title: "DSCC/CSC/TCS 462 Assignment 0"

output:

pdf\_document: default

html\_document:

df\_print: paged

subtitle: Due Thursday, September 8, 2022 by 3:59 p.m.

fontsize: 12pt

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = T, eval=T,tidy=TRUE, tidy.opts=list(width.cutoff=60))

```

This assignment will cover material from Lectures 1 and 2. You are expected to use the `ggplot2` library in `R` for completing all the graphics. To learn more about graphics using `ggplot2`, please read through the guide available here: http://www.cookbook-r.com/Graphs/. This is a wonderful open source textbook that walks through examples of many different graphics in `ggplot2`. If you have not done so already, start by installing the library. In the R console (i.e. NOT in your .RMD file), run the code `install.packages("ggplot2")`. Then, in your .RMD file, load the library as follows:

```{r, echo=T}

library(ggplot2)

```

For this first assignment, we will use the "car\_sales.csv" dataset, which includes information about 152 different cars. In particular, we will mainly focus on the selling price of cars throughout this assignment.

1. Getting familiar with the dataset via exploratory data analysiinstall.packages("ggplot2")s.

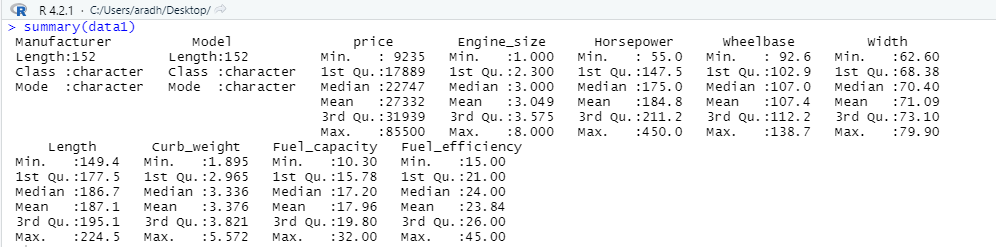
a. Read the data into RStudio and summarize the data with the `summary()` function.

library(readr)

library(ggplot2)

data1 <- read\_csv("car\_sales.csv")

summary(data1)



b. How many bins does Sturges' formula suggest we use for a histogram of `price`? Show your work.

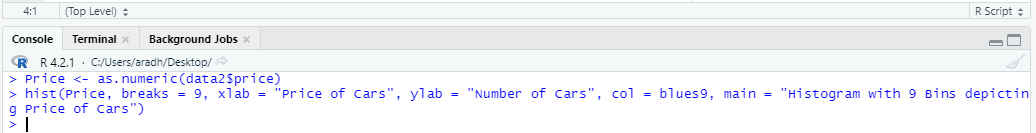
c. Create a histogram of `price` using the number of bins suggested by Sturges' formula in 1b. Make sure to appropriately title the histogram and label the axes. Comment on the center, shape, and spread.

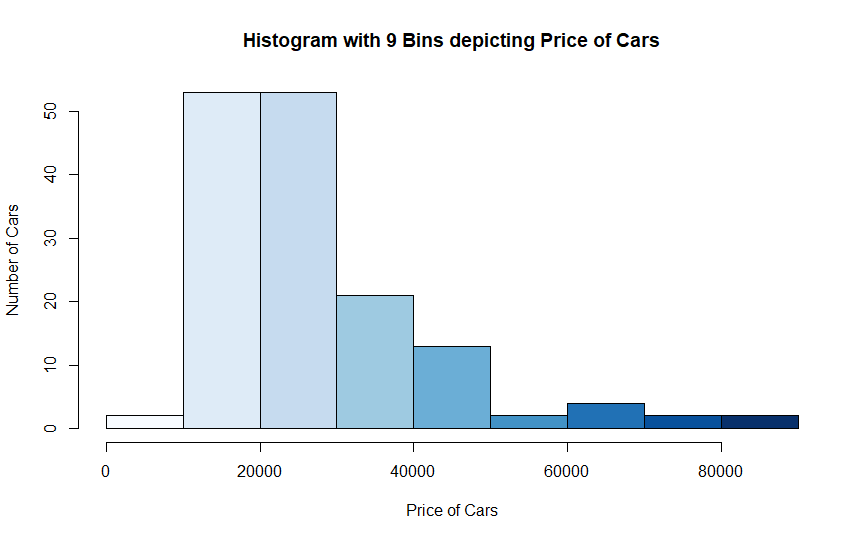
CODE:

data2 <- data1[c('price')]

Price <- as.numeric(data2$price)

hist(Price, breaks = 9, xlab = "Price of Cars", ylab = "Number of Cars", col = blues9, main = "Histogram with 9 Bins depicting Price of Cars")





2. Measures of center and spread for the selling price of cars.

