DSCC/CSC/TCS 462 Assignment 0

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

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:

```
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 analysis.

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

```
car_sales <- read.csv("car_sales.csv")
summary(car_sales)
```

```
##  Manufacturer          Model               price        Engine_size   
##  Length:152         Length:152         Min.   : 9235   Min.   :1.000  
##  Class :character   Class :character   1st Qu.:17889   1st Qu.:2.300  
##  Mode  :character   Mode  :character   Median :22747   Median :3.000  
##                                        Mean   :27332   Mean   :3.049  
##                                        3rd Qu.:31939   3rd Qu.:3.575  
##                                        Max.   :85500   Max.   :8.000  
##    Horsepower      Wheelbase         Width           Length     
##  Min.   : 55.0   Min.   : 92.6   Min.   :62.60   Min.   :149.4  
##  1st Qu.:147.5   1st Qu.:102.9   1st Qu.:68.38   1st Qu.:177.5  
##  Median :175.0   Median :107.0   Median :70.40   Median :186.7  
##  Mean   :184.8   Mean   :107.4   Mean   :71.09   Mean   :187.1  
##  3rd Qu.:211.2   3rd Qu.:112.2   3rd Qu.:73.10   3rd Qu.:195.1  
##  Max.   :450.0   Max.   :138.7   Max.   :79.90   Max.   :224.5  
##   Curb_weight    Fuel_capacity   Fuel_efficiency
##  Min.   :1.895   Min.   :10.30   Min.   :15.00  
##  1st Qu.:2.965   1st Qu.:15.78   1st Qu.:21.00  
```

```
##   Median :3.336   Median :17.20   Median :24.00
##   Mean   :3.376   Mean   :17.96   Mean   :23.84
##   3rd Qu.:3.821   3rd Qu.:19.80   3rd Qu.:26.00
##   Max.   :5.572   Max.   :32.00   Max.   :45.00
```

```
price <- car_sales$price
```

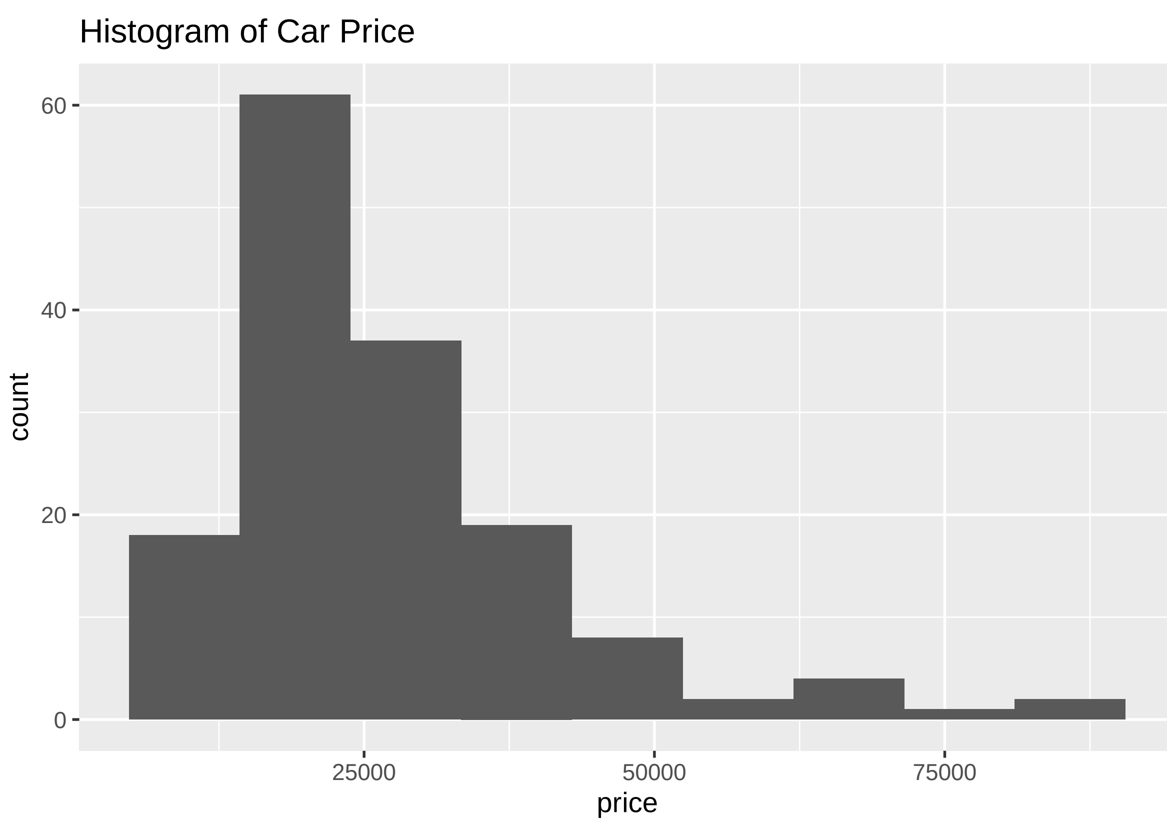
b. How many bins does Sturges' formula suggest we use for a histogram of `price`? Show your work.

