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_F24.xlsx file to your computer and then **import** into your RStudio session.
- Open a **new R script** to write your code in ***or*** try out the RMarkdown file provided if you're ready to streamline!
- Remember to **library(tidyverse)** so that you can use the ggplot function and pipes.

Variables

- **miles:** In miles, what is the longest roadtrip that you have taken in a car or bus before?
- **bones:** How many bones have you broken?
- **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
- **salary:** What do you think your annual salary will be 20 years from now?
- **screen**: Yesterday, approximately how many hours was your phone screen active?
- **sleep:** Approximately how many hours did you spend sleeping or napping in last 24 hours
- **rand_num**: Choose a random whole number from 1 to 20
- **alarm:** Did you wake up to an alarm this morning?
- **acad_level:** What academic level are you this semester?
- **car:** Do you have a car in town?
- **plans:** What is your plan after finishing your bachelor's program?
- **caffeine**: Did you drink a caffeinated beverage yesterday?
- **tiktok**: Do you have Tik Tok downloaded on your phone?
- **section**: Which section of the course are you in?
- **day**: What day are you filling this survey out?
- **val_sal**: Were you the valedictorian or salutatorian at your high school
- **rand_let**: Choose a letter below as "randomly" as you can

*If you are **not** using RMarkdown, remember to use an **R Script** to write, run, and save your code. Coding directly into the console is much harder and doesn't afford you the same ability to edit, reuse, and save code!*

Question 1 (5pts)**.** Is there any association between whether students have Tik Tok installed on their phone and how much time their phone screen was active? Create side by side boxplots to make the comparison.

Include an image of <u>side-by-side boxplots</u> representing these variables*. Sharing code is optional and may help with partial credit if your visualization is not correct.*

- 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:

- Briefly describe what you notice from this data visualization
- Is this the result you expected? Do you have any possible explanations for what you are seeing in the data?

Question 2 (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 across each of these grad plans groups. *Just these three statistics!*

***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. Check out the R tutorial where we learned about pipes!*

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

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

Briefly address these questions

- Based on the summary statistics, does it seem like expected salaries are associated with students' grad plans? Explain which statistics are helping you answer this.
- In context, do these differences make sense to you?
- Which grad plan group has the highest standard deviation, and what do you think is contributing to that? *Hint: look at the data viewer and sort by salary*

Question 3 (5pts)**.** Now, let's *visualize* students' expected salaries alongside their proposed grad plans.

Use overlapping densities to visualize the expected salaries and allow the plot to create separate density curves for each grad plan option. *Additionally*, let's filter the plot to only include proposed salaries below 2 million and only include responses from students who chose one of 1) A Job, 2) Graduate School, and 3) Medical School. Code your plot inside of a pipe to do this for full credit.

Include an image of your overlapping density curves here. *Sharing code is optional and may help with partial credit if your visualization is not correct.*

- Add an appropriate title
- Add at least some transparency to your overlapping densities using an alpha argument
- ***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)

TIP: This graph has a lot of pieces. Start with a simple visualization and add pieces one by one. Remember that plots inside pipes don't need the data argument listed again.

What do you notice about the association of these variables through this visualization?

Question 4 (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. *Sharing code is optional.*

- 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. *Sharing code is optional.*

- Filter out any numbers outside the range from 1 to 20
- Set your histogram to have 20 bins
- Choose a distinct fill color 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 5 (5pts)**.** One thing I wondered about is whether there might be a correlation between how much time students reported sleeping in the past 24 hours and how much time their phone screen was on yesterday. Can you make a plot that would help us see whether these two variables are associated?

Include an image of a plot for these variables here. *Sharing code is optional.*

- Add an appropriate title and axes labels
- Adding additional formatting or color is optional!

Do you see any association between these variables? Justify why or why not.

Question 6 (5pts)**.** Let's explore the relationship of two categorical variables: academic level and whether or not a student has consumed caffeine in the last 24 hours. *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. *Sharing code is optional.*

- 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 whether they consumed caffeine? 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 address what you found based on what you calculated or visualized.