Exploring Data Science Education: From Tutorials to Assessment

Duke Statistical Science | Graduation with Distinction

Evan Dragich supervised by Dr. Mine Çetinkaya-Rundel, PhD.

February 27, 2023

Table of contents

Abstract

Download PDF here

Abstract:

As data science (DS) continues to grow in popularity among university course offerings, it is becoming crucial to successfully measure students' learning outcomes in introductory courses. To do this requires an assessment designed to which could additionally be used to evaluate pedagogical techniques or curriculum interventions in data science curriculum

To develop a blueprint for the assessment, a multi-institutional team of statistics and data science education researchers identified common DS content (e.g., data wrangling, interpreting visualizations), drawing from published guidelines/recommendations and introductory DS syllabi. A draft of the assessment was written and used to conduct three think-aloud interviews with field-relevant faculty members.The interviews consisted of both open-ended brainstorming on the assessment's scope as well as individual examinations of each item for relevance, clarity, and efficacy in measuring the desired learning objective. Think-aloud interviews were also conducted with introductory DS students to gauge item clarity and gain insight into the reasoning for their responses.

This poster includes the blueprint developed, as well as example items, and results from the faculty and student think aloud interviews. We also present next steps for the project including plans for larger scale piloting and further analyses.

Goals

By sharing the work, we hope that participants will become familiar with an assessment they may use for designing intro data science curriculum or researching classroom innovations. We also hope that this instrument can serve as an inspiration or a starting point to be tailored by future researchers more specifically to their courses or to another discipline (eg. by adding more programming concepts instead of data visualization to better serve a computing-focused introductory data science class, etc.)

Introduction

next section to write after development: just like about the motivation for the project

Assessment Background

Background

An essential component of any educational research is a validated, relevant instrument to measure students' learning outcomes. Whether to award college credit like the College Board's AP and CLEP exams, to measure students' previously-held misconceptions, or just as a tool upon which to evaluate educational interventions, such assessments have been developed and statistically analyzed for a wide range of subjects such as Spanish, Psychology, Chemistry, and Calculus (Godfrey and Jagesic 2016; Solomon et al. 2021; Mulford 1996; Epstein 2013).

In the field of statistics, previous work on measuring students' reasoning skills led to the development of the Comprehensive Assessment of Outcomes in Statistics (CAOS). The revised CAOS 4, comprising 40 multiple-choice items on a variety of commonly-taught first-semester introductory concepts, was first administered in 2005 and allowed instructors to measure whether their courses were successfully resulting in their desired learning outcomes. However, many instructors noted that their findings reflected a much lower understanding than expected, specifically regarding the topics of data visualization and data collection (Delmas et al. 2007). However, a key feature of the CAOS was the lack of hard computation nor need to recall specific formulas or definitions, allowing greater accessibility for a variety of statistics-adjacent uses.

In 2022, as more universities begin to support the emerging field of data science via specific courses, concentrations, or even majors, there is a need to measure students' learning outcomes in these introductory classes analogously to the subjects named above (Swanstrom, n.d.). Lacking a clearly-defined scope, empirical studies of so called "data science" curricula suggest that the field can be thought of an augmentation of traditional statistical modeling concepts, with emphases on computing, data visualization and manipulation, as well as a consideration of ethics and the role data plays in society (Zhang and Zhang 2021).

good places to start: read DelMas 2007 paper, look at what theyve done to get it beginnning to end, at some point acknowledge that were not finishing with just my theiss (analyzing student data, subscale stuff that wont be part of my thesis)

refer to this for background: https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1804497 in on eo fthe earlier sections, summary of data science courses and what they cover (and compare and contrast

to our assessment). developed collectively based on the faculty and what they teach/what ive taken, but also capturing most of what's "out there"

are there any papers on: teaching data science with or without code, kind of justify our choice for language agnosticism

Just found this paper and thought you all would find it interesting: Emmanuel Schanzer, Nancy Pfenning, Flannery Denny, Sam Dooman, Joe Gibbs Politz, Benjamin S. Lerner, Kathi Fisler, and Shriram Krishnamurthi. 2022. Integrated Data Science for Secondary Schools: Design and Assessment of a Curriculum. In Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2022). Association for Computing Machinery, New York, NY, USA, 22–28. https://doi.org/10.1145/3478431.3499311 add one sentence on like: its in high schools too!

ask andy/chelsey–do they have other good references how to run think aloud interviews

Assessment Development

Phase 0: Initial Cleaning and Feedback

In January 2022, I inherited a repo with several documents: many were background information on what topics would be included, with one containing all currently-written questions. These questions were not yet organized into specific groups of stem and items, many sections were commented out or overwritten, and it was clearly something that had been written piecewise by a group of people. My first true task was to run through the current questions myself, answer them how I would, and provide feedback on clarity, wording, and reflect on the topic covered.

From this initial feedback, I created three pull requests attempting to solve some of these issues. These first fixes were minorly substantial edits: changing "most" to "the majority of" to make a question less ambiguous, adding a "fill-in-the-blank" slot to make even more clear what the question is asking, and background information on confidence intervals to an item that referenced them. This was my first exposure to what I would call "advanced" GitHub usage, or more than just typical cloning, pushing, and pulling from ordinary group projects. I used the very helpful `pr_*()` series of functions from the `usethis` package, which allowed me to easily create the pull requests for review.

While the team discussed my pull request proposals, I worked in parallel to clean up any "obvious" fixes to all items. These "no-brainer" edits were mainly typos, distorted or obscured plots and items that don't actually pose a question. It was also time to clean up the current document by converting it to a Quarto book, which would allow for easy webpage-like navigation. This marked my first exposure to Quarto, and I quickly grew to love its improvements to the RMarkdown workflow, like ease of rendering to different formats and intuitive structure of the index and YAML files. This was also when I received my first moment of creative liberty with the project, as I was tasked with dividing the ungrouped set of questions into discrete passages consisting of a stem and one or more corresponding items. Each of these passages–26 at the time–was given a title to identify it in the Quarto book sidebar.

At this point, the team came to a verdict on my initial pull request proposals: they approved and merged the slight text modification to "the majority" and the "fill-in-the-blank" title, but rejected the confidence interval language, asserting that the stem already provided sufficient motivation. The final step before presenting the assessment externally was to improve reproducibility, as many figures were not being rendered with each update, but rather embedded as images. In a throwback to my former STA 313 days, I used the current images as guides

to recreate a map of the US colored by region, movie-themed data tables, and pseudocode chunks. I had some slight HTML and CSS exposure in previous classes, but this new challenge of matching an existing format allowed me to expand my knowledge about web layout and styles.

By April 2022, concluding my first semester working on the project, we had a polished, reproducible, website-hosted prototype of the assessment ready to present externally and gather feedback.

Phase 1: Faculty Interviews

The remainder of the time spent working on the assessment–April 2022 to February 2023–was spent gathering feedback via interviews, iteratively updating items, and removing weaker questions based on group discussions. We conducted a series of three think-aloud informational interviews with various faculty nationwide who teach or have taught introductory data science courses. Participants were recruited from a shortlist of Dr. Çetinkaya-Rundel's data science assessment contacts, and were specifically chosen to represent a breadth of data science curricula. The three interviewees chosen were, in order of interview:

Name	Department Affiliation	Institution Type	Field of Ph.D. Dissertation
Professor X	Statistics	Liberal arts college	Biostatistics
Professor Y	Computer science	R1 research university	Computer science education
Professor Z	Computer and information science	Liberal arts school within university	Statistics

Each interview was scheduled for two hours long and consisted of three sections: open-ended introductory and concluding discussions, sandwiching an item-by-item run through of the assessment. In the initial discussion, we asked participants two big picture questions: What topics *must* be in an introductory data science course? And, what topics are *nice to have* in an introductory data science course? For each item in the assessment run-through, we asked interviewees to narrate out loud their thinking process from start to finish: any initial reactions, their process to arrive at an answer, and what said answer would be. We then asked for additional comments or suggestions, ranging from small- (formatting changes) to large-scale (removing the question entirely). We then concluded with three big-picture questions: What are the strengths of the current assessment? What topics are missing from the current assessment? And, what is in the current assessment, but doesn't belong?

While I scheduled and directed the flow of the interviews myself, I was joined by Dr. Çetinkaya-

Rundel for all three, Dr. Legacy for the second two, and Dr. Beckman for the third. Silently observing, these team members took notes in real-time while I was conversing with the interviewee as a supplement to the transcript.

Discipline-Specific Perspectives

After conducting each of three interviews, we met to discuss the new feedback and, when appropriate, made modifications and deletions to the current assessment. While each interviewee came from distinct backgrounds, there were some salient themes about their perspectives on introductory data science that emerged through patterns in their responses.

Professor X seemed to have a somewhat similar perspective on "what is data science" as the research team–chiefly, visualization and wrangling. He noted positively when items featured multivariate data, particularly when interpreting visualizations. He also notably brought a wealth of pedagogical experience with real-life data to our consideration: such as that movie budgets and revenue display less compelling of a relationship than one would think, that movie-themed data at all is less relevant to younger generations, and that county data is generally too heterogenous to be useful in most contexts. He also grounded our conceptions of what a pre-test student would know by claiming that residuals are now part of the Common Core in K-12 education. However, he thought residual analysis and other related modeling topics like supervised learning didn't align with his experience teaching introductory data science. But, as in the case of Realty Tree, he acknowledged that some topics may fall outside the scope of an introductory class, but could be reasoned through and provide motivation to learn more.

Professor Z held similar opinions on the scope of data science as did Professor X–visualization and wrangling. The main takeaway from her feedback was that we needed to consider carefully the scenarios posed in question stems. Hurricane data, like in Hurricane Andy, she claimed, may provoke discomfort in students from hurricane-prone areas. She noted that the former wording of He Said She Said, in which the verbs on the plot itself were in the past tense while those in the items were in the present, may present a burden to non-native English speakers. The SAT, referenced in Banana Conclusions, may need more explanation for international students who are less familiar with the US admissions process. She also, like Professor X, thought that some of our more technical questions breached the scope of introductory data science, such as residual analysis and the distinction between training and validation sets. She was similarly a fan of Realty Tree. However, while Professor X was in favor of summary statistic questions and mapping them to graphs, Professor Z called out language like "margin of error" as too statistical, and not as data science related.

As a computer science education representative, we knew Professor Y's interview would offer a unique perspective. This was evident to the start, when her first response to my question of "what is essential" was to ask me whether this course assumed prior coding experience or not. After I clarified that we are assuming no coding experience, she answered that learning coding "in the order that matters for data science", e.g. functions first, is a primary topic,

as is some unerstanding of transforming and cleaning data. But, she qualified, not too much as much cleaning can be done for them in an introductory class. It wasn't until the "what is nice to have" question that she mentioned data visualization. This CS-focused perspective continued when she pointed out our pseudocode didn't specify which type of join would be performed, that basic English vocabulary like "filter" and "select" might not be known in their data wrangling context, and generally dissuaded us from using pseudocode on something that would be a pre-test. Another notable pattern was that some of her suggested edits went against data visualization best practices: she suggested we change from a rainbow color scheme to a gradient one for categorical data, and to reorder the y-axis in He Said She Said to be alphabetical rather than allowing for easy comparison. Finally, she led us to remove the logarithmic transformation present in all Movie Budgets questions, explaining that since the transformation itself isn't a learning outcome, students seeing those words continually repeated may start losing sight of the item's objective.

