# Practice Problems 2

Solutions

# Instructions

To complete your assignment, please follow these steps:

0. **Download R and RStudio (if you haven't already).** See the *Useful Resources* section on Moodle for installation links and videos with step-by-step installation instructions. If you're having issues with this, you'll need to reach out to your instructor and/or come to office hours, so plan accordingly!
1. **Download `practice2.Rmd` from Moodle and save it some place on your computer that you can easily find again.** I strongly encourage you to create a new folder dedicated to homework assignments for this class. See the *File Structure and Organization* video on Moodle for tips on how to do this.
2. **Make sure that the file you downloaded is called `practice2.Rmd` and not `practice2.Rmd.txt`.** The latter often happens when you use Safari on a Mac—try downloading the file using a different browser (e.g., Chrome) instead, or edit the file name (as explained in the *File Structure and Organization* video).
3. **Open `practice2.Rmd` in RStudio.** See the *Intro to R and RStudio*, *R Data Types*, and *R Error Messages and Troubleshooting* videos.
4. **Update the section number, author, and due date** on the third, fourth, and fifth lines of the file.
5. **Make sure you've already installed the `dplyr`, `readr`, and `ggplot2` packages.** Open the `Packages` panel (usually in the bottom right corner) to see the list of all packages that are already installed. Look to see if `dplyr`, `readr`, and `ggplot2` are listed here. If they're on this list, that means they're already installed and you're good to go. If any packages are missing from this list, type `install.packages('packagename')` in the *Console* (usually in the bottom left corner) and hit enter. See the *R Packages* video.
6. **Try *Knitting* your document:** click the `Knit` button at the top of this screen (look for yarn and needle). See the *Intro to RMarkdown* video. A dialogue box may pop up asking you if you want to install some packages: click "Yes." If you encounter any error messages, get in touch with your instructor or the preceptors.
7. **Answer the questions in Parts 1–4.** Click `Knit` occasionally along the way to make sure everything looks okay.
8. Once you're done with Parts 1–4, **click `Knit` one final time.** This will turn your R Markdown document into a nicely formatted HTML file.
9. **Look at the HTML file to make sure it looks like you want it to**: graphs appearing, no error messages, no data print out that lasts 50 pages, etc.
10. Once your HTML file pops up and you've checked that it looks like what you want, **click `Open in Browser` and then Print and select "Save as pdf".**
11. **Turn in two files to Moodle:** this `.Rmd` file, and the knitted `.pdf` version you generated in Step 10.

# Part 1. Load and Prepare Data in R

# Project Data

Look through the information for the possible project data sets on Moodle. Choose one data set to work with (or get approval from Kelsey to use a different dataset). Then, use the code chunk provided below to read your chosen data set into R and print out the `names` of all the variables. Don't forget to load any packages you might need!

Note: you can change your mind about the dataset for the next assignment, but after that, you'll be working with that same data set for the remainder of the semester. So, make sure to choose a data set that seems interesting to you!

```
# load package first (if needed)
## if you're using data from a package, load that package first (e.g., library(NHANES))
## otherwise, if reading in data from a CSV file you'll need the readr package:
library(readr)

# then read in dataset
## read in the data using the read_csv function and save it
## for example:
project <- read_csv('projectdata.csv')

# then print out variable names
## run the names() function on your dataset
## for example:
names(project)
```

Note: in the example above, my code reads in a dataset and saves it as an object called `project`. You can give your data whatever* name you like! Just make sure to use the same name throughout. For example, if you named your data `kiva` (`kiva <- read_csv('...')`) then run the `names` function on kiva as well (`names(kiva)`).

*there are a few rules to naming objects in R. Names can't start with a number, can't include some special characters like @, /, &, or spaces, and shouldn't be the name of something that already exists (e.g., don't name your data "read_csv"). I recommend using a name that's short and informative, so it's easy to type and easy to remember what you called it.

Another note: the `read_csv()` function prints out quite a bit of information when it loads in the dataset. When we're working with a dataset for the first time, this information is somewhat useful for us to look at, but in future weeks I'll show you how to hide these messages so we don't have to look at them every time.