a. Calculate the mean, median, and 10\% trimmed mean of the selling price. Report the mean, median, and 10\% trimmed mean on the histogram. In particular, create a red vertical line on the histogram at the mean, and report the value of the mean in red next to the line using the form "$\bar{x}=$". Create a blue vertical line on the histogram at the median, and report the value of the median in blue next to the line using the form "$\tilde{x}=$". Create a green vertical line on the histogram at the 10\% trimmed mean, and report the value of the 10\% trimmed mean in green next to the line using the form "$\bar{x}\_{10}=$" (to get $\bar{x}\_{10}$ to print on the plot, use `bar(x)[10]` within the `paste()` function).

Code:

> mean(data2$price)

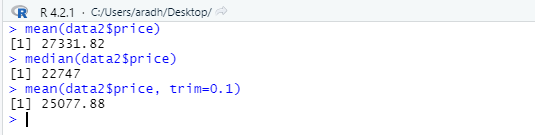
[1] 27331.82

> median(data2$price)

[1] 22747

> mean(data2$price, trim=0.1)

[1] 25077.88



hist(Price, breaks = 9, xlab = "Price of Cars", ylab = "Number of Cars", col = blues9, main = "Histogram with 9 Bins depicting Price of Cars")

> abline(v = mean(data2$price), col="red", lwd=3)

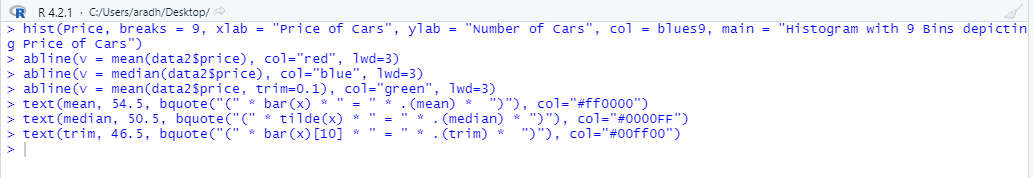
> abline(v = median(data2$price), col="blue", lwd=3)

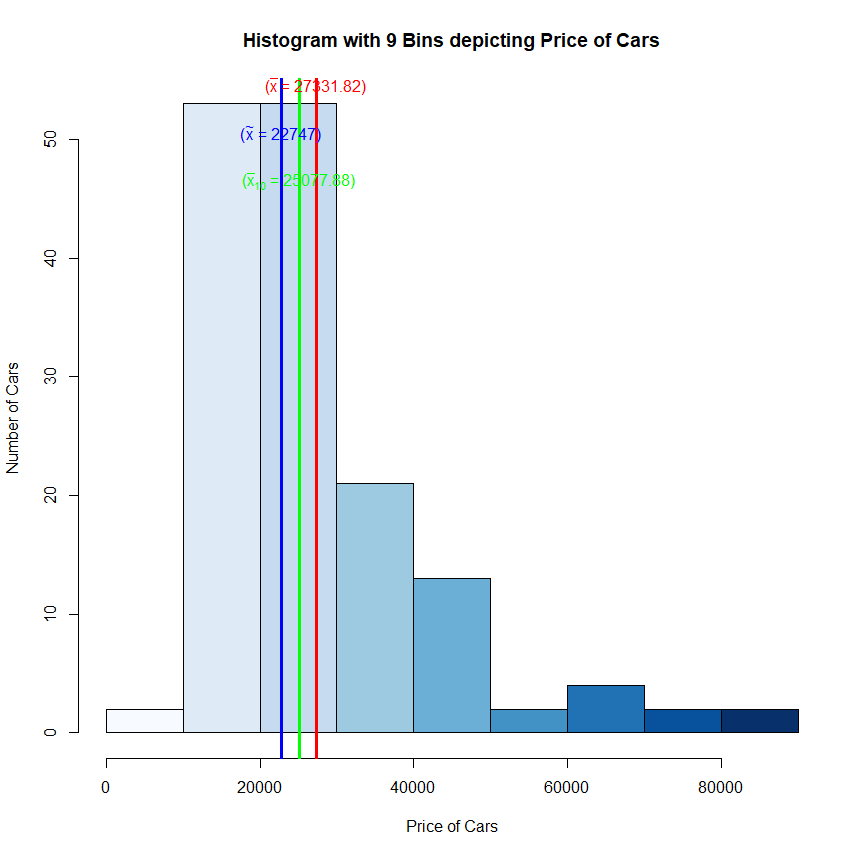
> abline(v = mean(data2$price, trim=0.1), col="green", lwd=3)

> text(mean, 54.5, bquote("(" \* bar(x) \* " = " \* .(mean) \* ")"), col="#ff0000")

> text(median, 50.5, bquote("(" \* tilde(x) \* " = " \* .(median) \* ")"), col="#0000FF")

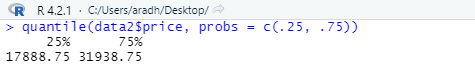
> text(trim, 46.5, bquote("(" \* bar(x)[10] \* " = " \* .(trim) \* ")"), col="#00ff00")





b. Calculate and report the 25th and 75th percentiles.

> quantile(data2$price, probs = c(.25, .75))



c. Calculate and report the interquartile range.

IQR(data2$price)



d. Calculate and report the standard span, the lower fence, and the upper fence.

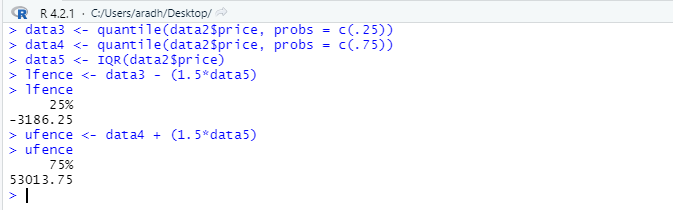
data3 <- quantile(data2$price, probs = c(.25))

data4 <- quantile(data2$price, probs = c(.75))

data5 <- IQR(data2$price)

lfence <- data3 - (1.5\*data5)

ufence <- data4 + (1.5\*data5)



e. Are there any outliers? Subset the outlying points. Use code based on the following:

```{r echo=T, eval=F}

car\_sales[car\_sales$price >= upper\_fence, ] #upper outliers

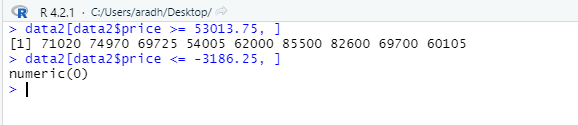
car\_sales[car\_sales$price <= lower\_fence, ] #lower outliers

# Use upper and lower fence values from part g.

```

data2[data2$price >= 53013.75, ]

data2[data2$price <= -3186.25, ]



f. Calculate and report the variance, standard deviation, and coefficient of variation of car prices.

Standard Deviation

> sd <- sd(data2$price)

> sd

[1] 14418.67

Variance

> var <- var(data2$price)

> var

[1] 207898012

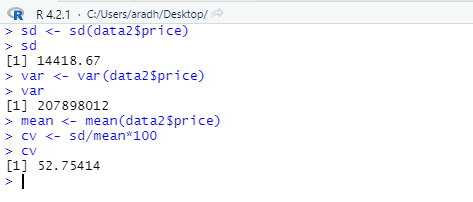
Coefficient of Variation

> mean <- mean(data2$price)

> cv <- sd/mean\*100

> cv

[1] 52.75414

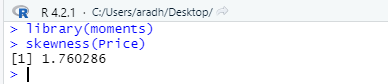


g. We have seen from the histogram that the data are skewed. Calculate and report the skewness. Comment on this value and how it matches with what you visually see in the histogram.

> library(moments)

> skewness(Price)

[1] 1.760286



3. Transforming the data.

a. Use a Box-Cox power transformation to appropriately transform the data. In particular, use the `boxcox()` function in the `MASS` library. Report the recommended transformation. Do not apply this transformation to the data yet. (Note: the `boxcox` function automatically produces a plot. You do NOT need to make this in `ggplot2`.)

b. Apply the exact Box-Cox recommended transformation (rounded to four decimal places) to the data (this transformation is hereon referred to as the Box-Cox transformed data). Use the `summary()` function to summarize the results of this transformation.

c. Create a histogram of the Box-Cox transformed data using the number of bins suggested by Sturges' formula. On this histogram, report the mean, median, and 10\% trimmed mean using the same formatting options as in part 2a above. Comment on the center, shape, and spread.

d. As an alternative to the Box-Cox transformation, let's also use a log transformation. Apply the log transformation to the original `price` data (this transformation is hereon referred to as the log transformed data). Use the `summary()` function to summarize the results of this transformation.

e. Create a histogram of the log transformed data using the number of bins suggested by Sturges' formula. On this histogram, report the mean, median, and 10\% trimmed mean using the same formatting options as in part 2a and 3c above. Comment on the center shape and spread.

f. Create a qqplot for the original data, a qqplot for the Box-Cox transformed data, and a qqplot of the log transformed data. Comment on the results.

g. Evaluate the empirical rule for the original data, the Box-Cox transformed data, and the log transformed data. In particular, make a table similar to that on slide 71 of the Chapter 2 notes. Comment on the results. Do either of the transformed data seem to be "better" to work with? Note, you can use code similar to the following to answer this question:

```{r eval=FALSE, echo=TRUE}

###Create a matrix named "mat" with 9 rows & 5 columns

mat <- matrix(NA, nrow=9, ncol=5)

### Set row names and column names

rownames(mat) <- c("Original","","","Box-Cox","","","Log","","")

colnames(mat) <- c("x","xbar-k\*s", "xbar+k\*s", "Theoretical %","Actual %")

### Fill in known quantities

mat[,1] <- c(1,2,3)

mat[,4]<-c(68,95,99.7)

### Fill in calculated values (I only give a preview of this and leave the remaining calculations for you).

### I use "orig" as the original data, "bcdat" as the Box-Cox transformed data,

### and "logdat" as the log-transformed data.

### Name your variables anything you'd like.

mat[3,2] <- mean(orig)-3\*sd(orig)

mat[2,3] <- mean(orig)+2\*sd(orig)

mat[4,5] <- sum(bcdat >=mean(bcdat)-1\*sd(bcdat)

& bcdat <= mean(bcdat)+1\*sd(bcdat))/length(bcdat)\*100

mat[9,5] <- sum(logdat >=mean(logdat)-3\*sd(logdat)

& logdat <= mean(logdat)+3\*sd(logdat))/length(logdat)\*100

### Create a table

library(knitr)

kable(x=mat, digits=2,row.names=T, format="markdown")

```

h. In your own words, provide some intuition about (1) why car price may not follow a normal distribution, and (2) why it may be useful to transform the data into a form that more closely follows a normal distribution.

Short Answers:

\* About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?

\* Who, if anyone, did you work with on this assignment?

\* What questions do you have relating to any of the material we have covered so far in class?

library(readr)

library(ggplot2)

library(moments)

#question1

data1 <- read\_csv("car\_sales.csv")

summary(data1)

data2 <- data1[c('price')]

Price <- as.numeric(data2$price)

hist(Price, breaks = 9, xlab = "Price of Cars", ylab = "Number of Cars", col = blues9, main = "Histogram with 9 Bins depicting Price of Cars")

#question2

mean <- mean(data2$price)

median <- median(data2$price)

trim <- mean(data2$price, trim=0.1)

abline(v = mean(data2$price), col="red", lwd=3)

abline(v = median(data2$price), col="blue", lwd=3)

abline(v = mean(data2$price, trim=0.1), col="green", lwd=3)

text(mean, 54.5, bquote("(" \* bar(x) \* " = " \* .(mean) \* ")"), col="#ff0000")

text(median, 50.5, bquote("(" \* tilde(x) \* " = " \* .(median) \* ")"), col="#0000FF")

text(trim, 46.5, bquote("(" \* bar(x)[10] \* " = " \* .(trim) \* ")"), col="#00ff00")

quantile <- quantile(data2$price, probs = c(.25, .75))

IQR <- IQR(data2$price)

percentile1 <- quantile(data2$price, probs = c(.25))

percentile2 <- quantile(data2$price, probs = c(.75))

lfence <- percentile1 - (IQR\*1.5)

ufence <- percentile2 + (IQR\*1.5)

outlier1 <- data2[data2$price >= 53013.75, ]

outlier2 <- data2[data2$price <= -3186.25, ]

var <- var(data2$price)

sd <- sd(data2$price)

cv <- sd/mean\*100

skewness(Price)

#question3

bxcx <- boxcox(lm(data2$price ~ 1))