Lab 4 – Class Data Visualization

Assignment Overview

- We'll be exploring our class survey data that we cleaned in Lab 3. This time, we'll focus on visualizations!
- Each row represents one student in the class, and each column is a variable/question from the survey.
- **Don't use your own Lab 3 file** for this assignment—use the cleaned data **provided in the Canvas instructions**.

Formatting Instructions

- Please write and save all code in an R script. Also be sure to save your written responses to upload to Gradescope when finished.

Step 0

- **Open RStudio** (or your RStudio workspace in Posit Cloud) to get started
- **Create a new script** to write and save your code and open/upload into your RStudio workspace. Click the small green plus sign in the top left corner of your RStudio window!
- Don't forget to run `library(tidyverse)` each time you open RStudio.
- Upload the Class_F25 dataset from the Canvas assignment page.

Variable names: Descriptions of question asked

- **phone_os:** Which phone operating system do you have?
- **tiktok**: Do you have Tik Tok downloaded on your phone?
- **instagram**: Do you have Instagram downloaded on your phone?
- **snapchat**: Do you have SnapChat downloaded on your phone?
- **threads**: Do you have Threads downloaded on your phone?
- **reddit**: Do you have Reddit downloaded on your phone?
- **youtube**: Do you have YouTube downloaded on your phone?
- **x**: Do you have X downloaded on your phone?
- **yikyak**: Do you have YikYak downloaded on your phone?
- **favorite_app:** What is your favorite app?
- **phone**: Yesterday, approximately how many minutes was your phone screen active?
- **messages:** Yesterday, approximately how many minutes did you spend with messaging apps on your screen?
- **study:** Yesterday, approximately how much time did you spend studying?
- **sleep:** In the last 24 hours, estimate how many hours you spent sleeping/napping
- **roadtrip:** In miles, what is the longest roadtrip that you have taken in a car or bus before?
- **salary:** What do you think your annual salary will be 20 years from now? Assuming 0 rate of inflation
- **hr_wage:** Consider a fast food restaurant near where you live. If you were looking for a job, what hourly wage would they need to offer before you would consider applying?
- **bpm**: Count how many times your heart beats in one minute
- **rand_number**: Choose a random whole number from 1 to 20
- **musician**: Which singer/band/musician have you been listening to the most in recent months?
- **day**: What day are you filling this survey out?
- **caffeine**: Did you drink a caffeinated beverage yesterday?
- **car:** Do you have a car in town?
- **alarm:** Did you wake up to an alarm this morning?
- **acad_group:** What academic level are you this semester?
- **season:** What is your favorite season?
- **football:** There are 7 home football games this fall. How many do you expect to attend?
- **extroverted:** On a scale of 1 to 4, how extroverted are you? (4 being most extroverted)
- **plans:** What is your plan after finishing your bachelor's program?
- **study_location:** If you had to choose from these options, where would you be most likely to study?

Generative AI policy: You are **encouraged** to use **Github Copilot** or another **external tool** to generate **code**. You should not use code directly from other people—you must be involved in the code generation.

Written responses should be in your own words unless otherwise stated—these questions are designed to get you thinking about interpretations and to develop your conceptual understanding to prepare you for the exams. **All questions in red** should ***not*** be fed to a generative AI tool and should be answered in ***your* own words *without* bullet points**.

Question 1 (4pts)**.** One of the questions on our survey asked students what their favorite social media app was.

Use the table function in R to see how many people selected each app. If you download the starter script, it's already written up for you! **Which 5 apps were selected most?**

Now, let's try visualizing the results while filtering to only include those 5 apps. When making your graph:

- Use ggplot, placed inside a pipe that first filters to only include these 5 apps.
    - *Hint: try using an %in% statement when writing your filter.*
    - *Tip: If you're using the tutorials to help you write code, be sure you checked the "Plotting with pipes" section of the "Data Wrangling with Pipes" tutorial.*
- Have an appropriate title and updated x-axis label
- Set each app to be represented in a different color
- **Export** and save the image to **upload to Gradescope** with your submission

Question 2 (6pts)**.** Are there differences in the amount of time students reported studying in a day based on their academic level group?

Let's first restructure the academic group variable to order the categories as Freshman, Sophomore, Junior Senior (as opposed to what it will do by default—order alphabetically). Use the code setup in the starter script! If set up correctly, note that you won't see a result after running the code. It's an internal restructuring.

Next, let's create a jitter plot to compare students' self-reported study time in a day based on their academic level. When making your graph:

- Add an appropriate title *and* appropriate axes labels
- Each group of dots should have its own color (if you're using an AI tool, feel free to add a line of code to choose the colors manually or choose a color palette!)
- Add `theme_minimal()` as a line of code in your ggplot to change the default background.
- *You do not need to save this plot—we're going to look at it first, change something, and remake it!*

If done correctly, you should see a handful of dots much higher than the others. Reread the question from the variable list. **What mistake do you think this handful of students made when answering?**

Now, recreate the plot, but put it inside a pipe and filter out the values that don't make sense to include.

