Lab 4 – Class Data Visualization

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

- Be sure that all **group members** are **added** in your submission to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.

Assignment Overview

- You'll be making several visualizations with the data and answering a few questions to learn more about the students taking STAT 200 & 212 this semester!
- We'll be using the class survey data we cleaned in Lab 1. Each row represents one student in our class, and each column is a variable/question that the survey asked.
- Don't use your Lab 1 file for this assignment—use the cleaned data provided in the Canvas instructions!
- When finished, upload a pdf of your report to Gradescope.

STEP 0

- Open the **starter script** on Canvas. *Or you can create your own script from scratch to work from!*
- **Import** the Class_Spring_2022.xlsx file into your RStudio via the files pane.
- 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

- **Home_Dis:** Approximately how many miles from Champaign is "home" for you? (where you lived before starting school here)
- **Bones:** How many bones have you broken?
- **Lifespan:** How many years old do you THINK you will be when you die?
- **Sleep:** How many hours of sleep did you get last night? (Round to 1 decimal place, Example: 7.5)
- **Salary:** What do you think your annual salary will be 20 years from now? (no $ or , please--just the number).
- **BPM**: Count how many times your heart beats in one minute.
- **Section**: Which section of the class are you enrolled in?
- **Residence**: Where did you sleep/stay last night?
- **Coffee**: Have you had coffee in the last 24 hours?
- **Overseas:** Have you traveled overseas before?
- **Academ_Level:** What academic level are you this semester?
- **Car:** Do you have a car in town?
- **Grad_Plans:** What is your plan after finishing your bachelor's program?
- **Top_Musician:** Name one musician/composer/artist/band you enjoy listening to right now (Please don't write "why" or other info here--just the artist!).

(4pts) Question 1. How much sleep did the class report getting the night before taking the survey?

a) Create a density curve of our class' previous nights' sleep. Your density curve should include…
   - An appropriate title
   - A fill color (any defined color is fine)

   Include the image of the density curve here. *No need to report your code unless you had issues making the graph and just want some partial credit!*

b) Report the **mean** amount of sleep the class reported getting, as well as the **standard deviation** in sleep.

(4pts) Question 2. Are students who reported having coffee in the last 24 hours reporting different amounts of sleep on average than the non-coffee drinkers?

a) Create side by side boxplots to compare these two groups. Your plot should include…
   - An appropriate title
   - Each box should be a different fill color
   - Add whiskers (errorbars) to your boxplots

   Include the image of the boxplots here. *No need to report your code unless you had issues making the graph and just want some partial credit!*

b) **Briefly address this question** (15 -30 words). Do you see evidence for a difference in sleep between these groups, or does each group have similar sleep distributions?

(6pts) Question 3. Do reported sleep levels vary by academic level?

a) Use the Academic Level variable to create a strip chart (with jittering) to compare these four groups. Your graph should include:

- An appropriate title
- Each group's data points should be a unique color (freshmen one color, sophomore one color, etc.)

But before you put that graph in your report...notice now that the academic level variable lists groups in alphabetical order. But we want to list these groups in a custom order! (Freshman, Sophomore, Junior, Senior). Run the following code to to "re-factor" this variable to have this specific order.

```
Class_Spring_2022$Academ_Level = factor(Class_Spring_2022$Academ_Level,
                                 levels = c("Freshman", "Sophomore", "Junior", "Senior"))
```

Note that when you run this, it will seem like nothing has happened! No output will come up. That's because this is internally restructuring the variable. You will only see the result of this change once you have used this variable again.

Now run your plot again with the re-ordered categories! **Include the image of the strip chart here.**

b) We don't have that many seniors, so let's look more closely at comparing the Freshmen and Juniors.
- Create a subset called "Freshmen" that only includes the rows of data of self-reported Freshmen. Do the same with the Juniors.
- *If done correctly, you will see a dataframe in your global environment for each group—Freshmen with 153 observations and Juniors with 62.*
- Next, use the summary function to report the 5-number summary + mean of sleep for each group.

Include the summary output for each subset here

c) **Briefly address this question** (15 – 30 words). Is there any difference between the Freshmen and Juniors regarding amount of sleep? Use your graph and/or numeric summaries to support your answer.

(4pts) Question 4: Let's now consider how sleep might associate with a numeric predictor, like students' reported heart rate.

a) *Create a scatterplot* to compare sleep across heart rate beats per minute (BPM). Your graph should include:
- An appropriate title
- Color all points the color of your choice!
- Sleep should be on the y axis here.

Include the image of the scatterplot here.

b) **Briefly address this question.** Do you see much association between students' BPM and their reported sleep? *Hint: Association here would mean that as BPM goes up, Sleep tends to also go up or go down. No association means that sleep has a similar mean across the range of BPM values.*

(4pts) Question 5: Are upper classmen more or less likely to have a car than lower classmen?

a) Create a 100% stacked barplot to compare the car ownership status of students by their academic level. Your graph should include…
   - An appropriate title
   - A white border color around the bars.
   - Academic level should be on the x or y axis, and car ownership should be represented through blocks of color

   Include the image of the barplot here.

b) **Briefly address this question.** How would you describe car ownership likelihood across academic level?

(3pts) Question 6: How do proposed salary values compare across each set of possible post-graduation plans?

Create a strip chart to compare people's expected Salary based on their proposed plan after graduation. Your graph should include:

- An appropriate title
- Plot your strips **horizontally**, use the **jitter** option, and choose a height of **0.2**
- Don't color the points by Graduation—the legend will take up valuable space, so let's not bother with it!

Due to some outlier values, it's difficult to make a good comparison visually. Let's subset out some of the very high salary values *purely for the sake of a clearer visual*. Create a subset called Salary that only includes rows where expected salary was less than 5 million dollars.

- *Hint: when writing 5 million in your code, record it as a number with no commas!*
- If done correctly, you will see this subset in your global environment with 319 observations

a) **Include your subset code here in your report**.

b) Create your plot again with this new subset Salary. **Include the image of this strip chart here.**

(5pts) Question 7: It might be difficult to compare these groups by points alone. Summary measures (like boxplots) could be additionally helpful. Additionally, the NA group is not meaningful to include (these are folks who left this question blank).

a) Create a subset called Grads that eliminates rows where Grad_Plans is equal to NA.
   - Your Grads subset should still only include rows with Salary below 5 million. You might find it easier to make a subset starting from your Salary subset!
   - Hint: Instead of selecting all other Grad Plans categires in your subset, just say Grad_Plans is not equal to NA. *Remember the logical symbol for not equals?*
   - If done correctly, you will see this subset in your global environment with 315 observations

   Include your subset code here in your report.

b) Now create side by side boxplots to compare Salary across grad plans. Your boxplot should be created from the Grads subset you just created. It should include:
   - An appropriate title
   - Still no color for this one! We need the space, so we want to avoid having an unnecessary legend.
   - Whiskers are optional for this plot—if you like them, you can add them!

   Include the image of the boxplots here.

c) **Briefly address this question**. What group (or groups) appear to expect the lowest median salaries? Which group (or groups) expect the highest median salaries?

Bonus Opportunity: *Go to the "Bonus! Create your own graph" assignment in the Chapter 9 module of Canvas and post your own graph with a short description. Do not post this here in your report—it needs to be posted in the canvas portal to receive bonus credit!*