Regrouping to synthesize feedback

While the team met once bewteen each faculty interview to make any minor changes before the next one, the bulk of the revision took place after all were completed, in the Fall 2022 semester. A common theme from all interviews was that our questions designed to test tricky, nuanced concepts (e.g. reading in data to statistical software) in a multiple choice format weren't landing as we hoped. While we thought we had written a former item, **?@sec-data-cleaning**, with enough language agnosticity, it ended up being too R-specific and baffling all faculty. To this item, and several others, Professor Z summed up the crux of our dilemma well, paraphrasing: "hmm.. I see what concept you're getting at, but this question doesn't really get there… Okay, well, this is something that's hard to write to be autogradable but also actually measure. I don't know how I would fix this but I really like the underlying idea."

Encouraged to keep these concepts on the assessment with modifications, we found it difficult to cut out any of the ~50 pilot questions we had at this point. One notable large-scale fix was combining the concepts tested on three former wrangling passages into a single context, hoping to further reduce students' cognitive load. We knew we needed to cut items down to the ~30 range, but found it very difficult to identify those items that were bringing the least to the current assessment. Interestingly, this involved almost no large disagreements between team members–there were no particularly polarizing items, like we observed during the faculty interviews–rather, we all genuinely couldn't identify any items we wanted to cut.

We did eventually make rounds of sacrificial cuts, choosing to prioritize culling questions that even slightly overlapped and those that were most straightforward. We acknowledged the tradeoff that this may lead to students taking the assessment in a pre-test context unable to confidently answer a majority of the questions. However, we were constrained by the need for an instrument that could reasonably be given in an hour, and cutting out several entire passages was the only way forward. Nevertheless, there were a handful of similar items that we decided to keep for the student interviews, such as two pseudocode chunks (in Movie

Wrangling) that varied by a single statement. Here, we were ambivalent on which would be more effective, and hoped to let student feedback dictate whether one "stuck" better.

Phase 2: Student Interviews

The chief purpose of Phase 2 was to see if the topics covered would be in the wheelhouse of students with DS exposure, as well as continuing to refine wording and pacing. With the big picture questions around "what is data science" previously discussed, a similar series of interviews with Duke Statistical Science students allowed us to start drilling down and seeing if items landed the way we thought they would.

To reflect this different type of feedback being solicited, we omitted the initial "big picture" questions while interviewing TAs. Combined with the fact that that the assessment was now about half the original length, these interviews were only one hour long. The final portion of big picture questions was kept, though, with slightly modified prompts: Are the pacing and length appropriate? Based on what you remember learning in intro data science, what topics are missing from the current assessment? Based on what you remember learning in intro data science, what is in the current assessment, but doesn't belong?

We first reached out in November 2022 to all TAs at that time for STA 199 or STA 198, its health-themed analog. However, we were unable to recruit any students so close to finals, and ended up reaching out to the same group in January. The interviews were conducted in early February, thus consisting of students who were at least TAs for STA199 in the previous Fall semester. The three students recruited were, in order of interview:

Name	Degree Year and Level	Program
Student A	2nd year Masters	Statistical Science
Student B	4th year Undergraduate	Economics major, Statistical Science minor
Student C	1st year Masters	Statistical Science

In general, while still engaging well and demonstrating a strong command of DS skills, the students tended to be much less verbose in their feedback. Across all interviews, a former item on **?@sec-realty-tree** that asked students to trace a non-trivial regression tree four times stood out as an interruption to the otherwise smooth flow of the assessment. Also, on a more cosmetic note, all students suggested that plots which had used `theme_minimal()` be modified to more clearly show which titles corresponded to which faceted subplots–this stood out to me, as I am someone who also has taken many Duke Statistical Science courses, which are often `theme_minimal()` heavy.

Other issues revealed during interviews was the question order. We originally arranged items near others testing a similar topic to facilitate our iterative editing. However, there was a clear

case of cognitive priming in Student A's interview. Having just correctly answered **?@sec-image-recognition**, he then immediately jumped to examining ethical implications in the following item, **?@sec-application-screening**, while the faculty members who had answered the questions in a different order required more consideration. As well, Student B was the first interviewee to answer incorrectly to the item I think is the single trickiest (#3 in **?@sec-he-said-she-said**), and best written in terms of getting students to think critically about what exactly plots represent. This stood out to me, as she was the only undergraduate interviewed; we observed a stratification of data science comfortability even among TAs. Finally, Student C was the quickest interviewee to run through the questions, answering all correctly and citing the logic we were looking for. As the most relative newcomer to the Duke Statistical Science program and having graduated from an undergraduate data science program that used Python, we interpreted his notable ease in getting through all questions in much less than an hour as a good sign that the current length might be ideal for students, who would be less experienced with the material but would also not have to think aloud or give feedback.

Regrouping to synthesize feedback/final pilot assessment

Not many changes were made following the student interviews in Phase B: a sign that we were converging on a viable prototype for distribution. To remedy the priming issue encountered in Student A's interview, we shuffled the question order between the student B and C interviews and landed on a solid layout: distributing topics well throughout and gradually ramping up our intended difficulty to prevent less-knowledgeable students from getting frustrated and giving up early on. As part of this pacing reform, we improved the flow of the assessment by reformatting **?@sec-movie-budgets-2** and **?@sec-disease-screening** from groups of several nearly-identical items into a single matrices.

Final assessment themes

The following table summarizes the learning objectives for each item in the most current version of the assessment.

Passage	Learning Objective(s)
Storm Paths	simulated data; interpreting uncertainty
Movie Budgets 1	comparing summary statistics visually
Movie Budgets 2	R^2; comparing trends visually
Application Screening	ethics; proxy variable
Banana Conclusions	causation; statistical communication
COVID Map	interpreting complex visualization; spatial data; time series data
	interpreting complex visualization; sophisticated scales

Passage	Learning Objective(s)
He Said She Said	interpreting basic visualization
	interpreting basic visualization
	interpreting basic visualization; sophisticated scales
Build-a-Plot	data to visualization process
Disease Screening	comparing classification diagnostics visually
Realty Tree	modeling; regression tree
	modeling; regression tree; variable selection
Website Testing	interpret trends visually; visualize uncertainty; time series data
	interpret trends visually; visualize uncertainty; time series data; extrapolation
	interpret trends visually; visualize uncertainty; time series data; extrapolation
Image Recognition	ethics; representativeness of training data
Data Confidentiality	ethics; data deidentification
Activity Journal	structuring data; storing data
Movie Wrangling	data cleaning; column-wise operations; string operations
	data cleaning; column-wise operations; string operations
	data cleaning; extrapolation
	data wrangling; pseudocode, joins
	data wrangling; pseudocode, joins
	data wrangling; pseudocode, joins

Phase 3: Large-scale Student Pilot

Looking ahead, the next step after individual item tweaking and refinement will be real-world measurements of pacing and length and the feasibility for introductory data science students. Collaborating with the professors of Duke's Introductory Data Science Course STA199, ____ students across three sections will be invited to take the assessment, on their own time, for a chance to earn extra credit in the course. We will not solicit general feedback like in the previous interviews, instead asking just a single question at the end to get a rough estimate of how long it took to complete, since they will be doing it on their own time. The extra credit points students receive will not reflect their performance on the assessment, but rather will be simply awarded proportionally to how much was completed. To ensure students are actively engaging with the questions, there will be ____ attention checks scattered throughout that must also be passed to earn credit.

In order to distribute the assessment at large and record students' responses, the Quarto book has been converted into a Qualtrics survey. Once complete, large-scale analysis will be

conducted, both to ensure that all questions were indeed appropriate for introductory level students and to explore potential instrument subscales.

Package Development

overview (half my thesis is finishing `dsbox` package). background on what the package attempts to do "The goal of dsbox is to supplement the Data Science Course in a Box project. The package contains the datasets that are used in the materials in Data Science Course in a Box as well as the learnr tutorials." - https://rstudio-education.github.io/dsbox/index.html

explain the larger Data Science in a Box project, https://datasciencebox.org, basically condensing all the curriculum material here into 10 freestanding, interactive, automatically-graded `learnr` tutorials

Phase 1: Initial tidying work

–first time ever with pull requests and learnr, ie more than just basics in the RStudio window –as well as roxygen and rds files –change question wording to be more clear, and feedback in places to get graduaally more scaffolded –chnage links in codebooks, notice an exercise wasnt making sense and looked at codebook and saw levels of a variable was coded the wrong way

Phase 2: Creating a new tutorial

–there were nine tutorials when it existed –the only HW assignment from DS in a box that hadnt been converted yet had to do with Dennys and LaQuinta data, and skeleton documents for the data already existed in dsbox –thought critically about how to convert that hw into a tutorial, changed pacing to make it better as a standalone and interactive

Phase 3: CRAN submission

–goal was to submit to CRAN, explain CRAN –learned a lot about package checks/github actions –had been away from it for so long, first thing this semester was to update the version numbers of all the github actions checks –also had to update to base pipe, it had been that long –that spurs conversation about DESCRIPTION and NAMESPACE files. did research on how other tutorial packages handled the "Depends" "Suggests" and "Requires" fields. –stuff like spell check, standardizing to American English, decreased sizes of cover photos for each

tutorial to fit on CRAN –eventually hitting final road block of `learnr` not being fully available on CRAN, thus our package that would want to Depend it can't

Discussion

also about future directions with 199 stuff

talk about how much quarto, html, css, yaml, qualtrics, github actions, CRAN etc ive learned

Bibliography

Çetinkaya-Rundel, Mine, and Victoria Ellison. 2021. "A Fresh Look at Introductory Data Science." *Journal of Statistics and Data Science Education* 29 (sup1): S16–26.

Delmas, Robert, Joan Garfield, Ann Ooms, and Beth Chance. 2007. "ASSESSING STUDENTS' CONCEPTUAL UNDERSTANDING AFTER A FIRST COURSE IN STATISTICS." *STATISTICS EDUCATION RESEARCH JOURNAL* 6 (2): 28–58.

Epstein, Jerome. 2013. "The Calculus Concept Inventory-Measurement of the Effect of Teaching Methodology in Mathematics." *Notices of the American Mathematical Society* 60 (8): 1018–27.

Godfrey, Kelly E., and Sanja Jagesic. 2016. *Validating College Course Placement Decisions Based on CLEP Exam Scores: CLEP Placement Validity Study Results. Statistical Report.* College Board.

Mulford, DouglasRobert. 1996. "An Inventory for Measuring College Students' Level of Misconceptions in First Semester Chemistry." *Unpublished Master's Thesis, Purdue University, IN.*

Solomon, Erin D., Julie M. Bugg, Shaina F. Rowell, Mark A. McDaniel, Regina F. Frey, and Paul S. Mattson. 2021. "Development and Validation of an Introductory Psychology Knowledge Inventory." *Scholarship of Teaching and Learning in Psychology* 7: 123–39.

