 geom\_node\_point() +  
 # this adds the links, or the edges; alpha = .2 makes it so that the lines are partially transparent  
 geom\_edge\_link(alpha = .2) +  
 # this last line of code adds a ggplot2 theme suitable for network graphs  
 theme\_graph()



Figure 36: Network Graph

Finally, let’s size the points based on a measure of centrality, typically a measure of how (potentially) influence an individual may be, based on the interactions observed.

g %>%   
 # this calculates the centrality of each individual using the built-in centrality\_authority() function  
 mutate(centrality = centrality\_authority()) %>%   
 ggraph(layout = "kk") +   
 geom\_node\_point(aes(size = centrality, color = centrality)) +  
 # this line colors the points based upon their centrality  
 scale\_color\_continuous(guide = 'legend') +   
 geom\_edge\_link(alpha = .2) +  
 theme\_graph()



Figure 37: Network Graph with Centrality

There is much more you can do with {ggraph} (and {tidygraph}); check out the {ggraph} tutorial here: <https://ggraph.data-imaginist.com/>

## Conclusion

In this chapter, we used social media data (from the #tidytuesday hashtag) to prepare and visualize social network data. This is a powerful technique; one that can reveal who is interacting with whom, and one that can begin to suggest why.

Behind these visualizations, though, there are also statistical models and methods that can help to understand what is going on with respect to particular relationships in a network in additional ways.

One way to consider these models and methods is in terms of two *processes* at play in our relationships (cite). These two processes are commonly (though not exclusively) the focus of statistical analyses of networks. In addition to not being exclusive, they do not interact independently: they affect each other reciprocally (Xu, Frank, & Penuel, 2018). They are:

* *Selection*: the processes regarding who chooses to have a relationship with whom
* *Influence*: the processes regarding how who we have relationships with affects our behavior

While these are complex, they can be studied with the type of data collected from asking people about their relationships (and possibly asking them about or studying their behavior–or measuring some outcome). Happily, the use of these methods has expanded along with R: many of the best tools for studying social networks are in the form of long-standing R packages. Additionally, while there are many potential nuances to studying selection and influence, these are models that can fundamentally be carried out with regression, or the linear model (or extensions of it). We describe these in the *Technical Appendix* for this chapter, as they do not use the tidytuesday dataset and are likely to be of interest to readers only after having mastered preparing and visualizing network data.

## Technical Appendix: Influence and Selection Models

As noted above, there is much more to understanding interactions, and network analysis, beyond creating edgelists and visualizing network data (through the use of an edgelist). Two processes that are particularly important (and able to be studied with network data using R) are for influence and selection.

### An Example of Influence

First, let’s look at an example of influence. To do so, let’s create three different data frames. Here is what they should, at the end of the process, contain:

* A data frame indicating who the *nominator* and *nominee* for the relation (i.e., if Stefanie says that José is her friend, then Stefanie is the nominator and José the nominee) - as well as an optional variable indicating the weight, or strength, of their relation.
* This data frame and its type can be considered the basis for many types of social network analysis and is a common structure for network data: it is an *edgelist*.
* Data frames indicating the values of some behavior - an outcome - at two different time points.

In this example, we create some example data that can be used to explore questions about how influence works.

Let’s take a look at the merged data. What this data now contains is the first data frame, data1, with each nominees’ outcome at time 1 (yvar1). Note that we will find each nominators’ outcome at time 2 later on.

data1 <-  
 data.frame(  
 nominator = c(2, 1, 3, 1, 2, 6, 3, 5, 6, 4, 3, 4),  
 nominee = c(1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6),  
 relate = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)  
 )  
  
data2 <-  
 data.frame(nominee = c(1, 2, 3, 4, 5, 6),  
 yvar1 = c(2.4, 2.6, 1.1, -0.5, -3, -1))  
  
data3 <-  
 data.frame(nominator = c(1, 2, 3, 4, 5, 6),  
 yvar2 = c(2, 2, 1, -0.5, -2, -0.5))

### Joining the Data

Next, we’ll join the data into one data frame. Note that while this is sometimes tedious and time-consuming, especially with large sources of network data, it is a key step for being able to carry out network analysis - often, even for creating visualizations that are informative.

data <-  
 left\_join(data1, data2, by = "nominee")  
  
data <-  
 data %>%   
 # this makes merging later easier  
 mutate(nominee = as.character(nominee))   
  
# calculate indegree in tempdata and merge with data  
tempdata <- data.frame(table(data$nominee))  
  
tempdata <-  
 tempdata %>%  
 rename(  
 # rename the column "Var1" to "nominee"   
 "nominee" = "Var1",   
 # rename the column "Freq" to "indegree"  
 "indegree" = "Freq"  
 ) %>%   
 # makes nominee a character data type, instead of a factor, which can cause problems  
 mutate(nominee = as.character(nominee))  
  
data <-   
 left\_join(data, tempdata, by = "nominee")

#### Calculating an Exposure Term

This is the key step that makes this model - a regression, or linear, model - one that is special. It is creating an exposure term. The idea is that the exposure term “captures” how your interactions with someone, over some period of time (between the first and second time points) impact some outcome. This model accounts for an individual’s initial report of the outcome, i.e., their time 1 prior value, so it is a model for *change* in some outcome.

# Calculating exposure  
data <-  
 data %>%   
 mutate(exposure = relate \* yvar1)  
  
# Calculating mean exposure  
mean\_exposure <-  
 data %>%  
 group\_by(nominator) %>%  
 summarize(exposure\_mean = mean(exposure))

What this data frame - mean\_exposure - contains is the mean of the outcome (in this case, yvar1) for all of the individuals the nominator had a relation with.

As we need a final data set with mean\_exposure,degree, yvar1, and yvar2 added, we’ll process the data a bit more.

data2 <-  
 data2 %>%   
 # rename nominee as nominator to merge these  
 rename("nominator" = "nominee")   
  
final\_data <-  
 left\_join(mean\_exposure, data2, by = "nominator")  
  
final\_data <-   
 # data3 already has nominator, so no need to change  
 left\_join(final\_data, data3, by = "nominator")

#### Regression (Linear Model)

Calculating the exposure term is the most distinctive and important step in carrying out influence models. Now, we can simply use a linear model to find out how much relations - as captured by the influence term - affect some outcome.

model1 <-  
 lm(yvar2 ~ yvar1 + exposure\_mean, data = final\_data)  
  
summary(model1)

##   
## Call:  
## lm(formula = yvar2 ~ yvar1 + exposure\_mean, data = final\_data)  
##   
## Residuals:  
## 1 2 3 4 5 6   
## 0.0295 -0.0932 0.0943 -0.0273 -0.0255 0.0222   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1161 0.0345 3.37 0.043 \*   
## yvar1 0.6760 0.0241 28.09 9.9e-05 \*\*\*  
## exposure\_mean 0.1254 0.0361 3.47 0.040 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0823 on 3 degrees of freedom  
## Multiple R-squared: 0.998, Adjusted R-squared: 0.997   
## F-statistic: 945 on 2 and 3 DF, p-value: 6.31e-05

So, the influence model is used to study a key process for social network analysis, but it is one that is useful, because you can quantify, given what you measure and how you measure it, *the network effect*, something that is sometimes not considered, especially in education (Frank, 2009). It’s also fundamentally a regression. That’s really it, as the majority of the work goes into calculating the exposure term.

### An Example of Selection

Selection models are also commonly used - and are commonly of interest not only to researchers but also to administrators and teachers (and even to youth and students).

Here, we briefly describe a few possible approaches for using a selection model.

At its core, the selection model is a regression - albeit, one that is a generalization of one, namely, a logistic regression (sometimes termed a generalized linear model, because it is *basically* a regression but is one with an outcome that consists just of 0’s and 1’s). Thus, the most straight-away way to use a selection model is to use a logistic regression where all of the relations (note the relate variable in data1 above) are indicated with a 1. But, here is the important and challenging step: all of the *possible relations* (i.e., all of the relations that are possible between all of the individuals in a network) are indicated with a 0 in an edgelist. Note that, again, an edgelist is the preferred data structure for carrying out this analysis. This step involves some data wrangling, especially the idea of widening or lengthening a data frame.

Once all of the relations are indicated with a 1 or a 0, then a simple linear regression can be used. Imagine that we are interested in whether individuals from the *same* group are more or less likely to interact than those from different groups; same could be created in the data frame based upon knowing which group both nominator and nominee are from:

m\_selection <-   
 glm(relate ~ 1 + same, data = edgelist1)

While this is a straightforward way to carry out a selection model, there are some limitations to it. Namely, it does not account for individuals who send more (or less) nominations overall–and not considering this may mean other effects, like the one associated with being from the *same* group, are not accurate. A few extensions of the linear model - including those that can use data for which relationships are indicated with weights, not just 1’s and 0’s, have been developed.

One type of selection model extends the logistic regression. The {amen} (Hoff et al. [2017](#ref-R-amen)) package can be used for data that is not only 1ss and 0s - like a logistic regression - but also data that is normally distributed. Another extension is an Exponential Random Graph Model, or ERGM. The {ergm} R package makes it easy to use these kinds of selection models. {ergm} (Handcock, Hunter, Butts, Goodreau, Krivitsky, and Morris [2019](#ref-R-ergm)) is itself a part of a powerful and often-used collection of packages for social network analysis, {statnet} (Handcock, Hunter, Butts, Goodreau, Krivitsky, Bender-deMoll, et al. [2019](#ref-R-statnet)). These packages collectively begin to illustrate the richness of the packages available within R for using social network analysis models and methods; moreover, as developments in social network analysis methods continue to be made, cutting-edge techniques are likely to continue to be available (through R packages) to those using R for these types of analyses.

# Walkthrough 7: The Role (and Usefulness) of Multi-Level Models

## Vocabulary

* dummy coding
* hierarchical linear model
* intra-class correlation
* multi-level model

## Chapter Overview

The *purpose* of this walkthrough is to explore students’ performance in these online courses. While this and the analysis in [Walkthrough 1/Chapter 7](#c07) focus on the time students spent in the course, this walkthrough is distinguished by a focus on the effects of being in a particular course. To do so, we explore the use of *multi-level models*, which are suited to addressing the fact that the students in our dataset shared classes. While the conceptual details underlying multi-level models can be complex, the basic problem that they address - cases, or observations, such as students, grouped within higher level units, such as classes or schools - is likely familiar to readers of this book.

### Background

Using multi-level models is a way to account for the way in which individual cases - like responses for individual students - are “grouped” together into higher-level units, like classes. As we describe later in this chapter, multi-level models do this by (still) estimating the effect of being a student in each group, but with a key distinction from a regression (or linear model), like those described in [Walkthrough 1/Chapter 7](#c07) and [Walkthrough 4/Chapter 10](#c10): the multi-level model “regularizes” the estimates for each group based upon systematically different the groups (classes) are, in terms of the dependent variable, from the overall (across all groups [classes]) values of the dependent variable.

These are the conceptual details underlying multi-level models, but, fortunately, fitting them is straightforward, and should be familiar if you have used R’s lm() function before. So, let’s get started!

### Data Source

We use the same data source on students’ motivation in online science classes that we processed in [Walkthrough 1](#c07).

### Methods

Are there course-specific differences in how much time students spend on the course as well as in how time spent is related to the percentage of points students earned? There are a number of ways to approach this question. Let’s use our linear model.

Specifically, we can dummy-code the groups. Dummy coding means transforming a variable with multiple categories into multiple, new variables, where each variable indicates the presence and absence of only one of the categories.

## Load Packages

We will load the tidyverse and a few other packages specific to using multi-level models: {lme4} (Bates et al. [2019](#ref-R-lme4)) and {performance} (Lüdecke, Makowski, and Waggoner [2020](#ref-R-performance)).

library(tidyverse)  
library(dummies)  
library(sjPlot)  
library(lme4)  
library(performance)  
library(dataedu)

## Import Data

### The Role of Dummy Codes

We can see how dummy coding works through using the {dummies} package, though, as we will see, you often do not need to manually dummy code variables like this.

Let’s consider the iris data that comes built into R. Since we are fans of the {tidyverse}, we will first change it into a tibble.

iris <- as\_tibble(iris)  
iris

## # A tibble: 150 x 5  
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## <dbl> <dbl> <dbl> <dbl> <fct>   
## 1 5.1 3.5 1.4 0.2 setosa   
## 2 4.9 3 1.4 0.2 setosa   
## 3 4.7 3.2 1.3 0.2 setosa   
## 4 4.6 3.1 1.5 0.2 setosa   
## 5 5 3.6 1.4 0.2 setosa   
## 6 5.4 3.9 1.7 0.4 setosa   
## 7 4.6 3.4 1.4 0.3 setosa   
## 8 5 3.4 1.5 0.2 setosa   
## 9 4.4 2.9 1.4 0.2 setosa   
## 10 4.9 3.1 1.5 0.1 setosa   
## # … with 140 more rows

As we can see above, the Species variable is a factor. If we consider how we could include a variable such as this in a linear model, things may become a little confusing. Species seems to be made up of, well, words, such as “setosa.” The common way to approach this is through dummy coding, where you create new variables for each of the possible values of Species (such as “setosa”). Then, these new variables have a value of 1 when the row is associated with that level (i.e., the first row in the data frame above would have a 1 for a column named setosa).

Let’s return to {dummies}.

How many possible values are there for Species? We can check with the levels function.

levels(iris$Species)

## [1] "setosa" "versicolor" "virginica"

When we run the dummy() function on the Species variable, we can see that it returns *three* variables, one for each of the three levels of Species - setosa, versicolor, and virginica.

dummies::dummy(iris$Species) %>%  
 head()

## {\n args = commandArgs(TRUE)\n bookdown:::source\_utf8(args[4])\n out = do.call(rmarkdown::render, c(args[1], readRDS(args[2]), list(run\_pandoc = FALSE, encoding = "UTF-8")))\n bookdown:::source\_utf8(args[5])\n out\_expected = xfun::with\_ext(args[1], ".md")\n if (out != out\_expected) {\n file.rename(out, out\_expected)\n attributes(out\_expected) = attributes(out)\n out = out\_expected\n }\n if (file.exists(args[3])) {\n res = readRDS(args[3])\n res[[args[1]]] = out\n saveRDS(res, args[3])\n }\n else saveRDS(setNames(list(out), args[1]), args[3])\n}setosa  
## [1,] 1  
## [2,] 1  
## [3,] 1  
## [4,] 1  
## [5,] 1  
## [6,] 1  
## {\n args = commandArgs(TRUE)\n bookdown:::source\_utf8(args[4])\n out = do.call(rmarkdown::render, c(args[1], readRDS(args[2]), list(run\_pandoc = FALSE, encoding = "UTF-8")))\n bookdown:::source\_utf8(args[5])\n out\_expected = xfun::with\_ext(args[1], ".md")\n if (out != out\_expected) {\n file.rename(out, out\_expected)\n attributes(out\_expected) = attributes(out)\n out = out\_expected\n }\n if (file.exists(args[3])) {\n res = readRDS(args[3])\n res[[args[1]]] = out\n saveRDS(res, args[3])\n }\n else saveRDS(setNames(list(out), args[1]), args[3])\n}versicolor  
## [1,] 0  
## [2,] 0  
## [3,] 0  
## [4,] 0  
## [5,] 0  
## [6,] 0  
## {\n args = commandArgs(TRUE)\n bookdown:::source\_utf8(args[4])\n out = do.call(rmarkdown::render, c(args[1], readRDS(args[2]), list(run\_pandoc = FALSE, encoding = "UTF-8")))\n bookdown:::source\_utf8(args[5])\n out\_expected = xfun::with\_ext(args[1], ".md")\n if (out != out\_expected) {\n file.rename(out, out\_expected)\n attributes(out\_expected) = attributes(out)\n out = out\_expected\n }\n if (file.exists(args[3])) {\n res = readRDS(args[3])\n res[[args[1]]] = out\n saveRDS(res, args[3])\n }\n else saveRDS(setNames(list(out), args[1]), args[3])\n}virginica  
## [1,] 0  
## [2,] 0  
## [3,] 0  
## [4,] 0  
## [5,] 0  
## [6,] 0

We can confirm that every row associated with a specific species has a 1 in the column it corresponds to. We can do this by binding together the dummy codes and the iris data and then counting, for each of the three species, how many of the rows for each dummy code were coded with a “1”.

For example, when the Species is “setosa”, the variable Speciessetosa always equals 1 - as is the case for the other species (for their respective variables). bind\_cols() is a useful tidyverse function for binding together data frames by column.

# create matrix of dummy-coded variables  
species\_dummy\_coded <- dummies::dummy(iris$Species)  
  
# convert matrix to tibble so we can use tidyverse functions  
species\_dummy\_coded <- as\_tibble(species\_dummy\_coded)  
  
# add dummy coded variables to iris  
iris\_with\_dummy\_codes <- bind\_cols(iris, species\_dummy\_coded)

Let’s look at the results.

iris\_with\_dummy\_codes %>%   
 count(Species, Speciessetosa, Speciesversicolor, Speciesvirginica)

Okay, this covers how dummy codes work. How do they work when used in a model, like with the linear model we have been using (through lm())?

In the context of using lm() (and many other functions in R) is that the number of levels to be created is always the number of different possible values minus one, because each group will be modeled in comparison to the group without a column, or what is commonly called the reference group.

Why can every group not simply have their own dummy-coded column? The reason has to do with how the dummy codes are used. The purpose of the dummy code is to show how different the dependent variable is for all of the observations that are in one group (i.e., all of the flowers that are setosa specimens). In order to represent how different those flowers are, they have to be compared to something else - and the intercept of the model usually represents this “something else.” However, if every level of a factor (such as Species) is dummy-coded, then there would be no cases available to estimate an intercept - in short, the dummy code would not be compared to anything else. For this reason, one group is typically selected as the reference group, to which every other group is compared.

We will load a built-in dataset from the {dataedu} package.

dat <- dataedu::sci\_mo\_processed

### Using Dummy Codes

Let’s return use online science class data and consider the effect (for a student) of being in a specific class in the data set.

First, let’s determine how many classes there are. We can use the count() function to see how many courses there are.

dat %>%   
 count(course\_id)

## # A tibble: 26 x 2  
## course\_id n  
## <chr> <int>  
## 1 AnPhA-S116-01 43  
## 2 AnPhA-S116-02 29  
## 3 AnPhA-S216-01 43  
## 4 AnPhA-S216-02 17  
## 5 AnPhA-T116-01 11  
## 6 BioA-S116-01 34  
## 7 BioA-S216-01 7  
## 8 BioA-T116-01 2  
## 9 FrScA-S116-01 70  
## 10 FrScA-S116-02 12  
## # … with 16 more rows

## Analysis

### Regression (Linear Model) Analysis with Dummy Codes

We will save this output to m\_linear\_dc, where the dc stands for dummy code. We will keep the variables we used in our last set of models - TimeSpent and course\_id - as independent variables, but will predict students’ final grade (a variable in the dataset), rather than the percentage\_earned variable that we created in (chapter 7)[#c07].

Since we will be using the final grade variable a lot, we can rename it to make it quicker (and easier) to type.

dat <-   
 dat %>%   
 rename(final\_grade = FinalGradeCEMS)

m\_linear\_dc <-   
 lm(final\_grade ~ TimeSpent\_std + course\_id, data = dat)

The output will be a bit, well, long, because each group will have its own intercept. We can see it using tab\_model() from {sjPlot}:

tab\_model(m\_linear\_dc,  
 title = "Table 13.1")

Table 13.1

final grade

Predictors

Estimates

CI

p

(Intercept)

73.20

67.20 – 79.20

<0.001

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [AnPhA-S116-02]

-1.59

-10.88 – 7.70

0.737

course\_id [AnPhA-S216-01]

-9.05

-17.44 – -0.67

0.034

course\_id [AnPhA-S216-02]

-4.51

-16.41 – 7.40

0.457

course\_id [AnPhA-T116-01]

7.24

-6.34 – 20.82

0.296

course\_id [BioA-S116-01]

-3.56

-12.67 – 5.55

0.443

course\_id [BioA-S216-01]

-14.67

-31.61 – 2.26

0.089

course\_id [BioA-T116-01]

9.18

-18.84 – 37.20

0.520

course\_id [FrScA-S116-01]

12.02

4.33 – 19.70

0.002

course\_id [FrScA-S116-02]

-3.14

-17.36 – 11.08

0.665

course\_id [FrScA-S116-03]

3.51

-5.43 – 12.46

0.441

course\_id [FrScA-S116-04]

5.23

-14.98 – 25.43

0.612

course\_id [FrScA-S216-01]

9.92

2.41 – 17.43

0.010

course\_id [FrScA-S216-02]

7.37

-2.70 – 17.45

0.151

course\_id [FrScA-S216-03]

2.38

-25.65 – 30.40

0.868

course\_id [FrScA-S216-04]

15.40

-2.92 – 33.72

0.099

course\_id [FrScA-T116-01]

8.12

-12.08 – 28.33

0.430

course\_id [OcnA-S116-01]

4.06

-5.67 – 13.79

0.413

course\_id [OcnA-S116-02]

2.02

-9.89 – 13.93

0.739

course\_id [OcnA-S116-03]

-18.75

-57.86 – 20.36

0.347

course\_id [OcnA-S216-01]

-6.41

-15.04 – 2.22

0.145

course\_id [OcnA-S216-02]

-2.76

-13.47 – 7.95

0.613

course\_id [OcnA-T116-01]

-2.05

-16.97 – 12.87

0.787

course\_id [PhysA-S116-01]

15.35

6.99 – 23.71

<0.001

course\_id [PhysA-S216-01]

5.40

-6.01 – 16.82

0.353

course\_id [PhysA-T116-01]

20.73

-7.23 – 48.70

0.146

Observations

573

R2 / R2 adjusted

0.252 / 0.216

Table 13.1

final grade

Predictors

Estimates

CI

p

(Intercept)

73.20

67.20 – 79.20

<0.001

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [AnPhA-S116-02]

-1.59

-10.88 – 7.70

0.737

course\_id [AnPhA-S216-01]

-9.05

-17.44 – -0.67

0.034

course\_id [AnPhA-S216-02]

-4.51

-16.41 – 7.40

0.457

course\_id [AnPhA-T116-01]

7.24

-6.34 – 20.82

0.296

course\_id [BioA-S116-01]

-3.56

-12.67 – 5.55

0.443

course\_id [BioA-S216-01]

-14.67

-31.61 – 2.26

0.089

course\_id [BioA-T116-01]

9.18

-18.84 – 37.20

0.520

course\_id [FrScA-S116-01]

12.02

4.33 – 19.70

0.002

course\_id [FrScA-S116-02]

-3.14

-17.36 – 11.08

0.665

course\_id [FrScA-S116-03]

3.51

-5.43 – 12.46

0.441

course\_id [FrScA-S116-04]

5.23

-14.98 – 25.43

0.612

course\_id [FrScA-S216-01]

9.92

2.41 – 17.43

0.010

course\_id [FrScA-S216-02]

7.37

-2.70 – 17.45

0.151

course\_id [FrScA-S216-03]

2.38

-25.65 – 30.40

0.868

course\_id [FrScA-S216-04]

15.40

-2.92 – 33.72

0.099

course\_id [FrScA-T116-01]

8.12

-12.08 – 28.33

0.430

course\_id [OcnA-S116-01]

4.06

-5.67 – 13.79

0.413

course\_id [OcnA-S116-02]

2.02

-9.89 – 13.93

0.739

course\_id [OcnA-S116-03]

-18.75

-57.86 – 20.36

0.347

course\_id [OcnA-S216-01]

-6.41

-15.04 – 2.22

0.145

course\_id [OcnA-S216-02]

-2.76

-13.47 – 7.95

0.613

course\_id [OcnA-T116-01]

-2.05

-16.97 – 12.87

0.787

course\_id [PhysA-S116-01]

15.35

6.99 – 23.71

<0.001

course\_id [PhysA-S216-01]

5.40

-6.01 – 16.82

0.353

course\_id [PhysA-T116-01]

20.73

-7.23 – 48.70

0.146

Observations

573

R2 / R2 adjusted

0.252 / 0.216

Wow! Those are a lot of effects. In addition to the time spent and subject variables, the model estimated the difference, accounting for the effects of being a student in a specific class. Let’s count how many classes there are.

