A Minimal Book Example

John Doe

2022-08-17

Contents

Chapter 1

About

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 2

Syllabus

Instructor: Tobin Turner

Office Hours: mutually convenient time arranged by email e-mail: jtturner@presby.edu

2.1 Course Objectives and Learning Outcomes

This course is designed to introduce to data science. Students will apply statistical knowledge and techniques to both business and non-business contexts.

At the end of this course students should be able to:

- Demonstrate mastery of the statistical software in R and the Rstudio IDE.
- Data wrangle (the process of cleaning and unifying messy and complex data sets for easy access and analysis)
- Demonstrate mastery of single and multiple regression.
- Demonstrate mastery of these dplyr functions: filter, select, mutate, group_by, summaize, and tally.
- Demonstrate mastery of how business analytics is related to other business functions and is important to the success of the business entity.

This course will be focused on both understanding and applying key business analytical concepts. Although the text serves as a useful foundation for the concepts covered in the class, simple memorization of the material in the text will not be sufficient. Class participation, discussion, and application are critical.

2.2 Text and Resources

- The course website (primary reource)
- An Introduction to Statistical Learning with Applications in R; by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
- R: A self-learn tutorial.
- Other free, publicly available datasets and publications.

2.3 Performance Evaluation (Grading)

- Quizzes and Assignments - 20%
- Exam 1 - 20%
- Exam 2 - 20%
- Exam 3 - 20%
- Final Exam - 20%

2.3.1 Exams

Exams will cover assigned chapters in the textbook, other assigned readings, lectures, class exercises, class discussions, videos, and guest speakers. I will typically allocate time prior to each exam to clearly identify the body of knowledge each test will cover and to answer questions about the format and objectives of the exam.

2.3.2 Quizzes – DON"T MISS CLASS

- The average of all quizzes and assignments will comprise the Quizzes and Assignments - 20% portion of your final grade
- Quizzes are designed to prepare you for your exams and to ensure you stay up wiht the course material
- Missed quizzes cannot be made up later. Be present.

Quizzes rule. **LISTEN.**

2.3.3 Final Average

- Final Average Grade
  - 90-100 A
  - 88-89 B+
  - 82-87 B+

- 80-81 B-
- 78-79 C+
- 72-77 C+
- 70-71 C-
- 60-69 D
- 59 and below F

2.4 Class Participation:

I will frequently give readings or assignments for you to complete prior to the next class meeting. I expect you to fully engage the material: answer questions, pose questions, provide insightful observations. Keep in mind that quality is an important component in "participation." Periodic cold calls will take place. I will also put students in the "hot seat" on occasion. In these class sessions, I may select a random group of students to lead us in the discussion and debate. Because the selection of participants will not be announced until class begins, everyone will be expected to prepare for the discussion. Reading the assigned chapters and articles are the best way to prepare for the discussion. If you have concerns about being called on in class, please see me to discuss. The purpose of the "hot seat" is not to stress or embarrass students, but to encourage students to actively engage the material.

2.5 Phones

Phones are not allowed to be used in class without the instructor's prior consent. If you have a need of a phone during class please let me know before class. Unauthorized use of electronic devices may result in the lowering of the grade or dismissal from the class. **I mean this.**

The phone thing? I mean this.

2.6 Attendance

You are expected to be regular and punctual in your class attendance. Students are responsible for all the material missed and homework assignments made. If class is missed, notes/homework should be obtained from another student. If I am more than 15 minutes late, class is considered cancelled. No more than 4 absences are allowed during a semester. Exceeding the absence policy may result in receiving an F for the course. The professors roll is the official roll and students not present when roll is taken will be counted as absent. If a student must miss an exam, she or he must work out an agreeable time with the instructor to take the test prior to the exam being given. If a student misses

a test due to an emergency, the student must inform the instructor as soon as is possible. In special cases, the instructor may allow the student to take a make-up exam.

2.7 Accommodations

Presbyterian College is committed to providing reasonable accommodations for all students with documented disabilities. If you are seeking academic accommodations under the Americans with Disabilities Act, you must register with the Academic Success Office, located on 5th Avenue (beside Campus Police). To receive these accommodations, please obtain the proper Accommodations Approval Form from that office, and then meet with me at the beginning of the semester to discuss how we may deliver your approved accommodations. I especially encourage you to meet with me well in advance of the actual accommodations being provided, as it may not be feasible to offer immediate accommodations without sufficient advance notice (such as in the case of tests). I can assure you that all discussions will remain confidential. Disability Services information is located at this link http://bit.ly/PCdisabilityservices

Additionally, it is the student's responsibility to give the instructor one week's notice prior to each instance where accommodation will be required.

2.8 Honor Code and Plagiarism:

All assignments/exams must be your own work. Any copying or use of unauthorized assistance will be treated as a violation of PC's Honor Code. If you are unsure of what resources are allowed, please ask. Please note that all text longer than 7 words taken from ANY other source must be placed in quotations and cited. Also, summarizing ANY other source must also be cited. Using ANY other source and showing work to be your own is a violation of plagiarism and the honor code.

2.9 First-Generation Version:

I am a Presby First+ Advocate. I am here to support our current first-generation students. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

2.10 Continuing Advocate Version

I am a Presby First+ Advocate. I am committed to supporting first-generation students at Presbyterian College. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me anytime or visit me during my office hours. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

Chapter 3

Our Class Rhythm

Monday: Wrap up previous topic and introduce what you've pre-read about. Chat. Play. Work some examples. Make sure the topics applies to real-life.

Wednesday: Work more examples. Chat as needed. **Live our best lives. :)**.

Friday: Apply what we've learned – demonstrate your mastery (typically in the form of a quiz, lab, or assignment). Rinse. Repeat.

Chapter 4

End in Mind

Dana Simmons: "Can you predict which students will enroll at PC?"

Christina Miller: ??? Well, can you? ???

Chapter 5

Schedule

This is a tentative schedule, **BUT** I will do my very best to stick to it, so that you may plan accordingly!

Spring 2022

Date	Topic
Date	A1
Monday, January 10, 2022	R basics and install
Wednesday, January 12, 2022	R basics and workflows
Friday, January 14, 2022	QUIZ 1
Monday, January 17, 2022	MLK Holiday
Wednesday, January 19, 2022	Objects and Arithmetic
Friday, January 21, 2022	QUIZ
Monday, January 24, 2022	Summaries and Subscripting
Wednesday, January 26, 2022	Matrices and `mtcars`
Friday, January 28, 2022	QUIZ
Monday, January 31, 2022	Class
Wednesday, February 2, 2022	Class
Friday, February 4, 2022	QUIZ
Monday, February 7, 2022	Social Dilemma and Review
Wednesday, February 9, 2022	**EXAM 1**
Friday, February 11, 2022	Class
Monday, February 14, 2022	Class
Wednesday, February 16, 2022	Online Class
Friday, February 18, 2022	Online QUIZ
Monday, February 21, 2022	Class
Wednesday, February 23, 2022	Class

Date	Topic
Friday, February 25, 2022	QUIZ
Monday, February 28, 2022	Class
Wednesday, March 2, 2022	Class
Friday, March 4, 2022	QUIZ
Monday, March 7, 2022	**EXAM 2**
Wednesday, March 9, 2022	Online Class
Friday, March 11, 2022	Online Class
Monday, March 14, 2022	SPRING BREAK
Wednesday, March 16, 2022	SPRING BREAK
Friday, March 18, 2022	SPRING BREAK
Monday, March 21, 2022	Class
Wednesday, March 23, 2022	Class
Friday, March 25, 2022	QUIZ
Monday, March 28, 2022	ADVISING WEEK
Wednesday, March 30, 2022	ADVISING WEEK
Friday, April 1, 2022	QUIZ
Monday, April 4, 2022	Class
Wednesday, April 6, 2022	Class
Friday, April 8, 2022	QUIZ
Monday, April 11, 2022	Class
Wednesday, April 13, 2022	**EXAM 3**
Friday, April 15, 2022	Easter Holidays
Monday, April 18, 2022	Easter Holidays
Wednesday, April 20, 2022	Class
Friday, April 22, 2022	QUIZ
Monday, April 25, 2022	Class
Wednesday, April 27, 2022	QUIZ
Friday, April 29, 2022	LAST DAY
Monday, May 2, 2022	**Final Exam** 5:30 p.m. – E period

Chapter 6

Elephant in the room: R

There are a variety of different applications for R. Yes, the more obvious ones would be things such as machine learning, artificial intelligence, and data mining. However, the possibilities with this program are honestly limitless.