- Don't forget to remove the dataframe name from the ggplot line as it should now be out front of your pipe!
- Keep the other features from our first draft of this plot
- **Export** and save the image to **upload to Gradescope** with your submission

Question 3 (6pts)**.** Are students' salary expectations associated with their post-graduation plans? To investigate this, we will make a summary table using a pipe.

- Reports the **mean**, **median**, and **standard deviation** in projected salary, with each row of the summary table representing a different potential post-graduation plan.
- Some people have no entry in the salary column. To avoid our statistics returning NA outputs, be sure to add an argument to your summary statistics calculations to **remove NAs**.
    - If you're using the tutorials to help you write code, be sure you checked the "Removing NAs" section of the "Data Wrangling with Pipes" tutorial.
- Have the table list the categories in **descending order by mean**.
- *If not trying the gt package option below,* **copy your console results for the table to Gradescope**

OPTIONAL: If using Copilot or a Gen AI tool, consider asking it to generate the table with the **gt package** to get a nicer, more professionally formatted table!

- Note that you'll need to run install.packages("gt") first. Then library the package to activate it.
- You might also try asking your AI tool to report the values in a financial format.
- **If trying this option, you can export** and save the image to **upload to Gradescope** with your submission.

According to your results, **which grad plan(s)** are associated with the **highest** expected salaries for the **typical, middle student** of the group?

Which grad plan group has the highest standard deviation, and **what do you think is contributing to that**?
Hint: Open the data viewer and sort by salary!

Question 4 (5pts)**.** How does phone screen time vary by day of the week?

First, let's internally structure the day variable to be in **day of the week** order *rather than alphabetical*. Use the template from Question 2, but update the days of the week and variable name appropriately.

Create side-by-side boxplots to compare phone screen time to day of the week. When making your graph:

- Place your plot inside a pipe.
- Add a filter to exclude rows where day is empty (listed as NA).
- Just to change it up, put your numeric variable on the x axis and the categorical variable on the y axis!
- Each boxplot should have its own color (if you're using an AI tool, feel free to add a line of code to choose the colors manually or choose a color palette!)
- Add a different default theme this time (see the list of options under "details" or scroll to visual examples at the bottom!)
- Have an appropriate title, x-axis, and y-axis label
- **Export and save the image to upload to Gradescope with your submission**

Look carefully again at the variable list from earlier to see how the phone screen time question was asked. With that in mind, **which day of the week tends to have the highest median phone screen time use?**

Question 5 (5pts)**.** Watch the video linked below for instructions on what you should make for this question!
url: **https://mediaspace.illinois.edu/playlist/dedicated/1_12nmg9tq/**

When making your graph:

- Place your plot inside a pipe and add a filter to exclude rows where the football games question was not answered (listed as NA).
- You may keep R's default colors, or choose manual colors or a color palette
- Add a default theme (see the list under "details" or scroll to visual examples at the bottom!)
- Have an appropriate title, x-axis label, and fill legend label
- **Export and save the image to upload to Gradescope with your submission**

Does there appear to be any association between these two variables? Briefly explain what you notice in your graph to support your answer!

Question 6 (4pts)**.** When asked to choose a number from 1 to 20 as randomly as they could, how well did the class do at generating a seemingly random set of values?

- Choose a way to represent this data that helps you draw some insight about this!
- Some people inputted numbers outside of the 1 to 20 range. Filter those out of your plot.
- Make sure you have exactly 1 bin or bar for each number option
- **Export** and save the image to **upload to Gradescope** with your submission.

Based on the results, how well do you think the class did at choosing numbers at random?

Propose explanations! **Why** might the results have turned out the way they did? *Any patterns you notice and ideas for why?*

Question 7 (5pts). What's a multivariate question that you have about the class data? Your question should pair two variables together. *You can reuse a variable that we have already used, but the variable* ***combinatio****n should be new.*

Pose a question involving two variables in our class dataset. You can reuse a variable that we

Create an appropriate graph OR summary table (or both!) that helps you address this question.

- Be proactive to filter out any outliers, non-sensical values, or NAs as needed
- Add basic formatting (titles, axes labels, color as appropriate)
- If making a summary table instead of a graph, add appropriate column headers

Briefly describe what you found and how it may help address your question (Do you notice any association?)

To avoid general penalties, be sure that your R script is well organized and concise before uploading. All question numbers should be commented as headings above the relevant question content. Add spaces between different code chunks or sections. Remove redundant code (e.g., librarying a package more than once).