If we count the number of classes, we see that there are 25 - and not 26! One has been automatically selected as the reference group, and every other class’s coefficient represents how different each class is from it. The intercept’s value of 0.74 represents the percentage of points that students in the reference group class, which is automatically the first level of the course\_id variable when it is converted to a factor, course\_idAnPhA-S116-01 (which represents an anatomy and physiology course from semester S1 (for the fall) of 2016 in the first section 01).

We can choose another class to serve as a reference group. For example, say that we want course\\_idPhysA-S116-01 (the first section of the physics class offered during this semester and year) to be the reference group. The fct\_relevel() function (which is a part of the {tidyverse} suite of packages) makes it easy to do this. This function allows us to re-order the levels within a factor, so that the “first” level will change. We’ll also use mutate() again here, which we introduced in the previous chapter.

dat <-  
 dat %>%  
 mutate(course\_id = fct\_relevel(course\_id, "PhysA-S116-01"))

We can now see that *that* group is no longer listed as an independent variable, or a predictor: every coefficient in this model is now in reference to it.

# Here we run a linear model again, predicting percentage earned in the course  
# The predictor variables are the (standardized) amount of time spent and the subject of the course (course\_id)  
m\_linear\_dc\_1 <-   
 lm(final\_grade ~ TimeSpent\_std + course\_id, data = dat)  
  
tab\_model(m\_linear\_dc\_1,  
 title = "Table 13.2")

Table 13.2

final grade

Predictors

Estimates

CI

p

(Intercept)

88.55

82.83 – 94.27

<0.001

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [AnPhA-S116-01]

-15.35

-23.71 – -6.99

<0.001

course\_id [AnPhA-S116-02]

-16.94

-26.20 – -7.67

<0.001

course\_id [AnPhA-S216-01]

-24.40

-32.77 – -16.04

<0.001

course\_id [AnPhA-S216-02]

-19.86

-31.71 – -8.01

0.001

course\_id [AnPhA-T116-01]

-8.11

-21.64 – 5.42

0.240

course\_id [BioA-S116-01]

-18.91

-27.72 – -10.09

<0.001

course\_id [BioA-S216-01]

-30.02

-46.80 – -13.24

<0.001

course\_id [BioA-T116-01]

-6.17

-34.09 – 21.75

0.664

course\_id [FrScA-S116-01]

-3.33

-10.76 – 4.10

0.379

course\_id [FrScA-S116-02]

-18.49

-32.58 – -4.39

0.010

course\_id [FrScA-S116-03]

-11.84

-20.59 – -3.08

0.008

course\_id [FrScA-S116-04]

-10.12

-30.32 – 10.08

0.326

course\_id [FrScA-S216-01]

-5.43

-12.62 – 1.75

0.138

course\_id [FrScA-S216-02]

-7.97

-17.85 – 1.90

0.113

course\_id [FrScA-S216-03]

-12.97

-40.89 – 14.95

0.362

course\_id [FrScA-S216-04]

0.05

-18.15 – 18.25

0.996

course\_id [FrScA-T116-01]

-7.22

-27.47 – 13.02

0.484

course\_id [OcnA-S116-01]

-11.29

-20.98 – -1.60

0.022

course\_id [OcnA-S116-02]

-13.33

-25.16 – -1.49

0.027

course\_id [OcnA-S116-03]

-34.10

-73.17 – 4.97

0.087

course\_id [OcnA-S216-01]

-21.76

-30.29 – -13.23

<0.001

course\_id [OcnA-S216-02]

-18.11

-28.66 – -7.56

0.001

course\_id [OcnA-T116-01]

-17.40

-32.22 – -2.58

0.021

course\_id [PhysA-S216-01]

-9.94

-21.16 – 1.28

0.082

course\_id [PhysA-T116-01]

5.39

-22.55 – 33.32

0.705

Observations

573

R2 / R2 adjusted

0.252 / 0.216

Table 13.2

final grade

Predictors

Estimates

CI

p

(Intercept)

88.55

82.83 – 94.27

<0.001

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [AnPhA-S116-01]

-15.35

-23.71 – -6.99

<0.001

course\_id [AnPhA-S116-02]

-16.94

-26.20 – -7.67

<0.001

course\_id [AnPhA-S216-01]

-24.40

-32.77 – -16.04

<0.001

course\_id [AnPhA-S216-02]

-19.86

-31.71 – -8.01

0.001

course\_id [AnPhA-T116-01]

-8.11

-21.64 – 5.42

0.240

course\_id [BioA-S116-01]

-18.91

-27.72 – -10.09

<0.001

course\_id [BioA-S216-01]

-30.02

-46.80 – -13.24

<0.001

course\_id [BioA-T116-01]

-6.17

-34.09 – 21.75

0.664

course\_id [FrScA-S116-01]

-3.33

-10.76 – 4.10

0.379

course\_id [FrScA-S116-02]

-18.49

-32.58 – -4.39

0.010

course\_id [FrScA-S116-03]

-11.84

-20.59 – -3.08

0.008

course\_id [FrScA-S116-04]

-10.12

-30.32 – 10.08

0.326

course\_id [FrScA-S216-01]

-5.43

-12.62 – 1.75

0.138

course\_id [FrScA-S216-02]

-7.97

-17.85 – 1.90

0.113

course\_id [FrScA-S216-03]

-12.97

-40.89 – 14.95

0.362

course\_id [FrScA-S216-04]

0.05

-18.15 – 18.25

0.996

course\_id [FrScA-T116-01]

-7.22

-27.47 – 13.02

0.484

course\_id [OcnA-S116-01]

-11.29

-20.98 – -1.60

0.022

course\_id [OcnA-S116-02]

-13.33

-25.16 – -1.49

0.027

course\_id [OcnA-S116-03]

-34.10

-73.17 – 4.97

0.087

course\_id [OcnA-S216-01]

-21.76

-30.29 – -13.23

<0.001

course\_id [OcnA-S216-02]

-18.11

-28.66 – -7.56

0.001

course\_id [OcnA-T116-01]

-17.40

-32.22 – -2.58

0.021

course\_id [PhysA-S216-01]

-9.94

-21.16 – 1.28

0.082

course\_id [PhysA-T116-01]

5.39

-22.55 – 33.32

0.705

Observations

573

R2 / R2 adjusted

0.252 / 0.216

Using dummy codes is very common - they are used in nearly every case in which you are using a model (such as a linear model, through lm()) and you have variables that are factors. A benefit of using lm() (and many other functions) in R for modeling, such as the lme4::lmer() function we discuss later, is that if you have variables which are not factors, but simply character strings, they will be automatically changed to factors when used in a model. This means, for instance, that if you have a variable for the subject matter of courses labeled “mathematics”, “science”, “english language” (typed like that!), “social studies”, and “art”, and you include this variable in an lm() model, then the function will automatically dummy-code these for you. The only essential step that is not taken for you is choosing which is the reference group.

We note that there are cases in which *not having a reference group* that the other, dummy-coded groups are compared to is desired. In such cases, no intercept is estimated. This can be done by passing a -1 as the first value after the tilde, as follows:

# specifying the same linear model as the previous example, but using a "-1" to indicate that there should not be a reference group  
m\_linear\_dc\_2 <-   
 lm(final\_grade ~ -1 + TimeSpent\_std + course\_id, data = dat)  
  
tab\_model(m\_linear\_dc\_2,  
 title = "Table 13.3")

Table 13.3

final grade

Predictors

Estimates

CI

p

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [PhysA-S116-01]

88.55

82.83 – 94.27

<0.001

course\_id [AnPhA-S116-01]

73.20

67.20 – 79.20

<0.001

course\_id [AnPhA-S116-02]

71.61

64.38 – 78.83

<0.001

course\_id [AnPhA-S216-01]

64.15

58.12 – 70.17

<0.001

course\_id [AnPhA-S216-02]

68.69

58.35 – 79.04

<0.001

course\_id [AnPhA-T116-01]

80.44

68.20 – 92.67

<0.001

course\_id [BioA-S116-01]

69.64

62.89 – 76.40

<0.001

course\_id [BioA-S216-01]

58.53

42.74 – 74.32

<0.001

course\_id [BioA-T116-01]

82.38

55.04 – 109.72

<0.001

course\_id [FrScA-S116-01]

85.22

80.46 – 89.98

<0.001

course\_id [FrScA-S116-02]

70.06

57.18 – 82.94

<0.001

course\_id [FrScA-S116-03]

76.71

70.08 – 83.34

<0.001

course\_id [FrScA-S116-04]

78.43

59.08 – 97.78

<0.001

course\_id [FrScA-S216-01]

83.12

78.72 – 87.52

<0.001

course\_id [FrScA-S216-02]

80.57

72.51 – 88.64

<0.001

course\_id [FrScA-S216-03]

75.58

48.23 – 102.92

<0.001

course\_id [FrScA-S216-04]

88.60

71.31 – 105.89

<0.001

course\_id [FrScA-T116-01]

81.32

61.94 – 100.71

<0.001

course\_id [OcnA-S116-01]

77.26

69.49 – 85.03

<0.001

course\_id [OcnA-S116-02]

75.22

64.88 – 85.56

<0.001

course\_id [OcnA-S116-03]

54.45

15.80 – 93.10

0.006

course\_id [OcnA-S216-01]

66.79

60.50 – 73.07

<0.001

course\_id [OcnA-S216-02]

70.44

61.57 – 79.31

<0.001

course\_id [OcnA-T116-01]

71.15

57.48 – 84.81

<0.001

course\_id [PhysA-S216-01]

78.60

68.94 – 88.27

<0.001

course\_id [PhysA-T116-01]

93.93

66.60 – 121.27

<0.001

Observations

573

R2 / R2 adjusted

0.943 / 0.940

Table 13.3

final grade

Predictors

Estimates

CI

p

TimeSpent\_std

9.66

7.91 – 11.40

<0.001

course\_id [PhysA-S116-01]

88.55

82.83 – 94.27

<0.001

course\_id [AnPhA-S116-01]

73.20

67.20 – 79.20

<0.001

course\_id [AnPhA-S116-02]

71.61

64.38 – 78.83

<0.001

course\_id [AnPhA-S216-01]

64.15

58.12 – 70.17

<0.001

course\_id [AnPhA-S216-02]

68.69

58.35 – 79.04

<0.001

course\_id [AnPhA-T116-01]

80.44

68.20 – 92.67

<0.001

course\_id [BioA-S116-01]

69.64

62.89 – 76.40

<0.001

course\_id [BioA-S216-01]

58.53

42.74 – 74.32

<0.001

course\_id [BioA-T116-01]

82.38

55.04 – 109.72

<0.001

course\_id [FrScA-S116-01]

85.22

80.46 – 89.98

<0.001

course\_id [FrScA-S116-02]

70.06

57.18 – 82.94

<0.001

course\_id [FrScA-S116-03]

76.71

70.08 – 83.34

<0.001

course\_id [FrScA-S116-04]

78.43

59.08 – 97.78

<0.001

course\_id [FrScA-S216-01]

83.12

78.72 – 87.52

<0.001

course\_id [FrScA-S216-02]

80.57

72.51 – 88.64

<0.001

course\_id [FrScA-S216-03]

75.58

48.23 – 102.92

<0.001

course\_id [FrScA-S216-04]

88.60

71.31 – 105.89

<0.001

course\_id [FrScA-T116-01]

81.32

61.94 – 100.71

<0.001

course\_id [OcnA-S116-01]

77.26

69.49 – 85.03

<0.001

course\_id [OcnA-S116-02]

75.22

64.88 – 85.56

<0.001

course\_id [OcnA-S116-03]

54.45

15.80 – 93.10

0.006

course\_id [OcnA-S216-01]

66.79

60.50 – 73.07

<0.001

course\_id [OcnA-S216-02]

70.44

61.57 – 79.31

<0.001

course\_id [OcnA-T116-01]

71.15

57.48 – 84.81

<0.001

course\_id [PhysA-S216-01]

78.60

68.94 – 88.27

<0.001

course\_id [PhysA-T116-01]

93.93

66.60 – 121.27

<0.001

Observations

573

R2 / R2 adjusted

0.943 / 0.940

This does not work in many cases, and it is much more common to dummy-code factors, and so we emphasized that in this walkthrough. However, we want you to be aware that it is possible (though uncommon) to estimate a model without an intercept.

### A Deep-Dive into Multi-Level Models

Dummy-coding is a very helpful strategy. It is particularly useful with a small number of groups (i.e., for estimating the effects of being in one of the five subjects in the online science data set, as in this walkthrough; we note that in addition to these five subjects, we also have multiple sections, or classes, for each subject). With effects such as being a student in a particular class, though, the output seems to be less useful: it is hard to interpret the 25 different effects (and to compare them to the intercept).

Additionally, analysts often have the goal not of determining the effect of being in a specific class, *per se*, but rather of accounting for the fact that students share a class. This is important because linear models (i.e., those estiated using lm()) have an assumption that the data points are - apart from sharing levels of the variables that are used in the model - independent, or not correlated. This is what is meant by the “assumption of independence” or of “independently and identically distributed” (*i.i.d.*) residuals (Field, Miles, & Field, 2012).

As we noted in the chapter overview, multi-level models are a way to deal with the difficulty of interpreting the estimated effects for each of many groups, like classes, and to address the assumption of independence. Multi-level models do this by (still) estimating the effect of being a student in each group, but with a key distinction from linear models: Instead of determining how different the observations in a group are from those in the reference group, the multi-level model “regularizes” (sometimes the term “shrinks” is used) the difference based on how systematically different the groups are. The reason why “shrinkage” is occasionally used is that the group-level estimates (e.g., for classes) that are obtained through multi-level modeling can never be larger than those from a linear model (regression). As described earlier, when there are groups included in the model, a regression effectively estimates the effect for each group independent of all of the others.

Through regularization, groups that comprise individuals who are consistently different (higher or lower) than individuals on average are not regularized very much - their estimated difference may be close to the estimate from a multi-level model - whereas groups with only a few individuals, or with a lot of variability within individuals, would be regularized a lot~ The way that a multi-level model does this “regularizing” is by considering the groups (and not the data points, in this case) to be samples from a larger population of classes. By considering the effects of groups to be samples from a larger population, the model is able to use information not only particular to each group (as the models created using lm()), but also information across all of the data.

Using multi-level models, then, means that the assumption of independence can be addressed; their use also means that individual coefficients for classes do not need to be included (or interpreted, thankfully!), though they are still included in and accounted for in the model. As we describe, the way that information about the groups is reported is usually in the form of the *intra-class correlation coefficient* (ICC), which explains the proportion of variation in the dependent variable that the groups explain. Smaller ICCs (such as ICCs with values of 0.05, representing 5% of the variation in the dependent variable) mean that the groups are not very important; larger ICCs, such as ICCs with values of 0.10 or larger (values as high as 0.50 are not uncommon!). ICCs that are larger would indicate that groups are important and that they have to do with a lot of the differences observed in the dependent variable (and that not including them may potentially ignore the assumption of independence in a case in which it may be important to recognize it - and lead to bias in the results).] In the former case, the multi-level model considers there to be strong evidence for a group effect, whereas in the latter, the model recognizes that there is less certainty about a group (class) effect for that particular group, in part because that group is small. Multi-level models are very common in educational research for cases such as this: accounting for the way in which students take the same classes, or even go to the same school (see Raudenbush & Bryk, 2002).

That was a lot of technical information about multi-level models; thank you for sticking with us through it!

We wanted to include this as multi-level models *are* common (and, we think, could usefully be even more common!).Consider how often the data you collect involves students nested (or grouped) in classes, or classes nested in schools (or even schools nested in districts - you get the picture!). Educational data is complex, and so it is not surprising that multi-level models may be encountered in educational data science analyses, reports, and articles.

### Multi-level model analysis

Fortunately, for all of the complicated details, multi-level models are very easy to use in R. This requires a new package. One of the most common for estimating these types of models is {lme4}. We use lme4::lmer() very similarly to the lm() function, but we pass it an additional argument about what the *groups* in the data are. This model is often referred to as a “varying intercepts” multi-level model. What is different between the groups is the effect of being a student in a class: the intercepts between groups vary.

You’ll only need to install {lme4} once to do the rest of this walkthrough. To install {lme4}, type this code in your console:

install.packages("lme4")

Now we can fit our multi-level model:

m\_course <-   
 lmer(final\_grade ~ TimeSpent\_std + (1|course\_id), data = dat)

To say *just* a bit more, there is a connection between multi-level models and Bayesian methods (Gelman and Hill ([2006](#ref-gelman2006data))); one way to think about the “regularizing” going on is that estimates for each group (class) are made taking account of the data across all of the groups (classes). The data for all of the classes can be interpreted as a *prior* for the group estimates.

In a way, what is going on above is straightforward (and similar to what we have seen with lm()), but, it is also different and potentially confusing. Parentheses are not commonly used with lm(); there is a term ((1|course\_id)) in parentheses. Also, the bar symbol - | - is not commonly used with lm().

As different as these (the parentheses and bar) are, they are used for a relatively straightforward purpose: to model the group (in this case, courses) in the data. With lmer(), these group terms are specified in parentheses - specifically, to the right of the bar. That is what the |course\_id part means - it is telling lmer that courses are groups in the data. The left side of the bar tells lmer that what we want to be specified are varying intercepts for each group (1 is used to denote the intercept).

There is potentially more to the story: in addition to the 1, variables which can be specified to have a different effect for each group can also be specified. These variables are referred to not as varying intercepts, but as varying slopes. We will not cover these in this walkthrough, but want you to be aware of them (we recommend the book by West, Welch, and Galecki [2014] provide an excellent walkthrough on how to specify varying slopes using lmer()). To say *just* a bit more, there is a connection between multi-level models and Bayesian methods (Gelman and Hill ([2006](#ref-gelman2006data))); one way to think about the “regularizing” going on is that estimates for each group (class) are made taking account of the data across all of the groups (classes). The data for all of the classes can be interpreted as a *prior* for the group estimates.

## Results

Let’s view the results using the tab\_model() function from {sjPlot} again.

tab\_model(m\_course,  
 title = "Table 13.4")

Table 13.4

final grade

Predictors

Estimates

CI

p

(Intercept)

75.63

72.41 – 78.84

<0.001

TimeSpent\_std

9.45

7.74 – 11.16

<0.001

Random Effects

σ2

385.33

τ00 course\_id

38.65

ICC

0.09

N course\_id

26

Observations

573

Marginal R2 / Conditional R2

0.170 / 0.246

Table 13.4

final grade

Predictors

Estimates

CI

p

(Intercept)

75.63

72.41 – 78.84

<0.001

TimeSpent\_std

9.45

7.74 – 11.16

<0.001

Random Effects

σ2

385.33

τ00 course\_id

38.65

ICC

0.09

N course\_id

26

Observations

573

Marginal R2 / Conditional R2

0.170 / 0.246

For lm() models, tab\_model() provides the output, including some fit statistics as well as coefficients and their standard errors and estimates. There are two things to note about lmer() output:

1. *p*-values are not automatically provided, due to debates in the wider field about how to calculate the degrees of freedom for coefficients[[2]](#footnote-377)
2. In addition to the coefficients, there are also estimates for how much variability there is between the groups.

As we mentioned earlier, a common way to understand how much variability is at the group level is to calculate the *intra-class* correlation. This value is the proportion of the variability in the outcome (the *y*-variable) that is accounted for solely by the groups identified in the model. There is a useful function in the {performance} package for doing this.

You can install the {performance} package by typing this code in your console:

install.packages("performance")

After that, try this function:

icc(m\_course)

## # Intraclass Correlation Coefficient  
##   
## Adjusted ICC: 0.091  
## Conditional ICC: 0.076

This shows that nearly 17% of the variability in the percentage of points students earned can be explained simply by knowing what class they are in.

There is much more to do with multi-level models. We briefly discuss a common extension to the model we just used, adding additional levels.

The data that we are using is all from one school, and so we cannot estimate a “two-level” model. Imagine, however, that instead of 26 classes, we had data from students from 230 classes, and that these classes were from 15 schools. We could estimate a two-level, varying intercepts (where there are now two groups with effects) model very similar to the model we estimated above, but simply with another group added for the school. The model will account for the way in which the classes are nested within the schools automatically (Bates, Maechler, Bolker, & Walker, 2015).

We don’t have a variable containing the name of different schools. If we did we could fit the model like this, where school\_id is the variable containing different schools:

# this model would specify a group effect for both the course and school  
m\_course\_school <-   
 lmer(final\_grade ~ TimeSpent + (1|course\_id) + (1|school\_id), data = dat)

Were we to estimate this model (and then use the icc() function), we would see two ICC values representing the proportion of the variation in the dependent variable explained by each of the two groups we added - the course *and* the school. A common question those using {lme4} have is whether it is necessary to explicitly nest the courses within schools; as long as the courses are unique labelled, this is not necessary to do.