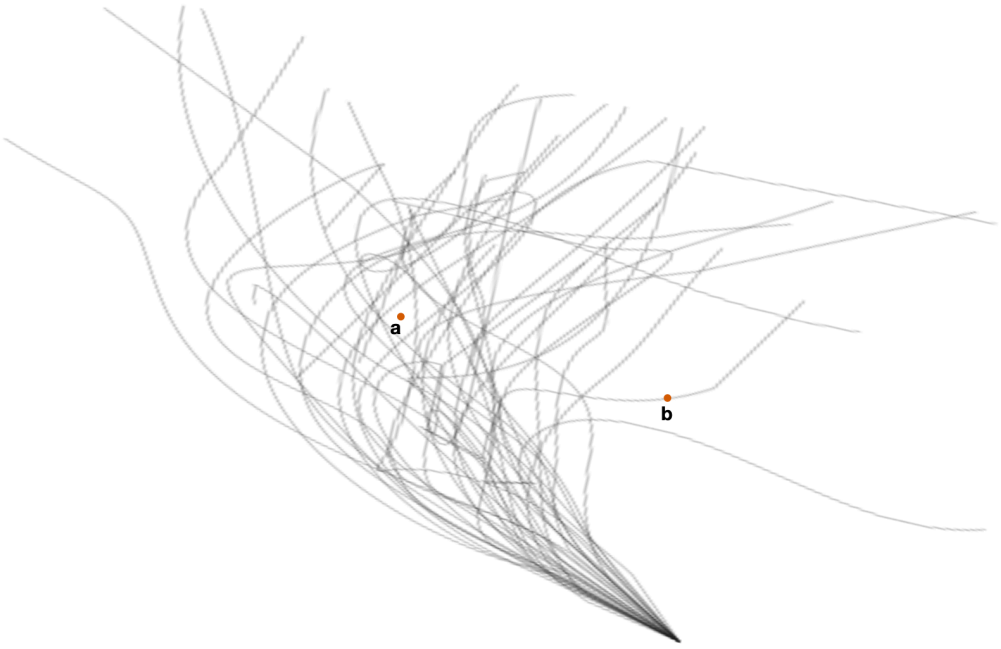
Swanstrom, Ryan. n.d. "Data Science Colleges and Universities." http://datascience.community/colleges.

Zhang, Zhiyong, and Danyang Zhang. 2021. "What Is Data Science? An Operational Definition Based on Text Mining of Data Science Curricula." *Journal of Behavioral Data Science* 1 (1): 1–16.

Appendix A: Assessment Prototype

Storm Paths

The figure below shows a forecast after simulating 50 potential paths for a large storm. The two points (a) and (b) represent two cities. Which city is more likely to be hit by the storm? Explain.

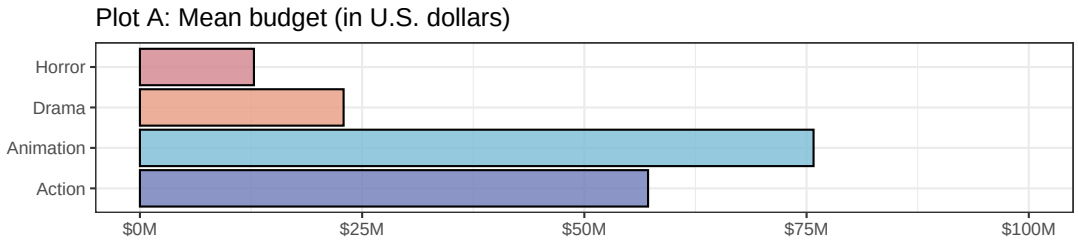


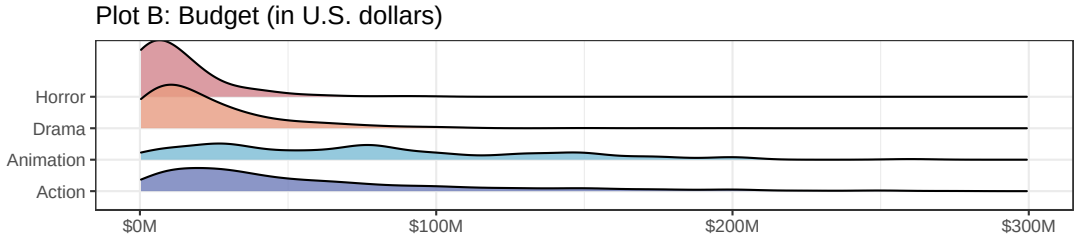

```
a. City a
b. City b
```

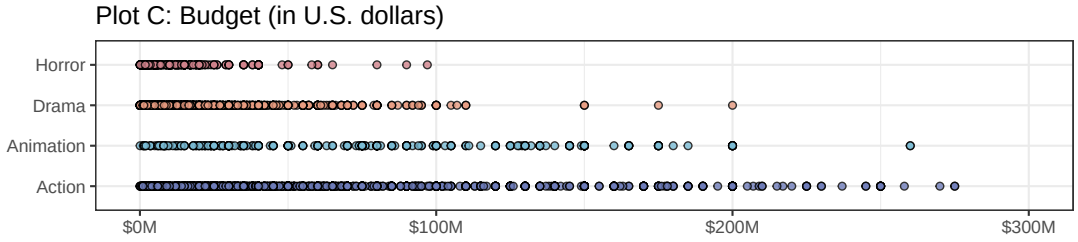
Movie Budgets 1

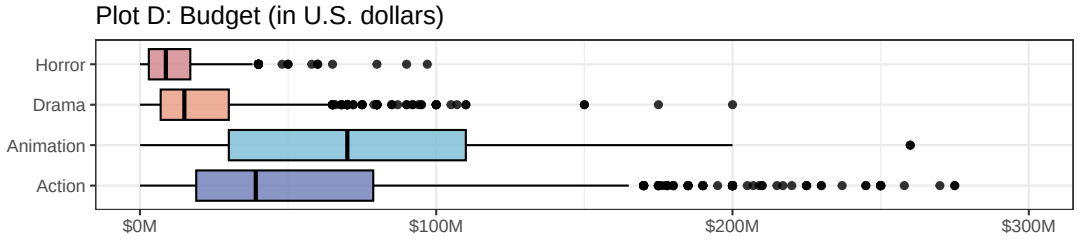
A data scientist at IMDb has been given a dataset comprised of the revenues and budgets for 2,349 movies made between 1986 and 2016.

Suppose they want to compare several distributional features of the budgets among four different genres—Horror, Drama, Action, and Animation. To do this, they create the following plots.







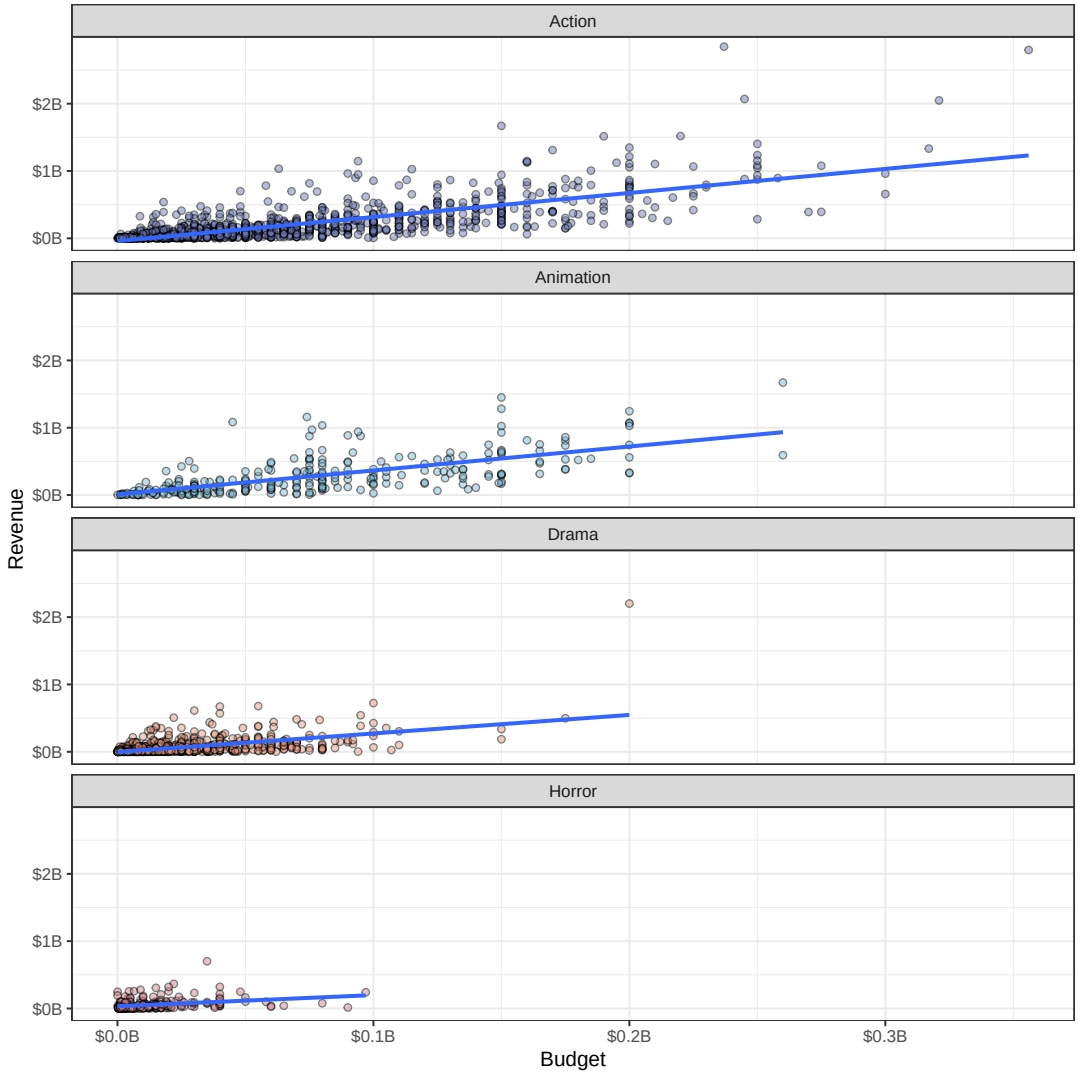


Fill in the following table by placing a checkmark in the cells corresponding to the attributes of the data that can be determined by examining each of the plots.

	Plot A	Plot B	Plot C	Plot D
Mean	[ ]	[ ]	[ ]	[ ]
Median	[ ]	[ ]	[ ]	[ ]
IQR	[ ]	[ ]	[ ]	[ ]
Shape	[ ]	[ ]	[ ]	[ ]

Movie Budgets 2

For each genre, the data scientist also fitted a regression line to model the relationship between movies' budgets and their revenues. A scatterplot of this relationship, along with the fitted regression line, is shown for each of the four genres below. For which genre would the fitted regression model produce the highest R^2 value? Explain.



Application Screening

You are working on a team that is making a deterministic model to quickly screen through applications for a new position at the company. Based on employment laws, your model may not include variables such as age, race, and gender, which could be potentially discriminatory.