```
ceiling(log(length(price), 2)) + 1
```

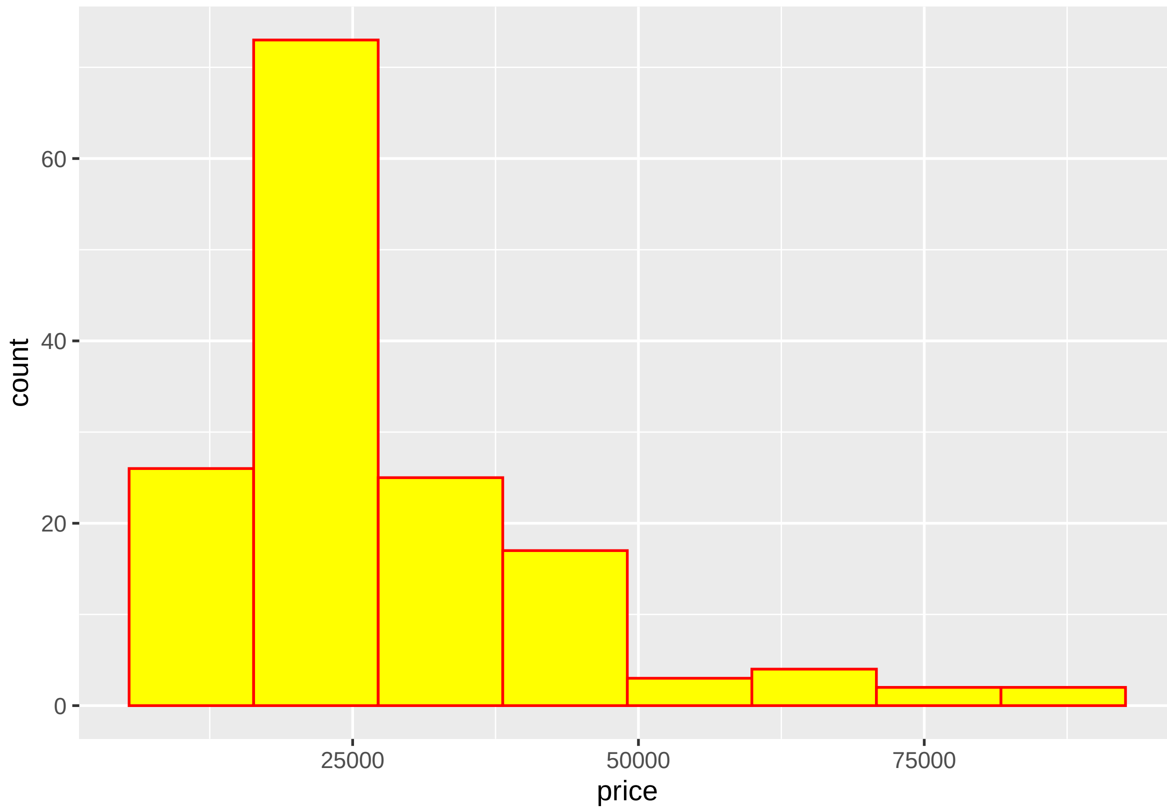
```
## [1] 9
```

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.