You can add further still levels to the model, as the {lme4} package was designed for complex multi-level models (and even those with not nested, but crossed random effects; a topic beyond the scope of this walkthrough, but which is described in West, Welch, & Galecki, 2015).

## Conclusion

In this example (and in many examples in educational research), the groups are classes. But, multi-level models can be used for other cases in which data is associated with a common group. For example, if students respond to repeated measures (such as quizzes) over time, then the multiple quiz responses for each student could be considered to be “grouped” within students. In such a case, instead of specifying the model with the course as the “grouping factor”, students could be.

Moreover, multi-level models can include multiple groups (as noted above), even if the groups are of very different kinds (i.e., if students from multiple classes responded to multiple quizzes).

We note that the groups in multi-level models do not need to be nested: they can also be “crossed”, as may be the case for data from, for example, teachers in different schools who attended different teacher preparation programs: not every teacher in a school necessarily (or even likely) attended the same teacher preparation program, and graduates from every teacher preparation program are highly unlikely to all teach in the same school!

There is much more that can be done with multi-level models; we have more recommendations in the [Additional Resources](#c18) chapter.

Finally, as noted earlier, multi-level models have similarities to the Bayesian methods which are becoming more common among some R users - and educational data scientists. There are also references to recommended books on Bayesian methods in the additional resources chapter.

# Walkthrough 8: Predicting Students’ Final Grades Using Machine Learning Methods with Online Course Data

## Vocabulary

* machine learning
* random forest
* resampling
* rsquared
* training data
* test data
* tuning parameter
* variable importance measures

## Chapter Overview

In this chapter, we explore the use of machine learning methods for predicting a student-level outcome, students’ final grade in online courses. We use the same dataset used in [Walkthrough 1/Chapter 7](#c07) and [Walkthrough 7/Chapter 13](#c13)], but pursue a different aim than in those chapters: We focus on *predicting* an outcome, final grade, more than *explaining* how variables relate to an outcome, such as how the amount of time students spend on the course relate to their final grade. We illustrate the use of a particular, common (yet powerful) machine learning method, random forests, choosing to explore their use in-depth, rather than providing a more general overview of other machine learning methods. Though we focus on the use of random forests, many of the ideas explored in this chapter will likely extend and prove useful with respect to the use of other machine learning methods; our goal is for you to finish this, the final walkthrough, with the confidence to explore using machine learning to answer a question or to solve a problem of your own with respect to teaching, learning, and educational systems.

### Background

One area of interest is the delivery of online instruction, which is becoming more prevalent: in 2007, over 3.9 million U.S. students were enrolled one or more online courses (Allen and Seaman [2008](#ref-allen2008)). With the dawn of online learning comes an abundance of new educational tools to facilitate that learning. Indeed, online learning interfaces are used to facilitate the submission of assignments and quizzes in courses in which students and instructor meet face-to-face, but these interfaces are also used in fully online courses to deliver all instruction and assessment.

In a face-to-face classroom, an educator might count on behavioral cues to help them effectively deliver instruction. However, one constraint of online education is that educators do not have access as readily to the behavioral cues that can be essential to effective face-to-face instruction. For example, in a face-to-face classroom, cues such as a student missing class repeatedly or many students seeming distracted during a lecture can trigger a shift in the delivery of instruction. While technology is rapidly developing, many educators find themselves looking for ways to understand and support students online in the same way that face-to-face instructors would. Educational technology affords unique opportunities to support student success online because it provides new methods of collecting and storing data.

Indeed, online learning management systems often automatically track several types of student interactions with the system and feed that data back to the course instructor. For example, an instructor might be able to quickly see how many students logged into their course on a certain day, or they might see how long students engaged with a posted video before pausing it or logging out. The collection of this data is met with mixed reactions from educators. Some are concerned that data collection in this manner is intrusive, but others see a new opportunity to support students in online contexts in new ways. As long as data are collected and utilized responsibly, data collection can support student success.

One meaningful perspective from which to consider students’ engagement with online courses is related to their motivation to achieve. More specifically, it is important to consider how and why students are engaging with the course. Considering the psychological mechanisms behind achievement is valuable because doing so may help to identify meaningful points of intervention for educators and for researchers and administrators in online *and* face-to-face courses interested in the intersection between behavioral trace measures and students’ motivational and emotional experiences in such courses.

In this walkthrough, we examine the educational experiences of students in online science courses at a virtual middle school in order to characterize their motivation to achieve and their tangible engagement with the course in terms of behavioral trace measures. To do so, we use a robust data set, which includes self-reported motivation as well as behavioral trace data collected from a learning management system (LMS) to identify predictors of final course grade. Our work examines the idea of educational success in terms of student interactions with an online science course.

We explore the following four questions:

1. Is motivation more predictive of course grades as compared to other online indicators of engagement?
2. Which types of motivation are most predictive of achievement?
3. Which types of trace measures are most predictive of achievement?
4. How does a random forest compare to a simple linear model (regression)?

### Data Sources

This dataset came from 499 students enrolled in online middle school science courses in 2015-2016. The data were originally collected for use as a part of a research study, though the findings have not been published anywhere yet.

The setting of this study was a public provider of individual online courses in a Midwestern state. In particular, the context was two semesters (Fall and Spring) of offerings of five online science courses (Anatomy & Physiology, Forensic Science, Oceanography, Physics, and Biology), with a total of 36 classes.

Specific information in the dataset included:

* a pre-course survey students completed about their self-reported motivation in science — in particular, their perceived competence, utility value, and interest
* the time students spent on the course (obtained from the learning management system (LMS), Blackboard
* students’ final course grades
* students’ involvement in discussion forums

For discussion board responses, we were interested in calculating the number of posts per student and understanding the emotional tone of the discussion board posts. We used the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015) tool to calculate the number of posts per student and to categorize the emotional tone (positive or negative) and topics of those posts. Those linguistic categorization was conducted after the data was gathered from the discussion posts, but is not replicated here to protect the privacy of the students’ posts. Instead, we present the already-categorized discussion board data, in its ready-to-use format. In the dataset used in this walkthrough, we will see pre-created variables for the mean levels of students’ cognitive processing, positive emotions, negative emotions, and social-related discourse.

At the beginning of the semester, students were asked to complete the pre-course survey about their perceived competence, utility value, and interest. At the end of the semester, the time students spent on the course, their final course grades, and the contents of the discussion forums were collected.

In this walkthrough, we used the R package {caret} to carry out the analyses.

### Methods

#### Defining a Research Question

When you begin a new project, there are often many approaches to analyzing data and answering questions you might have about it. Some projects have a clearly defined scope and question to answer. This type of project is characterized by 1) a defined number of variables (data inputs) and 2) specific directional hypotheses. For example, if we are studying the effect of drinking coffee after dinner on ability to quickly fall asleep, we might have a very specific directional hypothesis: we expect that drinking coffee after dinner would decrease the ability to fall asleep quickly. In this case, we might collect data by having some people drink coffee and having other people drink nothing or an herbal tea before bed. We could monitor how quickly people from each group fall asleep. Since we collected data from two clearly defined groups, we can then do a statistical analysis that compares the amount of time it takes to fall asleep for each group. One option would be a test called a t-test, which we could use to see if there is a significant difference in the average amount of minutes to fall asleep for the group. This approach works very well in controlled experimental situations, especially when we can change only one thing at a time (in our coffee example, the only thing we changed was the coffee-drinking behavior of our participants - all other life conditions were held equal for both groups). Rarely are educational data projects as clear-cut and simple.

For this walkthrough, we have many sources of data - survey data, learning management system data, discussion forum data, and academic achievement data as measured by final course grades. Luckily, having too much data is what we call a “good problem.” In our coffee example above, we had one really specific idea that we wanted to investigate - does coffee affect time taken to fall asleep? In this walkthrough we have many ideas we are curious to explore: the relationships among motivation, engagement in the course (discussion boards, time spent online in the course site), and academic achievement. If we wanted to tackle a simpler problem, we could choose just one of these relationships. For example, we could measure whether students with high motivation earn higher grades than students with low motivation. However, we are being a bit more ambitious than that here - we are interested in understanding the complex relationships among the different types of motivation. Rather than simply exploring whether A affects B, we are interested in the nuances: we suspect that *many* factors affect B, and we would like to see which of those factors has most relative importance. To explore this idea, we will use a machine learning approach.

#### Predictive Analytics and Machine Learning

A buzzword in education software spheres these days is “predictive analytics.” Administrators and educators alike are interested in applying the methods long utilized by marketers and other business professionals to try to determine what a person will want, need, or do next. “Predictive analytics” is a blanket term that can be used to describe any statistical approach that yields a prediction. We could ask a predictive model: “What is the likelihood that my cat will sit on my keyboard today?” and, given enough past information about your cat’s computer-sitting behavior, the model could give you a probability of that computer-sitting happening today. Under the hood, some predictive models are not very complex. If we have an outcome with two possibilities, a logistic regression model could be fit to the data in order to help us answer the cat-keyboard question. In this chapter, we’ll compare a machine learning model to another type of regression: multiple regression. We want to make sure to fit the simplest model as possible to our data. After all, the effectiveness in predicting the outcome is really the most important thing: not the fanciness of the model.

Data collection is an essential first step in any type of machine learning or predictive analytics. It is important to note here that machine learning only works effectively when (1) a person selects variables to include in the model that are anticipated to be related to the outcome and (2) a person correctly interprets the model’s findings. There is an adage that goes, “garbage in, garbage out.” This holds true here: if we do not feel confident that the data we collected are accurate, no matter what model we build, we will not be able to be confident in our conclusions. To collect good data, we must first clarify what it is that we want to know (i.e., what question are we really asking?) and what information we would need in order to effectively answer that question. Sometimes, people approach analysis from the opposite direction - they might look at the data they have and ask what questions could be answered based on that data. That approach is okay - as long as you are willing to acknowledge that sometimes the pre-existing dataset may *not* contain all the information you need, and you might need to go out and find additional information to add to your dataset to truly answer your question.

When people talk about “machine learning,” you might get the image in your head of a desktop computer learning how to spell. You might picture your favorite social media site showing you advertisements that are just a little too accurate. At its core, what machine learning really is is the process of “showing” your statistical model only some of the data at once, and training the model to predict accurately on that training dataset (this is the “learning” part of machine learning). Then, the model as developed on the training data is shown new data - data you had all along, but hid from your computer initially - and you see how well the model that you developed on the training data performs on this new testing data. Eventually, you might use the model on entirely new data.

#### Random Forest

For our analyses, we used Random Forest modeling (Breiman [2001](#ref-breiman2001)). Random forest is an extension of decision tree modeling, whereby a collection of decision trees are simultaneously “grown” and are evaluated based on out-of-sample predictive accuracy (Breiman [2001](#ref-breiman2001)). Random forest is random in two main ways: first, each tree is only allowed to “see” and split on a limited number of predictors instead of all the predictors in the parameter space; second, a random subsample of the data is used to grow each individual tree, such that no individual case is weighted too heavily in the final prediction.

One thing about random forest that makes it quite different from other types of analysis we might do is that here, we are giving the computer a large amount of information and asking it to find connections that might not be immediately visible to the naked human eye. This is great for a couple of reasons. First, while humans are immensely creative and clever, we are not immune to biases. If we are exploring a dataset, we usually come in with some predetermined notions about what we think is true, and we might (consciously or unconsciously) seek evidence that supports the hypothesis we privately hold. By setting the computer loose on some data, we can learn that there are connections between areas that we did not expect. We must also be ready for our hypotheses to not be supported! Random forest is particularly well-suited to the research questions explored here because we do not have specific directional hypotheses. Machine learning researchers talk about this as “exploring the parameter space” - we want to see what connections exist, and we acknowledge that we might not be able to accurately predict all the possible connections. Indeed, we expect - and hope - that we will find surprising connections.

Whereas some machine learning approaches (e.g., boosted trees) would utilize an iterative model-building approach, random forest estimates all the decision trees at once. In this way, each tree is independent of every other tree. Thus, the random forest algorithm provides a robust regression approach that is distinct from other modeling approaches. The final random forest model aggregates the findings across all the separate trees in the forest in order to offer a collection of “most important” variables as well as a percent variance explained for the final model.

500 trees were grown as part of our random forest. We partitioned the data before conducting the main analysis so that neither the training nor the testing data set would be disproportionately representative of high-achieving or low-achieving students. The training data set consisted of 80% of the original data (*n* = 400 cases), whereas the testing data set consisted of 20% of the original data (*n* = 99 cases). We built our random forest model on the training data set, and then evaluated the model on the testing data set. Three variables were tried at each node.

Note that the random forest algorithm does not accept cases with missing data, and so we deleted cases listwise if data were missing. This decision eliminated 51 cases from our original data set, to bring us to our final sample size of 499 unique students. If you have a very small dataset with a lot of missing data, the random forest approach may not be well suited for your goals – you might consider a linear regression instead.

A random forest is well suited to the research questions that we had here because it allows for nonlinear modeling. We hypothesized complex relationships between students’ motivation, their engagement with the online courses, and their achievement. For this reason, a traditional regressive or structural equation model would have been insufficient to model the parameter space we were interesting in modeling. Our random forest model had one outcome and eleven predictors.

One term you will hear used in machine learning is “tuning parameter.” People often think of tuning parameters as knobs or dials on a radio: they are features of the model that can be adjusted to get the clearest signal. A common tuning parameter for machine learning models is the number of variables considered at each split (Kuhn and others [2008](#ref-kuhn2008)); we considered three variables at each split for this analysis.

The outcome was the final course grade that the student earned. The predictor variables included motivation variables (interest value, utility value, and science perceived competence) and trace variables (the amount of time spent in the course, the course name, the number of discussion board posts over the course of the semester, the mean level of cognitive processing evident in discussion board posts, the positive emotions evident in discussion board posts, the negative emotions evident in discussion board posts, and the social-related discourse evident in their discussion board posts). We used this random forest model to address all three of our research questions.

To interpret our findings, we will consider three main factors: (1) predictive accuracy of the random forest model, (2) variable importance, and (3) variance explained by the final random forest model. In this walkthrough, we will use the R package {caret} (Kuhn [2020](#ref-R-caret)) to carry out the analysis. We also use the {tidylog} package (Elbers [2020](#ref-R-tidylog)) to help us to understand how the data processing steps we take have the desired effect.

## Load Packages

First, we will load the data. Our data is stored in the {dataedu} package that is part of this book. Within that package, the data is stored as an .rda file. We note that this data is augmented to have some other - and additional - variables that the sci\_mo\_processed data used in [Chapter 7](#c7) and [Chapter 14](#c14) does not.

## Import and View Data

#loading the data from the .rda file and storing it as an object named 'data'  
data <- dataedu::sci\_mo\_with\_text

It’s a good practice to take a look at the data and make sure it looks the way you expect it to look. R is pretty smart, but sometimes we run into issues like column headers being read as datapoints. By using the glimpse() function from the {dplyr} package, we can quickly skim our data and see whether we have all the right variables and data points. Remember that the {dplyr} package loads automatically when we load the {tidyverse} library, so there is no need to call the {dplyr} package separately. Now, we’ll glimpse the data.

glimpse(data)

## Observations: 606  
## Variables: 74  
## $ student\_id <dbl> 43146, 44638, 47448, 47979, 48797, 51943, 52326…  
## $ course\_id <chr> "FrScA-S216-02", "OcnA-S116-01", "FrScA-S216-01…  
## $ total\_points\_possible <dbl> 3280, 3531, 2870, 4562, 2207, 4208, 4325, 2086,…  
## $ total\_points\_earned <dbl> 2220, 2672, 1897, 3090, 1910, 3596, 2255, 1719,…  
## $ percentage\_earned <dbl> 0.677, 0.757, 0.661, 0.677, 0.865, 0.855, 0.521…  
## $ subject <chr> "FrScA", "OcnA", "FrScA", "OcnA", "PhysA", "FrS…  
## $ semester <chr> "S216", "S116", "S216", "S216", "S116", "S216",…  
## $ section <chr> "02", "01", "01", "01", "01", "03", "01", "01",…  
## $ Gradebook\_Item <chr> "POINTS EARNED & TOTAL COURSE POINTS", "ATTEMPT…  
## $ Grade\_Category <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ final\_grade <dbl> 93.5, 81.7, 88.5, 81.9, 84.0, NA, 83.6, 97.8, 9…  
## $ Points\_Possible <dbl> 5, 10, 10, 5, 438, 5, 10, 10, 443, 5, 12, 10, 5…  
## $ Points\_Earned <dbl> NA, 10.0, NA, 4.0, 399.0, NA, NA, 10.0, 425.0, …  
## $ Gender <chr> "M", "F", "M", "M", "F", "F", "M", "F", "F", "M…  
## $ q1 <dbl> 5, 4, 5, 5, 4, NA, 5, 3, 4, NA, NA, 4, 3, 5, NA…  
## $ q2 <dbl> 4, 4, 4, 5, 3, NA, 5, 3, 3, NA, NA, 5, 3, 3, NA…  
## $ q3 <dbl> 4, 3, 4, 3, 3, NA, 3, 3, 3, NA, NA, 3, 3, 5, NA…  
## $ q4 <dbl> 5, 4, 5, 5, 4, NA, 5, 3, 4, NA, NA, 5, 3, 5, NA…  
## $ q5 <dbl> 5, 4, 5, 5, 4, NA, 5, 3, 4, NA, NA, 5, 4, 5, NA…  
## $ q6 <dbl> 5, 4, 4, 5, 4, NA, 5, 4, 3, NA, NA, 5, 3, 5, NA…  
## $ q7 <dbl> 5, 4, 4, 4, 4, NA, 4, 3, 3, NA, NA, 5, 3, 5, NA…  
## $ q8 <dbl> 5, 5, 5, 5, 4, NA, 5, 3, 4, NA, NA, 4, 3, 5, NA…  
## $ q9 <dbl> 4, 4, 3, 5, NA, NA, 5, 3, 2, NA, NA, 5, 2, 2, N…  
## $ q10 <dbl> 5, 4, 5, 5, 3, NA, 5, 3, 5, NA, NA, 4, 4, 5, NA…  
## $ time\_spent <dbl> 1555.17, 1382.70, 860.43, 1598.62, 1481.80, 3.4…  
## $ TimeSpent\_hours <dbl> 25.9194, 23.0450, 14.3406, 26.6436, 24.6967, 0.…  
## $ TimeSpent\_std <dbl> -0.1805, -0.3078, -0.6933, -0.1484, -0.2347, -1…  
## $ int <dbl> 5.0, 4.2, 5.0, 5.0, 3.8, 4.6, 5.0, 3.0, 4.2, NA…  
## $ pc <dbl> 4.50, 3.50, 4.00, 3.50, 3.50, 4.00, 3.50, 3.00,…  
## $ uv <dbl> 4.33, 4.00, 3.67, 5.00, 3.50, 4.00, 5.00, 3.33,…  
## $ enrollment\_status <chr> "Approved/Enrolled", "Approved/Enrolled", "Appr…  
## $ enrollment\_reason <chr> "Course Unavailable at Local School", "Course U…  
## $ cogproc <dbl> 15.07, 7.11, 15.17, 14.51, 16.69, 11.98, 14.93,…  
## $ male <dbl> 0.5121, 0.0000, 0.1112, 0.0000, 0.0000, 0.5450,…  
## $ female <dbl> 0.1666, 0.0000, 0.1522, 0.0000, 0.4260, 0.0000,…  
## $ friend <dbl> 0.0000, 0.0000, 0.0127, 0.0000, 0.0000, 0.6600,…  
## $ family <dbl> 0.00605, 0.00000, 0.08488, 0.00000, 0.00000, 0.…  
## $ social <dbl> 6.20, 6.14, 5.05, 6.13, 7.53, 6.89, 7.37, 5.69,…  
## $ sad <dbl> 0.1808, 0.0000, 0.0910, 0.0000, 0.4350, 0.0000,…  
## $ anger <dbl> 0.4187, 0.0000, 0.1410, 0.1080, 0.1060, 0.0000,…  
## $ anx <dbl> 0.08000, 0.00000, 0.27537, 0.78800, 0.54500, 0.…  
## $ negemo <dbl> 1.136, 0.000, 1.419, 1.152, 1.282, 0.000, 0.562…  
## $ posemo <dbl> 3.56, 19.01, 2.91, 5.59, 3.79, 4.92, 5.69, 5.14…  
## $ affect <dbl> 4.76, 19.01, 4.33, 6.74, 5.08, 4.92, 6.25, 5.68…  
## $ quant <dbl> 2.05, 2.74, 3.25, 3.21, 2.55, 1.08, 1.92, 2.59,…  
## $ number <dbl> 0.913, 3.473, 2.307, 0.257, 0.213, 0.000, 1.508…  
## $ interrog <dbl> 1.286, 0.443, 1.787, 1.103, 1.715, 2.400, 1.729…  
## $ compare <dbl> 2.421, 4.147, 3.902, 2.699, 3.944, 1.085, 2.242…  
## $ adj <dbl> 5.11, 5.48, 5.61, 5.21, 4.62, 3.81, 4.79, 4.55,…  
## $ verb <dbl> 18.1, 11.0, 16.3, 16.3, 17.1, 17.9, 16.3, 18.0,…  
## $ negate <dbl> 1.206, 0.000, 1.681, 1.130, 0.748, 0.660, 0.684…  
## $ conj <dbl> 5.57, 6.66, 5.37, 6.20, 7.24, 6.44, 7.58, 5.17,…  
## $ adverb <dbl> 6.24, 6.66, 5.82, 5.31, 6.49, 5.34, 6.86, 7.25,…  
## $ auxverb <dbl> 11.30, 9.25, 10.23, 8.89, 9.49, 11.36, 9.71, 8.…  
## $ prep <dbl> 12.30, 11.85, 12.13, 13.63, 12.82, 14.19, 10.73…  
## $ article <dbl> 7.83, 2.22, 6.77, 9.12, 9.83, 5.66, 7.01, 7.66,…  
## $ ipron <dbl> 6.94, 2.74, 5.15, 4.33, 7.84, 4.49, 7.78, 5.30,…  
## $ they <dbl> 1.0103, 0.0000, 0.8434, 1.8630, 0.0980, 2.5200,…  
## $ shehe <dbl> 0.5434, 0.0000, 0.1695, 0.0000, 0.4260, 0.0000,…  
## $ you <dbl> 1.744, 3.473, 1.149, 2.049, 2.623, 0.000, 3.274…  
## $ we <dbl> 0.0658, 0.0000, 0.0332, 0.3020, 0.4910, 0.0000,…  
## $ i <dbl> 3.65, 7.99, 4.69, 3.45, 3.14, 7.00, 5.10, 3.83,…  
## $ ppron <dbl> 7.01, 11.47, 6.88, 7.66, 6.78, 9.52, 9.34, 6.60…  
## $ pronoun <dbl> 13.99, 14.21, 12.03, 12.22, 14.62, 14.02, 17.12…  
## $ `function` <dbl> 55.2, 44.6, 49.4, 53.1, 57.5, 55.4, 55.2, 51.7,…  
## $ Dic <dbl> 86.3, 86.3, 80.7, 86.5, 90.5, 83.0, 87.3, 83.3,…  
## $ Sixltr <dbl> 20.9, 22.2, 20.8, 21.8, 15.3, 25.0, 16.3, 20.5,…  
## $ WPS <dbl> 17.41, 9.83, 17.92, 18.82, 15.66, 16.80, 13.54,…  
## $ Tone <dbl> 56.6, 96.4, 49.4, 78.4, 55.4, 91.3, 59.4, 78.7,…  
## $ Authentic <dbl> 44.1, 70.3, 41.2, 49.0, 42.2, 34.1, 39.9, 35.1,…  
## $ Clout <dbl> 49.5, 53.6, 40.1, 53.1, 54.1, 40.0, 55.5, 52.5,…  
## $ Analytic <dbl> 55.7, 56.0, 59.0, 70.0, 55.8, 54.5, 41.8, 70.4,…  
## $ WC <dbl> 88.3, 34.7, 69.3, 61.2, 47.1, 84.0, 80.5, 59.2,…  
## $ n <dbl> 38, 3, 41, 10, 10, 2, 21, 18, 31, 37, 37, 18, 1…

Scanning the data we glimpsed, we see that we have 662 observations and 111 variables. Many of these variables - everything below WC except the variable n - are related to the text content of the discussion board posts. Our analysis here is not focused on the specifics of the discussion board posts, so we will select just a few variables from the LIWC analysis. If you’re interested in learning more about analyzing text, the text analysis walkthrough in this volume will be a good place to start.