6.1 The bad news

> **"The bad news is whenever you're learning a new tool, for a long time you're going to suck. It's going to be very frustrating. But, the good news is that that is typical, it's something that happens to everyone, and it's only temporary ... [T]here is no way to go from knowing nothing about a subject to knowing something about a subject and being an expert in it without going through a period of great frustration."**

– Hadley Wickham

6.2 About R

R is a software language for carrying out complicated (and simple) statistical analyses. It includes routines for data summary and exploration, graphical presentation and data modelling.

Chapter 7

R vs. Excel

7.1 Both are useful

Data analytics are increasingly important components of decision-making in any business. Whether you're a part of a marketing team that needs to generate visuals to highlight industry trends, or you're looking to generate financial statements, you will need an analytics program to help you develop your reports and effectively communicate your findings.

Both R and Excel are excellent data analytics tools, but they each have distinct functionality.

> **Please make sure you can explain the distinct functions of both R and Excel! YOU WILL NEED TO KNOW THIS!**

Excel is a well-known software program included in the Microsoft Office Suite. Used to create spreadsheets, execute calculations, produce charts, and perform statistical analysis, Excel is used by many professionals across a variety of industries. PC's **BADM 299** prepares you well for using Excel.

R is a free, open-source programming language and software environment that's frequently used in big data analysis and statistical computing. R has many advanced functions and capabilities.

7.2 Differences Between R and Excel

When choosing between R and Excel, it's important to understand how either software can get you the results you need. Here are some key differences between R and Excel to help you decide which makes the most sense to use.

7.3 Ease of Use & Learning the Software

Most people have likely already learned at least a few basic tips in Microsoft Excel. That's one substantial benefit of using Excel—the initial learning curve is quite minimal, and most analysis can be done via point-and-click on the top panel. Once a user imports their data into the program, it's not very hard to make basic graphs and charts.

R is a programming language, however, meaning the initial learning curve is steeper. It will take most at least a few weeks to familiarize themselves with the interface and master the various functions. Luckily, using R can quickly become second-nature with practice.

7.4 Replicating Analysis

R, while less user-friendly with a more intimidating user interface, has the capability to reproduce analyses repeatedly and with very different datasets. This can be incredibly helpful for large projects with multiple data sets, as you'll keep everything consistent and clean, without having to rewrite the script each time.

Since Excel's user interface is point-and-click, you'll need to rely on memory and repetition frequently. You cannot import codes and scripts as you would with R, so you'll have to "reinvent the wheel" to perform the same analysis across different data sets. This is not detrimental if you are doing basic statistics, but it may become time-consuming with more complicated analyses.

> For example, let's say you have thoroughly analyzed the analytics of 1 football season. How could R (vs Excel) help you quickly analyze a new season's data?

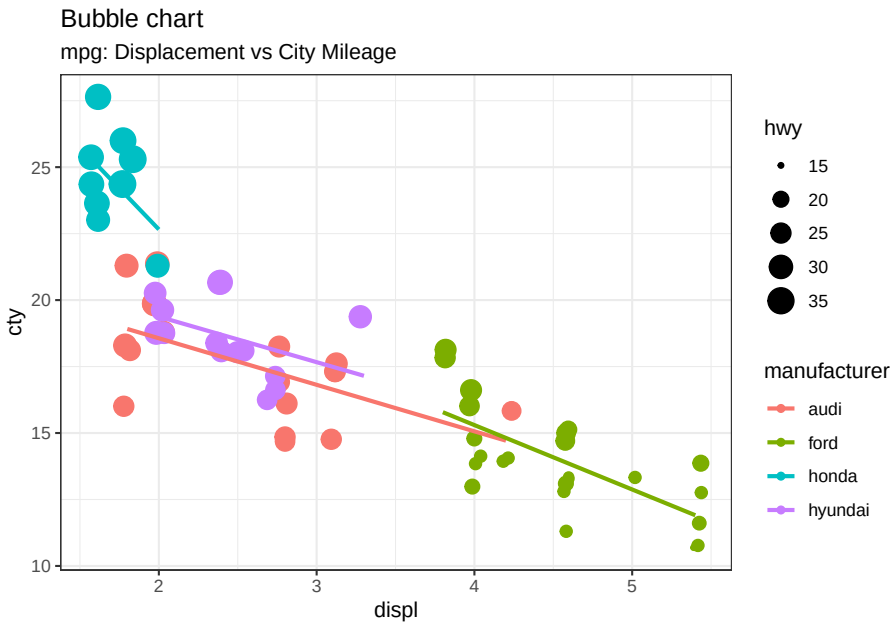
7.5 Visualization

When deciding between R and Excel, ask yourself, "How detailed do my visualizations need to be in order to achieve my goal(s)?" In Excel, for example, you can quickly highlight a group of cells and make a simple chart for PowerPoint. If you need a more comprehensive graph, however, R may be your best bet. R can produce incredibly attractive, detailed visuals that can help stakeholders understand your findings.

It all comes down to what you need your graphics to do. If you're just looking to cobble together a quick-and-dirty presentation to visualize data for your coworkers, then making simple straightforward charts in Excel will suffice. For those

planning to publish large amounts of complicated data to various stakeholders, spending the time in R to create impressive interactive visual representations will likely be worth your while.

For example, here's and example of a pretty easy visualization in R that would be challenging to do (and update) in Excel.



7.6 Packages

In R, the fundamental unit of shareable code is the package. A package bundles together code, data, documentation, and tests, and is easy to share with others. As of June 2019, there were over 14,000 packages available on the Comprehensive R Archive Network, or CRAN, the public clearing house for R packages. This huge variety of packages is one of the reasons that R is so successful: the chances are that someone has already solved a problem that you're working on, and you can benefit from their work by downloading their package.

But packages are useful even if you never share your code. As Hilary Parker says in her introduction to packages: "Seriously, it doesn't have to be about sharing your code (although that is an added benefit!). It is about saving yourself time." Organising code in a package makes your life easier because packages come with conventions. For example, you put R code in R/, you put tests in tests/ and you put data in data/. These conventions are helpful because:

- They save you time — you don't need to think about the best way to organise a project, you can just follow a template.
- Standardized conventions lead to standardized tools — if you buy into R's package conventions, you get many tools for free.

  We will chat more about packages, but for fun check out the links below…

7.7 Careers

Aptitude with Excel and R are incredibly valuable competencies that are in-demand across a variety of industries. Countless jobs are looking for applicants with at least some Excel experience (pivot tables look really good on a resumé), but R has a higher earning potential and is more in-demand than Excel.

