**Homework 1: Due Wednesday, February 8 at 11:59pm**

For this assignment, you will be analyzing the **College** data frame included in the **ISLR** package, which contains 777 observations on 18 features of U.S. colleges from the 1995 issue of U.S. News and World Report. The goal of this assignment is to become more familiar with R and data visualization, and therefore you will need to use resources such as the R **help()** function and **ggplot2** cheat sheet to learn how to create different types of plots. All analyses must be performed in R using **tidyverse** and other packages discussed in class. Provide your responses (including R code pasted in text format) in the designated spaces in this Word document, and then save it as a pdf and upload it to Canvas.

1. **[15%]** Generate jittered scatterplots of out-of-state tuition as a function of the percentage of new students who are from the top 10% of their high school classes that are faceted by public colleges (left facet) and private colleges (right facet). Overlay these scatterplots with straight lines (not smoothed lines, hint: **help(geom\_smooth)**) containing 95% confidence bands, with lines colored differently for each facet. Add rug plots to both facets, using the same colors as for the lines, and set the background of the plotting region to white rather than the default color of gray.

**Provide code below:**

**Provide figure below:**

1. **[15%]** Compute two correlation coefficients between the percentage of new students from the top 10% of their high school classes and out-of-state tuition – one for public colleges, and one for private colleges. What do both correlation coefficients say about the general relationship between these features? Is this relationship stronger for public or private colleges? Provide an explanation for this difference that is based on your examination of the plot from question 1 (hint: compare points, confidence intervals, and/or rug plots).

**Provide code and console output below:**

**Provide answer below:**

1. **[15%]** Generate a pair of box plots showing the distributions of out-of-state tuition for public and private colleges, using a different color to fill each box plot. Include notches and remove outliers (hint: **help(geom\_boxplot)**) for easier comparison of the distributions between the two groups, and set the background of the plotting region to white rather than the default color of gray. Non-overlapping notches of box plots tell us that the distributions of two groups are significantly different from one another. Based on this information, are the distributions of out-of-state tuition significantly different between public and private colleges? What can you conclude about the difference between out-of-state tuition at public and private colleges?

**Provide code below:**

**Provide figure below:**

**Provide answer below:**

1. **[15%]** Generate a pair of violin plots showing the distributions of out-of-state tuition for public and private colleges, using a different color to fill each violin plot. Set the background of the plotting region to white rather than the default color of gray. What is the difference in information displayed by box plots and violin plots? Based only on these violin plots, what can you say about the difference between out-of-state tuition at public and private colleges?

**Provide code below:**

**Provide figure below:**

**Provide answer below:**

1. **[15%]** Generate a pair of jittered strip plots (hint: same geoms as for scatter plots) showing the distributions of book costs for public and private colleges, with points colored in gray. Overlay these strip plots with box plots, using a transparency level of 0.5 and a different color to fill each box plot. Include notches, color outliers in red, and set the background of the plotting region to white rather than the default color of gray. Based on these plots, are book costs generally higher at public or private colleges? Is the college with the highest book costs public or private?

**Provide code below:**

**Provide figure below:**

**Provide answer below:**

1. **[15%]** Generate a pair of jittered strip plots showing the distributions of total expenses (out-of-state tuition, room and board, books, and personal spending) for public and private colleges, with points colored in gray. Overlay these strip plots with box plots, using a transparency level of 0.5 and a different color to fill each box plot. Include notches, color outliers in red, and set the background of the plotting region to white rather than the default color of gray. Based on these plots, are total expenses generally higher at public or private colleges? Are there more outliers in total expenses for public or private colleges?

**Provide code below:**

-

**Provide figure below:**

**Provide answer below:**

1. **[10%]** As discussed in class, an important part of being a data scientist is communicating your findings. Therefore, summarize the dataset and your findings from question 1-6 in a short paragraph. Do not simply copy your answers from above, but rather assume that you need to briefly explain your analysis to an audience with no prior knowledge of the dataset or data science.

**Provide answer below:**