As is the case with many datasets you’ll work with in education contexts, there is lots of great information in this dataset - but we won’t need all of it. Even if your dataset has many variables, for most analyses you will find that you are only interested in some of them. There are statistical reasons not to include twenty or more variables in a data analysis, and the quick explanation of the reason why is that at a certain point, adding more variables will *appear* to make your analysis more accurate, but will in fact obscure the truth from you. It’s generally a good practice to select a few variables you are interested in and go from there. As we discussed above, the way to do this is to start with the research questions you are trying to answer. Since we are interested in data from one specific semester, we’ll need to narrow down the data to make sure that we only include datapoints relevant to that semester.

Thus, we will *filter* the data to include only the data from one that semester, and then *select* variables of interest. For each step, we will save over the previous version of the “data” object so that our working environment doesn’t get cluttered with each new version of the dataset. Keep in mind that the original data will stay intact, and that any changes we make to it within R will not overwrite that original data (unless we tell R to specifically save out a new file with exactly the same name as the original file). Changes we make within our working environment are all totally reversible.

Below, we will *filter* to remove all the datapoints from the spring 2017 semester (indicated with a value of S217 for the semester variable). We use the ! to indicate that we want to keep all datapoints EXCEPT the datapoints that have a value of S217 for the semester variable. Then, we will *select* only the variables we are interested in: motivation, time spent in the course, grade in the course, subject, enrollment information, positive and negative emotions, cognitive processing, and the number of discussion board posts.

## Process Data

#filtering the data to only include 2016 data.   
  
data <-   
 data %>%   
 filter(semester != "S217")

## filter: no rows removed

#selecting only the variables we are interested in:   
data <-   
 data %>%  
 select(  
 int,  
 uv,  
 pc,  
 time\_spent,  
 final\_grade,  
 subject,  
 enrollment\_reason,  
 semester,  
 enrollment\_status,  
 cogproc,  
 social,  
 posemo,  
 negemo,  
 n  
 )

## select: dropped 60 variables (student\_id, course\_id, total\_points\_possible, total\_points\_earned, percentage\_earned, …)

## Analysis

### Use of {caret}

Here, we remove observations with missing data (per our note above about random forests requiring complete cases).

# checking how many rows are in our dataset  
# we see that we have 550 rows from spring 2017  
nrow(data)

## [1] 606

# calling the na.omit function to eliminate ANY rows that have ANY missing data  
data <- na.omit(data)

# checking whether our na.omit call worked as expected  
# after running the code above, we see that we now have 499 rows - this is as we expected  
nrow(data)

## [1] 464

First, machine learning methods often involve using a large number of variables. Oftentimes, some of these variables will not be suitable to use: they may be highly correlated with other variables, for instance, or may have very little - or no - variability. Indeed, for the data set used in this study, one variable has the same (character string) value for all of the observations. We can detect this variable and any others using the following function:

# swe run the nearZeroVar function to determine   
# if there are variables with NO variability  
nearZeroVar(data, saveMetrics = TRUE)

## freqRatio percentUnique zeroVar nzv  
## int 1.31 9.052 FALSE FALSE  
## uv 1.53 6.466 FALSE FALSE  
## pc 1.49 3.879 FALSE FALSE  
## time\_spent 1.00 98.707 FALSE FALSE  
## final\_grade 1.33 92.241 FALSE FALSE  
## subject 1.65 1.078 FALSE FALSE  
## enrollment\_reason 3.15 1.078 FALSE FALSE  
## semester 1.23 0.647 FALSE FALSE  
## enrollment\_status 0.00 0.216 TRUE TRUE  
## cogproc 1.33 83.190 FALSE FALSE  
## social 1.00 70.690 FALSE FALSE  
## posemo 1.00 66.164 FALSE FALSE  
## negemo 8.67 89.655 FALSE FALSE  
## n 1.33 10.129 FALSE FALSE

After conducting our zero variance check, we want to scan the zeroVar column to see if any of our variables failed this check. If we see any “TRUE” values for zeroVar, that means we should look more closely at that variable.

In the nearZeroVar() function we just ran, we see a result in the ZeroVar column of “TRUE” for the enrollment\_status variable. If we look at enrollment\_status, we will see that it is “Approved/Enrolled” for *all* of the students. When we use variables with no variability in certain models, it may cause some problems, and so we remove it first.

# Taking the dataset and re-saving it as the same dataset,   
# but without the enrollment status variable  
data <-   
 data %>%   
 select(-enrollment\_status)

## select: dropped one variable (enrollment\_status)

Note that many times you may wish to pre-process the variables, such as by centering or scaling them. Often the data will come to you in a format that is not ready for immediate analysis, as we have discussed elsewhere in the book. For our current dataset, we could work on pre-processing with code like you will see below. We set this next code chunk up to not run here (if you are viewing the book online), as we will do this analysis with the variables’ original values.

# example pre-processing step: manipulating the dataset 'data'   
# so that if a variable is numeric, its format will now be scale  
data <-   
 data %>%   
 mutate\_if(is.numeric, scale)

As another pre-processing step, we want to make sure our text data is in a format that we can then evaluate. To facilitate that, we want to make character string variables into factors.

# converting the text (character) variables in our dataset into factors  
data <-   
 data %>%   
 mutate\_if(is.character, as.factor)

## mutate\_if: converted 'subject' from character to factor (0 new NA)

## converted 'enrollment\_reason' from character to factor (0 new NA)

## converted 'semester' from character to factor (0 new NA)

Now, we will prepare the **train** and **test** datasets, using the caret function for creating data partitions. Here, the **p** argument specifies what proportion of the data we want to be in the **training** partition. Note that this function splits the data based upon the outcome, so that the training and test data sets will both have comparable values for the outcome. This means that since our outcome is final grade, we are making sure that we don’t have either a training or testing dataset that has too many good grades - or too many bad grades. Note the times = 1 argument; this function can be used to create *multiple* train and test sets, something we will describe in more detail later. Before we create our training and testing datasets, we want to initiate a process called “setting the seed.” This means that we are ensuring that if we run this same code again, we will get the same results in terms of the data partition. The seed can be any number that you like - some people choose their birthday or another meaningful number. The only constraint is that when you open the same code file again to run in the future, you do not change the number you selected for your seed. This ensures your code is reproducible. In fact, it ensures that anyone who runs the same code file on any computer, anywhere, will get the same result. With that background information, try running the code chunk below.

# First, we set a seed to ensure the reproducibility of our data partition.  
set.seed(2020)  
  
# we create a new object called trainIndex that will take 80 percent of the data  
trainIndex <- createDataPartition(data$final\_grade,  
 p = .8,   
 list = FALSE,  
 times = 1)  
  
# We add a new variable to our dataset, temporarily:  
# this will let us select our rows according to their row number  
# we populate the rows with the numbers 1:499, in order  
  
data <-   
 data %>%   
 mutate(temp\_id = 1:464)

## mutate: new variable 'temp\_id' with 464 unique values and 0% NA

# we filter our dataset so that we get only the   
# rows indicated by our "trainIndex" vector  
data\_train <-   
 data %>%   
 filter(temp\_id %in% trainIndex)

## filter: removed 92 rows (20%), 372 rows remaining

# we filter our dataset in a different way so that we get only the rows   
# NOT in our "trainIndex" vector   
# adding the ! before the temp\_id variable achieves the opposite of   
# what we did in the line of code above  
  
data\_test <-   
 data %>%   
 filter(!temp\_id %in% trainIndex)

## filter: removed 372 rows (80%), 92 rows remaining

# We delete the temp\_id variable from (1) the original data,   
# (2) the portion of the original data we marked as training, and   
# (3) the portion of the original data we marked as testing,   
# as we no longer need that variable  
  
data <-   
 data %>%   
 select(-temp\_id)

## select: dropped one variable (temp\_id)

data\_train <-   
 data\_train %>%   
 select(-temp\_id)

## select: dropped one variable (temp\_id)

data\_test <-   
 data\_test %>%   
 select(-temp\_id)

## select: dropped one variable (temp\_id)

Finally, we will estimate the models.

Here, we will use the train function, passing *all* of the variables in the data frame (except for the outcome, or dependent variable, final\_grade) as predictors.

The predictor variables include three indicators of motivation: interest in the course (int), perceived utility value of the course (uv), and perceived competence for the subject matter (pc). There are a few predictor variables that help differentiate between the different courses in the dataset: subject matter of the course (subject), reason the student enrolled in the course (enrollment\_reason), and semester in which the course took place (semester). We have a predictor variable that indicates the amount of time each student spent engaging with the online learning platform of the course (time\_spent).

We also have a number of variables associated with the discussion board posts from the course. Specifically, the variables include the average level of cognitive processing in the discussion board posts (cogproc), the average level of social (rather than academic) content in the discussion board posts (social), the positive and negative emotions evident in the discussion board posts (posemo and negemo), and finally, the number of discussion board posts in total (n). We are using all those variables discussed in this paragraph to predict the outcome of the final grade in the course (final\_grade).

Note that you can read more about the specific random forest implementation chosen [in the {caret} bookdown page](http://topepo.github.io/caret/train-models-by-tag.html#random-forest) (<http://topepo.github.io/caret/train-models-by-tag.html#random-forest>). To specify that we want to predict the outcome using every variable except the outcome itself, we use the formulation outcome ~ .. R interprets this code as: predict the outcome using all the variables except outcome itself. The outcome always comes before the ~, and the . that we see after the ~ means that we want to use all the rest of the variables. An alternative specification of this model would be to write (outcome ~ predictor1, predictor2). Anything that follows the ~ and precedes the comma is treated as predictors of the outcome.

Here, we set the seed again, to ensure that our analysis is reproducible. This step of setting the seed is especially important due to the “random” elements of random forest, because it’s likely that the findings would change (just slightly) if the seed were not set. As we get into random forest modeling, you might notice that the code takes a bit longer to run. This is normal - just think of the number of decision trees that are “growing”!

# setting a seed for reproducibility  
set.seed(2020)  
  
# we run the model here  
rf\_fit <- train(final\_grade ~ .,  
 data = data\_train,  
 method = "ranger")  
  
# here, we get a summary of the model we just built  
rf\_fit

## Random Forest   
##   
## 372 samples  
## 12 predictor  
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 372, 372, 372, 372, 372, 372, ...   
## Resampling results across tuning parameters:  
##   
## mtry splitrule RMSE Rsquared MAE   
## 2 variance 15.3 0.565 11.18  
## 2 extratrees 17.0 0.523 12.06  
## 10 variance 14.0 0.593 10.14  
## 10 extratrees 13.9 0.621 10.07  
## 19 variance 14.1 0.589 10.02  
## 19 extratrees 13.4 0.632 9.77  
##   
## Tuning parameter 'min.node.size' was held constant at a value of 5  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were mtry = 19, splitrule = extratrees  
## and min.node.size = 5.

We have some results! First, we see that we have 372 samples, or 372 observations, the number in the train data set. No pre-processing steps were specified in the model fitting, but note that the output of preProcess can be passed to train() to center, scale, and transform the data in many other ways. Next, in our example, a resampling technique has been used. This resampling is not for validating the model (per se), but is rather for selecting the tuning parameters - the options that need to be specified as a part of the modeling. These parameters can be manually provided, or can be estimated via strategies such as the bootstrap resample (or *k*-folds cross validation).

As we interpret these findings, we are looking to minimize the error (RMSE) and maximize the variance explained (rsquared).

It appears that the model with the value of the mtry tuning parameter equal to 19 seemed to explain the data best, the splitrule being “extratrees”, and min.node.size held constant at a value of 5. We know this model fits best because the RMSE is the lowest of the options (13.64) and the Rsquared is the highest of the options (.613).

The value of resampling here is that it allows for higher accuracy of the model (James et al. [2013](#ref-james2013)). Without resampling (bootstrapping or cross-validation), the variance would be higher and the predictive accuracy of the model would be lower.

Let’s see if we end up with slightly different values if we change the resampling technique to cross-validation, instead of bootstrap resampling. We set a seed again here, for reproducibility.

set.seed(2020)  
  
train\_control <-  
 trainControl(method = "repeatedcv",  
 number = 10,  
 repeats = 10)  
  
rf\_fit1 <-  
 train(final\_grade ~ .,  
 data = data\_train,  
 method = "ranger",  
 trControl = train\_control)  
  
rf\_fit1

## Random Forest   
##   
## 372 samples  
## 12 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 335, 334, 334, 336, 334, 334, ...   
## Resampling results across tuning parameters:  
##   
## mtry splitrule RMSE Rsquared MAE   
## 2 variance 14.6 0.590 10.97  
## 2 extratrees 16.2 0.585 11.79  
## 10 variance 13.3 0.618 9.83  
## 10 extratrees 13.0 0.660 9.66  
## 19 variance 13.3 0.617 9.68  
## 19 extratrees 12.6 0.664 9.36  
##   
## Tuning parameter 'min.node.size' was held constant at a value of 5  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were mtry = 19, splitrule = extratrees  
## and min.node.size = 5.

### Tuning the random forest model

When we look at this output, we are looking to see which values of the various tuning parameters were selected. We see at the bottom of the output above that the value of mtry was 19, the split rule was “extratrees,” and the minimum node size is 5. We let this model explore which value of mtry was best and to explore whether extra trees or variance was a better split rule, but we forced all iterations of the model to a minimum node size of five (so that minimum node size value in the output shouldn’t be a surprise to us). When we look at the bottom row of the output, it shows the final values selected for the model. We see also that this row has the lowest RMSE and highest Rsquared value, which means it has the lowest error and highest predictive power.

We won’t dive into the specifics of the statistics behind these decisions right now, but next we will try adjusting a few different parts of the model to see whether our performance improves. For a detailed statistical explanation of random forest modeling, including more about mtry and tuning a model, please see Chapter 8 in the book “An Introduction to Statistical Learning with Applications in R” (James et al. [2013](#ref-james2013)).

What would happen if we do not fix min.node.size to five? We’re going to let min.node.size change and let mtry change as well.

Let’s create our own grid of values to test for mtry and min.node.size. We’ll stick with the default bootstrap resampling method to choose the best model. We will randomly choose some values to use for mtry, including the three that were used previously (2, 10, and 19). Let’s try 2, 3, 7, 10, and 19.

# Create a grid of different values of mtry, different splitrules, and different minimum node sizes to test  
tune\_grid <-  
 expand.grid(  
 mtry = c(2, 3, 7, 10, 19),  
 splitrule = c("variance", "extratrees"),  
 min.node.size = c(1, 5, 10, 15, 20)  
 )  
  
# Set a seed  
set.seed(2020)  
  
# Fit a new model, using the tuning grid we created above  
rf\_fit2 <-  
 train(final\_grade ~ .,  
 data = data\_train,  
 method = "ranger",  
 tuneGrid = tune\_grid)  
  
rf\_fit2

## Random Forest   
##   
## 372 samples  
## 12 predictor  
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 372, 372, 372, 372, 372, 372, ...   
## Resampling results across tuning parameters:  
##   
## mtry splitrule min.node.size RMSE Rsquared MAE   
## 2 variance 1 15.2 0.564 11.12  
## 2 variance 5 15.3 0.564 11.18  
## 2 variance 10 15.5 0.554 11.35  
## 2 variance 15 15.6 0.550 11.45  
## 2 variance 20 15.7 0.543 11.57  
## 2 extratrees 1 16.9 0.522 11.96  
## 2 extratrees 5 17.0 0.518 12.08  
## 2 extratrees 10 17.3 0.506 12.27  
## 2 extratrees 15 17.4 0.502 12.39  
## 2 extratrees 20 17.7 0.489 12.56  
## 3 variance 1 14.7 0.574 10.73  
## 3 variance 5 14.7 0.572 10.80  
## 3 variance 10 14.8 0.567 10.94  
## 3 variance 15 15.0 0.557 11.08  
## 3 variance 20 15.2 0.550 11.22  
## 3 extratrees 1 15.7 0.560 11.20  
## 3 extratrees 5 15.9 0.556 11.31  
## 3 extratrees 10 16.2 0.545 11.58  
## 3 extratrees 15 16.4 0.539 11.75  
## 3 extratrees 20 16.6 0.530 11.90  
## 7 variance 1 14.1 0.589 10.27  
## 7 variance 5 14.1 0.590 10.26  
## 7 variance 10 14.2 0.584 10.37  
## 7 variance 15 14.2 0.581 10.46  
## 7 variance 20 14.4 0.576 10.56  
## 7 extratrees 1 14.3 0.607 10.31  
## 7 extratrees 5 14.4 0.603 10.37  
## 7 extratrees 10 14.5 0.600 10.54  
## 7 extratrees 15 14.7 0.596 10.68  
## 7 extratrees 20 15.0 0.586 10.87  
## 10 variance 1 13.9 0.593 10.12  
## 10 variance 5 14.0 0.593 10.13  
## 10 variance 10 14.0 0.591 10.18  
## 10 variance 15 14.1 0.587 10.26  
## 10 variance 20 14.1 0.585 10.33  
## 10 extratrees 1 13.8 0.623 10.00  
## 10 extratrees 5 13.9 0.619 10.08  
## 10 extratrees 10 14.1 0.613 10.22  
## 10 extratrees 15 14.3 0.606 10.38  
## 10 extratrees 20 14.4 0.603 10.49  
## 19 variance 1 14.0 0.590 9.99  
## 19 variance 5 14.0 0.588 10.01  
## 19 variance 10 14.1 0.588 10.03  
## 19 variance 15 14.1 0.588 10.07  
## 19 variance 20 14.1 0.587 10.08  
## 19 extratrees 1 13.4 0.634 9.73  
## 19 extratrees 5 13.4 0.633 9.75  
## 19 extratrees 10 13.5 0.629 9.83  
## 19 extratrees 15 13.6 0.626 9.90  
## 19 extratrees 20 13.7 0.621 10.01  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were mtry = 19, splitrule = extratrees  
## and min.node.size = 1.

The model with the same values as identified before for mtry (19) and splitrule (extratrees), but with min.node.size equal to 1 (not 5, as before) seems to fit best. We know this model fits best because the RMSE is lowest (13.027) and the variance explained is highest (0.684) for this model, though the improvement seems to be fairly small relative to the difference the other tuning parameters seem to make.

While the output above gives us a good summary of the model, we might want to look more closely at what we found with our rf\_fit2 model. The code below is a way for us to zoom in and look specifically at the *final* random forest model generated by our rf\_fit2.

In the code chunk below, you’ll notice we are selecting the “finalModel” output using a $ operator rather than the familiar select. We cannot use dplyr and the tidyverse here because of the structure of the rf\_fit2 object - we have stored a random forest model as a model, so it’s not a normal dataframe. Thus, we extract with a $. We want to select only the final model used, and not worry about the prior iterations of the model.

#Here, we select the "finalModel" output from the rf\_fit2 model  
rf\_fit2$finalModel

## Ranger result  
##   
## Call:  
## ranger::ranger(dependent.variable.name = ".outcome", data = x, mtry = min(param$mtry, ncol(x)), min.node.size = param$min.node.size, splitrule = as.character(param$splitrule), write.forest = TRUE, probability = classProbs, ...)   
##   
## Type: Regression   
## Number of trees: 500   
## Sample size: 372   
## Number of independent variables: 19   
## Mtry: 19   
## Target node size: 1   
## Variable importance mode: none   
## Splitrule: extratrees   
## Number of random splits: 1   
## OOB prediction error (MSE): 160   
## R squared (OOB): 0.66

In looking at this output, we see the same parameters we noted above: mtry is 19, the node size is 1, and the split rule is extra trees. We can also note the *OOB prediction error (MSE)*, of 154.332, and the proportion of the variance explained, or R squared, of 0.692. As before, we want the error to be low and the variance explained to be high.

Now that we understand how to develop a basic machine learning model, and how to use different tuning parameters (such as node size and the splitting rule), we can explore some other related themes. We might wonder about how we could examine the predictive accuracy of the random forest model we just developed.