R is one of the most popular programming languages and is an industry-standard for data analytics and data science. If you want to enter either field, there's a good chance you'll have a competitive advantage by knowing R. Entry-level jobs for those focusing on R also tend to make a high salary, frequently starting off earning more than $75,000.

Countless job listings also require Excel competency. From administrative assistants, marketers, academics, and more, everyone is expected to use Excel to some degree, whereas 10 to 15 years ago it was optional. Having a good background in Excel is still attractive on a resumé and will help to land a career with a high earning potential, but there are not many jobs looking for Excel skills alone.

7.8 Summary – Using R and Excel

R and Excel are beneficial in different ways. Excel starts off easier to learn and is frequently cited as the go-to program for reporting, thanks to its speed and efficiency. R is designed to handle larger data sets, to be reproducible, and to create more detailed visualizations. It's not a question of choosing between R and Excel, but deciding which program to use for different needs.

Chapter 8

R basics and workflows

8.1 Basics of working with R at the command line and RStudio goodies

Launch RStudio/R.

You will first intall R and then. RStudio.

- Installing R
- Installing RStudio
- Customizing RStudio
- RStudio Quick keys

8.2 In Rstudio - where we will live

Notice the default panes:

- Console (entire left)
- Environment/History (tabbed in upper right)
- Files/Plots/Packages/Help (tabbed in lower right)

8.3

Rstudio Console

FYI: you can change the default location of the panes, among many other things: Customizing RStudio.

Go into the Console, where we interact with the live R process.

You can make an object by assigning a value or statement to a letter or string. We use `<-` to assign objects meaning. Create and inspect the following object:

Your first analysis in R:

```
x <- 3 * 4
x
#> [1] 12
```

All R statements where you create objects – "assignments" – have this form:

```
objectName <- value
```

and in my head I hear, e.g., "x gets 12".

You will make lots of assignments and the operator `<-` is a pain to type. Don't be lazy and use `=`, although it would work, because it will just sow confusion later. Instead, utilize RStudio's keyboard shortcut: Alt + - (the minus sign).

I want to be your friend. As a friend, I implore you, learn this:

> In RStudio insert the <- assignment operator with Option + - (the minus sign)on a Mac, or Alt + - (the minus sign) on Windows.

Notice that RStudio automatically surrounds `<-` with spaces, which demonstrates a useful code formatting practice. Code is miserable to read on a good day. Give your eyes a break and use spaces.

RStudio offers many handy RStudio Quick keys. Also, Alt+Shift+K brings up a keyboard shortcut reference card.

Object names cannot start with a digit and cannot contain certain other characters such as a comma or a space. You will be wise to adopt a convention for demarcating words in names, but note that best practice is to choose ONE convention and stay true to it throughout your code.

```
i_use_snake_case
other.people.use.periods
evenOthersUseCamelCase
```

Make another assignment:

```
this_is_a_really_long_name <- 2.5
```

To inspect this, try out RStudio's completion facility: type the first few characters, press TAB, add characters until you disambiguate, then press return.

Make another assignment:

```
turner_rocks <- 2 ^ 3
```

When making assignments, it is best practice to keep the names brief, yet descriptive. For instance, while the name "this_is_a_really_long_name" is accurate, so is "long_name" and this is much more intuitive and easy to read/type over and over.

Let's try to inspect:

```
turnerrocks
#> Error in eval(expr, envir, enclos): object 'turnerrocks' not found
Turner_rocks
#> Error in eval(expr, envir, enclos): object 'Turner_rocks' not found
turner_rocks
#> [1] 8
```

Implicit contract with the computer / scripting language: Computer will do tedious computation for you. In return, you will be completely precise in your instructions. Typos matter. Case matters. Get better at typing.

R has a mind-blowing collection of built-in functions that are accessed like so:

```
functionName(arg1 = val1, arg2 = val2, and so on)
```

Let's try using `seq()` which makes regular sequences of numbers and, while we're at it, demo more helpful features of RStudio.

Type `se` and hit TAB. A pop up shows you possible completions. Specify `seq()` by typing more to disambiguate or using the up/down arrows to select. Notice the floating tool-tip-type help that pops up, reminding you of a function's arguments. If you want even more help, press F1 as directed to get the full documentation in the help tab of the lower right pane. Now open the parentheses and notice the automatic addition of the closing parenthesis and the placement of cursor in the middle. Type the arguments `1, 10` and hit return. RStudio also exits the parenthetical expression for you. IDEs are great.

```
seq(1, 10)
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

The above also demonstrates something about how R resolves function arguments. You can always specify in `name = value` form. But if you do not, R attempts to resolve by position. So above, it is assumed that we want a sequence `from = 1` that goes `to = 10`. Since we didn't specify step size, the default value of `by` in the function definition is used, which ends up being 1 in this case. For functions I call often, I might use this resolve by position for the first argument or maybe the first two. After that, I always use `name = value`.

Make this assignment and notice similar help with quotation marks.

```
yo <- "hello world"
```

If you just make an assignment, you don't get to see the value, so then you're tempted to immediately inspect.

```
y <- seq(1, 10)
y
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

This common action can be shortened by surrounding the assignment with parentheses, which causes assignment and "print to screen" to happen.

It is best practice to always attempt to "print" your assignments after creating them. This will help leviate the issue of searching 200+ lines of code for that one error causing argument.

```
(y <- seq(1, 10))
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

Not all functions have (or require) arguments:

```
date()
#> [1] "Wed Aug 17 19:34:48 2022"
```

Now look at your workspace – in the upper right pane. The workspace is where user-defined objects accumulate. You can also get a listing of these objects with commands:

```
objects()
#> [1] "this_is_a_really_long_name"
#> [2] "turner_rocks"
```

```
#> [3] "x"
#> [4] "y"
#> [5] "yo"
ls()
#> [1] "this_is_a_really_long_name"
#> [2] "turner_rocks"
#> [3] "x"
#> [4] "y"
#> [5] "yo"
```

If you want to remove the object named `y`, you can do this:

```
rm(y)
```

To remove everything:

```
rm(list = ls())
```

or click the broom in RStudio's Environment pane.

8.4 Workspace and working directory

One day you will need to quit R, go do something else and return to your analysis later (this is a very happy day).

One day you will have multiple analyses going that use R and you want to keep them separate (a not so happy day).

One day you will need to bring data from the outside world into R and send numerical results and figures from R back out into the world (the happiest of days).

To handle these real life situations, you need to make two decisions:

- What about your analysis is "real", i.e. will you save it as your lasting record of what happened?
- Where does your analysis "live"?

8.4.1 Workspace, `.RData`

As a beginning R user, it's OK to consider your workspace "real". *Very soon*, I urge you to evolve to the next level, where you consider your saved R scripts as "real". (In either case, of course the input data is very much real and requires

preservation!) With the input data and the R code you used, you can reproduce *everything*. You can make your analysis fancier. You can get to the bottom of puzzling results and discover and fix bugs in your code. You can reuse the code to conduct similar analyses in new projects. You can remake a figure with different aspect ratio or save is as TIFF instead of PDF. You are ready to take questions. You are ready for the future.

If you regard your workspace as "real" (saving and reloading all the time), if you need to redo analysis … you're going to either redo a lot of typing (making mistakes all the way) or will have to mine your R history for the commands you used. Rather than [becoming an expert on managing the R history][rstudio-command-history], a better use of your time and psychic energy is to keep your "good" R code in a script for future reuse.

Because it can be useful sometimes, note the commands you've recently run appear in the History pane.

But you don't have to choose right now and the two strategies are not incompatible. Let's demo the save / reload the workspace approach.

Upon quitting R, you have to decide if you want to save your workspace, for potential restoration the next time you launch R. Depending on your set up, R or your IDE, e.g. RStudio, will probably prompt you to make this decision.

Quit R/RStudio, either from the menu, using a keyboard shortcut, or by typing `q()` in the Console. You'll get a prompt like this:

> Save workspace image to ~/.Rdata?

Note where the workspace image is to be saved and then click "Save".

Using your favorite method, visit the directory where image was saved and verify there is a file named `.RData`. You will also see a file `.Rhistory`, holding the commands submitted in your recent session.

Restart RStudio. In the Console you will see a line like this:

```
[Workspace loaded from ~/.RData]
```

indicating that your workspace has been restored. Look in the Workspace pane and you'll see the same objects as before. In the History tab of the same pane, you should also see your command history. You're back in business. This way of starting and stopping analytical work will not serve you well for long but it's a start.

8.4.2 Working directory

Any process running on your computer has a notion of its "working directory". In R, this is where R will look, by default, for files you ask it to load. It also where, by default, any files you write to disk will go. Chances are your current working directory is the directory we inspected above, i.e. the one where RStudio wanted to save the workspace.

You can explicitly check your working directory with:

```
getwd()
```

It is also displayed at the top of the RStudio console.

As a beginning R user, it's OK let your home directory or any other weird directory on your computer be R's working directory. *Very soon*, I urge you to evolve to the next level, where you organize your analytical projects into directories and, when working on project A, set R's working directory to the associated directory.

Although I do not recommend it, in case you're curious, you can set R's working directory at the command line like so:

```
setwd("~/myCoolProject")
```

Although I do not recommend it, you can also use RStudio's Files pane to navigate to a directory and then set it as working directory from the menu: *Session > Set Working Directory > To Files Pane Location*. (You'll see even more options there). Or within the Files pane, choose "More" and "Set As Working Directory".

But there's a better way. A way that also puts you on the path to managing your R work like an expert.

8.5 Exercises

1. Create an object called "cool_object" and assign it the number 100.
2. Create a new object called "big_brain" and multiply the object from question one by 15.
3. Print both objects.
4. Use base R functions to return today's date and print it.
5. Create a sequence of numbers counting from 10 to 100 by 2.

6. Identify your working directory. What is it? Change it to where you want it.

7. Save the R script that answers questions 1 through 5 above. Save it; clean and close Rstudio; reopen your script and run it. Make sense?

Chapter 9

Objects and Arithmetic

9.1 Introduction

R stores information and operates on objects. The simplest objects are scalars, vectors and matrices. But there are many others: lists and dataframes for example. In advanced use of R it can also be useful to define new types of object, specific for particular application. We will stick with just the most commonly used objects here. An important feature of R is that it will do different things on different types of objects. For example, type:

```
4+6
#> [1] 10
```

So, R does scalar arithmetic returning the scalar value 10. (In actual fact, R returns a vector of length 1 - hence the [1] denoting first element of the vector. We can assign objects values for subsequent use. For example:

```
x<-6
y<-4
z<-x+y
```

would do the same calculation as above, storing the result in an object called z. We can look at the contents of the object by simply typing its name:

```
z
#> [1] 10
```

Storing things such as calculations as objects is extremely useful, especially in longer scripts. Mainly because you will likely call the same equation multiple times and if you need to revise it in any way you only need to change the initial assignment rather than every line.

9.2 Basic Functions

At any time we can list the objects which we have created:

```
ls()
#> [1] "x" "y" "z"
```

Notice that ls is actually an object itself. Typing ls would result in a display of the contents of this object, in this case, the commands of the function. The use of parentheses, ls(), ensures that the function is executed and its result - in this case, a list of the objects in the directory - displayed. More commonly a function will operate on an object, for example:

```
sqrt(16)
#> [1] 4
```

calculates the square root of 16. Objects can be removed from the current workspace with the rm function:

```
rm(x,y)
```

for example.

There are many standard functions available in R, and it is also possible to create new ones. Vectors can be created in R in a number of ways. We can describe all of the elements:

```
z<-c(5,9,1,0)
```

Note the use of the function c to concatenate or glue together individual elements. This function can be used much more widely, for example

```
x<-c(5,9)
y<-c(1,0)
z<-c(x,y)
```

would lead to the same result by gluing together two vectors to create a single vector.

Sequences can be generated as follows:

```
x<-1:10
```

while more general sequences can be generated using the seq command. For example:

```
seq(1,9,by=2)
#> [1] 1 3 5 7 9
seq(8,20,length=6)
#> [1]  8.0 10.4 12.8 15.2 17.6 20.0
```

These examples illustrate that many functions in R have optional arguments, in this case, either the step length or the total length of the sequence (it doesn't make sense to use both). If you leave out both of these options, R will make its own default choice, in this case assuming a step length of 1. So, for example,

```
seq(8,20,length=6)
#> [1]  8.0 10.4 12.8 15.2 17.6 20.0
x<-seq(1,10)
x
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

also generates a vector of integers from 1 to 10.

At this point it's worth mentioning the help facility. If you don't know how to use a function, or don't know what the options or default values are, type help(functionname) where functionname is the name of the function you are interested in. This will usually help and will often include examples to make things even clearer.

Another useful function for building vectors is the rep command for repeating things. For example:

```
rep(0,100)
#>   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#>  [28] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#>  [55] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
#>  [82] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
rep(1:3,6)
#>  [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
rep(1:3,c(6,6,6))
#>  [1] 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3
```

which we could also simplify cleverly as:

```
rep(1:3,rep(6,3))
#>  [1] 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3
```

As explained above, R will often adapt to the objects it is asked to work on. For example:

```
x<-c(6,8,9)
y<-c(1,2,4)
x+y
#> [1]  7 10 13
```

and

```
x*y
#> [1]  6 16 36
```

showing that R uses componentwise arithmetic on vectors. R will also try to make sense if objects are mixed. For example:

```
x<-c(6,8,9)
x+2
#> [1]  8 10 11
```

though care should be taken to make sure that R is doing what you would like it to in these circumstances.

Two particularly useful functions worth remembering are length which returns the length of a vector (i.e. the number of elements it contains) and sum which calculates the sum of the elements of a vector.

9.3 Statistics and Summaries

One of the most useful functions of R is its ability to perform statistical analysis on large pieces of data. Later in this book we will cover how this analysis can be used to build complex models and test their accuracy. For now, the focus will be on how to perform basic statistical analysis and the definitions of the more basic terms.

9.3.1 Statistical Analysis

You've undoubtedly have heard the terms mean, median, standard deviation and variance. However, if asked to define these terms and provide examples of their usefulness, what would you say? This section is going to answer the first part of that question.