```
library(ggplot2)
hist1 <- ggplot(car_sales, aes(x = price)) + geom_histogram(bins = 9)
hist1 <- hist1 + ggtitle("Histogram of Car Price")
hist1
```




```
hist0 <- ggplot(data = car_sales, aes(x = price)) + geom_histogram(color = "red",
    fill = "yellow", bins = 8)
hist0
```



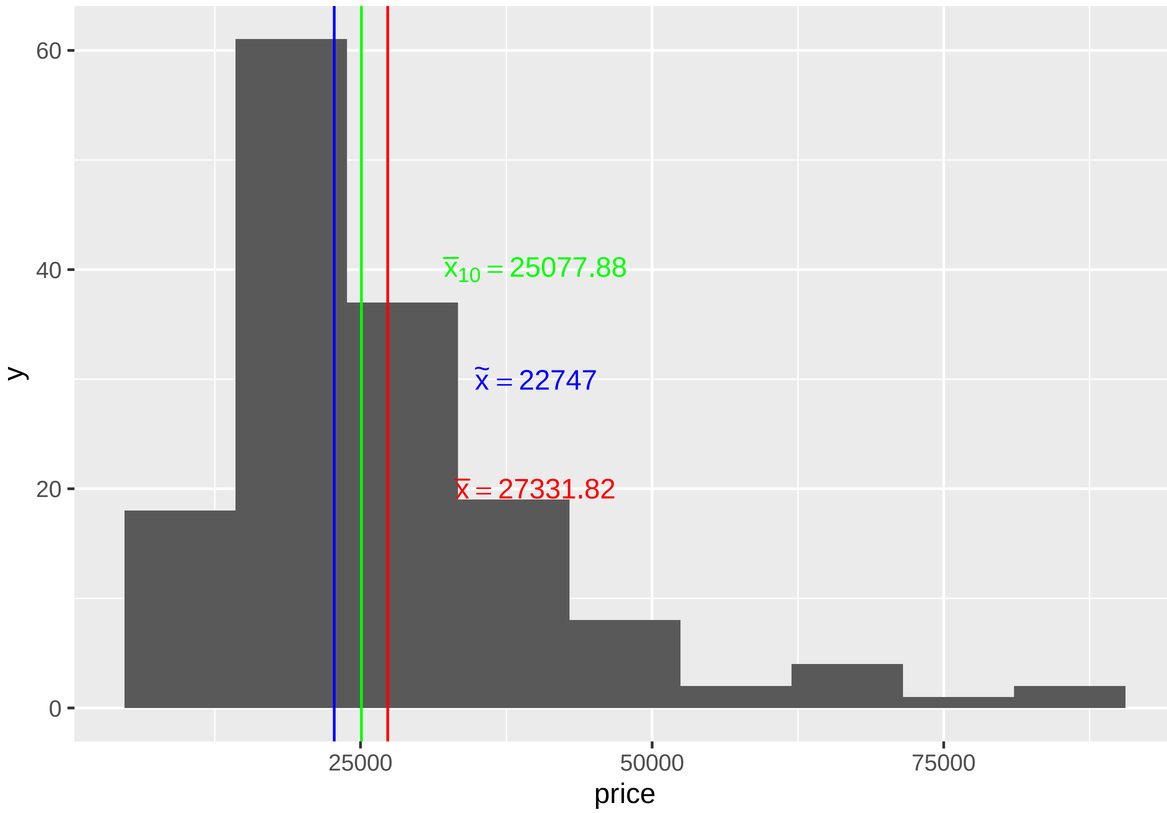
The histogram is unimodal, positively (right) skewed (with perhaps somewhat of a bell shape).

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).

```
hist1 <- ggplot(car_sales, aes(x = price)) + geom_histogram(bins = 9)
hist1 <- hist1 + geom_vline(aes(xintercept = mean(price)), color = "red")
hist1 <- hist1 + geom_vline(aes(xintercept = median(price)),
    color = "blue")
hist1 <- hist1 + geom_vline(aes(xintercept = mean(price, trim = 0.1)),
    color = "green")
hist1 <- hist1 + annotate("text", x = 40000, y = 20, label = paste("bar(x)==",
    round(mean(car_sales$price), 3)), parse = T, color = "red")
hist1 <- hist1 + annotate("text", x = 40000, y = 30, label = paste("tilde(x)==",
```

```r
    round(median(car_sales$price), 3)), parse = T, color = "blue")
hist1 <- hist1 + annotate("text", x = 40000, y = 40, label = paste("bar(x)[10]==",
    round(mean(car_sales$price, 0.1), 3)), parse = T, color = "green")
hist1
```



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

```r
quantile(price, c(0.25, 0.75))
```

```
##      25%      75%
## 17888.75 31938.75
```

c. Calculate and report the interquartile range.