### Examining predictive accuracy on the test data set

What if we use the test data set - data not used to train the model? Below, we’ll create a new object that uses the rf\_fit2 model we developed above. We will put our testing data through the model, and assign the predicted values to a row called “pred.” At the same, time, we’ll make a row called “obs” that includes the real final grades that students earned. Later, we’ll compare these predicted and observed values to see how well our model did.

set.seed(2020)  
  
##Create a new object for the testing data including predicted values   
data\_test\_augmented <-  
 data\_test %>%  
 mutate(pred = predict(rf\_fit2, data\_test),  
 obs = final\_grade)

## mutate: new variable 'pred' with 92 unique values and 0% NA

## new variable 'obs' with 90 unique values and 0% NA

###Transform this new object into a dataframe  
defaultSummary(as.data.frame(data\_test\_augmented))

## RMSE Rsquared MAE   
## 12.982 0.552 9.974

We can compare this to the values above to see how our model performs when given data that was not used to train the model. Comparing the RMSE values, we see that the RMSE is about the same when we use the model on the test data as it was on the training data. We get a value of 11.214 on the test data here, and it was 13.027 on the training data. The Rsquared value is 0.743 here, as compared to the 0.684 we got when we passed the training data through rf\_fit2 earlier.

While we might have expected that the model performance would be worse for the testing data as compared to the training data, we actually are seeing marginal improvements here: the model does better with the test data than with the training data. These results suggest to us that the model is fairly robust, as we get comparable - in fact, improved - results when running the model on data it has never “seen” before (the testing data). This is good news!

## Results

### Variable Importance Measures

One helpful characteristic of random forest models is that we can learn about which variables contributed most strongly to the predictions in our model, across all the trees in our forest.

We can examine two different variable importance measures using the **ranger** method in {caret}.

Note that importance values are not calculated automatically, but that “impurity” or “permutation” can be passed to the importance argument in train(). See more [here](https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html).

We’ll re-run the rf\_fit2 model with the same specifications as before, but this time we will add an argument to call the variable importance metric.

# Set a seed  
set.seed(2020)  
  
# Specify the same model as earlier in the chapter (rf\_fit2) with the addition of the variable importance metric  
rf\_fit2\_imp <-  
 train(  
 final\_grade ~ .,  
 data = data\_train,  
 method = "ranger",  
 tuneGrid = tune\_grid,  
 importance = "permutation"  
 )  
  
# Extract the variable importance from this new model  
varImp(rf\_fit2\_imp)

## ranger variable importance  
##   
## Overall  
## n 100.000  
## subjectFrScA 20.827  
## time\_spent 13.680  
## subjectPhysA 5.288  
## semesterS216 4.063  
## negemo 3.137  
## pc 2.558  
## social 2.360  
## posemo 1.724  
## int 0.933  
## cogproc 0.729  
## enrollment\_reasonScheduling Conflict 0.516  
## enrollment\_reasonLearning Preference of the Student 0.466  
## enrollment\_reasonOther 0.412  
## uv 0.387  
## enrollment\_reasonCredit Recovery 0.233  
## semesterT116 0.212  
## subjectOcnA 0.180  
## subjectBioA 0.000

Our results here give us a ranked order list of the variables in the order of their importance. Variables that appear at the top of the list are more important, and variables that appear at the bottom of the list are less important in the specification of our final random forest model. Remember that we are predicting final grade in the course, so this list will tell us which factors were most important in predicting final grade in online science courses. It can be a bit hard to visually scan a variable importance list, so we might be interested in doing a data visualization.

We can visualize this variable importance list with {ggplot2}.

varImp(rf\_fit2\_imp) %>%  
 pluck(1) %>%  
 rownames\_to\_column("var") %>%  
 ggplot(aes(x = reorder(var, Overall), y = Overall)) +  
 geom\_col(fill = dataedu\_colors("darkblue")) +  
 coord\_flip() +  
 theme\_dataedu()

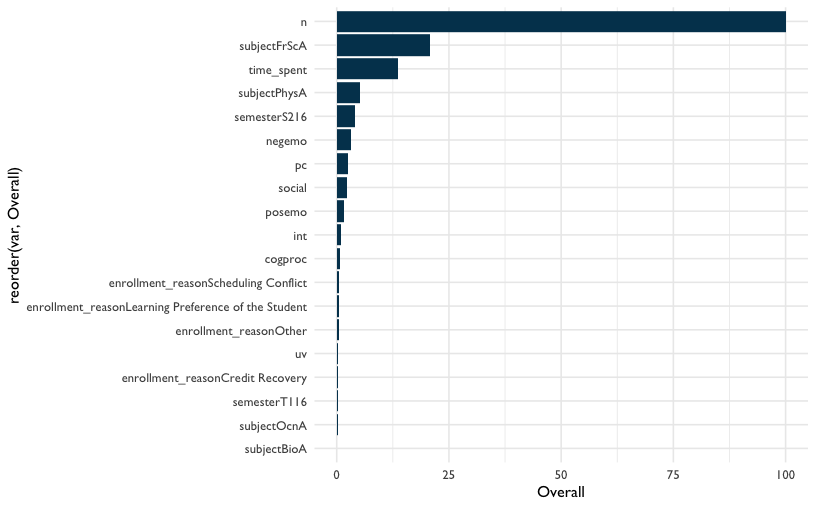


Figure 38: Variable Importance

Cool! We can now visualize which variables are most important in predicting final grade.

The first thing we notice is that the variable n is the most important. This variable indicates how much students write in their discussion posts. The second most important variable is the amount of time students spend in their course. The third most important variable is subjectFrScA. This is one of the course subjects: forensic science. Being enrolled in the forensic science course has a large impact on final grade. That would indicate to us that the forensic science course - more than the other science subjects in this dataset - is strongly correlated with students’ final course grades. We can keep scanning down the list to see the other variables that were indicated as less and less important for the model’s predictions. Variable importance can thus help us to better understand the inner workings of a random forest model.

Overall, there are some subject level differences in terms of how predictive subject is. Biology (subjectBioA) shows up pretty far down the list, whereas Physiology is in the middle (subjPhysA) and forensic science is towards the top (subjectFrScA). What this tells us is that the course students are in seems to have a different effect on final grade, depending on the course. Perhaps grades should be normalized within subject: would this still be an important predictor if we did that? We won’t dive into that question here, but you can see how the line of research inquiry might progress as you start to explore your data with a machine learning model.

As a quick statistical note: above, we selected our variable importance method to be “permutation” for our demonstrative example. There are other options available in the {caret} package if you would like to explore those in your analyses.

### Comparing a random forest to a regression

You may be curious about comparing the predictive accuracy of the model to a linear model (a regression). Below, we’ll specify a linear model and check out how the linear model performs in terms of predicting the real outcomes. We’ll compare this with the random forest model’s performance (rf\_fit2). Note that we are not actually re-running our random forest model here, but instead we are just making a dataset that includes the values that the rf\_fit2 model predicted as well as the actual rf\_fit2 values.

# Make sure all variables stored as characters are converted to factors  
data\_train\_lm <-   
 data\_train %>%   
 mutate\_if(is.character, as.factor)

## mutate\_if: no changes

# Create a linear regression model,   
# using the same formula approach as in the random forest: ~ .  
lm\_fit <-  
 train(final\_grade ~ .,  
 data = data\_train\_lm,  
 method = "lm")  
  
# Append the predicted values to the training dataset for the linear model,   
# so we can see both the predicted and the actual values  
data\_train\_lm <-  
 data\_train %>%  
 mutate(obs = final\_grade,  
 pred = predict(lm\_fit, data\_train\_lm))

## mutate: new variable 'obs' with 354 unique values and 0% NA

## new variable 'pred' with 372 unique values and 0% NA

# Append the predicted values to the training dataset for the random forest  
data\_train\_randomfor <-  
 data\_train %>%  
 mutate(pred = predict(rf\_fit2, data\_train),  
 obs = final\_grade)

## mutate: new variable 'pred' with 372 unique values and 0% NA

## new variable 'obs' with 354 unique values and 0% NA

# Summarize, as data frames, the training data with the predicted   
# and the actual values for both the linear model  
defaultSummary(as.data.frame(data\_train\_lm))

## RMSE Rsquared MAE   
## 14.434 0.556 10.848

# and the random forest  
defaultSummary(as.data.frame(data\_train\_randomfor))

## RMSE Rsquared MAE   
## 4.602 0.968 3.393

Our output will come in the order we wrote the code, so the linear model output shows up above the random forest output.

We can see that the random forest technique seems to perform better than regression. Specifically, the RMSE is lower for the random forest (4.605 as compared to 14.14 for the linear model). Second, the variance explained (Rsquared) is much higher in the random forest (0.97 as compared to 0.57 for the linear model). It may be interesting to compare the results from the random forest not to a more straightforward model, such as a regression, but to a more sophisticated model, like one for deep learning. As you expand your skills, you might be curious to do something like that.

## Conclusion

In this chapter, we introduced both general machine learning ideas, like a training and test data set and evaluating the importance of specific variables, and specific ideas, like how a random forest works and how to tune specific parameters so that the model is as effective as possible at predicting an outcome. Like many of the topics in this book - but, perhaps *particularly* so for machine learning - there is much more to discover on the topic, and we encourage you to consult the books and resources in the [Learning More chapter](#c17) to learn about further applications of machine learning methods.

# Introducing Data Science Tools To Your Education Job

## Chapter Overview

The purpose of this section is to explore what it is like to take newfound data science skills into your work place with the challenge of finding practical ways to use your skills, encouraging your coworkers to be better users of data, and develop analytic routines that are individualized to the needs of your organization. Whether you are helping an education institution as a consultant, an administrator leading teachers at a school, or a university department chair, there are things you can do to transform what you’ve learned in the abstract into more concrete learning objectives in the context of your education work place. We’ll discuss this topic using two areas of focus: bringing your organization the gift of speed and scale, and the importance of connecting well with others. We’ll close this chapter by discussing some of the ways that K-12 teachers in particular might engage a work culture that is bringing on data science as a problem-solving tool.

## The Gift of Speed and Scale

The power of doing data analysis with a programming language like R comes from two improvements over tools like Excel and Google Sheets: 1) a massive boost in the speed of your work and 2) a massive boost in the size of the size of the datasets you analyze. Here are some approaches to introducing data science to your education workplace that focus on making the most of these increases in speed and scale.

### Working With Data Faster

Data analysts who have have an efficient analytic process understand their clients’ questions and participate by rapidly cycling through analysis and discussion. They quickly accumulate skill and experience because their routines facilitate many cycles of data analysis. Roger Peng and Elizabeth Matsui discuss epicycles of analysis in their book, *The Art of Data Science* (Peng and Matsui [2015](#ref-peng2015)). In their book, *R for Data Science* Wickham and Grolemund ([2018](#ref-grolemund2018)) demonstrate a routine for data exploration. When the problem space is not clearly defined, as is often the case with education data analysis questions, the path to get from the initial question to analysis itself is full of detours and distractions. Having a routine that points you to the next immediate analytic step gets the analyst started quickly, and many quick starts results in a lot of data analyzed.

But speed gives us more than just an accelerated flow of experience or the thrill of rapidly getting to the bottom of a teacher’s data inquiry. It fuels the creativity required to understand problems in education and the imaginative solutions required to address them. Analyzing data quickly keeps the analytic momentum going at the speed needed to indulge organic exploration of the problem. Imagine an education consultant working with a school district to help them measure the effect of a new intervention on how well their students are learning math. During this process the superintendent presents the idea of comparing quiz scores at the schools in the district. The speed at which the consultant offers answers is important for the purposes of keeping the analytic conversation going.

When a consultant quickly answers a teacher’s analytic question about their students’ latest batch of quiz scores, the collaborative analytic process feels more like a fast-paced inspiring conversation with a teammate instead of sluggish correspondence between two people on opposite ends of the country. We’ve all experienced situations where a question like “Is this batch of quiz scores meaningfully different from the ones my students had six months ago?” took so long to answer that the question itself is unimportant by the time the answer arrives!

Users of data science techniques in education have wonderful opportunities to contribute in situations like this because speedy answers can be the very thing that sparks more important analytic questions. In our example of the education consultant presented with a superintendent’s curiosity about quiz score results, it is not too hard to imagine many other great questions resulting from the initial answers:

* How big was the effect of the new intervention, if any?
* Do we see similar effects across student subgroups, especially the subgroups we are trying to help the most?
* Do we see similar effects across grade levels?

The trick here is to use statistics, programming, and knowledge about education to raise and answer the right questions quickly so the process feels like a conversation. When there’s too much time between analytic questions and their answers, educators lose the momentum required to follow the logical and exploratory path towards understanding the needs of their students.

#### Example: Preparing quiz data to compute average scores

Let’s take our example of the education consultant tasked with computing the average quiz scores. Imagine the school district uses an online quiz system and each teacher’s quiz export looks like this:

library(tidyverse)

## ── Attaching packages ──────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.0 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.5  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

## ── Conflicts ─────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

set.seed(2020)  
  
quizzes\_1 <- tibble(  
 teacher\_id = 1,   
 student\_id = c(1:3),   
 quiz\_1 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_2 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_3 = sample(c(0:100), 3, replace = TRUE)  
)  
  
quizzes\_1

## # A tibble: 3 x 5  
## teacher\_id student\_id quiz\_1 quiz\_2 quiz\_3  
## <dbl> <int> <int> <int> <int>  
## 1 1 1 27 87 35  
## 2 1 2 86 64 41  
## 3 1 3 21 16 69

Tools like Excel and Google Sheets can help you compute statistics like mean scores for each quiz or mean scores for each student fairly quickly, but what if you’d like to do that for five teachers using the exact same method? First, let’s tidy the data. This will prepare our data nicely to compute any number of summary statistics or plot results. Using pivot\_longer() to separate the quiz number and its score for each student will get us a long way:

quizzes\_1 %>%   
 pivot\_longer(cols = quiz\_1:quiz\_3, names\_to = "quiz\_number", values\_to = "score")

## # A tibble: 9 x 4  
## teacher\_id student\_id quiz\_number score  
## <dbl> <int> <chr> <int>  
## 1 1 1 quiz\_1 27  
## 2 1 1 quiz\_2 87  
## 3 1 1 quiz\_3 35  
## 4 1 2 quiz\_1 86  
## 5 1 2 quiz\_2 64  
## 6 1 2 quiz\_3 41  
## 7 1 3 quiz\_1 21  
## 8 1 3 quiz\_2 16  
## 9 1 3 quiz\_3 69

Note now that in the first version of this dataset, each individual row represented a unique combination of teacher and student. After using pivot\_longer(), each row is now a unique combination of teacher, student and quiz number. This is often talked about as changing a dataset from “wide” to “narrow” because of the change in the width of the dataset. The benefit to this change is that we can compute summary statistics by grouping values in any of the new columns. For example, here is how we would compute the mean quiz score for each student:

quizzes\_1 %>%   
 pivot\_longer(cols = quiz\_1:quiz\_3, names\_to = "quiz\_number", values\_to = "score") %>%  
 group\_by(student\_id) %>%   
 summarise(quiz\_mean = mean(score))

## # A tibble: 3 x 2  
## student\_id quiz\_mean  
## <int> <dbl>  
## 1 1 49.7  
## 2 2 63.7  
## 3 3 35.3

Again, for one dataset this computation is fairly straight forward and can be done with a number of software tools. But what if the education consultant in our example wants to do this repeatedly for twenty five teacher quiz exports? Let’s look at one way we can do this fairly quickly using R. We’ll start by creating two additional datasets as an example. To make things feel authentic, we’ll also add a column to show if the students participated in a new intervention.

# Add intervention column to first dataset   
quizzes\_1 <- quizzes\_1 %>%   
 mutate(intervention = sample(c(0, 1), 3, replace = TRUE))  
  
# Second imaginary dataset  
quizzes\_2 <- tibble(  
 teacher\_id = 2,   
 student\_id = c(4:6),   
 quiz\_1 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_2 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_3 = sample(c(0:100), 3, replace = TRUE),   
 intervention = sample(c(0, 1), 3, replace = TRUE)  
)  
  
# Third imaginary dataset  
quizzes\_3 <- tibble(  
 teacher\_id = 3,   
 student\_id = c(7:9),   
 quiz\_1 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_2 = sample(c(0:100), 3, replace = TRUE),   
 quiz\_3 = sample(c(0:100), 3, replace = TRUE),   
 intervention = sample(c(0, 1), 3, replace = TRUE)  
)

The method we’ll use to compute the mean quiz score for each student is to:

1. Combine all the datasets into one big dataset: Use bind\_rows() to combine all three quiz exports into one dataset. Remember, this can be done because each teacher’s export uses the same imaginary online quiz system and export feature and thus use the same number of columns and variable names
2. Reuse the code from the first dataset on the new bigger dataset: Paste the code we used in the first example into the script so it cleans and computes the mean quiz score for each student

# Use `bind\_rows` to combine the three quiz exports into one big dataset  
all\_quizzes <- bind\_rows(quizzes\_1, quizzes\_2, quizzes\_3)

Note there are now nine rows, one for each student in our dataset of three teacher quiz exports:

all\_quizzes

## # A tibble: 9 x 6  
## teacher\_id student\_id quiz\_1 quiz\_2 quiz\_3 intervention  
## <dbl> <int> <int> <int> <int> <dbl>  
## 1 1 1 27 87 35 0  
## 2 1 2 86 64 41 0  
## 3 1 3 21 16 69 1  
## 4 2 4 55 79 2 1  
## 5 2 5 71 28 65 1  
## 6 2 6 41 97 92 1  
## 7 3 7 77 47 6 1  
## 8 3 8 77 46 83 0  
## 9 3 9 75 77 17 1

We’ll combine the cleaning and computation of the mean steps neatly into one this chunk of code:

# Reuse the code from the first dataset on the new bigger dataset  
all\_quizzes %>%   
 # Clean with pivot\_longer  
 pivot\_longer(cols = quiz\_1:quiz\_3, names\_to = "quiz\_number", values\_to = "score") %>%  
 # Compute the mean of each student  
 group\_by(student\_id, intervention ) %>%   
 summarise(quiz\_mean = mean(score))

## # A tibble: 9 x 3  
## # Groups: student\_id [9]  
## student\_id intervention quiz\_mean  
## <int> <dbl> <dbl>  
## 1 1 0 49.7  
## 2 2 0 63.7  
## 3 3 1 35.3  
## 4 4 1 45.3  
## 5 5 1 54.7  
## 6 6 1 76.7  
## 7 7 1 43.3  
## 8 8 0 68.7  
## 9 9 1 56.3

Note here that our imaginary education consultant from the example is thinking ahead by including the intervention column. By doing so she’s opened the possibility of collaboratively exploring any possible differences in the scores between the students who had the intervention and the students who did not when she reviews and discusses these results with the school staff. Adding these types of details ahead of time is one way to build conversation starters into your collaborations. It is also a way to get faster at responding to curiosities by anticipating useful questions from your clients.

The difference in time it takes to do this on three quiz exports using R versus non-programming tools is perhaps not significant. But the speed of computing means across larger volumes of data–say thirty quiz exports–is truly useful to an education consultant looking to help many educators.

*Summary*

While getting fast at answering analytic questions is not a silver bullet (but really, what is?), it does have a chain effect often leads to creative solutions. It works something like this:

1. Answering analytic questions faster helps more people
2. Helping more people creates opportunities for more data science practice
3. Helping more people also helps educate those same people about the solutions data science tools can offer
4. Lots of practice combined with a common understanding of the value of data science tools in the education workplace nurtures confidence
5. Confidence leads to the courage required to experiment with interesting solutions for designing the best solutions for students

Here are more ways to get faster at answering analytic questions:

* Recognize when you are using similar chunks of code to do repetitive operations. Store that code in an accessible place and reuse it
* Keep a notebook of the questions teachers and administrators ask to help you develop an instinct for common patterns of questions. Write your code to anticipate these questions
* Learn to use functions and packages like {purrr} to work on many datasets at once
* Install a prototyping habit by getting comfortable with quickly producing rough first drafts of your analysis. Your audience can give valuable feedback early and feel like you are quickly on the path to developing useful answers to their questions

### Working With More Data

Improving outcomes in education is about learning, obviously for the students, but just as importantly for the people teaching the students. The more data is available to examine, the more school staff learn about what is working for their students. Using R to prepare and analyze data so it is repeatable and easy to share increases the amount of data you can work with on an order of magnitude compared to tools like Google Sheets.

When cleaning and analyzing data is laborious, people tend to generate less data. This can be a problem because less data means less context for the data you do have. Without context, it is difficult to conduct one of the primary cognitive tasks of data analysis: making comparisons. For example, imagine a teacher whose students have an average quiz score of 75 percent. This information is helpful to the teacher because it shows her how close she is to some pre-determined average quiz score goal, say 95 percent. But that data alone doesn’t tell the teacher how unusual that class average is. For that, you need context. Say that line of code used to compute this teacher’s class average quiz score was applied to every classroom and she learned that the school average for the same quiz was 77 percent. From this information the teacher learns that her class average is not very different from everyone else’s. This is more information than just the knowledge that her class’s average was less than her pre-determined goal of 95 percent.

This is where using R for data analysis enters the conversation. Working with data past a certain size, say 10,000 rows, is difficult because you have to interact with each row through the graphical user interface. Instead, you can work with larger datasets like using programming languages like R to issue complex instructions for acting on the data rather than using a mouse and keyboard to act on what you can see on the screen.

#### Example: Replacing Many Student Names With Numerical IDs

Say, for example, an elementary school administrator wants to replace each student name in a classroom dataset with a unique numerical ID. Doing this in a spreadsheet using good old fashioned data entry is fairly straightforward. Doing this for a whole school’s worth of classrooms though, demands a different approach. Rather than hand enter a unique id into a spreadsheet, the administrator can write an R script that executes the following steps:

1. Use read\_csv() to store every classroom’s student list into the computer’s memory
2. Use bind\_rows() to combine the separate lists into one long list
3. Use mutate() to replace student names with a randomized and unique numerical ID
4. Use split() to separate the data into classrooms again
5. Use {purrr} and write\_csv() to create and rename individual spreadsheets to send back to teachers

With some initial investment into thoughtful coding on the front end of this problem, the administrator now has a script she can use repeatedly in the future when she needs to do this task again.

## Other Ways to Reimagine the Scale of Your Work

### Reflect on your current scale, then push to the next level

When you’ve been using the same data analysis tools and routines for a long time, it’s easy to forget to reflect on how you work. The analytic questions we ask, the datasets we use, and the scale of the analytic questions become automatic because for the most part they’ve delivered results. When you introduce data science techniques and R into your education analysis workflow, you also introduce an opportunity to ask yourself: How can I put this analytic question in context by analyzing on a larger scale?