Mean is a pretty straight forward concept for most. Commonly referred to as the "average", you find this number by adding all the numbers in your data set together and then divide by the total number of points you used. This is easily done on a calculator if you are working with less then 20 digits, however this

is almost never the case in data science. Often, especially with R, you will be dealing with data sets with hundreds, thousands, or maybe even hundreds of thousands of data points. Fortunately, R takes the pain away from this process and offers a very simple function that will do this for you. Start by creating an object that contains a list of numbers, and then simply use the mean function to calculate the average of your variable.

```
a <- c(3.8,4,3.1,2.1,12.6,17,8.43,11,2,3,9,5,3,0.5)

mean(a)
#> [1] 6.037857
```

The median of a data set is data point that fall in the middle of all the other points when they are ordered from lowest to highest. Again, something easily found even without a calculator if you have small data sets, however humans can miscount and some data sets are just too massive for us to do on our own. R never makes these counting mistakes and by using a simple function you save yourself hours of mundane counting. Start the same way you found mean by creating an object with various numbers and then simply use the `median` function.

```
b <- c(3.8,4,3.1,2.1,12.6,17,8.43,11,2,3,9,5,3,0.5)

median(b)
#> [1] 3.9
```

Many people have heard of standard deviation, but not many can provide an accurate definition. Standard deviation is the measure of variance (see next paragraph) between numbers in a data set and that data set's mean. Typically a lower standard deviation is what you want to see as this proves there is little variance in your data set and that typically means a higher correlation. If you wanted to find this by hand you would have to find the square root of the variance between each point and the mean. This is easy if you have all the numbers you need, but that is almost never the case and thus you would have to calculate all that variance by hand and that is too much work. Why take all those extra steps when R can do it for you? The `sd()` function does all the work without any hassle, all you have to do is identify the object you want it to find the standard deviation of.

```
d <- c(3.8,4,3.1,2.1,12.6,17,8.43,11,2,3,9,5,3,0.5)

sd(d)
#> [1] 4.821066
```

Finally, we can cover variance. Variance is the measure of distance between two numbers. This statistic is often used to measure how spread out your data is and

can be used to determine your model's accuracy. Typically, a higher variance means your model is inaccurate. So, when building models and reviewing your summary statistics, you want your variance to be as low as possible. There is a var() function that will find the variance of a variable for you, however typically you will create a confusion matrix (discussed in a later section) that will help you determine the variance, and thus accuracy of your model.

9.3.2 Anscombe's Quartet

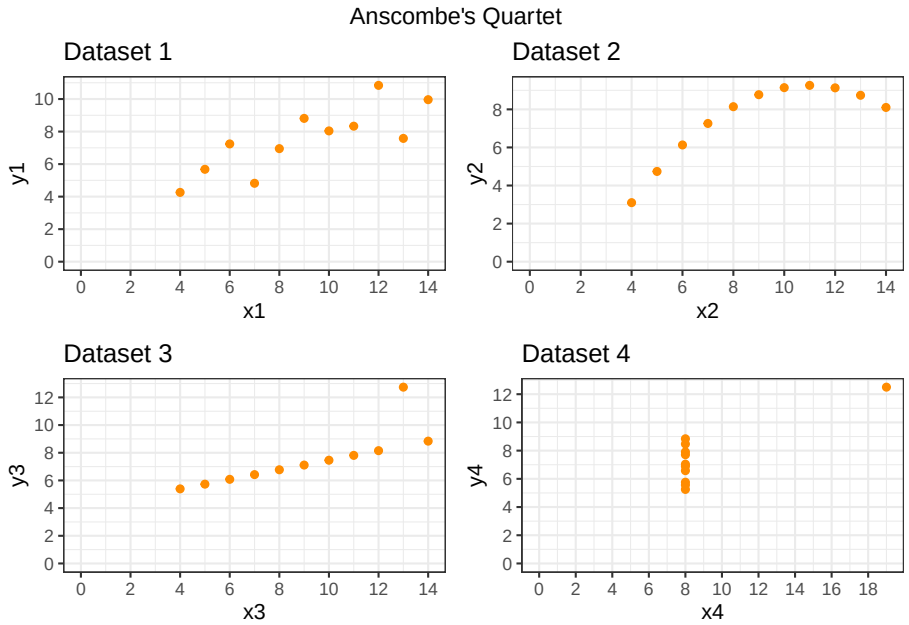
You may be asking now, why bother visualizing data if I have all these numbers that will tell me what I need to know about my data? You are correct in saying that statistical data is very useful, and in many cases essential, however visualization is equally as important. Anscombe's Quartet is a statistical phenomenon where four sets of data have very similar statistical properties, but very are completely different when graphed. Let's take a look at the data:

```
#>    x1 x2 x3 x4    y1   y2    y3    y4
#> 1  10 10 10  8  8.04 9.14  7.46  6.58
#> 2   8  8  8  8  6.95 8.14  6.77  5.76
#> 3  13 13 13  8  7.58 8.74 12.74  7.71
#> 4   9  9  9  8  8.81 8.77  7.11  8.84
#> 5  11 11 11  8  8.33 9.26  7.81  8.47
#> 6  14 14 14  8  9.96 8.10  8.84  7.04
#> 7   6  6  6  8  7.24 6.13  6.08  5.25
#> 8   4  4  4 19  4.26 3.10  5.39 12.50
#> 9  12 12 12  8 10.84 9.13  8.15  5.56
#> 10  7  7  7  8  4.82 7.26  6.42  7.91
#> 11  5  5  5  8  5.68 4.74  5.73  6.89
```

```
#>                    x1        x2        x3        x4
#> nobs        11.000000 11.000000 11.000000 11.000000
#> NAs          0.000000  0.000000  0.000000  0.000000
#> Minimum      4.000000  4.000000  4.000000  8.000000
#> Maximum     14.000000 14.000000 14.000000 19.000000
#> 1. Quartile  6.500000  6.500000  6.500000  8.000000
#> 3. Quartile 11.500000 11.500000 11.500000  8.000000
#> Mean         9.000000  9.000000  9.000000  9.000000
#> Median       9.000000  9.000000  9.000000  8.000000
#> Sum         99.000000 99.000000 99.000000 99.000000
#> SE Mean      1.000000  1.000000  1.000000  1.000000
#> LCL Mean     6.771861  6.771861  6.771861  6.771861
#> UCL Mean    11.228139 11.228139 11.228139 11.228139
#> Variance    11.000000 11.000000 11.000000 11.000000
#> Stdev        3.316625  3.316625  3.316625  3.316625
```

```
#> Skewness      0.000000  0.000000  0.000000  2.466911
#> Kurtosis     -1.528926 -1.528926 -1.528926  4.520661
#>                     y1        y2        y3        y4
#> nobs         11.000000 11.000000 11.000000 11.000000
#> NAs           0.000000  0.000000  0.000000  0.000000
#> Minimum       4.260000  3.100000  5.390000  5.250000
#> Maximum      10.840000  9.260000 12.740000 12.500000
#> 1. Quartile   6.315000  6.695000  6.250000  6.170000
#> 3. Quartile   8.570000  8.950000  7.980000  8.190000
#> Mean          7.500909  7.500909  7.500000  7.500909
#> Median        7.580000  8.140000  7.110000  7.040000
#> Sum          82.510000 82.510000 82.500000 82.510000
#> SE Mean       0.612541  0.612568  0.612196  0.612242
#> LCL Mean      6.136083  6.136024  6.135943  6.136748
#> UCL Mean      8.865735  8.865795  8.864057  8.865070
#> Variance      4.127269  4.127629  4.122620  4.123249
#> Stdev         2.031568  2.031657  2.030424  2.030579
#> Skewness     -0.048374 -0.978693  1.380120  1.120774
#> Kurtosis     -1.199123 -0.514319  1.240044  0.628751
```

As you can see from the table, the summary statistics for the four tables look very similar, however let's graph these and see what kind of visualizations we get.



