**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ B00#:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

In the steps below, you will get a chance to practice many of the different data visualizations we talked about in the previous class. There are two types of questions:

(a) **Show me your screen,** where your TA take a quick look at your screen to see if everything looks right. This is for situations where writing an answer in pen would be difficult (e.g., G Power outputs, graphs, database restructuring). If things look good, the TA will give you a stamp on your worksheet.

(b) **Write in the answer.** Usually, it will ask you to calculate some manner of statistic and report it on the sheet. This will verify whether or not you did the task correctly. Though you don’t have to submit syntax for labs, I *strongly recommend* you save all of your R syntax for your own benefit (you may need it for an assignment later)!

**Data Visualization**

**Exercise to Walk Through Together**

1. Create a scatter plot using **base R** using the built-in ‘iris’ dataset to visualize the relationship between Sepal Length and Sepal Width.

* Load the iris dataset that is built into R.
* Plot a scatterplot using Sepal Length on the x-axis and Sepal Width on the y-axis.
* Add a main title "Sepal Width vs. Sepal Length".
* Color the points based on the Species.
* Add a legend to the graph to label the different Species groups.

1. To see the difference between graphing in base R vs. tidyverse, let’s now do the exact same plot in ggplot2.

**Questions & Tasks**

-------------------SECTION 1: Practicing the Basics of Visualization--------------------------

You can begin your practice of plotting in the tidyverse with a basic scatter plot.

1. Use ggplot2 to create a scatter plot using the mtcars dataset (built into R), showing the relationship between wt (weight) and mpg (miles per gallon), colored by gear (number of gears).

* Load the mtcars dataset
* Plot wt on the x-axis and mpg on the y-axis, ensuring the point size is appropriately large and legible
* Color the points based on gear
* Add appropriate titles and labels

1. Now let’s create a boxplot with the same dataset for the variable ‘hp’ which represents horsepower.

* Load the mtcars dataset
* Create a boxplot that is slightly transparent for the variable ‘hp’
* Add appropriate titles and labels

1. Create a density plot for the mpg variable, colored by the number of cylinders (cyl). Since cylinders is not a factor in the dataset, make it into one (i.e., a categorical variable).

* Load the mtcars dataset
* Convert the cylinder variable into a categorical variable using a pipeline
* Feed the result into a density plot function with mpg on the x-axis and different cylinders represented by colour

1. Use the ‘patchwork’ package to combine your three plots into a single plot. Give it

appropriate titles and labels.

1. Save your final graph into a 8x10 inches png file with a pixel density of 700. Name it as you wish.

**SHOW ME YOUR SCREEN WITH THE PLOT.**

-------------------SECTION 2: Learning new graph types and customizations--------------------------

For this next section, you will need to download an SPSS dataset from Brightspace.

**Datafile Names on Brightspace:** qos.wide.sav

**Description:** Response-order effects refer to differences in participant responding based on the order of questionnaire response options. Theory predicts primacy effects (options presented earlier are more frequently endorsed). We predicted primacy effects for unipolar scales related to alcohol. Undergraduates (N = 791) completed questionnaires on drinking motives, personality, and alcohol problems. Participants were randomly assigned to ascending (highest-to-lowest) or descending (lowest-to-highest) response-order conditions. […] Overall, we found evidence for small primacy effects in undergraduate students.

**Number of cases:** 791

**Key Variables:**

1. **sex.r** (-1 = Female, 1 = Male)
2. **caucasian**: (-1 = Non-Caucasian, 1 = Caucasian)
3. **condition**: (0 = Ascending Order, 1 = Descending Order)
4. **HOP.total** – Hopelessness is a personality trait including bleak expectations about oneself and the future, perceived inability to change future outcomes, and chronic feelings of despondency. Numerical variable that can range from 7 to 28.
5. **AS.total**: Anxiety sensitivity. Anxiety sensitivity represents a fear of anxiety-related physical sensations due to an unrealistic expectation they could lead to catastrophic consequences or loss of control. Numerical variable that can range from 5 to 20.
6. **COP.total**: Drinking to cope motives (i.e., drinking alcohol to cope with negative emotions). Numerical variable that can range from 3 to 12.
7. **CON.total**: Conformity motives (i.e., drinking alcohol to fit in and/or avoid social sanctions) that can range from 3 to 12.

