Part 1. (18 Points)

On Canvas, you'll find the R script, `One_Variable_Display_and_Summary_Stats.R` and the ST314 student survey dataset, `st314_student_survey.csv`. You'll use both of these to explore one categorical and one quantitative variable from the survey. Download the R script and the dataset, open the R script and follow the command instructions. Then answer the following questions:

Categorical Variable

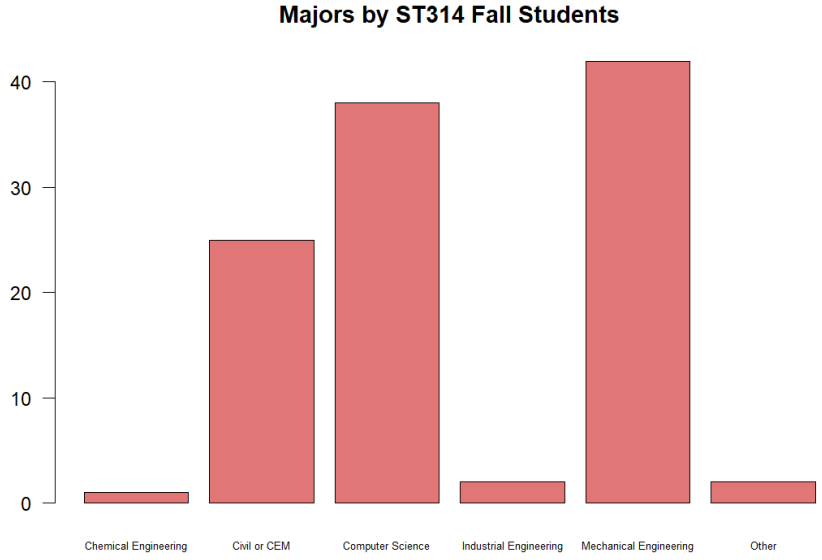
The variable "Major" describes the individual's major in school. The variable "Phone" identifies the type of phone the individual has (iOS, Android, other). Both of these variables are categorical. **Select one of the two categorical variables just mentioned and answer the following three questions.**

a. **(1 point) Choose a categorical variable to explore. Which variable did you choose?**
I chose to look at the Majors of the student in st314.

b. **(2 point) Paste the table of counts and bar chart for the categorical variable of your choosing. Include color and an appropriate title on your plot.**

```
> table(st314data$Major)

  Chemical Engineering            Civil or CEM       Computer Science Industrial Engineering
                     1                      25                      38                       2
Mechanical Engineering                   Other
                    42                       2
```

