# 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.
- 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 from the survey.
- **Don't use your own 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

- **Download** the Class_F22.xlsx file to your computer and then **import** into your RStudio session.
- 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? (where you lived before starting school here)
- **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?
- **hr_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.
- **travel**: Have you traveled overseas before?
- **class**: Which class and section are you enrolled in?
- **car:** Do you have a car in town?
- **coffee**: Have you had coffee in the last 24 hours?
- **academ_level:** What academic level are you this semester?
- **academ_year:** What year are you in school?
- **residence**: Where did you sleep/stay last night?
- **grad_plans:** What is your plan after finishing your bachelor's program?
- **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!).

**(5pts) Question 1.** How much sleep did students report getting the night before they filled out the survey?

Report the **numeric summary**, as well as the **standard deviation** in sleep for the class.

**Include the image of a density curve** for this variable here *(sharing your code is optional)*

- Add an appropriate title *and* an appropriate x axis label.
- Add a fill color (change the fill color from the default "white" option it currently has)
- Use a plot theme

**Briefly answer these questions**:

- In what range are the middle 50% of students reporting their sleep time to be?
- What is the sleep amount reported by the middle student?
- Would you say these results are consistent with what you would expect, or do they differ?

**(5pts) 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? Create side by side boxplots to compare these two groups.

**Include the image of the boxplots here.** *(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

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

- Do you think coffee drinking explains any variability in students' reported sleep?
- What do you think about these variables, and is this the result you expected?

**(5pts) Question 3.** Next, let's look at the values student reported as their expected salary in 20 years.

Report the **numeric summary**, as well as the **standard deviation** in salary expectation for the class.

**Include an image of a histogram** for this variable here *(sharing your code is optional)*

- Add an appropriate title and x axis label
- Add a border color and a fill color (change the fill color from the default "white" option it currently has)
- Use a plot theme

**Briefly address these questions**

- In what salary range did the middle 50% of students report?
- Why does this plot stretch so high? Is this range reflective of most class responses?

**(5pts) Question 4.** Let's try comparing students' expected salaries based on whether or not they have traveled overseas before.

Create a strip chart for these two variables, and place salary on the y axis.

**Include an image of a <u>strip chart</u>** for these variables here *(sharing your code is optional)*

- Color the points based on which travel group they are in
- **jitter** your points at a width of **0.05**
- Use the limits argument to set the y axis to only span from 0 to 2 million dollars (not necessarily because higher responses are not valid, but because it's difficult to **visualize** those responses).
- Add an appropriate title and axes labels
- Use the `theme_bw` plot theme.

**(5pts) Question 5.** 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 10 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 (either directly copied, or screenshot)** of your summary table

**Copy (or screenshot) the <u>*code*</u>** you used to create that table

**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?

**(5pts) Question 6.** Let's explore how students' expected salary might relate to both their post-graduation plans and whether they have previously traveled overseas.

While there are several categories in the survey for post-graduation plans, let's focus on the three categories represented by about 85% of the class: A Job, Graduate School, and Medical School.

Start by creating a subset called `Grads` that just includes the students who said one of those three categories for that question. Then create side by side boxplots using this subset to compare graduation plans against projected salary.

**Copy (or screenshot) the *code*** you used to create that subset

**Include an image of your side-by-side boxplots**

- Color the fill of the boxes based on grad plans **AND** use a custom or pre-built color palette!
- Use the limits argument to set the y axis to only span from 0 to 2 million dollars
- Add an appropriate title and axes labels
- Use a plot theme of your choice that **wasn't** used in the tutorial for this section (no `theme_classic` or `theme_bw`)
- Hide the legend

**Copy (or screenshot) the *code*** you used to create that plot

**Briefly address this question:** Does there appear to be any association between students' graduation plans and projected salaries?

**(5pts) Question 7.** Finally, 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.

*But first*, the academic level variable will list the categories *alphabetically*, rather than in order of *seniority*. Check the "Re-ordering Categories" section of the "Customizing with ggplot2" tutorial for help with how to restructure the variable. Identify your data frame name and variable name correctly, then run this code to restructure the variable—then you can make your plot!

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

**Include an image of your plot**

- Use color appropriately
- Add an appropriate title and an appropriate axis label for any axis a variable is assigned to
- Use a plot theme of your choice that **wasn't** used in the tutorial for this section (no `theme_classic` or `theme_bw`)
- Hide the legend
- Use the theme function to center and bold the plot title

**Copy (or screenshot) the *code*** you used to create that plot

**Briefly address this question:** Does there appear to be any association between students' academic level and car ownership status?

**Bonus Opportunity:** *Go to the "Bonus! Create your own graph" assignment in the Chapter 9 module of Canvas and post your own multivariate 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!*