When an education client or coworker asks for help answering an analytic question, consider the following:

1. At what level is this question about, student, classroom, school, district, regional, state, or federal?
2. What can we learn by answering the analytic question at the current level, but also at the next level of scale up?

If a teacher asks you to analyze the attendance pattern of one student, see what you learn by comparing to the the attendance pattern of the whole classroom or the whole school. If a superintendent of a school district asks you to analyze the behavior referrals of a school, analyze the behavior referrals of every school in the district. One of the many benefits of using programming languages like R to analyze data is that once you write code for one dataset, it can be used with many datasets with a relatively small amount of additional work.

### Look for lots of similarly structured data

Train your eyes to be alert to repositories that contain many datasets that have the exact same structure, then design ways to act on all those datasets at once. Data systems in education generate standardized data tables all the time. It’s one of the side effects of automation. Software developers design data systems to automatically generate many datasets for many people. The result is many datasets that contain different data, but all have the same number of columns and the same column names. This uniformity creates the perfect condition for R scripts to automatically act on these datasets in a way that is predictable and repeatable. Imagine a student information system that exports a list of students, their teacher, their grade level, and the number of school days attended to date. School administrator’s that have a weekly routine of exporting this data and storing it in a folder on their laptop will generate many uniformly structured datasets. When you train your eyes to see this as an opportunity to act on a lot of data at once, you will find an abundance of chances to transform data on a large scale so school staff can freely explore and ask questions aimed at improving the student experience.

### Cleaning data

Folks who work in education want to look at data about their students with tools like Excel, but the data is frequently not ready for analysis. You can empower these folks to explore data and ask more questions by being alert to opportunities to prepare lots of data for analysis. Offer to clean a dataset! Then do it again and do it fast. When you get into this habit, you not only train your data cleaning skills but you also train your education client’s expectations for how quickly you can prepare data for them.

## Solving Problems Together

Steven Spielberg said, “When I was a kid, there was no collaboration; it’s you with a camera bossing your friend around. But as an adult, filmmaking is all about appreciating the talents of the people you surround yourself with and knowing you could never have made any of these films by yourself” (Murphy [2011](#ref-nytimes2011)).

Data science techniques are a powerful addition to an educational organization’s problem-solving capacity. But when you’re the only person who codes or fits statistical models, it’s easy to forget that the best solutions magically arrive when many perspectives come crashing together. Here are some things to think about as you challenge yourself to introduce data science to your education workplace in a lasting and meaningful way.

### Data Science in Education and Empathy

One definition of empathy is seeing things as others do, which points to a barrier to our mission of discovering ways to use our data science skills to improve the experience of learners–it is all too easy to assume that our coworkers will be inspired by possibilities of data science as you are. In 1990 Elizabeth Newton, then a Stanford University graduate, asked research subjects to “tap” out well-known songs with their fingers and estimate how many people would recognize the songs (Newton [1991](#ref-newton1991), @hbr2006). She found that they overestimated every time! When we know a subject well, we tend to forget the experience of not knowing that subject. So how do we make use of this knowledge?

First, listen carefully to your coworkers as they work with data. As you listen, aim to understand the thinking process they use when making sense of reports, tables, and graphs. This will help you understand the problems and solutions they gravitate towards.

Second, ask them if you can “borrow the problem” for a bit. “Borrowing a problem” is not solving it for them, it’s using a little data science magic to get them unstuck so they can continue solving the problem the way they want to. If they’re struggling to make a scatter plot from their pivot table data, offer to help by cleaning and summarizing the dataset before they try again.

Third, if your first attempt at borrowing the problem didn’t help, make an effort to learn more. Doing data science together is a conversation, so ask them how it went after you cleaned the dataset. Then listen, understand, and try again. After many rounds of this process, you may find your coworkers willing to try new methods for advancing their goals.

A workplace going from not using data science to using data science regularly is a process that takes longer than you think. Responses to new ideas might include excitement and inspiration, but they might just as likely include resistance and fear. Changing the way an organization works requires new skills which often take years to learn. But here we are talking about one part of this change that is easily missed: listening to people and the system and using empathy to determine the unique place in your education organization that your data science skills will help students the most. Introducing data science techniques to your system is as much about having good people skills and empathy as it is about learning how to code and fit models.

Data scientists and non-data scientists in education are similar in this regard–they both get excited and inspired by solving meaningful problems for their students. Once we recognize that that is the unifying goal, the exploration of how we do that with a diversity of expertise and tools begins. When we use empathy to connect with our coworkers about the common problems we are solving, we open the door to all kinds of solutions. Data science in education becomes a tool for a student-centered common cause, not an end in and of itself.

Here are some reflection questions and exercise to use to inspire connection in your education workplace. Practice these questions both as personal reflections and also as questions you ask your coworkers:

1. What does data analysis in our organization look like today?
2. How do I wish data analysis will look like in the future?
3. What is the hardest challenge I face in building my vision of student learning?
4. What is one story about a rewarding experience I had with a student?

### Create a Daily Practice Commitment That Answers Someone Else’s Question

In his book *Feck Perfuction*, designer Victore ([2019](#ref-victore2019)) writes “Success goes to those who keep moving, to those who can practice, make mistakes, fail, and still progress. It all adds up. Like exercise for muscles, the more you learn, the more you develop, and the stronger your skills become” (p. 31). Doing data science is a skill and like all skills, repetition and mistakes are their fuel for learning. But what happens if you are the first person to do data science in your education workplace? When you have no data science mentors, analytics routines, or examples of past practice, it can feel aimless to say the least. The antidote to that aimlessness is daily practice.

Commit to writing code everyday. Even the the simplest three line scripts have a way of adding to your growing programming instincts. Train your ears to be radars for data projects that are usually done in a spreadsheet, then take them on and do them i R. Need the average amount of time a student with disabilities spends in speech and language sessions? Try it in R. Need to rename the columns in a student quiz dataset? Try it in R. The principal is hand assembling twelve classroom attendance sheets into one dataset? You get the picture.

Now along the path of data science daily practice you may discover that your non-data science coworkers start kindly declining your offers for help. In my experience there is nothing mean happening here, but rather this is a response to imagining what it’s like to do what you are offering to do using the more commonly found spreadsheet applications. As your programming and statistics skills progress, some of the tasks you offer to help with will be the kind that, if done in a spreadsheet app, are overwhelmingly difficult and time intensive. So in environments where programming is not used for data analysis, declining your offers of help are more perceived acts of kindness to you and probably not statements about the usefulness of your work. As frustrating as these situations might be, they are necessary experiences as an organization learns just how available speed and scale of data analysis are when you use programming as a tool. In fact, these are opportunities you should seize because they serve both as daily practice and as demonstrations of the speed and scale programming for data analysis provides.

### Build Your Network

It is widely accepted that participating in personal and professional networks is important for survival, thriving, and innovation. The path to connecting to a data science in education network is apparent if your education workplace has an analytics department, but it will take a little more thought if you are the lone data scientist. When looking for allies that will inspire and teach you, the mind immediately searches for other programmers and statisticians and to be sure, these are relationships that will help you and the organization grow in its analytic approach.

What the authors argue here is that data science in education is not just about bringing programming and statistics, but in the broader view is about evolving the whole approach to analytics. When viewed that way, members of a network broaden beyond just programmers and statisticians. It grows to include administrators and staff who are endlessly curious about the lives of students, graduate students fascinated with unique research methodologies, and designers who create interesting approaches to measurement.

Networks for growing data science in education are not limited to the workplace. There are plenty of online and in real life chances to participate in a network that are just as rewarding as the networks you participate in during regular work hours. Here are a few to check out:

* Communities on Twitter like #R-Ladies and #rstats
* Local coding communities
* Conferences like rstudio::conf and useR!
* Online forums like RStudio Community

## For K-12 Teachers

We’ve used almost all of this chapter to explore what to think about and what to do to help you bring your data science skills to your education workplace. So far the discussion has been from the data scientist’s point of view, but what if you are one of the many who have an interest in analytics but very little interest in programming and statistics? Teachers in elementary and high schools are faced with a mind boggling amount of student data. A study by Campaign ([2018](#ref-dataqualitycampaign2018)) estimated that “95 percent of teachers use a combination of academic data (test scores, graduation rates, etc.) and nonacademic data (attendance, classroom, behavior, etc.) to understand their students’ performance”. 57 percent of the teachers in the study said a lack of time was a barrier to using the data they have. Data literacy is also increasingly important within teacher preparation programs (Mandinach and Gummer [2013](#ref-mandinach2013)).

Yet the majority of teachers aren’t interested in learning a programming language and statistical methods as a way to get better at analytics, and both time and professional development with respect to working with data are necessary (Datnow and Hubbard [2015](#ref-datnow2015)). After all, most teachers chose their profession because they love teaching, not because they enjoy cleaning datasets and evaluating statistical model output. But to leave them out feels like a glaring omission in a field where perhaps the most important shared value is the effective teaching of students.

If you do happen to be an elementary or high school teacher who wants use programming and statistics to improve how you use data, you will find the approaches in this book useful. But if you are not that person, there is still much to explore that will lead to a rewarding experience as you grow your analytic skill. This book lacks the scope to explore this topic thoroughly, but there are many ways to improve how you use data without requiring a programming language or deep knowledge of statistics.

For example, you can explore what is perhaps the most important element of starting a data analysis: asking the correct question. Chapter three of *The Art of Data Science* (Peng and Matsui [2015](#ref-peng2015)) provides a useful process for getting better at asking data questions.

Given how often data is served to us through data visualizations, it is important to learn the best ways to create and consume these visualizations. Chapter one of book *Data Visualization: A Practical Introduction* (Healy [2019](#ref-healy2019)) explores this topic using excellent examples and writing.

For practical applications of a data-informed approach, *Learning to Improve: How America’s Schools Can Get Better at Getting Better* (Bryk et al. [2015](#ref-bryk2015)) offers a thorough explanation of the improvement science process. The book is filled with examples of how data is used to understand problems and trial solutions.

The final recommendation for elementary and secondary teachers wanting to get better at analysis is this: find, and partner with, someone who can help you answer the questions you have about how to serve your students better. You have the professional experience to come up with the right ideas and the curiosity to see what these ideas look like in the classroom. Inviting someone who can collaborate with you and help you measure the success of your ideas can be a rewarding partnership for you and your students.

# Teaching Data Science

## Chapter Overview

This book is focused on the application of data science to education. In other words, this book focuses on how to apply data science to questions of teaching, learning, and educational systems. The previous chapters have addressed this topic through narrative and walkthroughs for common questions (or problems) and the types of data encountered in education. In this way, much of the book has focused on *applying data science methods*. However, for a book on data science in education, we think it is important to not only discuss the application of data science methods, but also to consider what we know about *how to teach about data science*. In recognition of these dual meanings of data science in education, A recent chapter by one of our authors refers to the application of data science methods as “data science *in* education\*, and the teaching and learning of data science as”data science \*for" education (Rosenberg et al. [2020](#ref-rosenberg2020mdsc)).

Thus, an important consideration for data science is how to teach others about data science and how to do it: Naturally, we think educators who do data science are positioned well to try to describe how to do so. In addition, we expect readers of this book - many who will also be involved in education - will be interested in teaching others about data science.

This chapter, then, focuses on teaching data science to others. This chapter is organized around three topics, which progress from concrete and specific (with reference to pedagogical approach we used to guide how we wrote this book) to more general ideas and findings from educational research about teaching and learning.

1. The pedagogical principles this book is based upon
2. Strategies for teaching data science
3. General strategies related to teaching and learning

## 1. The pedagogical principles this book is based upon

As the authors of a book about data science in education - and readers of books that taught us about data science - we considered what would make it effective for our readers when we set out to write it. The result of this process was a pedagogical framework that consists of two principles: problem-based learning, differentiation, and working in the open. We consider each of these in turn.

### Problem-based learning

Problem-based learning (PBL) is a method of instruction that presents learners with a real-world challenge in which they must apply their skills and knowledge to solve. We applied this principle to the design of this book through including walkthroughs for common questions with respect to data science and education. Whereas some topics may benefit less from such an approach (and the inclusion of walkthroughs), for data science, we believe this is important because we do not have all of the right answers in this text. Moreover, there is not one right statistical model or algorithm, a technique to write code or even software and tools to utilize.

Thus, the text features walkthroughs that reflect the types of challenges that educational data scientists may encounter in the course of their work: All of the data (as well as the code) is available, and readers may choose to go about approaching the analysis of the data used in each walkthrough differently. Moreover, the walkthroughs are structured in such a way that readers return to some of the analytic challenges, but with different aims, over the course of the book: [Walkthrough 1/Chapter 7](#c07), [Walkthrough 7/Chapter 13](#c13), and [Walkthrough 8/Chapter 14](#c14) all use the same dataset on online science learning, but Walkthrough 7 expands Walkthrough 1 by its focus on modeling the effects of courses, and Walkthrough 8 takes a *predictive*, rather than an explanatory, goal, through the use of machine learning. Other challenges, such as processing and preparing data, are introduced in the first walkthrough, and - reflective of the importance and ubiquity of these steps of the data science pipeline - returned to in each of the subsequent chapters.

### Differentiation

Differentiation is a method for providing multiple pathways for learners to engage with, understand, and ultimately apply new content and new skills. To differentiate this text, we first created personas for who we expected to be common groups of readers of the book (see Wilson ([2009](#ref-wilson2009)) for an example of this approach).

The objective of this approach was for us to write in a way that helped readers see themselves in the scenarios. The personas were a way to imagine our audience and guide who we interviewed to prepare for the writing. The interviews equipped us to go beyond what we imagined the needs of our readers were and include their voices in the way we presented the content.

We then aimed to differentiate the book by recognizing and providing background knowledge (either explicitly or through references to other resources) and recommendations for where to begin based upon prior expertise. We also provided screenshots–particularly in [Chapter 5/Getting Started with R and RStudio](#c05) and [Chapter 6/Foundational Skills](#c06)–that are annotated and reflective of the content in the text to help to show readers how they are able to use what they are reading about.

Lastly, a part of differentiating this book concerns for whom we are aiming to differentiate it. We consider inclusivity (in terms of who belongs as a part of the audience for this text and how this broader view of who participates in data science implies the types of challenges, topics, and data that we include in the book) and accessibility (technically, in terms of how a wide audience of readers is able to access and use the book, as well as in terms of the ways in which the content is written based on the unique assets that those in education bring) along with how we differentiate the book.

### Working in the Open

We started writing this book in the open, on GitHub. This allowed us to share the book as it developed. Writing the book in the open also allowed others from the wider educational data science and data science community to contribute. These contributions included writing sections of the book in which contributors had specific expertise, asking clarifying questions, and, even, creating a logo for the book which informed our choice of a color palette. We decided to write this book in the open after witnessing the success of other books on data science (such as Wickham ([2019](#ref-wickham2019advr)) *Advanced R* (<https://adv-r.hadley.nz/>) book.

### Universal Design

In our original proposal for this book (see (**???**)), we noted that Universal Design (McTighe and Willis [2019](#ref-mctighe2019upgrade), @wiggins2005understanding) was a part of our pedagogical framework. As we worked toward completing the book, we recognized that we did not fully meet the aims we had laid out. Here is what we wrote in the proposal:

Universal Design is a series of principles which guide the creation of spaces that are inclusive and accessible to individuals from all walks of life regardless of age, size, ability, or disability. While traditionally applied to physical spaces, we have extended these principles to the creation of a data science text in such a way that the text and accompanying materials will be designed for individuals from all walks of life, regardless of educational level, background, ability, or disability. Many of the seven guiding principles of Universal Design are readily transferable to the creation of a text, such as equitable use, flexibility in use (aided in large part through differentiation), simple and intuitive use, perceptible information, and tolerance for error.

While we did not adequately address these in the book, they remain important to us, and we hope to address them in a future edition of the book.

## 2. Strategies for teaching data science

You may be interested in teaching others’ data science. You may be doing this informally (such as by teaching a colleague with whom you work in your school district or organization), in a formal environment (such as a class on data science for educational data scientists or analysts), or in some setting in-between (such as a workshop you are asked to provide). There is some research on teaching data science, as well as practical advice from experienced instructors that can inform these efforts, which we detail in this section.

### Provide a home base for learners to access resources (and to learn more)

As we discuss in the next section, along with other important factors (such as learners’ motivation and having a supportive atmosphere), learning strategies can make a difference for learners. Especially when it comes to learning to do data science, there are many tools and resources to keep track of, such as:

* How to download and install R
* How to download and install R Studio
* How to install packages
* How to access resources related to the workshop or course (or simply other resources you wish to share)
* How to contact the instructor
* How to get help and learn more

Having a “home base,” where you can remind learners to first look to for resources, can help to lower some of learners’ demands in terms of remembering how these tools and resources can be accessed. One way to do this is through a personal website. Another is through GitHub pages. For some organizations, a proprietary learning management system - such as Desire2Learn, Blackboard, Moodle, or Canvas - can be helpful (especially if your learners are accustomed to using them).

### When it comes to writing code, think early and often

It is important to get learners to start writing code early and often. It can be tempting to teach classes or workshops that front-load content about data science and using R. While this information is doubtlessly important, it can mean that those you are teaching do not have the chance to do the things that they want to do, including installing R (and R Studio) and beginning to run analyses. Because of this, we recommend starting with strategies that lower the barrier to writing code for learners. Ways to do this in teaching settings include:

* Using R Studio Cloud
* Providing an R Markdown document for learners to work through
* Providing a data set and ideas for how to begin exploring it

While these strategies are especially helpful for courses or workshops, they can be translated to teaching and learning R in tutoring (or “one-on-one”) opportunities for learners. In these cases, being able to work through and to modify an existing analysis (perhaps in R Studio Cloud) can be a way to quickly begin to run analyses - and to use the analysis as a template for analyses associated with other projects. Also, having a data set associated with a project or analysis - and a real need to analyze it using R - can be an outstanding way for an individual to learn to use R.

### Don’t touch that keyboard!

Resist helping learners to the point of hindering their learning. Wilson ([2009](#ref-wilson2009)) writes about the way in which those teaching others about R - or to program, in general - can find it easier to correct errors in learners’ work. But, by fixing errors, learners’ may perceive themselves to not be capable of carrying out all of the steps needed in an analysis on their own.

This strategy relates to a broader issue, as well: issues that have to do with writing code in a way that runs correctly (e.g., with the correct capitalization and syntax) can be minor to those with experience programming, but can be major barriers to using R in an independent way for those new to it. For example, becoming comfortable with where arguments to functions belong and how to separate them, how to use brackets belong in functions or loops, and when it is necessary to use an assignment operator can be *completely new* to beginners: doing these steps for learners may push their learning later when they do not have as many resources available to help them than when you are teaching them. So, consider taking the additional time needed to help learners to navigate minor issues and errors in their code: it can pay off in increased motivation on their part in the longer-term.

### Anticipate issues (and sacrifice accuracy for clarity)

Don’t worry about being perfectly accurate early on, especially if doing so would lead to learners who are less interested in the topic you are teaching. Especially in cases for which additional details may not be helpful to beginning learners, it can be valuable to not only anticipate these questions, but to have responses or answers that provide more clarity, rather than confusion.

For example, there are complicated issues at the heart of why data that is built-in to packages or to R (such as the iris dataset) appear in the environment after they are first used in an R session (see the section on “promises” in Wickham ([2019](#ref-wickham2019advr))). Similarly, there are complicated issues that pertain to how functions are evaluated that can explain why it is important to provide the name of packages installed via install.packages() (whereas the names of arguments to other functions, such as dplyr::select() do not need to be quoted).

### Start lessons or activities with visualizing data

There are examples from data science books Wickham and Grolemund ([2018](#ref-grolemund2018)) and past research (e.g. Lehrer and Schauble ([2015](#ref-lehrer2015))) that suggests that starting with visualizing data can be beneficial in terms of learners’ ability to work with data. Wickham and Grolemund ([2018](#ref-grolemund2018)) write that they begin their book, *Data Science Using R*, with a chapter on visualization, because doing so allows learners to create something that they can share immediately, whereas tasks such as loading data can be rife with issues - and does not immediately lead learners to have a product they can share. Lehrer, Kim, and Schauble ([2007](#ref-lehrer2007)) show how providing students with an opportunity to invent statistics by displaying the data in new ways. This led to (productive) critique among fifth- and sixth-grade students and their teacher.

### Consider representation and inclusivity in the data and examples you use

One way to think about data is that it is objective, free of decisions about what to value or prioritize. Another is to consider data as a process that is value-laden, from deciding what question to ask (and what data to collect) to interpreting findings with attention to how others will make sense of them (e.g., O’Neil ([2016](#ref-oneill2016))’s *Weapons of Math Destruction*, and Lehrer, Kim, and Schauble ([2007](#ref-lehrer2007))’s description of data modeling). From this broader view, choosing representative data is a choice, like others, that teachers can make. For example, instructors can choose data that directs attention to issues–equity-related issues in education, for example–that she or he believes would be valuable for students to analyze.

We think this consideration is important - particularly when it comes to data-related issues that we consider to be objective, such as how variables are assumed and constructed to be dichotomous (such as variables for individual’s gender) or categorical (such as variables for individual’s race), when the truth may be that such variables are based on decisions that analysts or those collecting data made - decisions that may benefit from questioning. This consideration is also important when it comes to what data is used for teaching and learning. If names of individual’s in data exclusively from individuals from majority racial or ethnic group, for example, some learners may implicitly perceive the content being taught to be designed for others. While we may think that such issues are better left to those we are teaching to decide later on, setting the stage in classes, courses, and other contexts in which data science is taught and learned can set an important precedent for the data our learners use and how they use it.

### Draw on other resources

In this section of this chapter, we presented some strategies for teaching data science. There are others that go more into depth on this topic from different perspectives, such as the following:

* [GAISE Guidelines](https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx): guidelines for teaching statistics
* [Data Science for Undergraduates](https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options): a report on undergraduate data science education
* R Studio Education: <https://education.rstudio.com/>

There are also a number of data science-related curricula (for the K-12 level) which may be helpful:

* [Bootstrap Data Science](https://www.bootstrapworld.org/blog/index.shtml)
* [Exploring CS, unit 5](http://www.exploringcs.org/curriculum)
* Chromebook Data Science: <http://jhudatascience.org/chromebookdatascience/>
* [Oceans of Data Institute Curricula](http://oceansofdata.org/our-work/ocean-tracks-high-school-learning-modules)

Last, there are also books that emphasize the importance–for teachers–of understanding their students–every student. These books include (**???**), Kozol ([2012](#ref-kozol2012savage)), and will likely be valuable for teachers of data science who wish to understand and honor the diversity of their students. Moore Jr, Michael, and Penick-Parks ([2017](#ref-moore2017guide)) and Emdin ([2016](#ref-emdin2016)) may be helpful for data science educators who aim to be aware and intentional about teaching students from backgrounds other than their own.