# Intro Survey Data

Enter code in the code chunk below to load the `introsurveyfinal.csv` dataset into R, and then print out the first six rows of the dataset. Make sure that you've downloaded the `.csv` file from Moodle (see Activity 4) and it's saved in the same folder as this RMD. Don't forget to load any packages that you might need first!

```
# load package first (if needed)
## we'll need the readr package for this
## note: if you already loaded the package above, you don't need to do it again here!
library(readr)

# then read in dataset
intro <- read_csv('../activities/introsurveyfinal.csv')
```

```
## Rows: 53 Columns: 10
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): Timestamp, What is your declared or potential major? (If you are a ...
## dbl (6): How many hours of sleep did you get last night?, How many cups of c...
```

```
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# then print out first six rows
head(intro)
```

```
## # A tibble: 6 × 10
##   Timestamp  `How many hours… `How many cups … `What is your d… `What is your a…
##   <chr>                 <dbl>            <dbl> <chr>                        <dbl>
## 1 8/26/2021…              6.5                3 math                          2014
## 2 9/4/2021 …              7                  0 Computer Science              2025
## 3 9/6/2021 …              8                  0 econ                          2025
## 4 9/6/2021 …              8                  0 math                          2024
## 5 9/7/2021 …              7                  0 Math                          2022
## 6 9/8/2021 …              7                  0 Biology                       2023
## # … with 5 more variables:
## #   How many stats courses have you taken in the past? <dbl>,
## #   Have you used R or RStudio before? <chr>,
## #   On a scale of 1 to 10, how excited are you about this class? <dbl>,
## #   When is your birthday? <chr>,
## #   How many unread emails do you have in your inbox right now? <dbl>
```

> Remember: when you read in your dataset, you can call it whatever you want! In Activity 4 we called these data "results" but I decided to call them "intro" here. It's okay to use different names as long as you are consistent throughout the same R Markdown document.

# Preparing and Cleaning Data

Enter code in the code chunk below to filter the Intro Survey data to only include the *students* taking Stat 155 this semester, and save that new filtered dataset as an object called `students`.

> The `filter` function and the pipe (`%>%`) both live inside the `dplyr` package, so we need to remember to load that package first!

```
# load the dplyr package first!
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# filter dataset to keep only students
students <- intro %>%
  filter(`What is your anticipated graduation year?` >= 2021)
```

> The code above removes Kelsey's responses by keeping only students with an anticipated graduation year of 2021 or later. But there's more than one way to do this! Kelsey is also an outlier with respect to the number of previous stats courses she has taken, so we could also do the filtering this way:

```
# fiter dataset to keep people with < 2 prior stats courses
students <- intro %>%
  filter(`How many stats courses have you taken in the past?` < 2)
```

# Part 2. Data Context

Answer the following questions about the context of your chosen project dataset.

# Who & What

Use the code chunk below to figure out the dimensions of your data set: how many rows? how many columns?

```
# get number of rows and columns
nrow(project) # get number of rows
ncol(project) # get number of columns
dim(project) # or get both at once using dim
```

Then, describe the *who* — how many cases are in the data set, what do the cases represent — and the *what* — how many variables are in the data set, what do the variables generally represent (don't list them all, but provide a general summary of what they represent, e.g., demographic information, socioeconomic indicators, measures of physical health, etc.).

Cases (Who): make sure to include both how many cases **and** what the cases represent

- NHANES = 10,000 people who filled out the survey
- Scorecard = 3,676 US colleges
- Election = 3,111 US counties
- Kiva Loans = 97,183 individual loans
- Powerlifting = 1,423,354 competition lifts (note that some lifters are in this dataset more than once, so each case represents what that person lifted at that a particular competition. since this is such a big dataset, you'll probably want to think about filtering out some of these cases to make the project more manageable!)

Variables (What): how many variables **and** a summary of what types of information was collected (e.g., demographic, health, socioeconomic, …)

- NHANES = 76
- Scorecard = 93
- Election = 37
- Kiva Loans = 75
- Powerlifting = 37

# Where, When, Why, By Whom, How

For your chosen data set, use the information provided to describe the context surrounding the data (where were these data collected, when were these data collected, why were these data collected, by whom were these data collected, how were these data collected).