Your colleague suggests including a rule that eliminates candidates with more than 20 years of previous work experience, because they may have high salary expectations. Are there ethical implications of using this variable to select candidates? Explain your answer.

Banana Conclusions

Data scientists at FiveThirtyEight administered a food frequency questionnaire. With 54 complete responses they found that people who ate bananas tended to score higher on the SAT verbal section than the SAT math section ($p = 0.0073$). An article reporting the results of this study has the headline, *"Eat more bananas to score higher on the SAT verbal section"*. Is this headline accurate, or could it be misleading? Explain.

COVID Map

The visualization below displays the 14-day rolling average of new COVID-19 cases January 1 - August 31, 2021 in the United States. Each plot represents a state or Washington, D.C., and is labeled using the state's abbreviation (e.g., MA = Massachusetts). The shaded area under each curve represents the increase in new cases since the state's minimum point in 2021. This is a recreation of a similar plot that originally appeared in the New York Times.

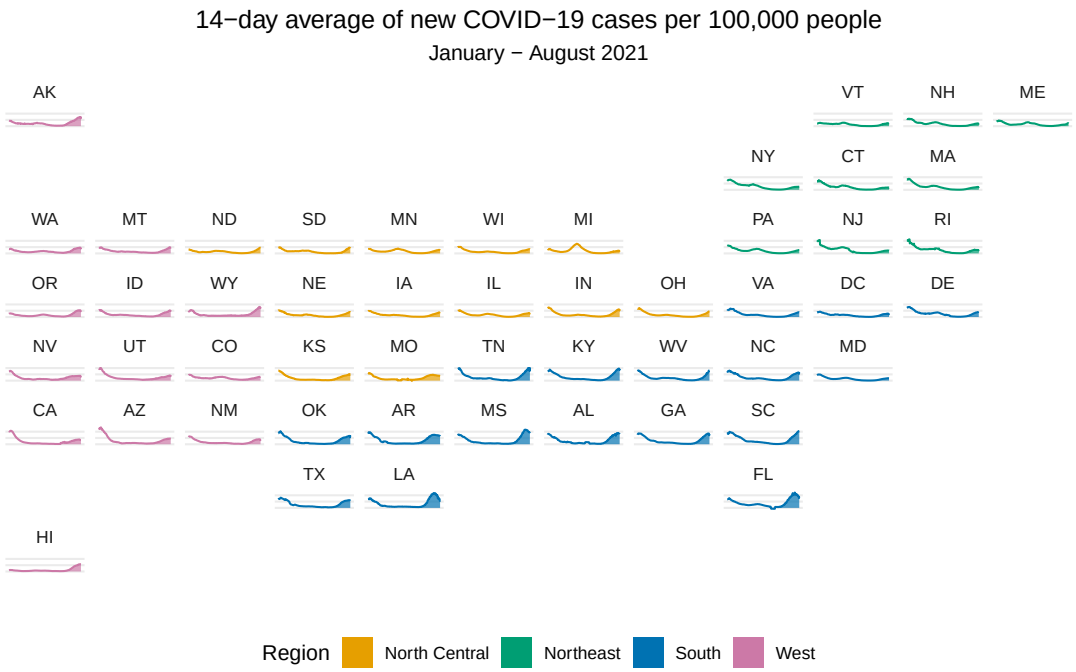
What do we learn from this plot about COVID-19 cases in the US?

Compare KY (Kentucky) in the South region to CA (California) in the West region. Based on this plot, can we conclude there was a difference in overall number of COVID cases in KY and CA in August 2021? Explain.

He Said She Said

For each of the following items, indicate whether the statement is TRUE, FALSE, or whether you would need additional information to determine this. If you can determine the statement is true/false, indicate the evidence that you used to make that determination. If you need additional information to make that determination, indicate what else you would need.

Men in Austen's novels are more likely to have 'dared', 'expected', and 'ran' than women.



a. True
b. False
c. Need additional information to determine this

Women in Austen's novels are more likely to have 'remembered', 'felt', and 'cried' than men.

a. True
b. False
c. Need additional information to determine this

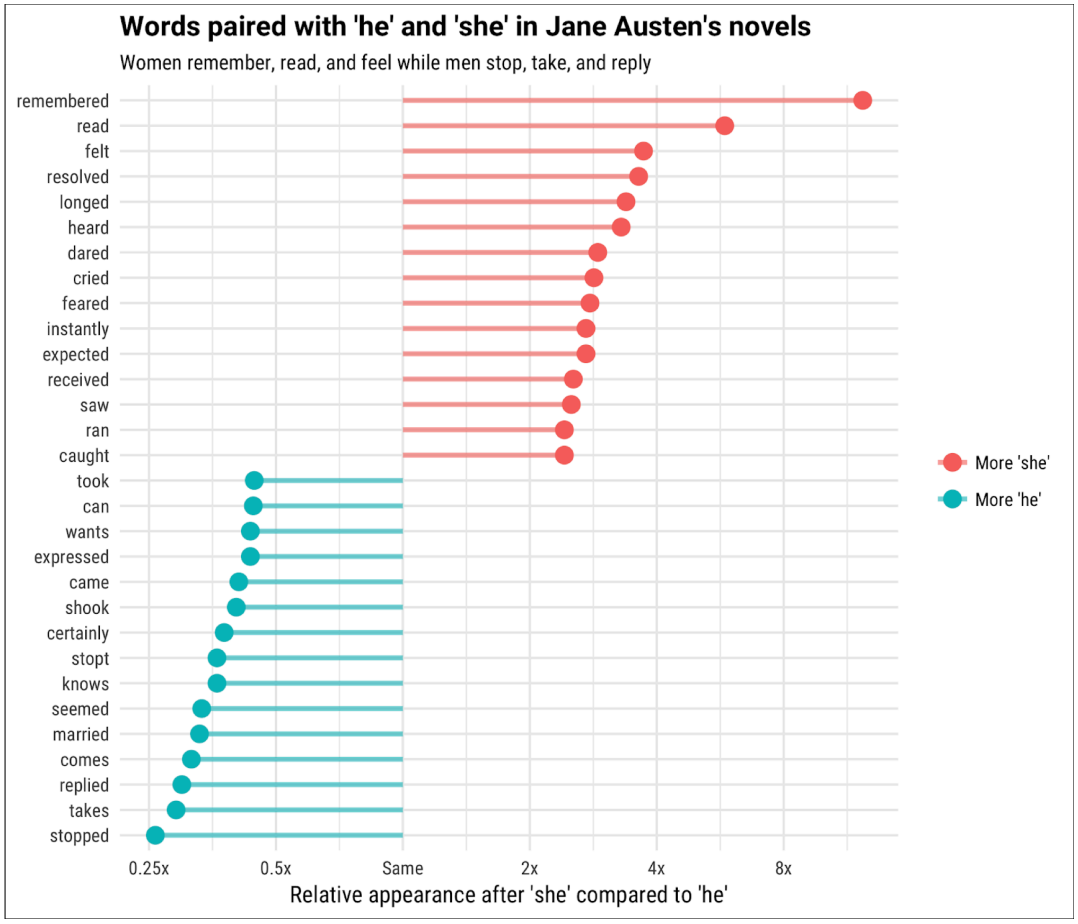
Women in Austen's novels are more likely to have 'remembered' than 'feared'.

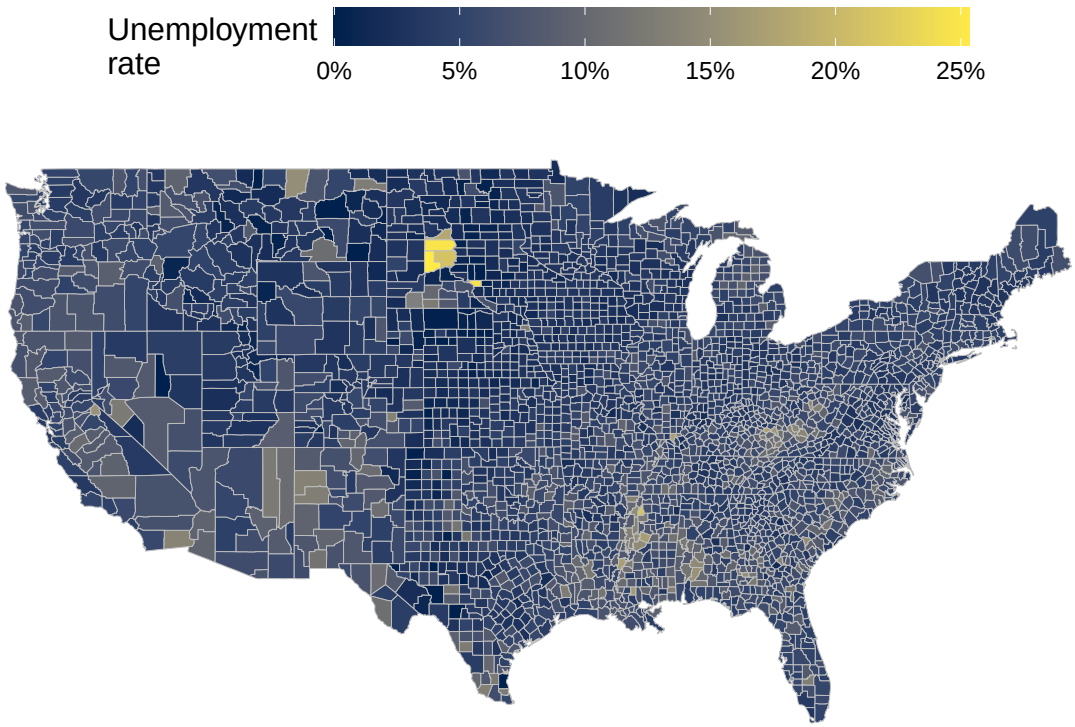
a. True
b. False
c. Need additional information to determine this

Build-a-Plot

The following is an intensity map of the unemployment rate among adults in the counties in the United States (based on data from 2019).

Indicate which of the following data you need to recreate this map? (Select all that apply.)


Words paired with 'he' and 'she' in Jane Austen's novels
Women remember, read, and feel while men stop, take, and reply
remembered
read
felt
resolved
longed
heard
dared
cried
feared
instantly
expected
received
saw
ran
caught
took
can
wants
expressed
came
shook
certainly
stopt
knows
seemed
married
comes
replied
takes
stopped
0.25x
0.5x
Same
2x
4x
8x
Relative appearance after 'she' compared to 'he'
More 'she'
More 'he'



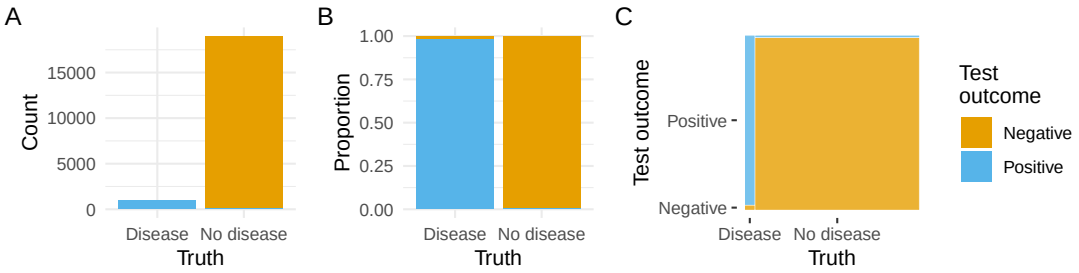
- [ ] County boundaries
- [ ] Unemployment rate in each county
- [ ] Number of adults living in each county
- [ ] Number of unemployed adults living in each county
- [ ] Total population of the county

Disease Screening

COVID screening tests are not 100% accurate. It's possible to have COVID but not test positive or not have COVID but test positive for it. The following three visualizations display the outcomes of a COVID screening test with a sensitivity (true positive rate) of 98.1% and specificity (true negative rate) of 99.6% in a population where 5% of the individuals have COVID.

