Lab 5 – 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 1. 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 1 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 5 first: https://stat212-learnr.stat.illinois.edu/
- **Download** the Class_S23.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

- **dist:** Approximately how many miles from Champaign is "home" for you?
- **bones:** How many bones have you broken?
- **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?
- **sleep:** How many hours of sleep did you get last night?
- **bpm**: Count how many times your heart beats in one minute.
- **shower**: How many minutes do spend in the shower during a typical showering period?
- **salary:** What do you think your annual salary will be 20 years from now?
- **travel**: Have you traveled overseas before?
- **academ_level:** What academic level are you this semester?
- **academ_year:** What year are you in school?
- **car:** Do you have a car in town?
- **grad_plans:** What is your plan after finishing your bachelor's program?
- **coffee**: Have you had coffee in the last 24 hours?
- **residence**: Where did you sleep/stay last night?
- **section**: Which section of the course are you in?
- **day**: What day are you filling this survey out?
- **letter**: Choose a letter below as "randomly" as you can

Question 1 (5pts)**.** Are students who reported having coffee in the last 24 hours reporting different amounts of sleep as compared to non-coffee drinkers?

Include an image of side-by-side boxplots representing these variables**.** *(sharing your code is optional)*

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

Briefly address these questions (suggested: 30-50 words):

- Do you think drinking coffee in the last 24 hrs explains much variability in students' reported sleep?
- 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 density curve for this variable here *(sharing your code is optional)*

- Add an appropriate title
- Add a fill color (change the fill color from the default "white" option it currently has)
- Use a plot theme
- *All other features optional!*

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? *Hint: sort the salary variable and scroll to the bottom!*

Question 3 (5pts)**.** Is there a difference in the salary expectations of students who have been overseas as compared to those who haven't?

Include an image of a jitter plot for these variables here, with salary placed on the y-axis *(sharing your code is optional)*

- Color the points based on which travel group they are in (you can use the default colors or choose custom colors)
- **jitter** your points at a width of **0.05**
- Use the **limits** argument to set the y axis to only span from 0 to 1 million dollars (this will leave out the 4 highest values and make the consensus data much easier to visualize!)
- Set the y axis breaks to be in increments of 100 thousand dollars
- Add an appropriate title and axes labels
- *All other features optional!*
- *OPTIONAL: If you're curious how to turn off scientific notation and report comma form, try librarying the* `scales` *package and add* `labels = comma` *to your scale function. https://www.geeksforgeeks.org/change-formatting-of-numbers-of-ggplot2-plot-axis-in-r/*

Question 4 (5pts)**.** Using a dplyr pipe, create a summary table that calculates the mean and median salary by travel. Add a filter option to only include salary levels below 1 million dollars (we will set a cut-off there so that our mean values aren't too susceptible to crazy high values). When you are done, you should have 4 values in a table style output, showing the mean and median salary of the "no" responses, and those for the "yes" responses.

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

- Do you think travel status explains any variability in students' projected salaries?
- What ideas or explanations do you have for any association or lack of association you see in the data?

Question 5 (5pts)**.** Are students' minimum wage expectations associated with academic level?

Intermediate step*: First*, 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 pipe output that the order is correct! *If you make a mistake and accidentally messed up something with the data, try re-importing the data again.*

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

To investigate this, we will again make a summary table using dplyr that reports the mean, median, and standard deviation in wage based on academic level.

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 academic level and minimum wage expectations for fast food jobs? How might you explain this result in context?
- Why might the senior/grad students have such a high standard deviation compared to other groups? *Hint: sort the wage column and scroll to the bottom!*

Question 6 (5pts)**.** When asked to choose a letter at random, how did the class do? Create a univariate barplot showcases the results of the random letter question.

Include an image of your barplot *(sharing your code is optional)*

- Fill each bar a different color
- Use a color palette (or custom colors) for this plot
- Use a plot theme of your choice
- Add an appropriate title

Briefly comment on the plot. What do you notice? Is this what you expected? *I would ask you if you have any explanation ideas, but I honestly have none! I've replicated this two semesters now and still don't know why this is what results. So if you have one, tell me!*

Question 7 (5pts)**.** Let's explore the relationship of two categorical variables: academic level and whether or not a student owns a car. Create the appropriate graph to represent these two variables.

Include an image of your plot *(sharing your code is optional)*

- Use a color palette (or custom colors) for this plot.
- Add an appropriate title and an appropriate axis label for any axis a variable is assigned to
- Use a plot theme of your choice
- Use the theme function to center and **bold** the plot title

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.

Bonus Opportunity: *Go to the "Bonus! Create your own graph" assignment in the Chapter 8 module of Canvas and post your own graph with a short description.*

- *Your graph should be multivariate (at least 2 variables represented)*
- *Your graph should be based on the Class data we used in this lab*
- *Your graph should be different from any graphs requested in this assignment.*
- *Your graph should Include a title (and any other formatting you wish to use)*
- *Your graph should be accompanied by a brief description/interpretation of what you notice.*

*Do not post this in your report—it needs to be posted in the **bonus portal** on **canvas** to receive bonus credit.*