Pretty rad right? So yes, statistical data is vital, however it should not be the only thing you look at, sometimes a visualization can help you understand the

data more! Later in this book, and in the upper level data analytics courses, we will cover visualizations more and how they can be useful.

9.4 Exercises

1. Define

```
x<-c(4,2,6)
y<-c(1,0,-1)
```

2. Decide what the result will be of the following:
   - length(x)
   - sum(x)
   - sum(x^2)
   - x+y
   - x*y
   - x-2
   - x^2
3. Use R to check your answers.
4. Decide what the following sequences are and use R to check your answers:
   - 7:11
   - seq(2,9)
   - seq(4,10,by=2)
   - seq(3,30,length=10)
   - seq(6,-4,by=-2)
5. Determine what the result will be of the following R expressions, and then use R to check if you are right:
   - rep(2,4)
   - rep(c(1,2),4)
   - rep(c(1,2),c(4,4))
   - rep(1:4,4)
   - rep(1:4,rep(3,4))
6. Use the rep function to define simply the following vectors in R.
   - 6,6,6,6,6,6
   - 5,8,5,8,5,8,5,8
   - 5,5,5,5,8,8,8,8

Chapter 10

Summaries and Subscripting

10.1 Introduction

Let's suppose we've collected some data from an experiment and stored them in an object x:

```
x<-c(7.5,8.2,3.1,5.6,8.2,9.3,6.5,7.0,9.3,1.2,14.5,6.2)
```

Some simple summary statistics of these data can be produced:

```
mean(x)
#> [1] 7.216667
var(x)
#> [1] 11.00879
sd(x)
#> [1] 3.317949
summary(x)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   1.200   6.050   7.250   7.217   8.475  14.500
```

which should all be self explanatory.

It may be, however, that we subsequently learn that the first 6 data points correspond to measurements made on one machine, and the second six on another machine.

This might lead us to want to summarize the two sets of data separately, so we would need to extract from x the two relevant subvectors. This is achieved by subscripting:

```
x[1:6]
#> [1] 7.5 8.2 3.1 5.6 8.2 9.3
x[7:12]
#> [1]  6.5  7.0  9.3  1.2 14.5  6.2
summary(x[1:6])
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   3.100   6.075   7.850   6.983   8.200   9.300
summary(x[7:12])
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   1.200   6.275   6.750   7.450   8.725  14.500
```

Other subsets can be created in the obvious way. For example:

```
x[c(2,4,9)]
#> [1] 8.2 5.6 9.3
```

Negative integers can be used to exclude particular elements. For example

```
x[-(1:6)]
#> [1]  6.5  7.0  9.3  1.2 14.5  6.2
```

has the same effect as

```
x[7:12]
#> [1]  6.5  7.0  9.3  1.2 14.5  6.2
```

10.2 Exercises (Summaries and Subscripting)

1. If x<- c(5,9,2,3,4,6,7,0,8,12,2,9) decide what each of the following is and use R to check your answers:

(a) x[2]
(b) x[2:4]
(c) x[c(2,3,6)]
(d) x[c(1:5,10:12)]
(e) x[-(10:12)]

2. The data y<-c(33,44,29,16,25,45,33,19,54,22,21,49,11,24,56) contain sales of milk in gallons for 5 days in three different shops (the frst 3 values are for shops 1,2 and 3 on Monday, etc.) Produce a statistical summary of the sales for each day of the week and also for each shop.

Chapter 11

Matrices

11.1 CBind and RBind

Matrices can be created in R in a variety of ways. Perhaps the simplest is to create the columns (just a couple of objects) and then glue them together with the command cbind. For example,

```
x<-c(5,7,9)
y<-c(6,3,4)
z<-cbind(x,y)
z
#>      x y
#> [1,] 5 6
#> [2,] 7 3
#> [3,] 9 4
```

The dimension of a matrix can be checked with the dim command:

```
dim(z)
#> [1] 3 2
```

[1] 3 2 i.e., three rows and two columns. There is a similar command, rbind, for building matrices by gluing rows together.

The functions cbind and rbind can also be applied to matrices themselves (provided the dimensions match) to form larger matrices. For example,

```
rbind(z,z)
#>      x y
```

```
#> [1,] 5 6
#> [2,] 7 3
#> [3,] 9 4
#> [4,] 5 6
#> [5,] 7 3
#> [6,] 9 4
```

11.1.1 Review Questions

1) Create a matrix made up of two columns showing the GPAs and number of hours studied by seven students.

2) Recreate the following matrix in R:

```
#>      x   y
#> [1,] 5 3.4
#> [2,] 7 4.0
#> [3,] 2 2.5
#> [4,] 3 3.2
#> [5,] 8 2.8
#> [6,] 4 3.1
#> [7,] 2 3.6
```

3) Using the appropriate function, combine the two matrices you created above.

11.2 Matrix Function

Matrices can also be built by explicit construction via the function matrix. For example,

```
z<-matrix(c(5,7,9,6,3,4),nrow=3)
```

results in a matrix z identical to z above. Notice that the dimension of the matrix is determined by the size of the vector and the requirement that the number of rows is 3, as specified by the argument nrow=3. As an alternative we could have specified the number of columns with the argument ncol=2 (obviously, it is unnecessary to give both). Notice that the matrix is 'flled up' column-wise. If instead you wish to fill up row-wise, add the option byrow=T. For example,

```
z<-matrix(c(5,7,9,6,3,4),nr=3,byrow=T)
z
#>      [,1] [,2]
#> [1,]    5    7
#> [2,]    9    6
#> [3,]    3    4
```

Notice that the argument nrow has been abbreviated to nr. Such abbreviations are always possible for function arguments provided it induces no ambiguity - if in doubt always use the full argument name.

As usual, R will try to interpret operations on matrices in a natural way. For example, with z as above, and

```
y<-matrix(c(1,3,0,9,5,-1),nrow=3,byrow=T)
y
#>      [,1] [,2]
#> [1,]    1    3
#> [2,]    0    9
#> [3,]    5   -1
```

we obtain

```
y+z
#>      [,1] [,2]
#> [1,]    6   10
#> [2,]    9   15
#> [3,]    8    3
```

and

```
y*z
#>      [,1] [,2]
#> [1,]    5   21
#> [2,]    0   54
#> [3,]   15   -4
```

Other useful functions on matrices are to transpose a matrix:

```
z
#>      [,1] [,2]
#> [1,]    5    7
#> [2,]    9    6
#> [3,]    3    4
```

```
t(z)
#>      [,1] [,2] [,3]
#> [1,]    5    9    3
#> [2,]    7    6    4
```

As with vectors it is useful to be able to extract sub-components of matrices. In this case, wemay wish to pick out individual elements, rows or columns. As before, the [ ] notation is used to subscript. The following examples should make things clear:

```
z[1,1]
#> [1] 5
```

```
z[c(2,3),2]
#> [1] 6 4
```

```
z[,2]
#> [1] 7 6 4
```

```
z[1:2,]
#>      [,1] [,2]
#> [1,]    5    7
#> [2,]    9    6
```

So, in particular, it is necessary to specify which rows and columns are required, whilst omitting the integer for either dimension implies that every element in that dimension is selected.