## 3. General strategies related to teaching and learning

The National Academy of Science commissioned a report, *How People Learn* (Bransford et al. [2000](#ref-nrc2000)), that aimed to summarize research from the educational psychology and the learning sciences on teaching and learning. In 2018, the book was updated in *How People Learn II* (National Academies of Sciences, Medicine, and others [2018](#ref-nrc2018)), which aimed to emphasize the social and cultural aspects of teaching and learning, which were not as much the focus of the earlier report. These may be helpful to those teaching data science for the general strategies they reference related to teaching and learning

In addition to these reports, there are some books that are more practically-oriented, including Hattie ([2012](#ref-hattie2012visible)), Lemov ([2015](#ref-lemov2015teach)), and (**???**). These can help to provide some answers to the question of how to teach data science–and to mitigate some of the anxiety that those teaching data science may feel about how to teach others to know about and do data science: There are existing resources on teaching that are highly-relevant, and it is important to not feel like you must recreate the wheel!

In this section, then, we highlight general strategies related to teaching and learning, drawing from the above reports and books, with an emphasis on strategies applicable to teaching and learning data science. These general strategies are more conceptual than those described in the last section, and are likely more useful as starting points for further research or reflection, instead of as specific techniques that can be brought to the next workshop, class, or peer-to-peer teaching session one is leading.

### Teaching and learning are complex

One principle that *HPL2* begins with is that learning is complex. in short, learning is not just about what learners know, or think, but is also about developmental, cultural, contextual, and historical factors - and distinctions between individuals with respect to each of these factors. This is an asset to teachers, as the authors of the report state: Learners bring resources that can serve as a starting point for their learning trajectory when it comes to data science. These distinctions also mean that educators need to be concerned with - and to consider within their purview - factors well beyond what learners know, but also what their prior educational experiences have been and what resources and other individuals they have access to at work and at home, for example.

### Learners learn many different things (consciously and unconsciously)

We often think of learning in terms of objectives for specific lessons, but learners learn many different things at different times. The authors of *HPL2* point out that individuals learn in response to different challenges and circumstances, including those in formal learning environments, such as workshops or classes. This principle implies a strategy that involves supporting learners to do data science, however and whenever they learn it. This means that it is both okay - and even to be expected - that learners may take away more from a problem they try to solve on their own, than what they do from a workshop or class (or even a degree!). This also suggests that learners may learn things that we do not anticipate, such as how instructors try to solve problems that arise in class.

### Meta-cognition is important (even though it sounds more sophisticated than it is!)

Educators and educational researchers often talk about meta-cognition, or thinking (and ideas) about thinking, as if it is something only very sophisticated learners do, but it is truly something much more commonplace, as people (and learners) are thinking about what they are learning and doing regularly. One strategy for instructors to support meta-cognition is to include moments wherein learners are asked to consider what they learned and what they would like to learn more about (exit tickets can be a great way to do this, but brief period in-class can also be used). Another strategy is to help learners to recognize when it is important to ask for help: Often, when doing data science, the right question to the right person (or community) can save hours of work.

### Learning strategies matter

While teachers are responsible for designing learning opportunities, learners also play an important role - in their own learning! Learning strategies, according to the authors of *HPL2*, matter, including those that help students to retrieve information and summarize and explain what they have learned (for themselves and others). There are many specific strategies documented in [chapter four](https://www.nap.edu/read/24783/chapter/6#72) of *HPL2*. What is important for teachers of data science to know is less the specific strategies, and more the commitment to teaching learners how to learn.

In addition to strategies for learners, teaching strategies, such as how content is spaced and sequenced, can also help learners. Dirksen ([2015](#ref-dirksen2015))’s *Design for How People Learn* presents these strategies, based largely about instructional design research, that may be helpful to those teaching data science.

### Educators can help students to learn

Educators know that how motivated learners are matters. According to the authors of *HPL2*, teachers can make an impact on how motivated learners are. Some specific things that educators can do to support learners include helping learners to set (and to work toward) goals, selecting content that is valuable and interesting to learners, helping learners to have choices and showing them how they are in-control of their learning value, and supporting learners to feel good at and supported in what they are doing. In short, motivation matters, and may be especially important when teaching learners who do not see the value of data science (initially!).

## Summary

In this section, we described the pedagogical principles for this book and strategies for teaching data science. We also directed attention to more general strategies for teaching and learning. Teaching data science is a relatively new area, but data science educators are not alone, and do not need to reinvent the wheel. There are many already-existing resources and those that can be adapted from other disciplines and the wider body of educational research.

# Learning More

## Introduction

If you’re reading this book cover to cover, you’ve been through quite a journey! So far, you’ve:

* Learned about the challenges of doing data science in education
* Practiced some basic coding and statistics techniques
* Worked through examples of analytic routines using education datasets
* Reflected on introducing data science to your education organization over time
* Learned about teaching data science to others

We hope this book sparked an interest in data science that you want to nurture. We’ve talked to many people in your shoes - folks who care about educating students and want to help by using their data skills. Indeed we’ve found the common thread in our audience is wanting to use data to improve the experience of learners. It’s important to nurture this passion by keeping the learning going.

Surrounding yourself with continuous learning experiences can turn this spark into a specialization that makes a real contribution to the lives of students. There are three reasons we feel these ongoing learning experiences are essential to realizing your vision for data in education. First, developing technical skills is a continuous process. The learning mindset is the same whether you’re taking your first steps toward using data science techniques or you’re a seasoned data scientist trying to make a bigger impact in education: there is always something new to learn about programming and statistics. Setting regular time aside to evolve your craft is a commitment to this mindset.

Second, education and data science are like most industries–they are constantly evolving. That means today’s tools and best practices might be tomorrow’s outdated techniques. To keep up with changes, it is important to develop a learning routine that exposes you to the pulse of these two fields. Sometimes this means learning a new technique, sometimes it means deepening expertise in a technique you haven’t mastered, and other times it means revisiting a skill you’ve mastered long ago.

And last, when you surround yourself with learning experiences, you inevitably surround yourself with others who are learning. Along your journey, you’ll interact with folks who are struggling through the same concepts as you, folks who are struggling through more complex concepts, and folks who are struggling with concepts you’ve already mastered. Participating in a community of learners has magical properties–it’s a place to learn, teach, inspire, and get inspired all at once. In Chase Javis’ book [Creative Calling](@jarvis2019) touches on this very point:

Whether online or in person, connecting with a community will support your learning efforts. It will also expose you to a diverse set of ideas that will dramatically enrich your perspective on what you’re learning. If you weren’t in love with your new skill before, this step can tip the balance. Passion is infectious.

You’ll need to use your intuition to find the areas where you want to deepend your knowledge. When you feel it, go there and dive in. Remember that the learning experience includes all kinds of activities. It’s a combination of reading, doing, discussing, walking away, and coming back. Here are some activities to include in your practice. We hope you take these and construct your own system of rewarding learning experiences.

## Adopt a Growth Mindset

It’s normal to feel overwhelmed while learning skills like R and data science. This is particularly true when these fields themsleves are learning and growing. The R, data science, and education communities are constantly developing new techniques to move the field forward. It’s part of the beauty of this work!

When you’re feeling overwhemed by everything you’re trying to learn, consider adopting a growth mindset. Carol Dweck argues that we think of ourselves as *being* or *not being* a type of person. For example, we might think of ourselves as “math people” or “reading people”. What matters is whether or not this state is changeable. When we believe we can change, we adopt a desire to learn, choose to be around people who help us learn, and make the effort to learn. When we move from a *fixed* mindset to a *growth* mindset, we create the possibility of mastering new techniques and realizing our vision for using data in education.

The nuances of the growth mindset (Dweck [2015](#ref-dweck2015)) are beyond the scope of this book, but we do encourage the general belief that we can learn how to apply these techniques. We encourage you to adopt a growth mindset as a way to inspire learning and belief that you can introduce data science in your education job. In doing so, you’ll be joining other data scientists who created a way to contribute to their fields.

## Discover New Information

The content you surround yourself with matters. You can learn a lot and stay inspired by high quality books, blog posts, journals, journalism, and talks. In his book *Steal Like An Artist* Kleon ([2012](#ref-kleon2012steal)) encourages people to surround themselves with great content:

There’s an economic theory out there that if you take the incomes of your five closest friends and average them, the resulting number will be pretty close to your own income. I think the same thing is true of our idea incomes. You’re only going to be as good as the stuff you surround yourself with.

In our Resources chapter, we share books and online resources that inspire us and help us learn. Use these as a starting point and build on them by seeking out authors, data scientists, and educators that inspire you to learn and master your craft. There are lots of ways to do this. Some folks follow data scientists on social media and take note of articles or talks that are getting attention. Others read data informed publications like FiveThirtyEight (<https://fivethirtyeight.com/>), The Economist (<https://www.economist.com/>), or The Upshot (<https://www.nytimes.com/section/upshot>) in the New York Times. Whichever you choose, make sure to stick with something that you’re drawn to and you just might find yourself with a new learning habit that is rewarding and fun.

## Ask for Help

So far, we’ve discussed learning activities you can do on your own. Data science is a team sport, so at eventually your learning will lead you to others in the data science community. You can do this in many ways, both virtual and in real life. Here are a few examples you can try online. Try these and learn about what you’re comfortable with. Then build on that to surround yourself with many ways to ask and answer questions.

### Discussion forums

Visiting discussion forums is a common way to learn and participate in the R community. Websites like R Studio Community (<https://community.rstudio.com/>) and Stack Overflow (<https://stackoverflow.com/>) are very popular ways to do this. On these forums you’ll find many years worth of discussion about R and statistics. It’s quite unusual to search these and not find a way to get unstuck. Many discussions include a reproducible example of code that you can copy and paste into your own R console. This is a fantastic way to learn!

Consider learning best practices for asking forum questions. Including a reproducible example, or “reprex”, to communiate problems is a widely-accepted norm. Bryan ([2019](#ref-bryan2019))’s video about making reproducible examples is a great place to learn more.

### GitHub repositories

When you want to learn more about how a package works or engage a package’s online community, consider visiting the its GitHub repository. {dplyr}’s repository (<https://github.com/tidyverse/dplyr>) is a great example. You can start with the README then dive deeper in the vignettes, which contain demonstrations of the package’s functions. You can even browse the code on GitHub to learn more about how the packages work. Don’t worry, you won’t break anything!

When you’re ready to see how the community engages a package’s authors, you can read through the Issues page. Each respository’s Issues page contains questions, feature requests, and bugs submitted by the programming community. Visit this page when you want to see if someone’s already submitted the coding challenge you’re working through. If you find you’re working on something that’s *not* a known problem, you can contribute by adding an issue. And finally, you can contribute to the development of packages by submitting code to the respository–this process is called a pull request. To learn more about contributing to packages, check out Woo ([2018](#ref-woo18)) talk.

## Share what you’ve learned

If you keep asking questions and finding solutions, you will soon find yourself ready to help others who are just getting started. The adage of learning by teaching applies here–answering someone else’s question also helps you deepen your learning and build empathy for new learners.

Adopting a regular sharing routine is a great way to start helping others. A sharing routine encourages participation in the community, invites feedback for improvement, and calls on you to build your craft in a way that others can understand it.

So what can you share? Really, what can’t you share? If you’ve built a cool function or visualization that took your project to the next level, you just might help or inspire someone else by sharing it. Maybe you’ve found an R package that really helped you–chances are it will help others.

Sharing isn’t always about the output of your work, it can also be about *how* you work. Consider sharing a workflow you’ve developed or your experience at a recent data science conference. Anything that you learned or found interesting will be relevant to others too!

What you share doesn’t have to be perfect. You can decide when you’re ready to share. Some data scientist’s blogs are polished and others are ideas-in-progress or shorter posts. You never know when someone will find value in your work, regardless of whether your work is in a refined state or not.

Laslty, you can select your best work from all your sharing and use it as an online work portfolio.

### Where to share

There are many ways to share your work online. For rapid fire conversational sharing, Twitter. Be sure to use the hashtag #rstats to reach more data scientists. For long form sharing, consider posting to a data science blog. Robinson ([2018](#ref-robinson18))’s blog post *Advice to aspiring data scientists: start a blog* is wonderful inspiration for getting started.

If you decide to post to a blog, there are tools to help you post data science content regularly. As noted earlier, Xie, Thomas, and Hill ([2019](#ref-xie2019blogdown))’s {blogdown} is designed to help you create websites using R Markdown and a static website creator called Hugo. Blogdown makes it easy to create, run, and publish code directly from R Studio. Hill ([2017](#ref-hill2017)) has a great introduction on getting started with with Blogdown.

When you do share a blog post or a tweet, broadcast what you have to say! On Twitter, use hashtags or “at” other community members to include them in the Tweet. On your blog, use blog aggregators that help share your posts to a wider audience. Here are two aggregators to get you started:

* R Weekly newsletter (<https://rweekly.org/>)
* R Bloggers (<https://www.r-bloggers.com>)

Finally, share the love by engaging your fellow data scientists! Retweet others, leave comments, and interact with the vibrant data science and R communities online.

## Welcome others

If you find yourself becoming an envangelist for R and data science in education–that’s what happened to us!–welcome folks who are curious and ready to learn. The strength of any community comes from its inclusiveness, safe learning environment, and capacity to welcome new members. The data science community is no exception–many members work hard to create an environment with active participants, engaging conversations, and celebrations for little and big data science wins. Our call to action is this: continue growing this inclusive and positive environment by being the community member you’d want in your own network.

Data science in education is a wonderful Venn diagram of communities, with new members joining every day. Welcoming, helping, and teaching new members is a great way to contribute to a positive community and to continue your own learning. What better way to inspire new members than to share your work and how it has impacted the lives of students!

# Additional Resources

## Chapter Overview

In this chapter, we provide references and links to additional resources related to data science in education. They are organized around the following headings:

* Data science courses
* Workshop materials
* Data visualization
* Books on data science in education
* Articles on data science in education
* Programming with R
* Package vignettes and descriptions of packages
* Statistics
* R package and statistical software development
* A career in data science
* Places to share your work
* Cheat sheets
* Learning communities

## Data science courses

Anderson, D. J. (2019). University of Oregon Data Science Specialization for the College of Education. <https://github.com/uo-datasci-specialization>

A series of courses that emphasize the use of R on data science in education (graduate-level).

Landers, R. N. (2019). Data science for social scientists. <http://datascience.tntlab.org/>

A data science course for social scientists.

R Studio. (2019). Data Science in a Box. <https://datasciencebox.org/hello/>

A complete course, including a curriculum and teaching materials, for data science.

## Workshop materials

Staudt Willet, B., Greenhalgh, S., & Rosenberg, J. M. (2019, October). Workshop on using R at the Association for Educational Communications and Technology. <https://github.com/bretsw/aect19-workshop>

Contains slides and code for a workshop carried out at an educational research conference, focused on how R can be used to analyze Internet (and social media) data.

Anderson, D. J., and Rosenberg, J. M. (2019, April). Transparent and reproducible research with R. Workshop carried out at the Annual Meeting of the American Educational Research Association, Toronto, Canada. <https://github.com/ResearchTransparency/rr_aera19>

Slides and code for another workshop carried out at an educational research conference, focused on reproducible research and R Markdown.

## Data visualization

Tufte, E. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press LLC.

A classic text on data visualization.

Healy, K. (2018). *Data visualization: A practical introduction*. Princeton, NJ: Princeton University Press.

A programming- (and R-) based introduction to data visualization.

Chang, W. (2013). *R graphics cookbook*. Sebastopol, CA: O’Reilly.

Wilke, C. (2019). *Fundamentals of data visualization*. O’Reilly. <https://serialmentor.com/dataviz/>

A fantastic (though more conceptual than practical, i.e., there is no R code or other software implementation ror creating the plots) introduction to data visualization.

## Books related to data science in education

Geller, W., Cratty, D., & Knowles, J. (2020). *Education data done right: Lessons from the trenches of applied data science.* Leanpub. <https://leanpub.com/eddatadoneright>

Krumm, A., Means, B., & Bienkowski, M. (2018). *Learning analytics goes to school: A collaborative approach to improving education.* Routledge.

Lawson, J. (2015). *Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers.* CreateSpace.

Powers, K., & Henderson, A. E. (Eds.). (2018). *Cultivating a data culture in higher education.* Routledge.

Swing, R. L. (2018). *The Analytics Revolution in Higher Education: Big Data, Organizational Learning, and Student Success*. Stylus Publishing, LLC.

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice.* Sage.

## Articles related to data science in education

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991-16005.

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects.* Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, D.C..

Liu, M. C., & Huang, Y. M. (2017). The use of data science for education: The case of social-emotional learning. *Smart Learning Environments, 4*(1), 1.

Rosenberg, J. M., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher, 87*(5).

Rosenberg, J. M., Lawson, M. A., Anderson, D. J., Jones, R. S., & Rutherford, T. (2020). Making data science count in and for education. In E. Romero-Hall (Ed.), *Research Methods in Learning Design & Technology*. Routledge.

Schneider, B., Reilly, J., & Radu, I. (2020). Lowering Barriers for Accessing Sensor Data in Education: Lessons Learned from Teaching Multimodal Learning Analytics to Educators. *Journal for STEM Education Research*, 1-34. <https://link.springer.com/article/10.1007/s41979-020-00027-x>

Williamson, B. (2017). Who owns educational theory? Big data, algorithms and the expert power of education data science. *E-Learning and Digital Media, 14*(3), 105-122.

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165-181.

## Programming with R

Wickham, H. & Grolemund, G. (2017). *R for data science*. O’Reilly.

“You have data but have no idea on how to make sense off it?”. If this statement resonates to you, then look no further. Introducing R for data analysis. At it’s core, R is a statistical programming language. It helps to derive useful information from the data deluge. This book assumes your a novice at data analytics and will subtly introduce you to the nuances of R, RStudio, and the tidyverse (which is a collection of R packages designed to ensure your learning curve is minimal).

Teetor, P. (2011). *R cookbook*. Sebastopol, CA: O’Reilly.

This book provides over 200 practical solutions for analysing data using R.

Bryan, J. & Hestor, J. *Happy git and github for the useR*. Retrieved from <https://happygitwithr.com>

A fantastic and accessible introduction to using git and GitHub.

## Package vignettes and descriptions of packages

*Introduction to dplyr*. Retreived from <https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html>

dplyr is a R package that helps to manipulate and understand data. It provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate your thoughts into code. Anyone with no programming experience can easily understand the verbs and use them for data analytics with ease.

*A short introduction to the caret package*. Retrieved from <https://cran.r-project.org/web/packages/caret/vignettes/caret.html>

The caret package is an abbreviated form of Classification And REgression Trees. It consist of methods that solve regression and classification problems. It also consist of methods related to the initial data preprocessing activities like data cleaning, feature selection and model tuning for predictive analytics.

*tidy data*. Retrieved from <https://tidyr.tidyverse.org/articles/tidy-data.html>

Wickham et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43). 1686-1691.

The philosophy of ensuring data is clean such that each column contains data of just one type (or atomicity) and each row is an observation (made up of several atomic columns) is made possible in this paper. An atomic data type simply means that the data type cannot be broken down further. For example, consider a person address to be like, “127, Napier Street, Kuala Lumpur 50603”. Suppose we save this address to a variable called house-address. If you look at this variable, it contains mixed-type data with both numeric and text values. So its important to break down the house-address variable such that it will contain data of just one type only. So it can be broken down into sub-variables like, street-number, street-name, city-name, post-code. This is known as data atomicity. Such atomicity leads to ensuring data is tidy and clean. For more details on this subject, please read this wonderful paper by Hadley Wickham (Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23.).

## Statistics

### Introductory Statistics

Open Intro. (2019). Textbooks. <https://www.openintro.org/>

Three open-source textbooks for statistics, one for high school students.

Bruce, P. & Bruce, A. (2017). *Practical statistics for data scientists*. O’Reilly.

Navarro, D. (2019). *Learning Statistics With R*. <https://learningstatisticswithr.com/>

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Covers the foundations thoroughly and in an entertaining way.

Ismay, C., & Kim, A. Y. (2019). *ModernDice: Statistical inference via data science.* CRC Press. <https://moderndive.com/>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning with applications in R*. Springer.

This is an introductory (and R-based) version of a classic book on machine learning by Hastie, Tibshirani, and Friedman ([2009](#ref-hastie2009)).

Peng, R. D. (2019). *R programming for data science*. Leanpub. <https://leanpub.com/rprogramming>

Peng, R. D., & Matsui, E. (2018). *The art of data science*. Leanpub. <https://leanpub.com/artofdatascience>

### Advanced Statistics

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

A fantastic introduction not only to regression (and multi-level/hierarchical linear models, as well as Bayesian methods), but also to statistical analysis in general.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

A classic text on machine learning.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). Linear mixed models: a practical guide using statistical software. Chapman and Hall/CRC.

A solid introduction to multi-level/hierarchical linear models, including code in R (with an emphasis on the lme4 R package).

McElreath, R. (2018). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

A new classic, accessible introduction to Bayesian methods. We note that this book has been “translated” into tidyverse code by Kurz ([2019](#ref-kurz2019)).

## R packages and statistical software development

Peng, R. D. (2019). *Mastering software development in R*. Leanpub. <https://leanpub.com/msdr>

Wickham, H. (2015). *R packages: Organize, test, document, and share your code*. O’Reilly. <http://r-pkgs.had.co.nz/>

A comprehensive introduction to (and walkthrough for) creating your own R packages.

## A career in data science

Robinson, E., & Nolis, J. (2020). *Building a career in data science*. Manning. <https://www.manning.com/books/build-a-career-in-data-science?a_aid=buildcareer&a_bid=76784b6a>

## Places to share your work

Twitter: <twitter.com>

Especially through the hashtags we mentioned below.

LinkedIn: <linkedin.com>

Can be a place not only to share career updates, but also data science-related works-in-progress.