We are also interested in the false positive (individuals classified as with COVID, who don't actually have it) and false negative (individuals classified as without COVID, but who do actually have it) rates.

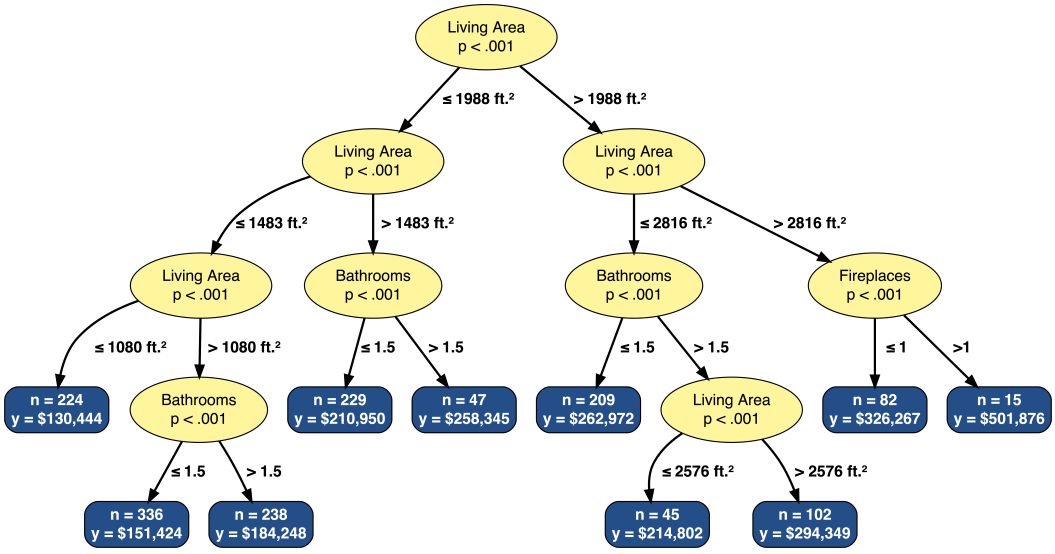
Fill in the following table by placing a checkmark in the cells corresponding to the attributes of the data that can be determined by examining each of the plots.



	Plot A	Plot B	Plot C
Sensitivity	[ ]	[ ]	[ ]
Specificity	[ ]	[ ]	[ ]
False positive rate	[ ]	[ ]	[ ]
False negative rate	[ ]	[ ]	[ ]

Realty Tree

A realtor has trained a regression tree to predict the price of a house from features such as number of bedrooms, number of bathrooms, number of fireplaces, and size of the living area.



What price would the tree predict for a house with 3200 ft.^{2} of living area, 1.5 bathrooms, and 1 fireplace?

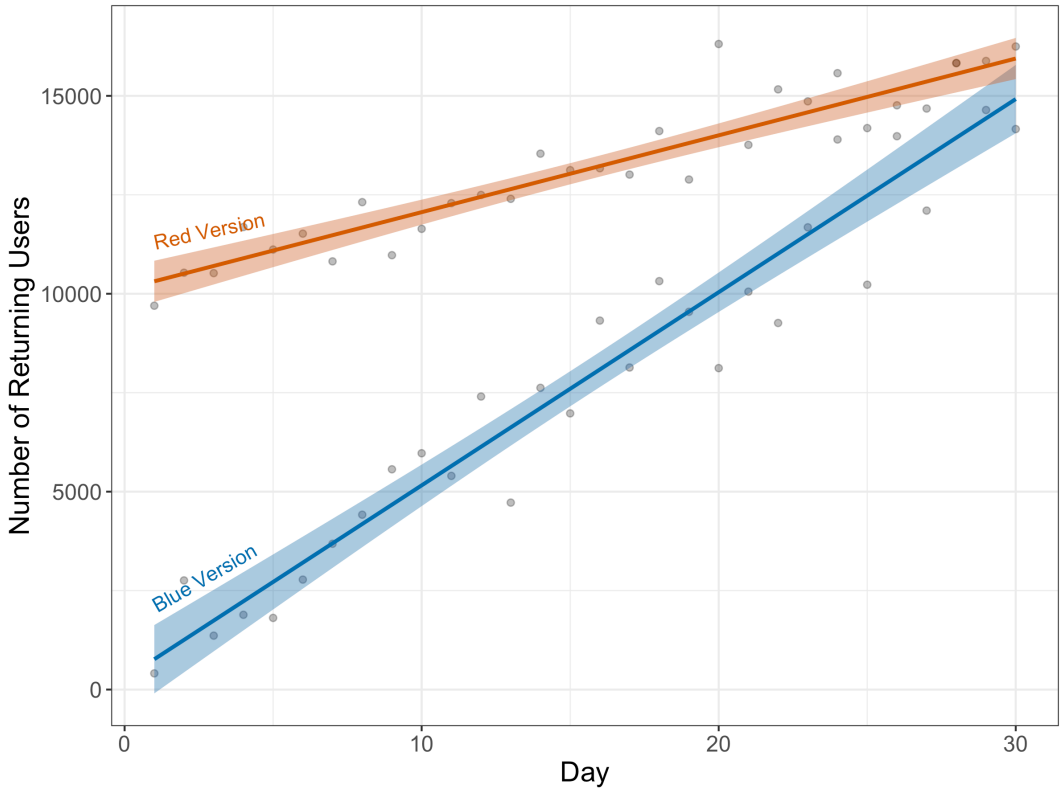
- ☐ $262,972
- ☐ $326,267
- ☐ $501,876
- ☐ Can't be determined from the information given.

What price would the tree predict for a house with 1200 ft.^{2} of living area and 1.5 bathrooms?

- ☐ $151,424
- ☐ $184,248
- ☐ $210,950
- ☐ Can't be determined from the information given.

Website Testing

An e-commerce company is working on their website design and is interested in knowing whether having the website mainly in blue or red would lead to better business outcomes. One outcome they are measuring is the number of returning users to the website. They design two versions of the website one in blue and the other in red. A random half of the visitors see the website in blue and the other half see it in red. The plot shows the number of returning users per day for the two different versions of the website.



Indicate whether each of the following conclusions are valid. Explain.

Over time the company is getting more returning users regardless of the version of the website.

a. Valid
b. Invalid
c. Cannot determine this from the plot.

On the 31st day, the blue version of the website is expected to have higher number of returning users.

a. Valid
b. Invalid
c. Cannot determine this from the plot.

On the 60th day, the blue version of the website is expected to have higher number of returning users.

a. Valid
b. Invalid
c. Cannot determine this from the plot.

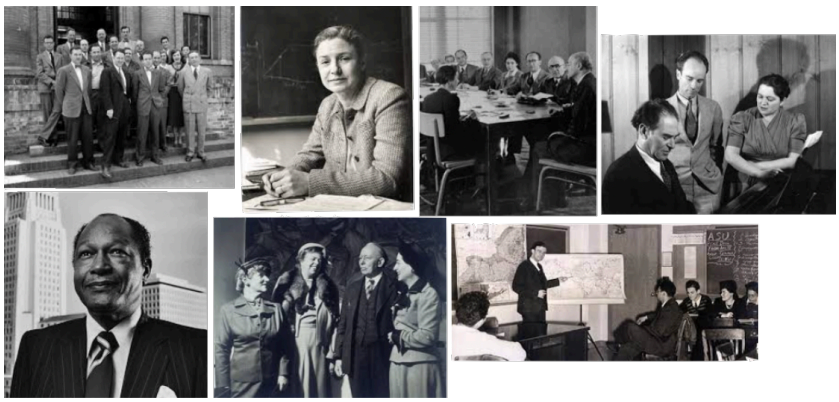
Image Recognition

A data science student wants to create an image recognition algorithm to identify whether a university professor belongs to a department in the sciences or not. To do this, she collects data by scraping several university photo archives of university faculty. She labels faculty in the photos as "Sciences" or "Not sciences". The images below depict a small representative sample of her data.

Sciences

Not sciences

The data science student plans to use these photos of current university faculty to predict whether they are scientists. What concerns might you have about the predictions from this algorithm? Explain.



Data Confidentiality

A newspaper reports on the results of a survey from a small (<2000 student) university. The university agrees to have the data released to the public so long as the students' identities and academic standing information are kept confidential. Select the safe combinations of variables that are unlikely to identify any individual students. Explain.

a. Class year and sports played
b. Student ID and dorm ZIP code
c. GPA and major
d. Birth date and phone number
e. None of the above

Activity Journal

Below is data that was recorded in an activity journal.

A data scientist reformats the data into a table so that each variable represented in the data is recorded in a single column. Describe what each of the columns of this table will contain, as well as what each row or observation of the table will represent.

Movie Wrangling

The table below provides data about 10 movies released in the United States. It provides data on the movie's title (`title`), the movie's director (`director`), the date the movie was released (`release_date`), the season the movie was released (`season`), the worldwide gross intake in U.S. dollars (`gross`), the cleaned version of the worldwide gross intake in U.S. dollars (`gross_clean`), and whether or not the movie won the Best Picture Oscar (`best_picture`).