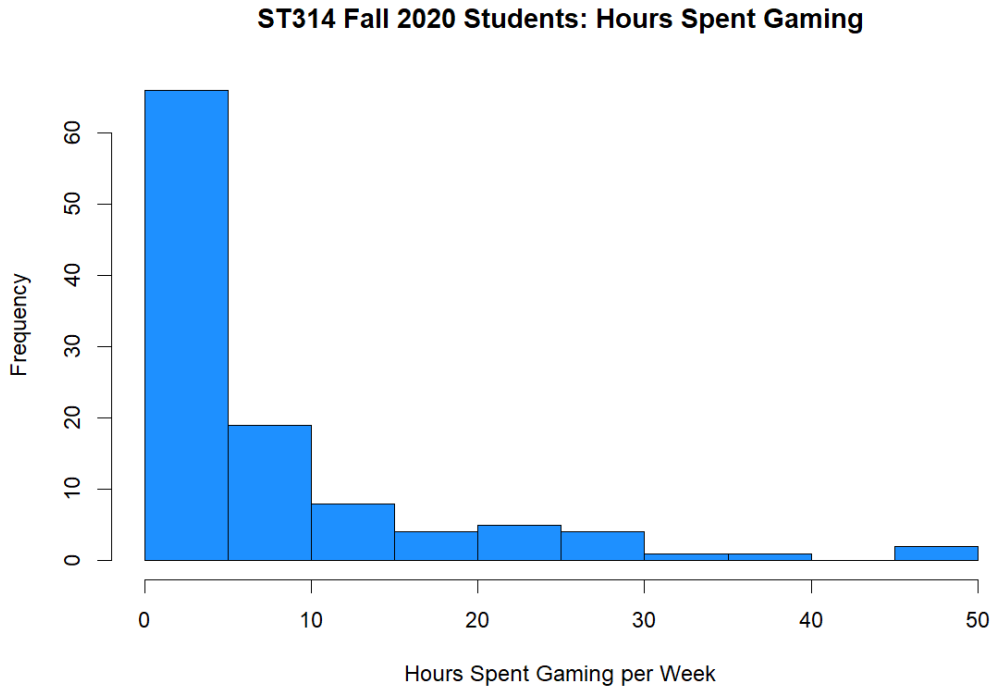


c. **(2 point) Briefly, describe the distribution in context. Recall, categorical variables are summarized by counts and/or percents.**
This data shows that mechanical engineering and computer science are the most prevalent majors in our st314 class, with civil engineering being not too far behind them. Industrial and chemical engineering are at the opposite end with counts of only 1 and 2.

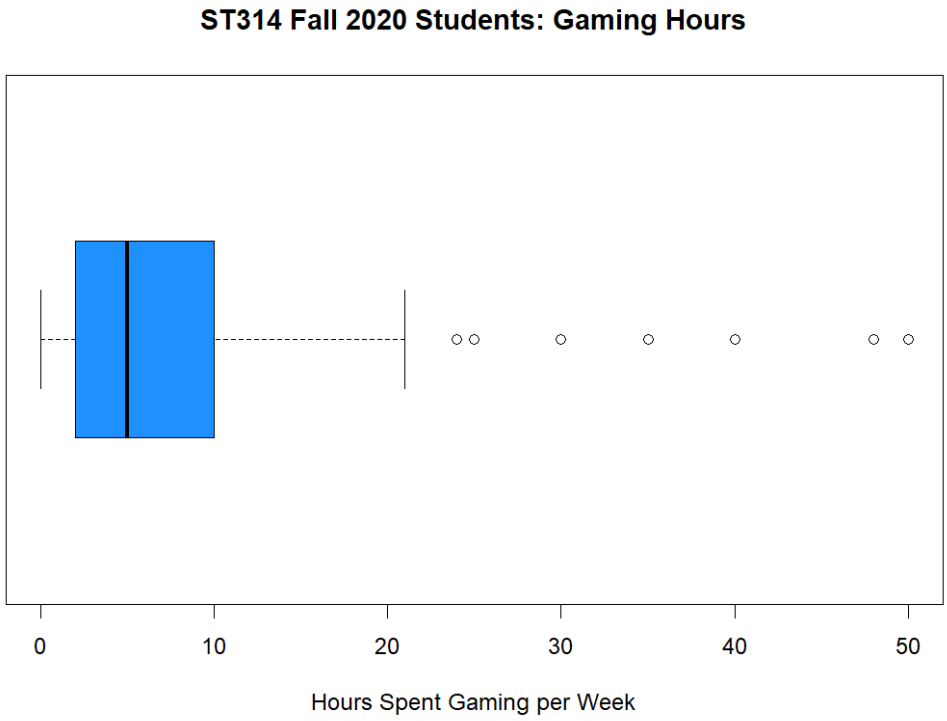
Quantitative Variable
The variable "Credit Hours" indicates the number of credit hours the individual was enrolled in during the term the survey was completed. The variable "Gaming Hours" describes approximately how many hours a week the survey participant games. Both of these variables are quantitative. **Select one of the two quantitative variables just mentioned and answer the following three questions.**

a. **(1 point) Choose a quantitative variable to explore. Which variable did you choose? Is the variable discrete or continuous?**
I chose to look at hour spent gaming as I spend a lot of time gaming. This variable is continuous as it is a measurement of time.

b. **(2 point) Create a histogram of the variable. Include color and an appropriate title on your plot. Paste plot.**



c. **(2 point) Create a boxplot of the variable. Include color and an appropriate title on your plot. Paste plot.**



d. **(1 point) Which plot do you prefer (histogram or boxplot) to visualize the variable? Why?**
In this case, I would prefer the histogram. It gives more insights into our data and is easier to read compared to the compressed box plot.

e. **(2 points) Give a table that includes the mean, standard deviation, minimum, 1st quartile, median, 3rd quartile, maximum and IQR.**