```r
IQR(price)
```

```
## [1] 14050
```

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

```r
ss <- 1.5 * IQR(price)
ss
```

```
## [1] 21075
```

```r
lf <- quantile(price, 0.25) - ss
lf
```

```
##      25%
## -3186.25
```

```r
uf <- quantile(price, 0.75) + ss
uf
```

```
##      75%
## 53013.75
```

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

```r
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.
```

```r
sum(price < lf)
```

```
## [1] 0
```

```r
sum(price > uf)
```

```
## [1] 9
```

```r
car_sales[price >= uf | price <= lf, ]
```

```
##     Manufacturer          Model price Engine_size Horsepower Wheelbase Width
## 46       Porsche  Carrera Coupe 71020         3.4        300      92.6  69.5
## 55       Porsche Carrera Cabrio 74970         3.4        300      92.6  69.5
## 82         Dodge          Viper 69725         8.0        450      96.2  75.7
## 123        Lexus          LS400 54005         4.0        290     112.2  72.0
## 125         Audi             A8 62000         4.2        310     113.0  74.0
## 136   Mercedes-B          CL500 85500         5.0        302     113.6  73.1
## 138   Mercedes-B       SL-Class 82600         5.0        302      99.0  71.3
## 139   Mercedes-B        S-Class 69700         4.3        275     121.5  73.1
## 151        Lexus          LX470 60105         4.7        230     112.2  76.4
##     Length Curb_weight Fuel_capacity Fuel_efficiency
## 46   174.5       3.032          17.0              21
## 55   174.5       3.075          17.0              23
## 82   176.7       3.375          19.0              16
## 123  196.7       3.890          22.5              22
## 125  198.2       3.902          23.7              21
## 136  196.6       4.115          23.2              20
## 138  177.1       4.125          21.1              20
## 139  203.1       4.133          23.2              21
## 151  192.5       5.401          25.4              15
```

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

```
var(price)
```

```
## [1] 207898012
```

```
sd(price)
```

```
## [1] 14418.67
```

```
sd(price)/mean(price)
```

```
## [1] 0.5275414
```

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)
```