Where: where were the data collected?

- NHANES & Scorecard & Election = US
- Kiva Loans = 65 countries
- Powerlifting = 96 countries

When: when were these data collected?

- NHANES = 2009–2012
- Scorecard = last updated June 2020
- Election = 2012–2016
- Kiva Loans = 2005–2012
- Powerlifting = snapshot of the OpenPowerlifting database as of April 2019; includes competitions going back all the way to 1964

Why: what was the original purpose for which these data were collected?

- NHANES = assess the health and nutritional status of adults and children in the United States
- Scorecard = designed to increase transparency, putting the power in the hands of students and families to compare how well individual postsecondary institutions are preparing their students to be successful. This project provides data to help students and families compare college costs and outcomes as they weigh the tradeoffs of different colleges, accounting for their own needs and educational goals
- Election = to improve the scholarship and practice in election science through high-quality data gathering and analysis
- Kiva Loans = to track what happened to loans posted on their website
- Powerlifting = to create an archive of powerlifting history

By whom: who collected these data?

- NHANES = US National Center for Health Statistics
- Scorecard = US Dept of Education
- Election = MIT curated the data
- Kiva Loans = the non-profit called Kiva
- Powerlifting = meet organizers and powerlifting federation post competition results, and the database of all results is curated by OpenPowerlifting

How: how were the data collected? make sure to discuss sampling methods (was it a random sample? was it a convenience sample? was it a census? more here (https://bcheggeseth.github.io/Stat155Notes/sampling.html))) and study design (experiment or observational study?)

- NHANES = used somewhat complicated sampling scheme to make sure certain groups were represented in the study; observational study; interviews in homes + health examination in mobile examination center
- Scorecard = close to a census of all US colleges (but not entirely clear why some schools are not included); observational study; college report to federal government
- Election = close to a census of all US counties (just a few are missing, and Alaska is excluded, for unknown reasons); observational study; county election results & census surveys
- Kiva Loans = inferred to be close if not entirely a census of all loans posted on Kiva (details unclear); observational study; loan requests posted on Kiva website
- Powerlifing = seems to be a census of all powerlifting competitions; observational study; recorded by competition organizers and then posted to the OpenPowerlifting database; Kaggle took a snapshot of these data in 2019

Consider the context above. Can you think of any implications that this context might have on the way you will be able to analyze your data or the types of conclusions you will be able to draw? (e.g., if you are working with dollar amounts in the past, note that the purchasing power of a dollar changes over time; if some variables are only collected on a subset of the cases, this limits your conclusions)

There are many possible “correct” answers to this question. Most important is that you’ve spent some time thinking about how the context of your dataset might impact your analysis and/or conclusions. For example, in the election data, what are the implications of the fact that data are reported on the county level rather than for individual voters? Similarly, in the scorecard data, what are the implications of the fact that data are reported on the college level rather than for individual students? What are the implications, for many of these datasets, that they were collected nearly ten years ago?

It’s also important to remember the impact of sampling methods and study design. If random sampling techniques were not used, do you think that the sample is still representative of the larger population? If not, we’ll need to be careful not to generalize our findings too broadly. The study design also impacts our analysis and conclusions: we’ll need to give careful consideration to possible confounding variables, and take care not to draw conclusions that sound too causal.

# Part 3. Data Limitations

Answer the following questions about your chosen project dataset.

## Sampling and Information Biases

Consider the ideal population of interest that you might want to draw conclusions about, and then think about the difference between that ideal population and the data available. In a couple of sentences, describe the ideal population of interest and any sampling biases (who is included and who is excluded?) or information biases (do the values represent the truth?) that might be present in the data available. Wherever possible, use the terminology presented in the online notes and videos.

Sampling biases: comment on whether any sampling bias (https://bcheggeseth.github.io/Stat155Notes/sampling.html) might exist; who is included and who is excluded? are any groups over- or under-represented in this sample?

Information biases: comment on whether any information bias (https://bcheggeseth.github.io/Stat155Notes/information-bias.html) might exist; do the values represent the truth? could there have been any measurement error? or social pressures to answer a question in a particular way? are there questions involved that might have been hard for participants to recall exact answers?