11.3 Exercises

1. Create this matrix in R

```
#>      [,1] [,2] [,3] [,4] [,5]
#> [1,]    1    7    8   11   -5
#> [2,]    3    8    6    3   -9
#> [3,]    0   11   14   14   14
```

2. Create in R these matrices:

```
x
#>      [,1] [,2]
#> [1,]    1    7
#> [2,]    8   11
#> [3,]    5    9
y
#>      [,1] [,2]
#> [1,]    6    8
#> [2,]    2    1
#> [3,]    1   -7
```

3. Calculate the following and check your answers in R:

(a) 2*x
(b) x*x
(c) t(y)

```
#>      [,1] [,2]
#> [1,]    2   14
#> [2,]   16   22
#> [3,]   10   18
#>      [,1] [,2]
#> [1,]    1   49
#> [2,]   64  121
#> [3,]   25   81
#>      [,1] [,2] [,3]
#> [1,]    6    2    1
#> [2,]    8    1   -7
```

4. With x and y as above, calculate the effect of the following subscript operations and check your answers in R.

(a) x[1,]
(b) x[2,]
(c) x[,2]
(d) y[1,2]
(e) y[,2:3]

Chapter 12

Preloaded data and `mtcars`

R comes with several built-in data sets, **which are generally used as demo data for playing with R functions.**

To see the datasets type:

```
data()
```

12.1 Practicing with `mtcars` data set

This demonstration is based on the datasset `mtcars`.

1. Read in `mtcars`

```
data(mtcars)
```

2. View first few rows and last few rows of mtcars dataframe using functions `head()` and `tail()`

```
head(mtcars)
#>                   mpg cyl disp  hp drat    wt  qsec vs am
#> Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1
#> Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1
#> Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1
#> Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0
#> Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0
#> Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0
#>                   gear carb
#> Mazda RX4            4    4
```

```
#> Mazda RX4 Wag        4    4
#> Datsun 710           4    1
#> Hornet 4 Drive       3    1
#> Hornet Sportabout    3    2
#> Valiant              3    1
tail(mtcars)
#>                  mpg cyl  disp  hp drat    wt qsec vs am
#> Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1
#> Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1
#> Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1
#> Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1
#> Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1
#> Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1
#>                gear carb
#> Porsche 914-2     5    2
#> Lotus Europa      5    2
#> Ford Pantera L    5    4
#> Ferrari Dino      5    6
#> Maserati Bora     5    8
#> Volvo 142E        4    2
```

3. Some info about mtcars dataframe using function `colnames()`, `rownames()`, `summary()`and `dim()`

```
colnames(mtcars)
#>  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"
#>  [9] "am"   "gear" "carb"
rownames(mtcars)
#>  [1] "Mazda RX4"           "Mazda RX4 Wag"
#>  [3] "Datsun 710"          "Hornet 4 Drive"
#>  [5] "Hornet Sportabout"   "Valiant"
#>  [7] "Duster 360"          "Merc 240D"
#>  [9] "Merc 230"            "Merc 280"
#> [11] "Merc 280C"           "Merc 450SE"
#> [13] "Merc 450SL"          "Merc 450SLC"
#> [15] "Cadillac Fleetwood"  "Lincoln Continental"
#> [17] "Chrysler Imperial"   "Fiat 128"
#> [19] "Honda Civic"         "Toyota Corolla"
#> [21] "Toyota Corona"       "Dodge Challenger"
#> [23] "AMC Javelin"         "Camaro Z28"
#> [25] "Pontiac Firebird"    "Fiat X1-9"
#> [27] "Porsche 914-2"       "Lotus Europa"
#> [29] "Ford Pantera L"      "Ferrari Dino"
#> [31] "Maserati Bora"       "Volvo 142E"
summary(mtcars)
```

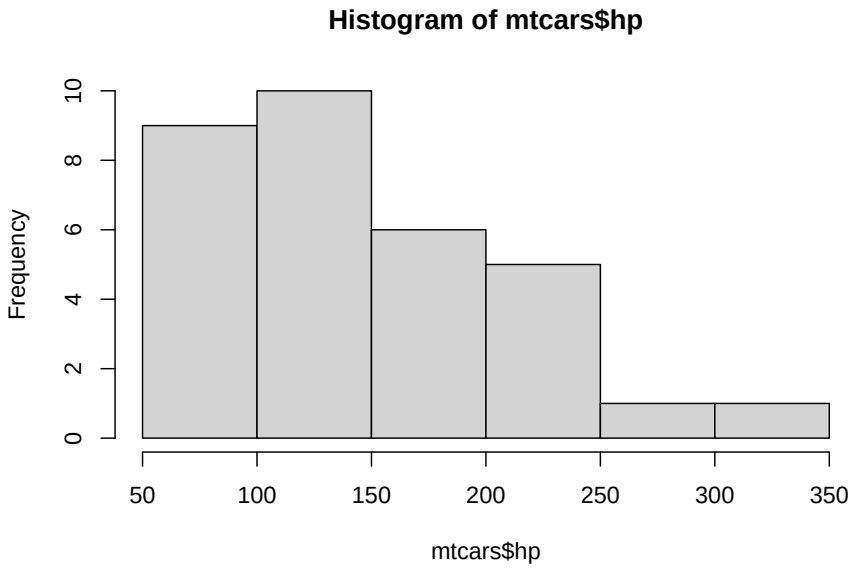
```
#>       mpg             cyl             disp
#>  Min.   :10.40   Min.   :4.000   Min.   : 71.1
#>  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8
#>  Median :19.20   Median :6.000   Median :196.3
#>  Mean   :20.09   Mean   :6.188   Mean   :230.7
#>  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0
#>  Max.   :33.90   Max.   :8.000   Max.   :472.0
#>        hp             drat             wt
#>  Min.   : 52.0   Min.   :2.760   Min.   :1.513
#>  1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581
#>  Median :123.0   Median :3.695   Median :3.325
#>  Mean   :146.7   Mean   :3.597   Mean   :3.217
#>  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610
#>  Max.   :335.0   Max.   :4.930   Max.   :5.424
#>       qsec             vs              am
#>  Min.   :14.50   Min.   :0.0000   Min.   :0.0000
#>  1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000
#>  Median :17.71   Median :0.0000   Median :0.0000
#>  Mean   :17.85   Mean   :0.4375   Mean   :0.4062
#>  3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000
#>  Max.   :22.90   Max.   :1.0000   Max.   :1.0000
#>       gear            carb
#>  Min.   :3.000   Min.   :1.000
#>  1st Qu.:3.000   1st Qu.:2.000
#>  Median :4.000   Median :2.000
#>  Mean   :3.688   Mean   :2.812
#>  3rd Qu.:4.000   3rd Qu.:4.000
#>  Max.   :5.000   Max.   :8.000
dim(mtcars)
#> [1] 32 11
```

4. To calculate the variance of weight:

```
var(mtcars$wt)
#> [1] 0.957379
```

5. To get the histogram of hp, the code below will produce a histogram:

```
hist(mtcars$hp)
```



6. To calculate the quantiles by percent:

```
quantile(mtcars$wt, c(.2, .4, .8))
#>   20%   40%   80%
#> 2.349 3.158 3.770
```