Min	1st Qtr	Median	Mean	3rd Qtr	Max	SD
0	2	5	8.291	10	50	10.30

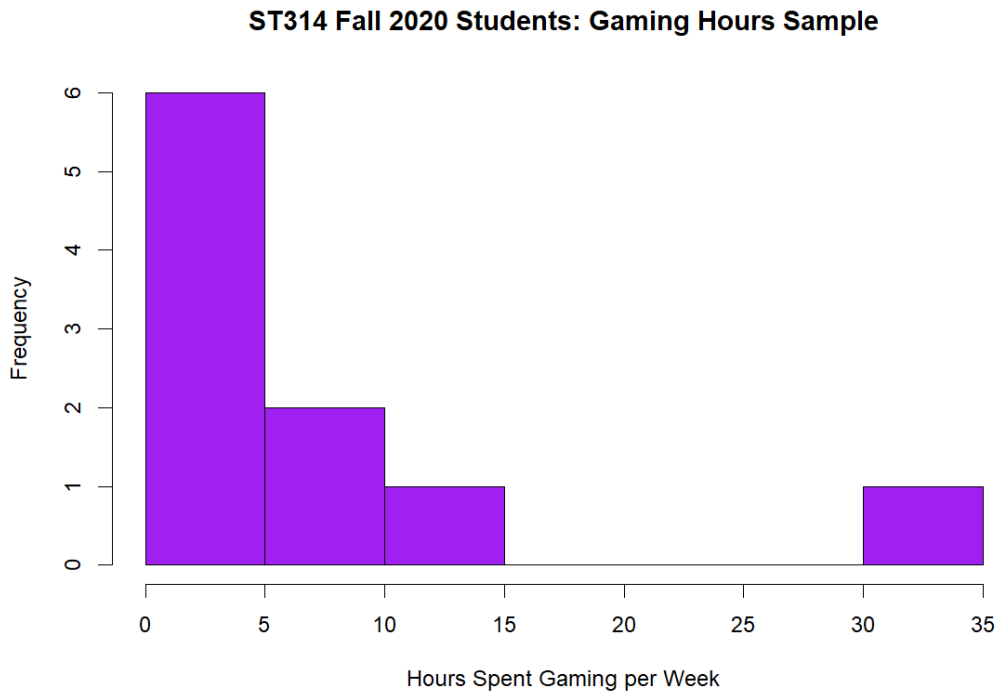
f. **(3 points) Use the plots and summary statistics to describe the data in the context of the problem. Include the shape, center and spread in your description. State whether there are any outliers.**
Both the histogram and the box plot show the data as being right skewed. This means less people play more hours of games per week. The center in the box plot is around 5 hours and in the histogram it is the 0-5 hours range. The spread in the box plot can be seen as 0-20 hours, where most of the data lies. The spread can be seen similarly in the histogram where most of the data can be seen between 0 and 20 hours. Outliers are clearly visible in the box plot, with there being around 7. Although you can not explicitly see outlies in a histogram, you can tell that there is data that does not fit.

g. **(2 points) Given the shape of the data which measure, the mean, median or either, would be a more appropriate to represent the center of the data? Explain your reasoning.**
I would say the median is more appropriate to represent the center of the data. With a skewed distribution, the mean is pulled towards the direction of the tail. This makes the mean less representative of the center than the median.

Part 2. (5 points)
Using the code provided in the `One_Variable_Display_and_Summary_Stats.R` script, take a simple random sample of 10 students from the class data. Consider the same quantitative variable you chose in part 1.

a. **(1 point) Calculate the mean and standard deviation of your sample.**
The mean and standard deviation of my sample are 8.0 and 10.42 respectively.

b. **(2 point) Make a boxplot or histogram of the sample data. Give the plot a title and change the color of the plot from the plot in part 1. Paste plot.**



c. **(2 points) How different are your sampled statistics in comparison to the overall class mean and standard deviation? How different is the distribution?**
The sample obviously has less data to work with, so it looks more empty and less smooth. However, you can still tell that there is a skew to the data. With the sample, it is more obvious that there are outliers as there is an empty space between the main data and the outlier piece of data. The mean and standard deviation were very close, with minimal difference.

Gradescope Page Matching (2 points)
When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page "Submitting Assignments in Gradescope". Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item "Gradescope Page Matching".