Next, describe how these biases may impact the way that you can analyze these data and/or what conclusions you might be able to draw. (e.g., Even if you analyze these data, what limitations might there be in how we can generalize results to our target population of interest? are there any variables that you might avoid, or interpret differently, because you think they could be affected by information bias?)

> Your response should describe how the sampling and information biases might impact your conclusions (e.g., who you can/can't generalize to; what/how strong of conclusions you'll be able to draw; if values don't represent the truth, are they lower or higher than you should see?) and analyses (e.g., whether you'll need to filter any inaccurate responses; how many cases with missing values you'll need to remove; whether you'll need to avoid certain variables).

# Part 4. Data Viz

Answer the following questions using the dataset that includes *only students' responses* to the Stat 155 Intro Survey.

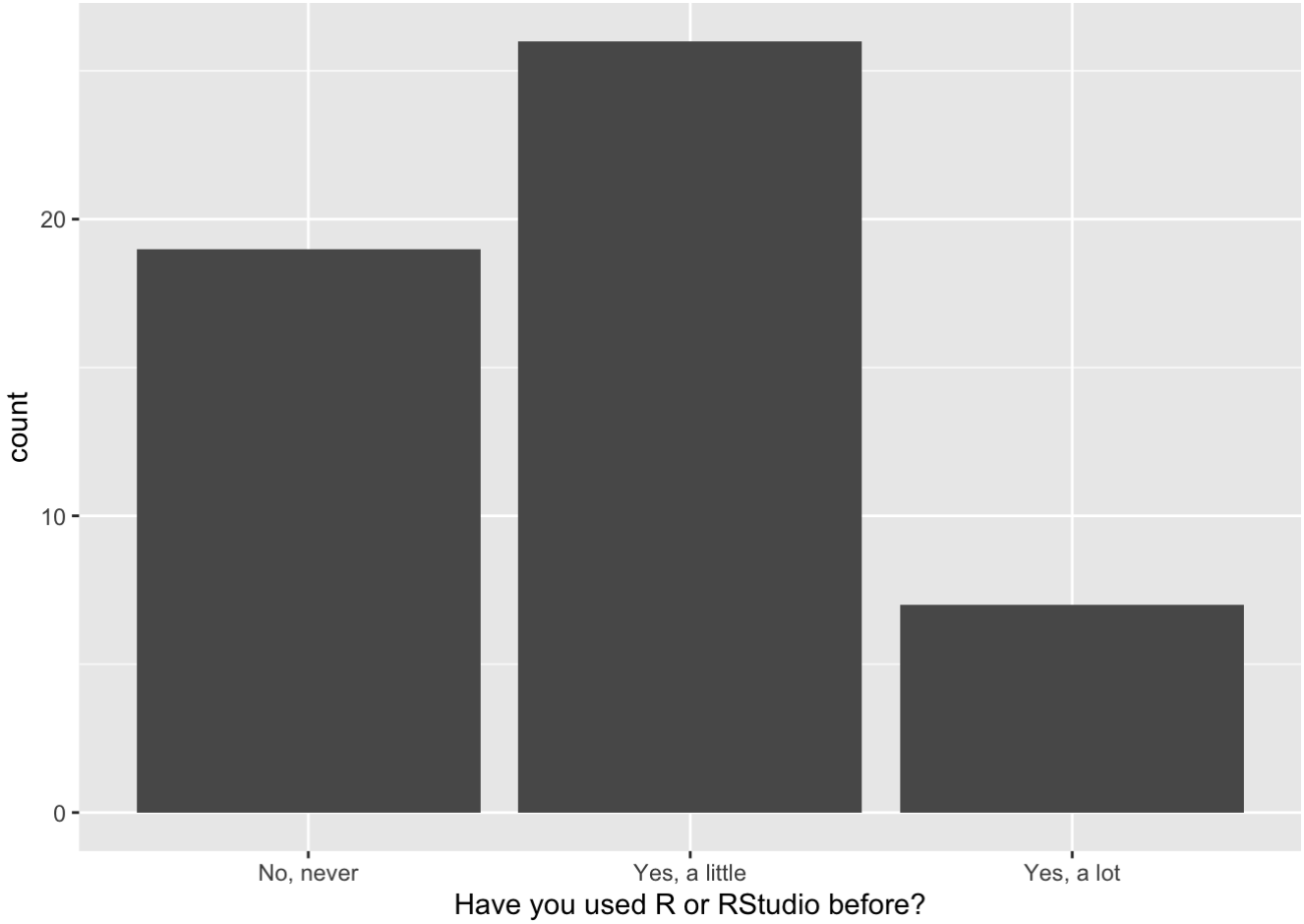
## Design and Interpretation - One Categorical Variable