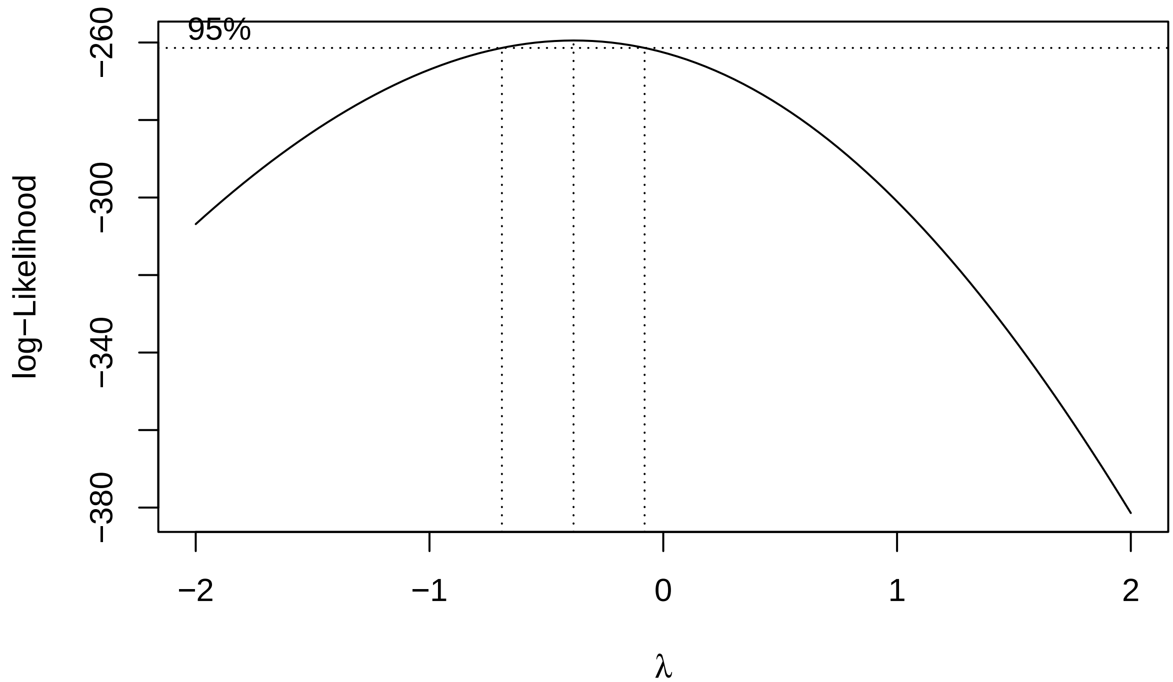
```
## [1] 1.760286
```

The histogram showed a positive skew, which is verified by the skewness being a positive value.

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`.)

```
library(MASS)
bc1 <- boxcox(price ~ 1)
```




```
bc1$x[bc1$y == max(bc1$y)]
```

```
## [1] -0.3838384
```

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.

```
car_sales$bcprice <- (price^(-0.3838) - 1)/(-0.3838)
summary(car_sales$bcprice)
```

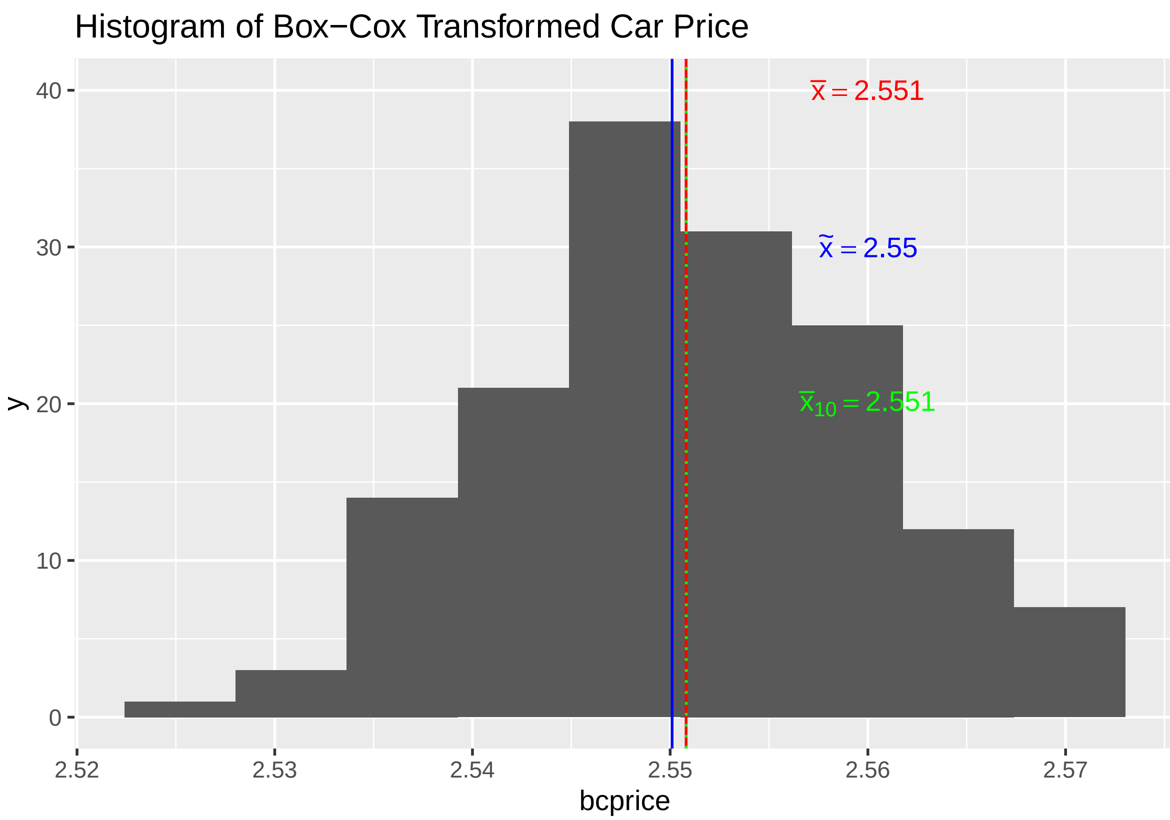
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.527   2.545   2.550   2.551   2.557   2.572
```

```
bcprice <- car_sales$bcprice
```

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.