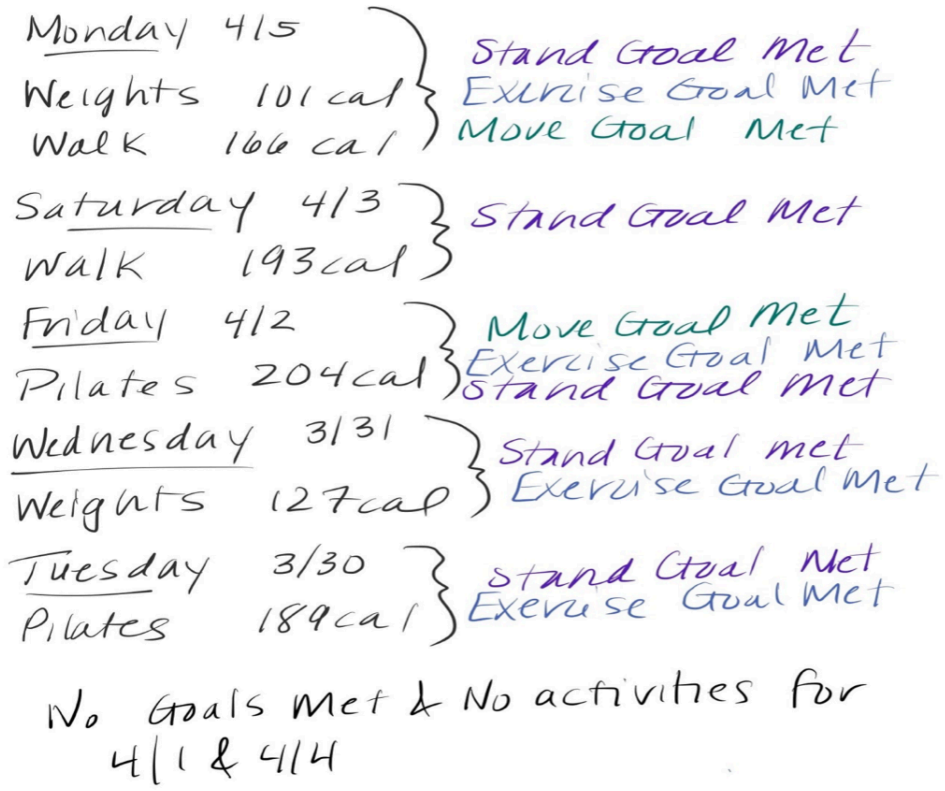


Table 0.6: Movies Table

title	director	release_date	season	gross	gross_clean	best_picture
Almost Famous	Cameron Crowe	22 September 2000	Fall	$47.39M	47.39	No
CODA	Sian Heder	13 August 2021	Summer	$1.61M	1.61	Yes
E.T. the Extra-Terrestrial	Steven Spielberg	11 June 1982	Summer	$792.91M	792.91	No
Luca	Enrico Casarosa	18 June 2021	Summer	$49.75M	49.75	No
Middle of Nowhere	Ava DuVernay	1 September 2014	Fall	$0.24M	0.24	No
Moonlight	Barry Jenkins	18 November 2016	Fall	$65.34M	65.34	Yes
Parasite	Bong Joon Ho	8 November 2019	Fall	$262.69M	262.69	Yes
Say Anything	Cameron Crowe	14 April 1989	Spring	$21.52M	21.52	No
Selma	Ava DuVernay	9 January 2015	Winter	$66.79M	66.79	No
We Bought a Zoo	Cameron Crowe	23 December 2011	Winter	$120.08M	120.08	No

23. Describe a process that you could use to generate the data in the `season` column using the information in the `release_date` column.

24. Describe a process that you could use to generate the data in the `gross_clean` column using the information in the `gross` column.

25. You have been tasked with adding a new column called `nominated_for_best_picture` which indicates whether or not each movie was nominated for the Best Picture Oscar ("Yes" if it was, "No" if it was not). Is there sufficient information in this dataset to generate this new column? Explain.

The table below provides data about 10 movie directors. It provides data on the director's name (`director`), the number of Oscars the movie's director has been nominated for (`nominations`), and the number of Oscars the director has won (`oscars`).

Table 0.7: Directors Table

director	nominations	oscars
Ava DuVernay	1	0
Barry Jenkins	3	1
Bong Joon Ho	3	3
Cameron Crowe	3	1
Enrico Casarosa	2	0
Loveleen Tandan	0	0
Nora Ephron	3	0
Penny Marshall	0	0
Sian Heder	1	1
Steven Spielberg	19	3

Use the data in the Movies Table and in the Directors Table to answer the following questions. For each question, what is the result of carrying out the given pseudocode (ie. code recipe)?

26.

START_WITH(the Movies table) then

KEEP_ROWS_WHERE(the `season` value is `Fall`) then

COUNT(the number of rows)

27.

START_WITH(the Movies table) then

KEEP_ROWS_WHERE(the `season` value is `Fall`) then

COUNT(the number of rows) WHERE(`best_picture` value is `Yes`)

28.

START_WITH(the Movies table) then

KEEP_ROWS_WHERE(the `season` value is `Fall` then

ADD_COLUMNS_FROM(the Director Table) MATCHING_BY(the `director` column) then

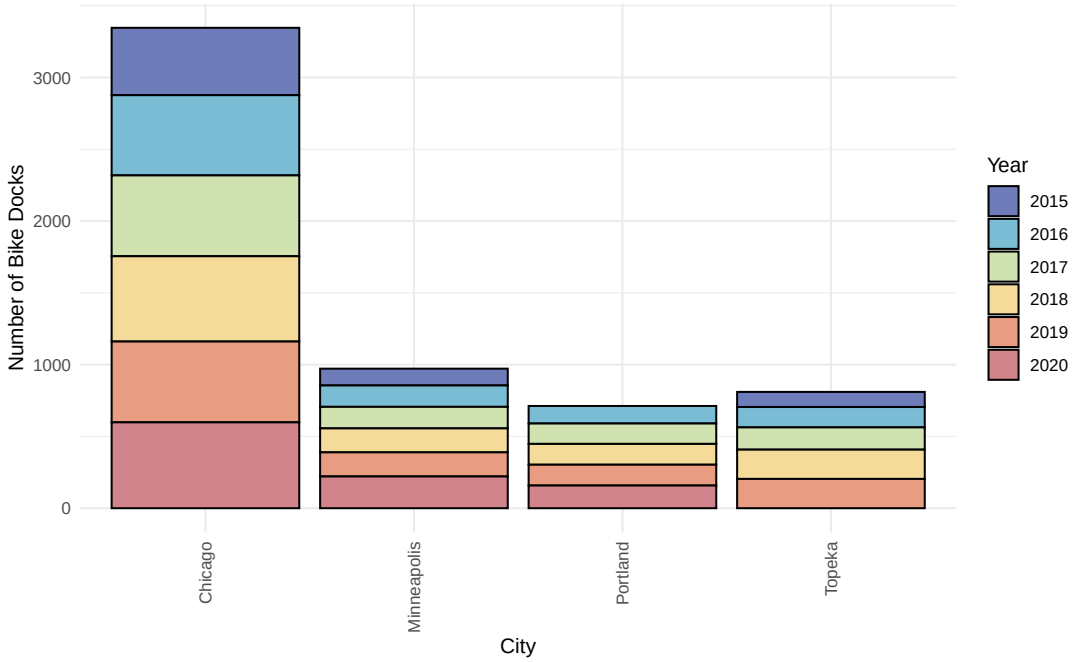
COUNT(the number of rows) WHERE(`oscars` value is `3`) AND(`best_picture` value is `No`)

Appendix B: Item Graveyard

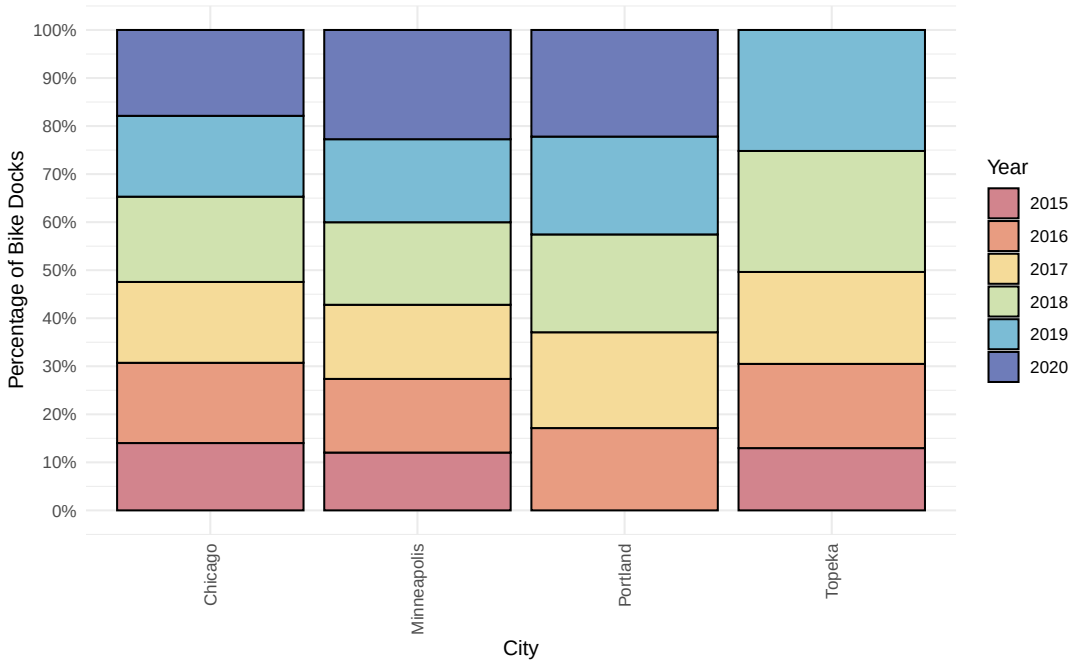
Bikes and Scooters 1

As a way to help the environment some cities in the U.S. are adding bike and e-scooter share stations which allow people to rent a bike or e-scooter for commuting or pleasure. The bikes and scooters are often kept at electronic docking stations at multiple locations around the cities. The following graphs were created using data from the Department of Transportation Statistics[1] about public use of these shared dock systems in four U.S. cities—Chicago, Minneapolis, Portland, and Topeka.

1. Which of the four cities had the most bike docks in 2020? Explain how you determined this. Or if you cannot answer it from the visualization, explain why not.



2. The bar chart shows the the percentage of docks in each city that were bike docks for each year from 2015–2020.



The pie chart was created by plotting the percentage of bike docks for each year in one particular city. Unfortunately, the data scientist has forgotten which city this is. Using the information in the bar chart, identify the city. Explain how you determined this or if you cannot answer it from the visualization, explain why not.

3. Which of the four cities had the biggest increase in the number of bike docks from 2018 to 2020? Explain how you determined this or if you cannot answer it from the visualization, explain why not.

Bikes and Scooters 2

The map below shows the number of cities in each region of the United States that have docked bikes, dockless bikes, or e-Scooters in both 2018 and 2020. Use that information to answer each of the following questions. For each question, explain how you determined your answer, or if you cannot answer it from the visualization, explain why not.

	2018			2020		
	Docked Bikes	Dockless Bikes	e-Scooters	Docked Bikes	Dockless Bikes	e-Scooters
West	17	14	7	13	7	3

[1]Data from: https://data.bts.gov/d/7m5x-ubud/visualization

Southwest	11	7	5	8	6	4
Midwest	27	5	8	16	6	6
Northeast	17	18	5	10	3	1
Southeast	29	13	9	14	11	11

4. How many cities in the Southeast had e-Scooters in 2018?
5. In 2020, the Southwest region has more docked bike stations than dockless bike stations.
6. The majority of regions decreased the number of e-scooter stations from 2018 to 2020.
7. Across the majority of regions, the trend is that over time, cities tend to be adopting dockless bikes rather than docked bikes.
8. Across the majority of regions, the trend is that over time, there are fewer cities that are making docked bikes, dockless bikes, and e-scooters available. Explain how you determined this or if you cannot answer it from the visualization, explain why not.