Make a bar plot and calculate relevant numerical summaries to summarize the responses to the "Have you used R or RStudio before?" question. Then, write up a short paragraph describing what you learn from those summaries. Don't forget to consider the context of how the data were collected and why it might be relevant to the class.

> Remember that in order to make plots using `ggplot()`, we need to load the `ggplot2` package first!

```
# load ggplot2 package
library(ggplot2)

# create a bar plot
students %>%
  ggplot(aes(x = `Have you used R or RStudio before?`)) +
  geom_bar()
```




```
# calculate numerical summaries
students %>%
  count(`Have you used R or RStudio before?`)
```

```
## # A tibble: 3 × 2
##   `Have you used R or RStudio before?`     n
##   <chr>                                <int>
## 1 No, never                               19
## 2 Yes, a little                           26
## 3 Yes, a lot                               7
```

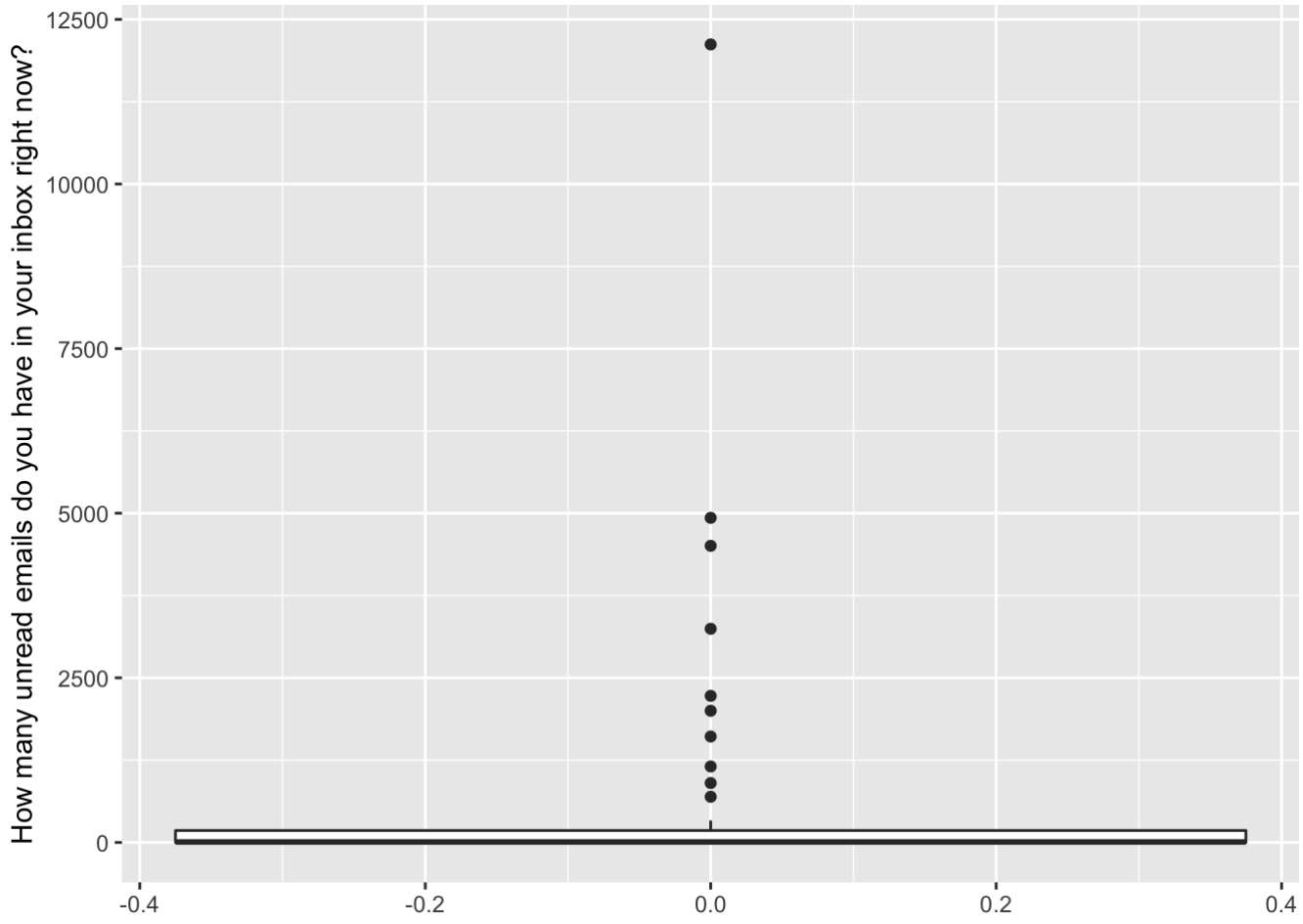
There are 56 students enrolled in Stat 155 with Professor Grinde in Fall 2021, and 52 filled out this survey (so we have some nonresponse bias). Among those that filled out the survey, about one third (36.5%) have never used R before. The vast majority of those students who *have* used R before (26 out of 33) have used R just a little. This survey was a convenience sample and may not be representative of the larger Macalester body. In particular, students with a little prior R experience in other STEM courses may be more likely to sign up to take STAT 155, and thus we should probably not generalize our findings to the larger population and conclude that two thirds of the Macalester student body has used R before.