**Before starting:** Make sure to install and load the following packages: tidyverse, ggthemes, foreign, and ggExtra. You will need to read in the SPSS dataset as previously taught.

1. a) Create a scatterplot with hopelessness predicting coping motives, split by sex (male vs. female), either with color or facets. Make sure to include a line of best fit with standard error. Do you notice anything unusual about the graph? **WRITE IN YOUR ANSWER BELOW.**

**Not only is the plot graphing NA’s on it separately from the male/female, it’s the graph is also very grid like which is odd.**

b) Because there are missing values for sex, it will plot the NA values separately, which is undesirable (try it out!). Thus, you will need to omit the missing values. You can create a new dataset with these three missing values removed with the following code using the ‘dplyr’ package:

mydata2 <- mydata %>%

filter(!is.na(sex.r))

This will remove all rows where the variable “sex.r” has missing values. However, we still have the problem of what we call ‘overplotting’ where many points overlap. This is often due to discrete responses being in the dataset (e.g., Likert scale responses where many people choose the same answer). In order to take care of something like this we can use a “jitter” graph which is a scatterplot that slightly moves or “jitters” the points to shift them away from the original position. It essentially creates random noise in the data. The points are not exactly representative of their value but it removes overlap and allows us to see exactly how the points are spread.

Since you didn’t learn this in the lecture, this is an opportunity to extrapolate your knowledge and learn from a help file. Use the following link to access the manual for ggplot2 and figure out how to create a jitter graph (hint: it’s not very complicated).

Create the scatterplot again without the missing values and having jittered it. Include a Loess line and SE in your graph. **SHOW ME YOUR SCREEN WITH THE PLOT.**

<https://cloud.r-project.org/web/packages/ggplot2/ggplot2.pdf>

2. Create a histogram for “Conformity Motives.” Make sure to play with the bin numbers and scale limits to get an accurate graph. Answer the following questions based on this histogram: (a) What is the optimal number of bins for these data? (b) What is minimum and maximum? (c) Do these data approximate a “normal” z distribution? If not, how does it deviate from that shape? **WRITE IN YOUR ANSWER BELOW**

**The optimal number of bins for this data is 1 bin for the binwidth or 10 bins. The minimum number appearing on my graph is 2.5 and the max is 12.5. This is not a normal distribution at all, it is very skewed positively.**

3. Next, I want you to calculate an error bar plot comparing Anxiety Sensitivity for men and women. **SHOW ME YOUR SCREEN WITH THE PLOT.**

-------------------SECTION 3: Bonus--------------------------

1. In the next two questions, I’m going to give you a general question to answer with data visualization, but I’m not going to say exactly what kind of plot you should use.

A researcher is conducting a statistical test, and wants to compare the medians of the Caucasian participants to the non-Caucasian participants on the “Anxiety Sensitivity” variable. Create a data visualization that highlights the comparison between the medians. Does it look like there is a difference between the medians in the two groups? How do you know? **WRITE IN YOUR ANSWER BELOW.**

Note that this question requires you to adapt one of the plots you learned about to include a special function called “notch”. I won’t tell you what graph it is, but consider whether the ggplot2 manual might help you find a type of graph where a “notch” can be created.

**It doesn’t look like there is a meaningful difference between Caucasian and non-caucasian median scores of anxiety sensitivity, the two notches overlap each other greatly.**

2. Imagine that you want to look at the spread or distribution of the variable in different groups. You want to see if the spread (i.e., the variance) is about the same in every group (i.e., homogeneity of variances). Create a plot that highlights the spread / variance of the “COP.total” variable split across 4 groups: (a) male, ascending condition; (b) male, descending condition, (c) female, ascending condition, (d) female, descending condition. **SHOW ME YOUR SCREEN WITH THE PLOT.**