Bikes and Scooters 3

9. An electric bike, also known as an e-bike, is a bicycle with a battery-powered "assist" that comes via pedaling. An online product recommendation service that tests and reviews products has gathered a representative sample of 15 e-bikes from a single manufacturer and measured their ranges (how far they can go on a full battery without recharging). Based on this sample, they calculated an average range of 60 kilometers, plus or minus 10 kilometers. Suppose you're in the market for an e-bike and during your research you come across the following two items:

- An e-bike with a range of 85 kilometers.
- A report from a different product recommendation service that has also gathered data from a different, but also representative sample of 15 e-bikes from this same manufacturer, with a mean range of 85 kilometers.

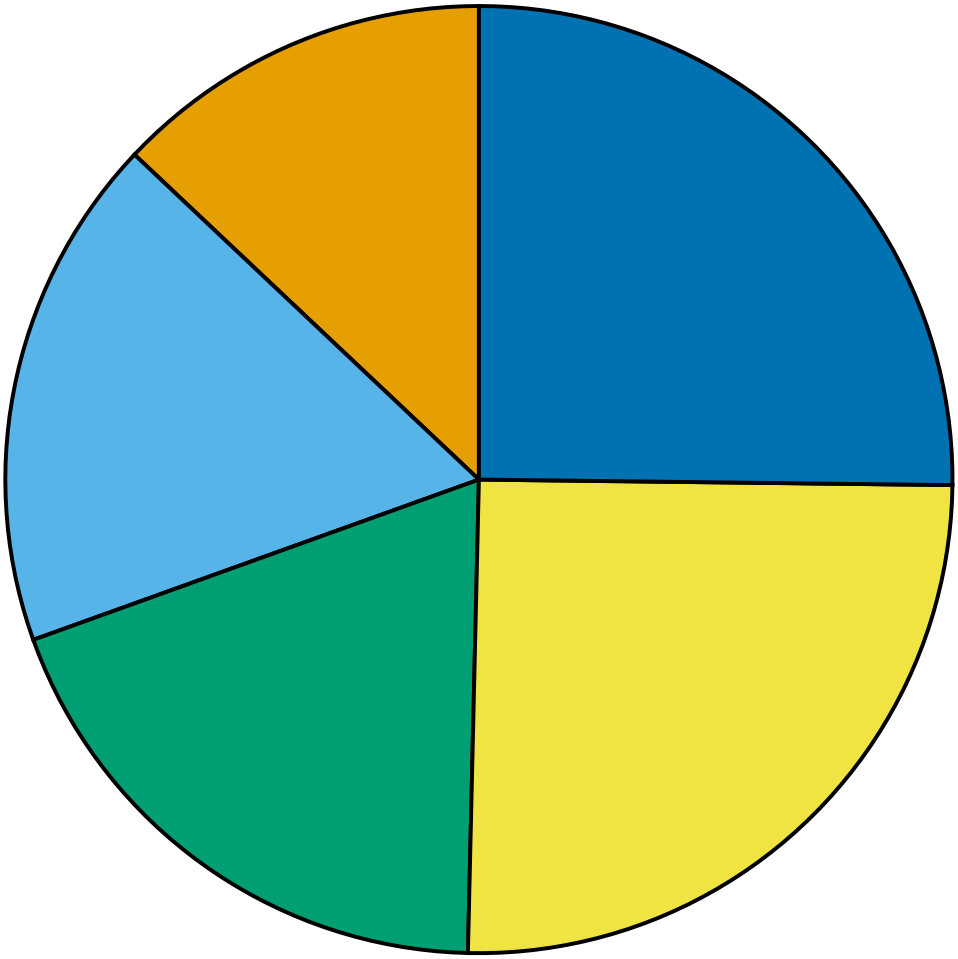
  Which one of these make you doubt the original report more?

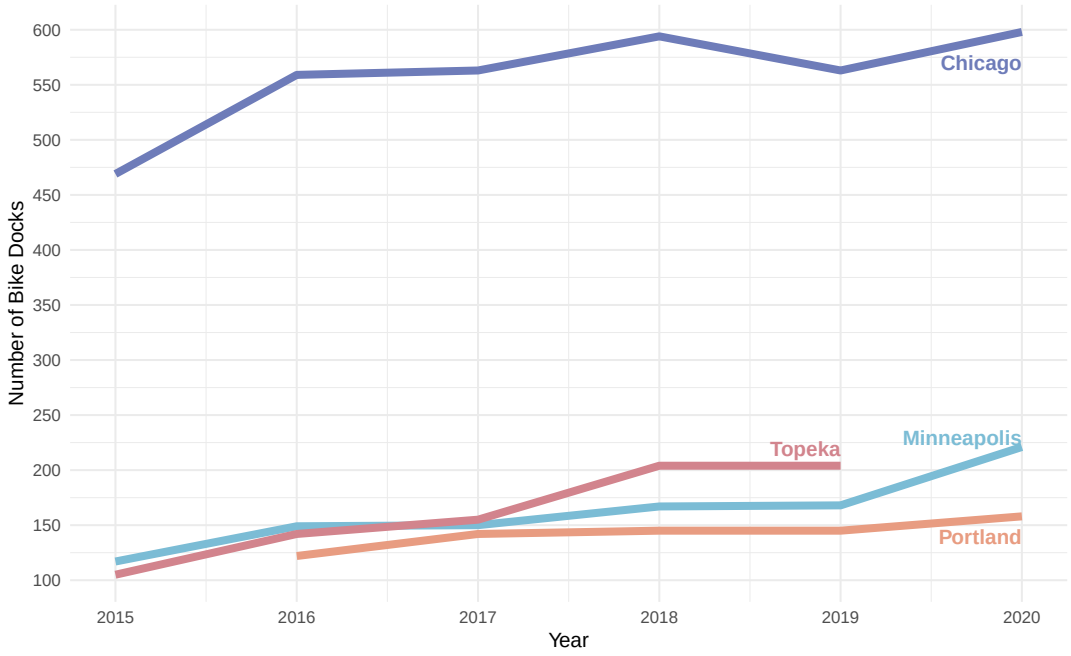
Births per Day

A data scientist for a large urban hospital examined a sample of data to estimate the mean number of births that took place on Fridays and Saturdays. The plots below show the number of births that took place on either a Friday or Saturday for that sample.

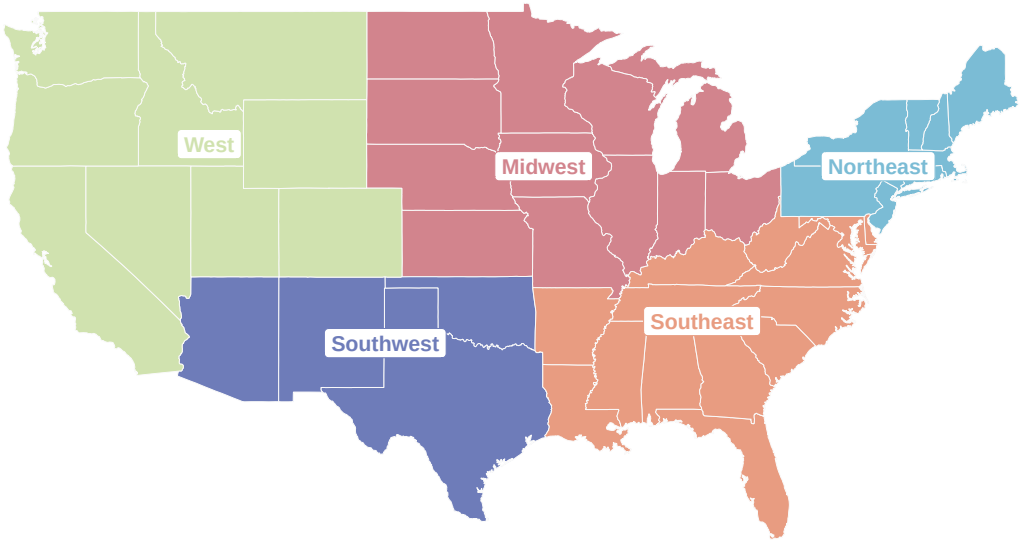
Percentage of Bike Docks per Year

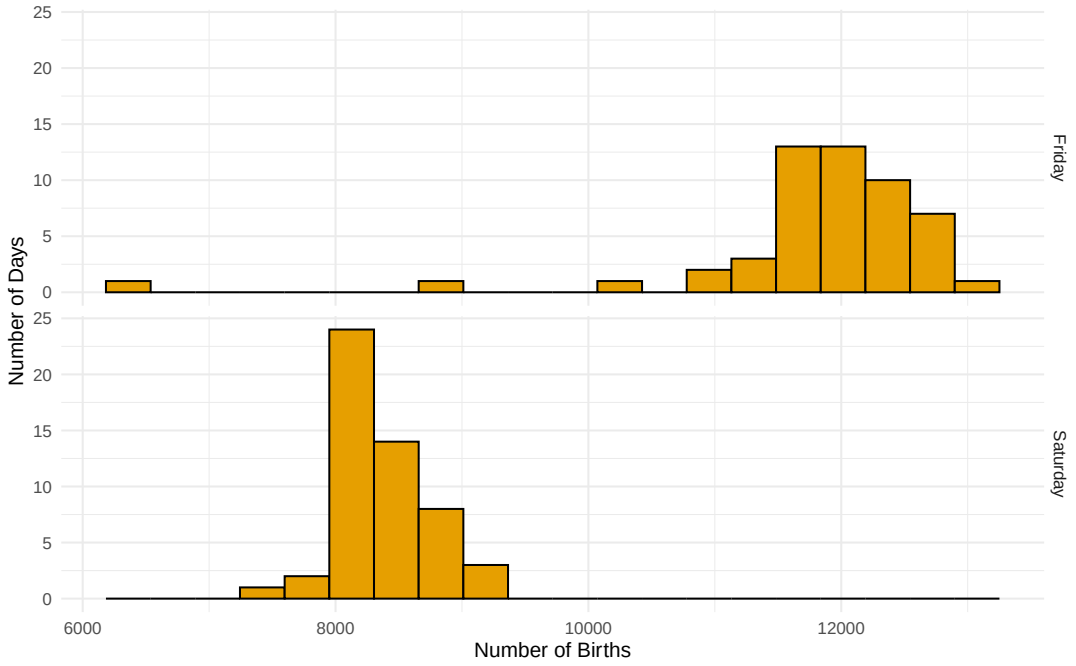
In _____





Number of cities with docked bikes, dockless bikes, and e−scooters in 2018 and 2020.





To estimate the mean number of births that took place on Fridays and Saturdays, the data scientist computed confidence intervals (mean $\pm$ margin of error) for both days. Unfortunately they forgot which mean and margin of error was associated with each day.