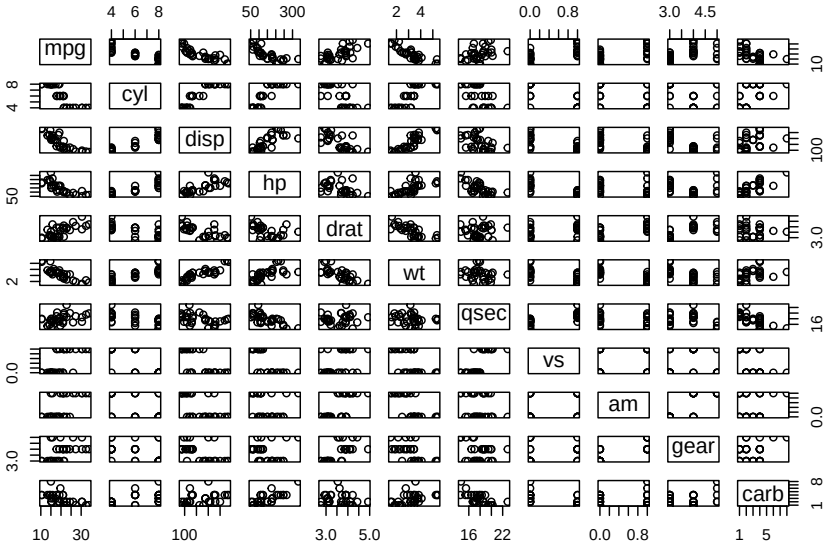
12.2 Excerises for you:

1. Find the minimum and maximum value of `mpg`

2. Find the mean and standard deviation of data variable `mpg`

3. What variable has a 3rd quartile value of 180.0?

4. Create and explain what this means



5. Create and explain what this means

```
#>              mpg        cyl       disp         hp
#> mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684
#> cyl  -0.8521620  1.0000000  0.9020329  0.8324475
#> disp -0.8475514  0.9020329  1.0000000  0.7909486
#> hp   -0.7761684  0.8324475  0.7909486  1.0000000
#> drat  0.6811719 -0.6999381 -0.7102139 -0.4487591
#> wt   -0.8676594  0.7824958  0.8879799  0.6587479
#> qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234
#> vs    0.6640389 -0.8108118 -0.7104159 -0.7230967
#> am    0.5998324 -0.5226070 -0.5912270 -0.2432043
#> gear  0.4802848 -0.4926866 -0.5555692 -0.1257043
#> carb -0.5509251  0.5269883  0.3949769  0.7498125
#>              drat         wt        qsec         vs
#> mpg   0.68117191 -0.8676594  0.41868403  0.6640389
#> cyl  -0.69993811  0.7824958 -0.59124207 -0.8108118
#> disp -0.71021393  0.8879799 -0.43369788 -0.7104159
#> hp   -0.44875912  0.6587479 -0.70822339 -0.7230967
#> drat  1.00000000 -0.7124406  0.09120476  0.4402785
#> wt   -0.71244065  1.0000000 -0.17471588 -0.5549157
#> qsec  0.09120476 -0.1747159  1.00000000  0.7445354
#> vs    0.44027846 -0.5549157  0.74453544  1.0000000
#> am    0.71271113 -0.6924953 -0.22986086  0.1683451
```

```
#> gear  0.69961013 -0.5832870 -0.21268223  0.2060233
#> carb -0.09078980  0.4276059 -0.65624923 -0.5696071
#>               am       gear        carb
#> mpg   0.59983243  0.4802848 -0.55092507
#> cyl  -0.52260705 -0.4926866  0.52698829
#> disp -0.59122704 -0.5555692  0.39497686
#> hp   -0.24320426 -0.1257043  0.74981247
#> drat  0.71271113  0.6996101 -0.09078980
#> wt   -0.69249526 -0.5832870  0.42760594
#> qsec -0.22986086 -0.2126822 -0.65624923
#> vs    0.16834512  0.2060233 -0.56960714
#> am    1.00000000  0.7940588  0.05753435
#> gear  0.79405876  1.0000000  0.27407284
#> carb  0.05753435  0.2740728  1.00000000
```

6. Create a variable called `efficiency` which is mpg divided by weight. Which car has the max `efficiency` and what is this value?
7. Which variable in this dataset has the greatest standard deviation?
8. How many cars have 3 gears?
9. How many cars get more than 17 mpg?

Chapter 13

More simple data wrangling

13.1 a nice, fun little matrix for you

```
#>      [,1] [,2] [,3] [,4] [,5]
#> [1,]    1    4    7   10   13
#> [2,]    2    5    8   11   14
#> [3,]    3    6    9   12   15
```

1. Write the code that creates this matrix:
2. Write DIFFERENT code that creates this matrix in an alternate way:
3. In the matrix above, what does `[,4]` mean?
4. What code would return the value in the 3rd column and 3rd row?
5. What single line of would give you the average of the all the numbers in columns 2, 4, and 5 and in rows 1 and 3?
6. turn `x` into a data frame.
7. How do you **know** you have turned `x` into a data frame?

13.2 More fun (this class is really awesome isn't it?)

```
df
#>    X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
#> 1   1 11 21 31 41 51 61 71 81  91
#> 2   2 12 22 32 42 52 62 72 82  92
#> 3   3 13 23 33 43 53 63 73 83  93
#> 4   4 14 24 34 44 54 64 74 84  94
#> 5   5 15 25 35 45 55 65 75 85  95
#> 6   6 16 26 36 46 56 66 76 86  96
#> 7   7 17 27 37 47 57 67 77 87  97
#> 8   8 18 28 38 48 58 68 78 88  98
#> 9   9 19 29 39 49 59 69 79 89  99
#> 10 10 20 30 40 50 60 70 80 90 100
```

1. Consider the dataframe above called `df`. What would running this code return `sum(df[7,7:10])`
2. How can you tell if an object in R is a dataframe?
3. How could you create the dataframe above called `df`?
4. What code would return the average of row 2 of `df`?
5. Consider `mtcars` dataset that comes preloaded with R that looks like this:

```
head(mtcars)
#>                    mpg cyl disp  hp drat    wt  qsec vs am
#> Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1
#> Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1
#> Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1
#> Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0
#> Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0
#> Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0
#>                   gear carb
#> Mazda RX4            4    4
#> Mazda RX4 Wag        4    4
#> Datsun 710           4    1
#> Hornet 4 Drive       3    1
#> Hornet Sportabout    3    2
#> Valiant              3    1
```

6. Why do I get this error when I run the code below: `Error in plot(hp, mpg) : object 'hp' not found`?

```
plot(hp,mpg)
```

```
Error in plot(hp, mpg) : object 'hp' not found
```

Bonus: What is a topic that you find confusing at this point in class?

Chapter 14

GDH Ice Cream

14.1 Problem Introduction

GDH provides ice cream for its wonderful customers. I LOVE GDH. Do you love it as much as me (let's discuss)?

In the last three years GDH used ice cream, in pounds, by month, as shown in the attached file.

```
#>  Month.Name year1 year2 year3
#>         Jan    60    67    64
#>         Feb    68    67    72
#>         Mar    83    62    61
#>         Apr   102    95   107
#>         May    95   105   101
#>         Jun    57    89    75
#>         Jul    61    57    81
#>         Aug   109   109   104
#>         Sep    56    86    88
#>         Oct    53    53    65
#>         Nov    74    72    72
#>         Dec    73    64    65
```

14.2 Assignment

Please answer the following questions *using R*:.

GDH provides ice cream cones for its customers. In the last three years GDH used ice cream, in pounds, by month, as shown in the attached file.

1. In R, create the above data frame and name it ice.cream
2. What is another way you could have created the same data set?
3. Using R, what is the mean using for the months of Feb and Oct?
4. Create a chart showing ice cream use over time.
5. Which year used the most ice cream?
6. Which month has the highest standard deviation of ice cream use?
7. Which year has the highest standard deviation of ice cream use?
8. Also, you May want to check out this link to look at something called dataframes that may help with this assignment (but is not absolutely necessary) https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/data.frame
9. Can you transpose your matrix?
10. Can you add meaningful row names and column names?