The students in our class have a range of prior experience with R and RStudio! While about two thirds of the class has used R before in some capacity, most of those students have used it only a little. And a large part of the class has never used R before! This is important for us to remember: if you're new to R, you're not alone! And if you have some experience with R already, remember that not everyone that you'll be working with has the same background as you. Be patient and work to support one other as we all work to build experience and comfort with using R.

# Design and Interpretation - One Quantitative Variable

Make a boxplot and calculate relevant numerical summaries to summarize the responses to the "How many unread emails do you have in your inbox right now?" question. Then, write up a short paragraph describing what you learn from those summaries. Try to write in a way that tells a story rather than sounding like a check list.

```
# create a boxplot
students %>%
  ggplot(aes(y = `How many unread emails do you have in your inbox right now?`)) +
  geom_boxplot()
```




```
# calculate numerical summaries
students %>%
  summarize(Mean = mean(`How many unread emails do you have in your inbox right now?`),
            Median = median(`How many unread emails do you have in your inbox right now?
`),
            Minimum = min(`How many unread emails do you have in your inbox right now?
`),
            Maximum = max(`How many unread emails do you have in your inbox right now?
`),
            SD = sd(`How many unread emails do you have in your inbox right now?`),
            IQR = IQR(`How many unread emails do you have in your inbox right now?`))
```

```
## # A tibble: 1 × 6
##    Mean Median Minimum Maximum    SD   IQR
##   <dbl>  <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1  674.    7.5       0   12120 1945.   180
```

There is a huge amount of variation in unread emails among our class members. The numbers of unread emails range from as low as zero to over twelve thousand. The distribution of unread emails is heavily right-skewed, with most students having a (relatively) manageable inbox, centered around a median of 7.5 unread emails.

Notice that my summary commented on all the key features of this plot: shape (right-skewed), center (median = 7.5), and spread (range = 0 to twelve thousand). However, I tried to write this summary in a way that doesn't just sound like a checklist. There is a narrative with (I hope!) some clear takeaway messages: most students have manageable inboxes, but there is a lot of variation.

# Part 5. Knit

Once you've answered all the questions above, you'll need to `Knit` this document and then save that knitted document as a PDF. See the assignment *Instructions* at the top of this document and on Moodle for more details.