9. Which mean is associated with Friday? Explain.

    a. 8350
    b. 11,800

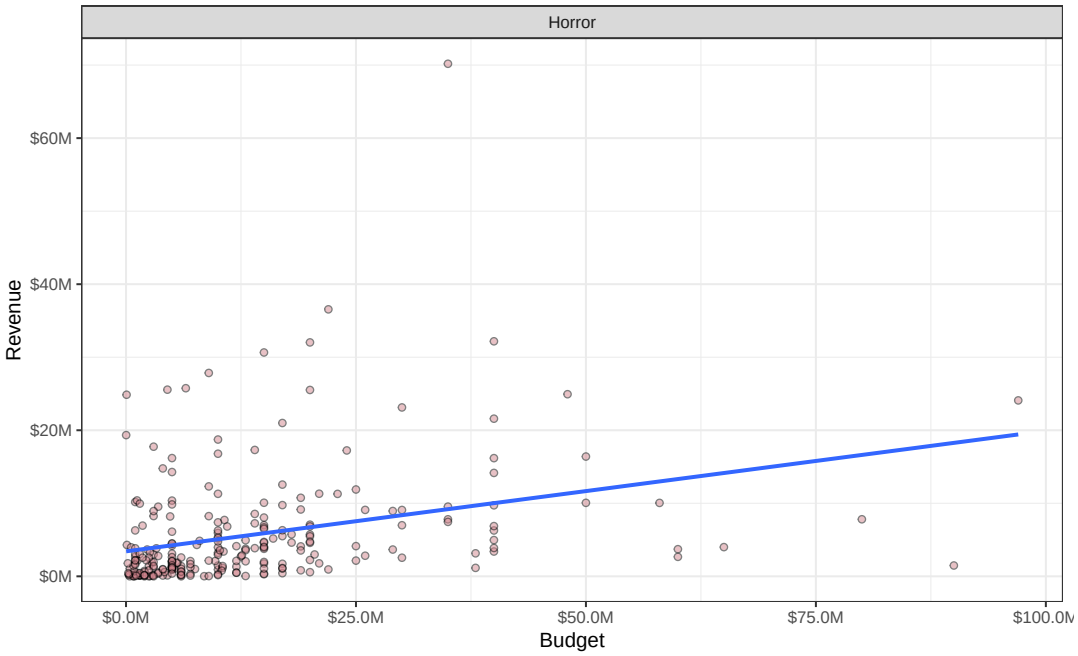
10. Which margin of error is associated with Friday? Explain.

    a. 100
    b. 280

Movie Budgets 3

The data scientist was asked to use the fitted regression model to make a prediction for the revenue for a horror movie using two different potential budgets; a budget of \$25M and a budget of \$50M They were also asked to compute a prediction interval for these two predictions to estimate the uncertainty in the prediction. The scatterplot and fitted regression line for Horror movies is displayed below.

8. Which of the predictions would have a greater predicted revenue associated with it? Explain



a. budget = \$25M
b. budget = \$50M
c. They are the same.
d. Not enough information to determine this.

Model Comparison

A data scientist has trained four different classification models (null model, Naive Bayes model, k Nearest Neighbors (kNN) model, and random forest model) on a set of data. The observed responses and the model predictions for a set of 10 observations from a validation set of data are shown in the table below.

	Prediction			
Observed Responses	Null	Naive Bayes	kNN	Random Forest
No	No	No	No	No
No	No	Yes	Yes	No
No	No	No	No	No
No	No	No	No	No
Yes	No	No	No	No
No	No	Yes	Yes	No
No	No	No	No	No
No	No	No	No	No
Yes	No	Yes	Yes	Yes
Yes	No	Yes	Yes	Yes

28. Are the predictions from the kNN model more, less, or equally as accurate as the results from the null model? Explain.

Training or Validation

29. The figure shows a plot of prediction error as a function of model complexity for a training and validation sample. Which sample (training or validation) is associated with the *orange, solid* line? Explain.

The following three items were reworked into one context, as the current Movie Wrangling

TV Show Wrangling

The two tables below provide data about several TV shows.

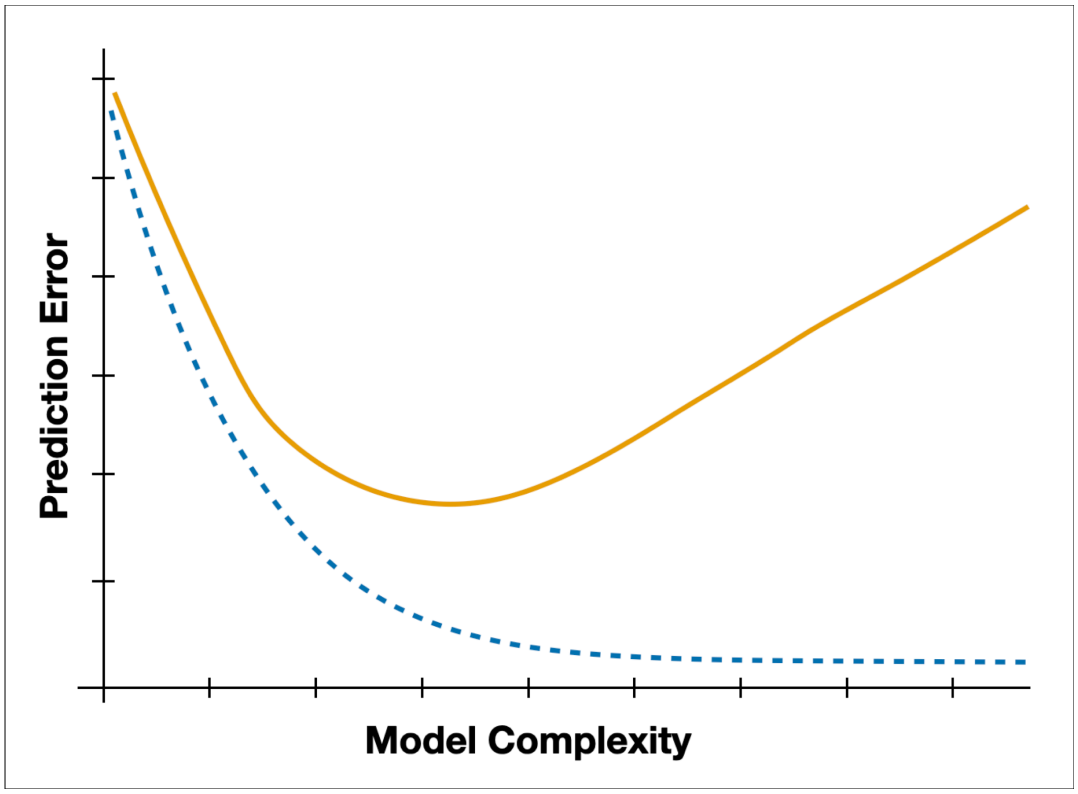

Prediction Error
Model Complexity

Table 0.9: Creator Table

Creator	TV_Show	TV_Show_ID
Aguirre-Sacasa, Roberto	Riverdale	I
Blair, April	All-American	A
Dunham, Lena	Girls	L
Glover, Donald	Atlanta	B
Levitan, Steven	Modern Family	F
Lloyd, Christopher	Modern Family	F
Murphy, Kevin	Hellcats	E
Scheuring, Paul T.	Prison Break	H
Sherman-Palladino, Amy	Bunheads	C
Sherman-Palladino, Amy	Gilmore Girls	D
Star, Darren	Sex and the City	M
Star, Darren	Younger	K
Watson, Sarah	The Bold Type	J

33. Consider the two following sets of pseudocode (ie. code recipe). Would they produce the same results? Explain.

    1.

[H]
start_with(the Creator table) and_then
add_columns_from(the TV Show table matching_by the TV Show column) and_then
count_of(the number of rows for the CW network)

Table 0.10: TV Show Table

TV_Show	TV_Show_ID	Network	Seasons
All-American	01	CW	4
Atlanta	02	FX	4
Bunheads	03	ABC Family	2
Gilmore Girls	04	WB	7
Hellcats	05	CW	1
Modern Family	06	ABC	11
Ozark	07	Netflix	4
Prison Break	08	Fox	6
Riverdale	09	CW	6
The Bold Type	10	Freeform	5
Younger	11	TV Land	7

    2.

start_with(the TV Show table) and_then
add_columns_from(the Creator table matching_by the TV Show column) and_then
count_of(the number of rows for the CW network)

34. Consider the two following sets of pseudocode (ie. code recipe). Would they produce the same results? Explain.

    1.

start_with(the Creator table) and_then
add_columns_from(the TV Show table matching_by the TV Show column) and_then

count_of(the number of rows for the CW network)

2.

start_with(the TV Show table) and_then

add_columns_from(the Creator table matching_by the TV Show ID column) and_then

count_of(the number of rows for the CW network)

Shopping Wrangling

The dataset below contains information on 8 people; we know their names and how many items they purchased online today.

name	number_of_items	visited_online_retailer
Miriam	10	
Marcel	2	
Ayesha	0	
Rebecca	3	
Lola	0	
Laurence	1	
Tomos	9	
Abdul	0	

37. You have been tasked with adding a new column called `visited_online_retailer` which indicates whether or not each person visited the website of an online retailer ("yes" if they did, "no" if they did not). Is there sufficient information in this dataset to generate this new column? Explain.

Park Wrangling

The data set `park_visits` contains the number of annual visitors to 376 national park sites in the United States from 1904–2016. The data were originally collected from the National Park Service. There are 20,920 total records in the data set, since the parks were open for the entire date range. A few rows of `park_visits` data are shown below.

year	state	park_site	visitors
1904	AR	Hot Springs National Park	101000
1904	CA	Kings Canyon National Park	1000

```
1904  OR     Crater Lake National Park            1500
1904  SD     Wind Cave National Park              2900
...   ...    ...                                  ...
2016  WY     Devils Tower National Monument       496210
2016  WY     Fort Laramie National Historic Site  57444
```

A data scientist would like to find the most popular park in each state in 2016. To do so, they decided to create a new data table named `most_visited_2016` that includes the national park site in each state with the most visitors in 2016. The final table will include 51 rows (one for each state and Washington D.C.) and the columns `year`, `state`, `park_site` and `visitors`. Six of the 51 rows of the table are shown below.

```
year  state park_site                                  visitors
2016  AK    Klondike Gold Rush National Historical Park  912351
2016  AL    Little River Canyon National Preserve        462700
2016  AR    Buffalo National River                       1785359
2016  AS    National Park of American Samoa              28892
2016  WV    Harpers Ferry National Historical Park       335691
2016  WY    Yellowstone National Park                    4257177
```

38. Arrange the steps to get from the original data set `park_visits` to the final table `most_visited_2016`.

Steps:

- Start with `park_visits`
- FILTER(year == 2016): Filter for observations in 2016
- GROUP_BY(state) : Group by state / Perform subsequent lines of code within each state.
- ARRANGE(DESC((visitors)): Sort the number of visits in descending order
- SLICE(1) : Take the first observation.
- End with `most_visited_2016`

Data Cleaning

A researcher randomly selects 10 students in a school and collects data about their age and number of siblings. They enter the data into a spreadsheet and are interested in calculating descriptive statistics. The software that they are using displays the data as shown below.

```
row   X1     X10     X3
1     2      11      1
```

```
2     3      12      3
...   ...    ...     ...
8     9      12      2
9     10     8       0

10    Total  113     21
```

43. Would it be safe to assume that the average age of children in the sample is 11.3? Explain.

44. If you could change the way that the data are displayed in this software, would you change anything? If yes, then list the thing(s) that you would change.

Appendix C: dsbox in action

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).