Lab 4 – Class Data Visualization

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please submit your lab report as a **pdf** to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.
- Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- We'll be exploring our class survey data that we cleaned in Lab 3. This time, we'll focus on visualizations!
- Note that each row represents one student in our 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**.

STEP 0

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 4 first: https://stat212-learnr.stat.illinois.edu/
- **Download** the Class_F23.xlsx file to your computer and then **import** into your RStudio session.
- Open a **new R script** to write your code in—this is much easier than trying to code directly in the console!
- Remember to **library(tidyverse)** so that you can use the ggplot function.
- Coding Tip: Remember that R is CaSe AnD sYmBoL_SeNsItIvE. As you code, type in your variable names **exactly** as they appear in the data frame. sleep ≠ Sleep. Grad Plans ≠ Grad_Plans

Variables

- **miles:** Approximately how many miles from Champaign is "home" for you?
- **bones:** How many bones have you broken?
- **phone:** Approximately how many hours did you spend on phone in last 24 hours
- **screen:** Approximately how many hours did you spend on screen other than phone in last 24 hours
- **exercise:** Approximately how many hours did you spend exercise or actively moving in last 24 hours
- **sleep_total:** Approximately how many hours did you spend sleeping or napping in last 24 hours
- **salary:** What do you think your annual salary will be 20 years from now?
- **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
- **day**: What day are you filling this survey out?
- **section**: Which section of the course are you in?
- **coffee**: Have you had coffee in the last 24 hours?
- **residence**: Where did you sleep/stay last night?
- **sleep_qual**: On a scale of 1 to 5, how would you rate the quality of sleep you had in the last 24 hours
- **car:** Do you have a car in town?
- **academ_level:** What academic level are you this semester?
- **grad_plans:** What is your plan after finishing your bachelor's program?
- **rand_letter**: Choose a letter below as "randomly" as you can

Question 1 (5pts)**.** Is there any association between students' heart rates (measured in beats per minute) and whether they drank coffee in the last 24 hours? Create side-by-side boxplots to make the comparison.

Include an image of <u>side-by-side boxplots</u> representing these variables *and please share your code in your report*

- Add an appropriate title *and* appropriate axes labels
- Each box should be a different fill color
- Add whiskers (errorbars) to your boxplots

Briefly address these questions:

- Do you see any difference in BPM between these two groups of students? If so, which group seems to have higher BPM values and by how much?
- Is this the result you expected?

Question 2 (5pts)**.** Next, let's look at the values students reported as their expected salary in 20 years.

Report the **numeric summary** (min, Q1, Q2, mean, Q3, max) for salary expectation for the class.

Include an image of a <u>density curve</u> for this variable here *and please share your code in your report. <u>If you do this correctly, your graph might look empty. Check the questions below to figure out why!</u>*

- Add an appropriate title
- Add a fill color
- ***OPTIONAL:*** If you're curious how to turn off scientific notation and report values in normal notation, you can run library(scales) and add the following line to your ggplot code: scale_x_continuous(labels = comma)

Briefly address these questions

- The middle 50% of students reported expected salary levels between what two values?
- Why does the scale of this plot stretch so high? Are class responses scattered evenly across this range, or more concentrated in one numeric range of this plot? What might be the reason why the graph looks empty? *Hint: sort the salary variable and scroll to the bottom!*

Question 3 (5pts)**.** Are students' salary expectations associated with their plans after graduation?

To investigate this, we will make a summary table using a pipe that reports the mean, median, and standard deviation in projected salary based on students' graduation plans.

Hint*: some people have no entry in the salary column (which creates a default response of "NA"). You'll need to program in a response to remove the NAs when telling R to calculate the statistics.*

Include an image of your summary table *(screenshot or copy+paste the output)*

Include the *code* you used to create that table *(screenshot or copy+paste)*

Briefly address these questions

- Based on the summary stats, does there seem to be any association between these two variables? How might you explain this result in context?
- Which grad plan group has the highest standard deviation, and what do you think is contributing to that? *Hint: sort the salary column and scroll to the bottom!*

Question 4 (5pts)**.** Is there any association between how much time students reported using a screen other than their phone in the last 24 hours and how much time they spent exercising or actively moving?

Include an image of a scatterplot for these variables here, with exercise placed on the x-axis *and please share your code in your report.*

- Build your plot inside a pipe that only includes students whose daily activities (phone time, other screen time, exercise time, and sleep time) ***add up*** to less than or equal to 24 hours
- Add an appropriate title and axes labels

Do you see any association between these variables? How might you explain this relationship?

Question 5 (5pts)**.** When asked to choose a letter or number at random, how did the class do?

Create a univariate barplot that showcases the results of the random letter question *and please share your code in your report.*

- Fill each bar with a different color
- Add an appropriate title and x axis label

Create a histogram that showcases the results of the random number question *and please share your code in your report.*

- Filter out any numbers outside the range from 1 to 20
- Set your histogram to have 20 bins
- Choose distinct fill colors and border color for your histogram bins
- Add an appropriate title and x axis label

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

Question 6 (5pts)**.** Let's explore the relationship of two categorical variables: academic level and whether or not a student owns a car. *Consider whether these are categorical or numeric variables and choose an appropriate visualization to represent them!*

Intermediate step: Before creating the graph, note that the academic level variable will list the categories *alphabetically*, rather than in order of *seniority*. Use the following template to complete a custom re-ordering of the levels. Identify your data frame name and variable name correctly and plug that into each slot. Then run this code to restructure the variable. Nothing will output—but you'll see in your graph that the order is correct! If you make a mistake and accidentally messed up something with the data, try re-importing the data again. *This part is only worth 1 point, so even if you fail to re-order the categories, just move on to the graph!*

```
Data$variable = factor(Data$variable, levels = c("Freshman", "Sophomore", "Junior",
"Senior or grad student"))
```

Include an image of your plot *and please share your code in your report.*

- Add an appropriate title and an appropriate axis label for any axis a variable is assigned to
- You are welcome to add or adjust any other features if appropriate

Briefly address this question: Does there appear to be any association between students' academic level and car ownership status? Briefly explain what you notice in your graph to make this conclusion.

Question 7: What's a multivariate question that *you* have about the class data?

Pose a question involving two variables in our class dataset.

Create an appropriate visualization and/or summary table that helps you address this question.

- Be proactive to filter out any outliers as needed
- Please format any visualizations (titles, axes labels, color as appropriate)
- If making a summary table, please add appropriate column headers

Briefly answer your question based on your findings in the data