Medium : <medium.com>

## Cheat Sheets

R Studio Cheat Sheets. <https://rstudio.com/resources/cheatsheets/>

See especially the dplyr, tidyr, purrr, ggplot2, and other cheat sheets

## Learning communities

Here some online communities and community resources we recommend:

* #rstats
* #tidyverse
* #RLadies
* #tidytuesday

Here are two resources by co-author Mostipak related to the #r4ds community (from which #tidytuesday came):

Mostipak, J. (2017). *R4DS: the next iteration*. Retrieved from <http://bit.ly/r4ds-intro>

Mostipak, J. (2019). *R4DS online learning community Improvements to self-taught data science & the critical need for diversity, equity, and inclusion*. Retrieved from <http://bit.ly/r4ds-talk>

# Conclusion: Where to Next?

To start closing our journey together, let’s recap what we’ve learned so far. When we started writing this book, we set out to create a learning experience that had recognizable education examples as its foundation. We used these examples to explore the role of a data scientist in the education field. Building on that context, we introduced basic R tools. In the analysis walkthroughs, we learned to apply data science techniques to datasets and scenarios we’ve seen in our education jobs. Our goal was to help you learn data science using datasets, functions, language, and analytic approaches that you’ll keep using in your own education job. Finally, we discussed using these technical skills to positively influence how your education organization uses data.

After all that we hope that you feel more prepared to take on the data-related questions and problems that matter to your students. We want you to feel excited about choosing your next steps. But there’s one more thing we’d like to share about data science in education: how do we all keep our efforts going in the long-term?

As exciting as the promise of data science can be, over time there are inevitable bumps in the road that all educators encounter. As you master doing data science with education datasets, your learning challenges will evolve beyond the problems of coding syntax and into the realm of larger questions about education systems. For example, you may take an interest in new coding or statistics techniques but struggle to find the right application in your education job. Or you may find yourself working alone on a data project when you need a collaborator as a thought partner. Or maybe you’re wondering how a project you’re working on aligns with your values as an educator.

When these questions come up, it can be inspiring and reinvigorating to reflect on the bigger picture. Try thinking a little less about what you learn and a little more about how you learn, why you learn, and the ways your work can positively affect the students and staff in your education system. Think of this as a strategic move to bring back the excitement and hope you feel when you solve problems that truly make the lives of your students better. In this final chapter, we’ll discuss ways to think about your journey that ground you in your service to people learning in your education system. Let’s kick off the next stage of your learning!

## Learn in the Context of Education

The mental models and technical skills we learn are separate from the places and people we practice them with. If you intentionally learn new skills in the context of your daily work, you’ll naturally gravitate towards the skills that are right for the problems that you are trying to solve. It makes sense for an auto mechanic to learn the science of diagnostics while working on automobiles. It makes sense for a surgeon to learn about anatomy while mastering surgical methods. It makes sense for a farmer to learn about the ecosystem while growing crops. In all these examples, the practitioner concerns themselves with mastering skills in service of solving meaningful problems for people.

We’ve already started learning the basics of data science in education by introducing tools like R. Mastering R will help you analyze data at scale. It will also help you share your work while your organization’s data practices grow. But as these skills develop and become muscle memory, your time and attention can focus more and more on what this data tells us about the students and staff that actually generated the numbers to begin with. This is why we believe so strongly in teaching data science using the language, problems, and datasets commonly found in education. It helps you make the cognitive jump from simply working on datasets to working on datasets on service to students.

We hope that you’re excited to take what you’ve learned and share it with others. Remember that what you share is not just a way to analyze data, but a way to contribute to someone’s school experience.

And speaking of sharing, let’s talk about one of the most powerful creative tools we have available to us: each other.

## Learn to Collaborate With Others

In the last section we talked about data in the service of people. But working with data is as much about working with people as it is about working for people. To understand why collaboration in data science is so important, we first have to see data analysis as a fundamentally creative process. We don’t mean this in the same sense that art is creative. First, data science is creative because practitioners create a process that extracts meaning from data. Then second, they create output like writing, visualizations, and conversations that convey this meeting to an audience.

In most creative endeavors, collaboration is the magical ingredient that evolves an individual idea into something truly unique and responsive to the needs of an audience. Daniel Kahneman and Amos Tversky brought the world new knowledge about cognitive biases (see Kahneman ([2011](#ref-kahneman2011))). Ben Cohen and Jerry Greenfield collaborated to create Ben and Jerry’s Ice Cream. Data journalism like the kind you find at FiveThirtyEight and The Economist use collaborative work to produce many visualizations and analyses that consistently delivery high quality information.

When we started writing this book together, we knew early on that the best product would come from a truly collaborative experience. When you work with others to write words, write code, and think analytically, you’ll learn practices that create inspiration and excitement about the ideas you want to bring to life. Here are some lessons that we learned in our journey:

* Building trust with your collaborators will lead to productive experimentation with new ideas. Build trust slowly with your collaborators by actively listening to feedback and taking risks with new ideas from your teammates
* Adopting an experimental mindset will maximize the opportunities to fail fast and find the solutions that are right for the problem. If you and your collaborators brainstorm an idea that feels promising but ambiguous, try it out and evaluate the results. Do this together often.
* Asking for feedback and giving feedback when asked will lead to a refined end product. Ask your collaborators for their reaction to a visualization you’ve made or a report you’ve written regularly. Does their feedback provide evidence that you’ve made what you intended to make?
* Starting a draft of a project then releasing it to a collaborator is an exercise that builds trust. During our editing process, Josh reminded us of his collaborative spirit by saying “I no longer consider anything I’ve written in this book mine. Change anything you want!”
* Collaboration is contagious. There’s a really easy way to make a work environment more collaborative: approach someone and invite them to collaborate. You might be surprised at how being an active collaborator inspires similar behavior from others

## Learn Every Time You Code

Whether you’re working on a solo project or on a collaborative project, you’ll often find that completing it requires learning something new. Learning requires you to get comfortable with not knowing how to do something because it liberates you from the pressure to know all the time. And when we don’t feel pressure to be perfect, our minds are free to enjoy the challenges of learning.

When some of us first started learning to program in R, the amount of things we didn’t know was glaring. We’d never typed a line of R code in our lives. The thought of writing code that worked, much less using it to do data analysis, felt like a distant goal. After lots of practice and patience, we find ourselves writing code and doing data analysis every day in our education jobs. Yet we still have a long list of things we want to learn so we can push for new ways to understand the lives of our students. Our vocabulary of R syntax has grown, but we still regularly enjoy the experience of learning a new function, discovering a new package, or learning about a analytic routine. The difference between our early learning experiences with R and more recent experiences is we’ve embraced learning as a necessary part of enjoying this craft.

It’s daunting to begin learning a function, concept, or statistical technique. We know it will require sustained discomfort, trial and error, and some frustration before we experience the sweet thrill of a well-executed code chunk. Have you ever noticed how hard it is to get started? But simply beginning the learning process is like strengthening a muscle through exercise: it’s really difficult at first but with repetition, patience, rest, and kindness towards yourself, it gets easier.

So we encourage you to just start. Set a small goal to open up that book about machine learning you’ve been avoiding and get through that first paragraph. Or fire up RStudio and copy and paste that first code chunk from the GitHub repository you’ve been trying to understand. Or run the first example from the documentation of the package you’ve been trying to learn. Trust us–you get used to it and eventually you’ll start to enjoy all that comes with learning.

But even the most motivated of us can’t sustain high effort and challenges indefinitely. When you’ve hit a wall trying to learn a new concept or you’ve tried to fix your code one too many times, it may be time to take a strategic break.

## Learn to Take Strategic Breaks to Help Solve Problems

Taking breaks is one of the most strategic moves you can make when you’re trying to break through to the other side of a programming challenge. And it’s not just true of R programming. Ryan talks about one of the first times he saw taking breaks in action:

I grew up in the 1980s and 1990s, when completing Super Mario Brothers was a monk-like endeavor. I spent hours on helping Mario navigate across green-colored plumbing that unexplainably stuck out from the ground. I practiced combinations of jumps and runs to avoid Koopas to eventually reach the end of a two dimensional level.

Looking back, there were moments when playing this game reached levels of frustration comparable to pressing an impossibly small Lego piece into a complex Lego structure that fell apart in my hands. When these frustrating Mario moments came, there was only one strategy: attempting the challenge over and over again until the inevitable throwing of the controller.

After throwing the controller, I’d engage some other activity. I’d take a nap, go outside, hang out with my sister, or watch TV. Strangely, when I returned to the game console to play again, I’d often progress through the same frustrating challenge on the first or second try. Some moments I succeeded so quickly it was hard to believe the level was ever a challenge to begin with.

I learned a valuable lesson from indulging this exercise repeatedly for many years: during moments of frustration, the mind and body need a break. Taking breaks is a functional activity. It gives you time to synthesize all the learning that happened during repeated attempts to solve a problem. During the breaks the mind and body work to replenish energy to take on problems again. Taking strategic breaks is not failure, it’s the mark of a professional who understands the most efficient way to get to viable solutions.

So when you need it, take a break and live to code another day. We need as many data scientists working in different corners of education as possible, and we don’t want to lose you to burn-out!

## Learn More Meaningfully By Knowing Your Why

Our last bit of advice for staying connected to the bigger picture: Make time to regularly reflect on why you’re doing this work to begin with. Doing data science in education without reflecting on the cause you’re trying to positively affect can make you feel a little like you’re just spinning your wheels. Fortunately, working in education has a common built-in purpose: to provide the highest quality learning experience for students with the resources available.

To be clear, getting familiar with the “why” of your work is an ongoing process. As we learn and grow, our motivations evolve. But still, better to have this evolution be an intentional and self-aware journey. Here are some reflection questions that are useful to prompt the kinds of thinking that activate your “why”:

* Think of what your own learning experience was like. Were there things you wish were different? Were there positive experiences you hope more people will have?
* Is there a teacher or school leader that you’ve worked with or that positively influenced your life as a student? What was it about them that you’d like to see more of in schools?
* Is there an education-related topic, like diversity, special education, or curriculum and instruction, that you have a natural passion for?

Most people don’t have immediate answers to questions like these. It’s good enough to just ask them regularly and reflect on the thoughts and feelings that come up. Finding your “why” and being able to talk about it is more like an exploration and less like a singular “a-ha” moment. But we believe asking these questions is a way to intentionally design meaning into your data science work. When there is a clear purpose and personal connection to why you use data science in education, you’ll be reminded that your work affects the lives of learners, sometimes in profoundly meaningful ways.

That brings our time together to a close, but only for a short while. We hope to see you in the data science in education community, offering value, learning from others, bonding over challenges, and inspiring each other to the best we can for our students.

# References

Allen, I Elaine, and Jeff Seaman. 2008. *Staying the Course: Online Education in the United States, 2008.* ERIC.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2019. *Lme4: Linear Mixed-Effects Models Using ’Eigen’ and S4*. <https://CRAN.R-project.org/package=lme4>.

Betebenner, Damian W. 2019. *RandomNames: Generate Random Given and Surnames*. <https://CRAN.R-project.org/package=randomNames>.

Bovee, Emily, Ryan Estrellado, Jesse Mostipak, Joshua Rosenberg, and Isabella Vel’asquez. 2020. *Dataedu: Package for Data Science in Education Using R*. <https://github.com/data-edu/dataedu>.

Bransford, John D, Ann L Brown, Rodney R Cocking, and others. 2000. *How People Learn*. Vol. 11. Washington, DC: National academy press.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Bryan, Jenny. 2019. “Reproducible Examples and the ‘Reprex‘ Package.” <https://community.rstudio.com/t/video-reproducible-examples-and-the-reprex-package/14732>.

———. 2020. *Happy Git with R*. <https://happygitwithr.com/>.

Bryk, Anthony S, Louis M Gomez, Alicia Grunow, and Paul G LeMahieu. 2015. *Learning to Improve: How America’s Schools Can Get Better at Getting Better*. Harvard Education Press.

Campaign, Data Quality. 2018. “Teachers See the Power of Data - but Don’t Have the Time to Use It.” <https://dataqualitycampaign.org/wp-content/uploads/2018/09/DQC_DataEmpowers-Infographic.pdf>.

Conway, Drew. 2010. “The Data Science Venn Diagram.” *Drew Conway* 10. <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>.

Datnow, Amanda, and Lea Hubbard. 2015. “Teachers’ Use of Assessment Data to Inform Instruction: Lessons from the Past and Prospects for the Future.” *Teachers College Record* 117 (4): n4.

Dirksen, Julie. 2015. *Design for How People Learn*. New Riders.

Dweck, Carol. 2015. “Carol Dweck Revisits the Growth Mindset.” *Education Week* 35 (5): 20–24.

Education Statistics U.S. Department of Education, National Center for. 2019. “Concentration of Public School Students Eligible for Free or Reduced-Price Lunch.” *The Condition of Education 2019*. <https://nces.ed.gov/fastfacts/display.asp?id=898>.

Elbers, Benjamin. 2020. *Tidylog: Logging for ’Dplyr’ and ’Tidyr’ Functions*. <https://CRAN.R-project.org/package=tidylog>.

Emdin, Christopher. 2016. *For White Folks Who Teach in the Hood... And the Rest of Y’all Too: Reality Pedagogy and Urban Education*. Beacon Press.

Firke, Sam. 2020. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://CRAN.R-project.org/package=janitor>.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, Skye Bender-deMoll, and Martina Morris. 2019. *Statnet: Software Tools for the Statistical Analysis of Network Data*. <https://CRAN.R-project.org/package=statnet>.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. 2019. *Ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. <https://CRAN.R-project.org/package=ergm>.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hattie, John. 2012. *Visible Learning for Teachers: Maximizing Impact on Learning*. Routledge.

Healy, Kieran. 2019. *Data Visualization: A Practical Introduction*. Princeton University Press.

Heath, Chip Heathand Dan. 2006. “The Curse of Knowledge.” *Harvard Business Review*. <https://hbr.org/2006/12/the-curse-of-knowledge>.

Hill, Alison. 2017. “Up and Running with Blogdown.” <https://alison.rbind.io/post/2017-06-12-up-and-running-with-blogdown/>.

Hoff, Peter, Bailey Fosdick, Alex Volfovsky, and Yanjun He. 2017. *Amen: Additive and Multiplicative Effects Models for Networks and Relational Data*. <https://CRAN.R-project.org/package=amen>.

Ismay, Chester, and Albert Y. Kim. 2019. *Statistical Inference via Data Science*. CRC Press.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*.

Kearney, Michael W. 2016. “Rtweet: Collecting Twitter Data.” *Comprehensive R Archive Network. Available at: Https://Cran. R-Project. Org/Package= Rtweet*.

Kleon, Austin. 2012. *Steal Like an Artist: 10 Things Nobody Told You About Being Creative*. Workman Publishing.

Kozol, Jonathan. 2012. *Savage Inequalities: Children in America’s Schools*. Broadway Books.

Kuhn, Max. 2020. *Caret: Classification and Regression Training*. <https://CRAN.R-project.org/package=caret>.

Kuhn, Max, and others. 2008. “Building Predictive Models in R Using the Caret Package.” *Journal of Statistical Software* 28 (5): 1–26.

Kurz, Solomon. 2019. *Statistical Rethinking with Brms, Ggplot2, and the Tidyverse*. <https://bookdown.org/ajkurz/Statistical_Rethinking_recoded/>.

Leeper, Thomas J. 2018. *Tabulizer: Bindings for ’Tabula’ Pdf Table Extractor Library*. <https://CRAN.R-project.org/package=tabulizer>.

Lehrer, Richard, Min-joung Kim, and Leona Schauble. 2007. “Supporting the Development of Conceptions of Statistics by Engaging Students in Measuring and Modeling Variability.” *International Journal of Computers for Mathematical Learning* 12 (3): 195–216.

Lehrer, Richard, and Leona Schauble. 2015. “The Development of Scientific Thinking.” *Handbook of Child Psychology and Developmental Science*, 1–44.

Lemov, Doug. 2015. *Teach Like a Champion 2.0: 62 Techniques That Put Students on the Path to College*. John Wiley & Sons.

Lüdecke, Daniel. 2020. *SjPlot: Data Visualization for Statistics in Social Science*. <https://CRAN.R-project.org/package=sjPlot>.

Lüdecke, Daniel, Dominique Makowski, and Philip Waggoner. 2020. *Performance: Assessment of Regression Models Performance*. <https://CRAN.R-project.org/package=performance>.

Mandinach, Ellen B, and Edith S Gummer. 2013. “A Systemic View of Implementing Data Literacy in Educator Preparation.” *Educational Researcher* 42 (1): 30–37.

McTighe, Jay, and Judy Willis. 2019. *Upgrade Your Teaching: Understanding by Design Meets Neuroscience*. ASCD.

Mohammad, Saif M, and Peter D Turney. 2013. “Crowdsourcing a Word–Emotion Association Lexicon.” *Computational Intelligence* 29 (3): 436–65.

Moore Jr, Eddie, Ali Michael, and Marguerite W Penick-Parks. 2017. *The Guide for White Women Who Teach Black Boys*. Corwin Press.

Murphy, Mekado. 2011. “The Adventures of Spielberg: An Interview.” <https://carpetbagger.blogs.nytimes.com/2011/12/20/the-adventures-of-spielberg-an-interview/>.

Müller, Kirill. 2017. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.

National Academies of Sciences, Engineering, Medicine, and others. 2018. *How People Learn Ii: Learners, Contexts, and Cultures*. National Academies Press.

Newton, Elizabeth Louise. 1991. “The Rocky Road from Actions to Intentions.” PhD thesis, Stanford University.

of Education, United States Department. 2020. <https://www2.ed.gov/programs/osepidea/618-data/state-level-data-files/index.html#bccee>.

O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Pedersen, Thomas Lin. 2019. *Tidygraph: A Tidy Api for Graph Manipulation*. <https://CRAN.R-project.org/package=tidygraph>.

———. 2020. *Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. <https://CRAN.R-project.org/package=ggraph>.

Peng, Roger D, and Elizabeth Matsui. 2015. “The Art of Data Science.” *A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC*.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

“Reachable: Data Collection Methods for Sexual Orientation and Gender Identity.” 2016. <https://williamsinstitute.law.ucla.edu/wp-content/uploads/Reachable-Data-collection-methods-for-sexual-orientation-gender-identity-March-2016.pdf](https://williamsinstitute.law.ucla.edu/wp-content/uploads/Reachable-Data-collection-methods-for-sexual-orientation-gender-identity-March-2016.pdf>.

Robinson, David. 2018. “Advice to Aspiring Data Scientists: Start a Blog.”

Robinson, David, and Julia Silge. 2019. *Tidytext: Text Mining Using ’Dplyr’, ’Ggplot2’, and Other Tidy Tools*. <https://CRAN.R-project.org/package=tidytext>.

Rosenberg, Joshua M, Spencer P Greenhalgh, Matthew J Koehler, Erica R Hamilton, and Mete Akcaoglu. 2016. “An Investigation of State Educational Twitter Hashtags (Seths) as Affinity Spaces.” *E-Learning and Digital Media* 13 (1-2): 24–44.

Rosenberg, Joshua M., Michael Lawson, Daniel Anderson, Teya Rutherford, and Ryan Seth Jones. 2020. “Making Data Science ‘Count’: Data Science and Learning, Design, and Technology Research.” In *Research Methods in Learning Design & Technology.*, 1–13. Springer.

RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. <http://www.rstudio.com/>.

Schools, Great. n.d. “Aggregate Data.” *Ed Glossary*. <https://www.edglossary.org/aggregate-data/>.

Siemens, George, and Ryan SJ d Baker. 2012. “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration.” In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252–54. ACM.

Silge, Julia, and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. " O’Reilly Media, Inc.".

Spillane, James P, Chong Min Kim, and Kenneth A Frank. 2012. “Instructional Advice and Information Providing and Receiving Behavior in Elementary Schools: Exploring Tie Formation as a Building Block in Social Capital Development.” *American Educational Research Journal* 49 (6): 1112–45.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2018. *Lubridate: Make Dealing with Dates a Little Easier*. <https://CRAN.R-project.org/package=lubridate>.

Stanley, David. 2018. *ApaTables: Create American Psychological Association (Apa) Style Tables*. <https://CRAN.R-project.org/package=apaTables>.

Trust, Torrey, Daniel G Krutka, and Jeffrey Paul Carpenter. 2016. “‘Together We Are Better’: Professional Learning Networks for Teachers.” *Computers & Education* 102: 15–34.

Urbanek, Simon. 2019. *RJava: Low-Level R to Java Interface*. <https://CRAN.R-project.org/package=rJava>.

Victore, James. 2019. *Feck Perfuction: Dangerous Ideas on the Business of Life*. Chronicle books.

Wickham, Hadley. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (10). <https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf](https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf>.

———. 2015. *R Packages*. O’Reilly Media.

———. 2019. *Advanced R*. 2nd ed. <https://adv-r.hadley.nz/>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.

Wickham, Hadley, and Garrett Grolemund. 2018. *R for Data Science*. O’Reilly.

Wiggins, Grant, Grant P Wiggins, and Jay McTighe. 2005. *Understanding by Design*. Ascd.

Wikipedia. 2020. “Reproducible Research.” <https://en.wikipedia.org/wiki/Reproducibility#Reproducible_research>.

Wilson, Greg. 2009. *Teaching Tech Together*. <https://teachtogether.tech/>.

Woo, Kara. 2018. “Anyone Can Play Git/R: Tips for First-Time Contributions to R Packages.” <https://speakerdeck.com/karawoo/r-tips-for-first-time-contributions-to-r-packages>.

Xie, Yihui. 2019. *Bookdown: Authoring Books and Technical Documents with R Markdown*. CRC Press. <https://bookdown.org/yihui/bookdown/>.

Xie, Yihui, Amber Thomas, and Alison Presmanes Hill. 2019. *Blogdown: Creating Websites with R Markdown*. CRC Press. <https://bookdown.org/yihui/blogdown/>.

1. The documentation for the dataset is available here: <https://www2.ed.gov/programs/osepidea/618-data/collection-documentation/data-documentation-files/part-b/child-count-and-educational-environment/idea-partb-childcountandedenvironment-2017-18.pdf> [↑](#footnote-ref-270)
2. Run ?lme4::pvalues to see a discussion of the issue as well as solutions; we have found the lmerTest to be helpful as an easy solution, though we note that some of the recommendations available through ?lme4::pvalues may be preferable, as the technique lmerTest implements has some known issues. [↑](#footnote-ref-377)