```
hist2 <- ggplot(car_sales, aes(x = bcprice)) + geom_histogram(bins = 9)
hist2 <- hist2 + ggtitle("Histogram of Box-Cox Transformed Car Price")
hist2 <- hist2 + geom_vline(aes(xintercept = mean(bcprice)),
    color = "red")
hist2 <- hist2 + geom_vline(aes(xintercept = median(bcprice)),
    color = "blue")
hist2 <- hist2 + geom_vline(aes(xintercept = mean(bcprice, trim = 0.1)),
```

```
    color = "green", linetype = "dotted")
hist2 <- hist2 + annotate("text", x = 2.56, y = 40, label = paste("bar(x)==",
    round(mean(bcprice), 3)), parse = T, color = "red")
hist2 <- hist2 + annotate("text", x = 2.56, y = 30, label = paste("tilde(x)==",
    round(median(bcprice), 3)), parse = T, color = "blue")
hist2 <- hist2 + annotate("text", x = 2.56, y = 20, label = paste("bar(x)[10]==",
    round(mean(bcprice, 0.1), 3)), parse = T, color = "green")
hist2
```



The histogram is unimodal, fairly symmetric, and fairly bell-shaped.

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.

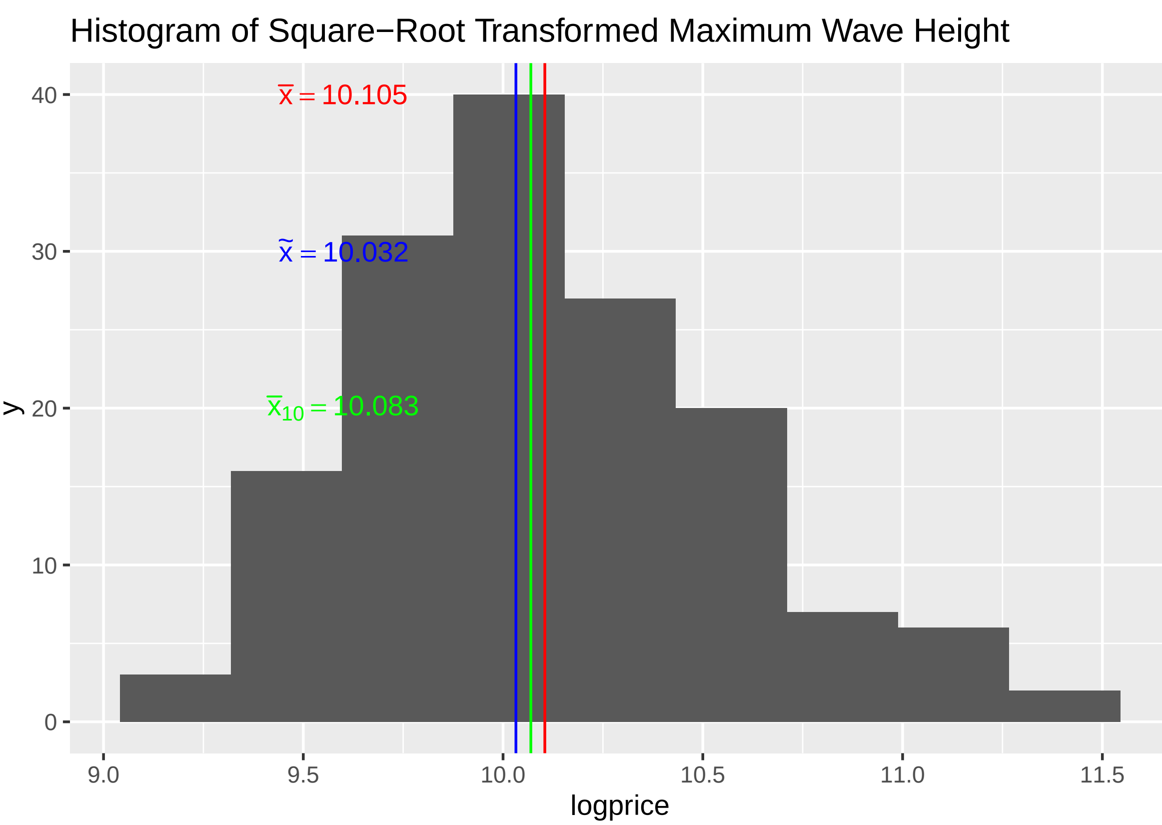
```
logprice <- car_sales$logprice <- log(price)
summary(logprice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.131   9.792  10.032  10.105  10.372  11.356
```

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.

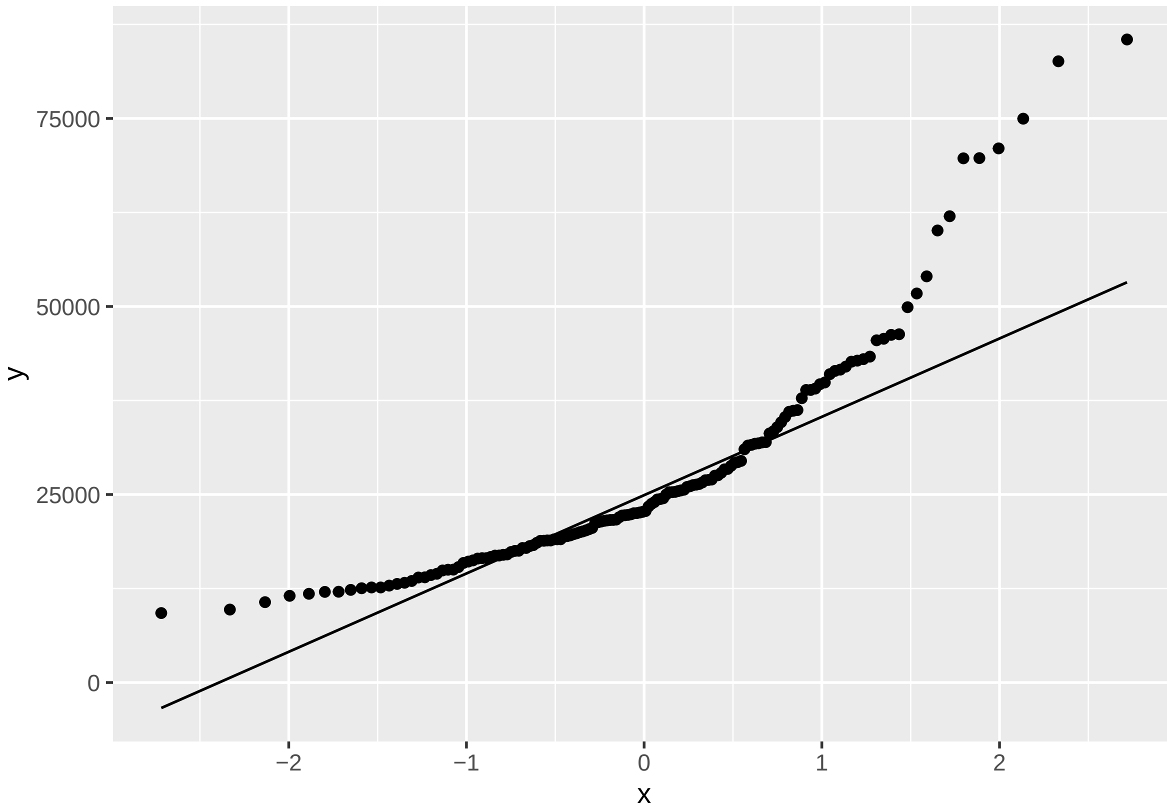
```
hist3 <- ggplot(car_sales, aes(x = logprice)) + geom_histogram(bins = 9)
hist3 <- hist3 + ggtitle("Histogram of Square-Root Transformed Maximum Wave Height"
hist3 <- hist3 + geom_vline(aes(xintercept = mean(logprice)),
    color = "red")
hist3 <- hist3 + geom_vline(aes(xintercept = median(logprice)),
    color = "blue")
hist3 <- hist3 + geom_vline(aes(xintercept = mean(logprice, trim = 0.2)),
    color = "green")
hist3 <- hist3 + annotate("text", x = 9.6, y = 40, label = paste("bar(x)==",
    round(mean(logprice), 3)), parse = T, color = "red")
hist3 <- hist3 + annotate("text", x = 9.6, y = 30, label = paste("tilde(x)==",
    round(median(logprice), 3)), parse = T, color = "blue")
hist3 <- hist3 + annotate("text", x = 9.6, y = 20, label = paste("bar(x)[10]==",
    round(mean(logprice, 0.1), 3)), parse = T, color = "green")
hist3
```



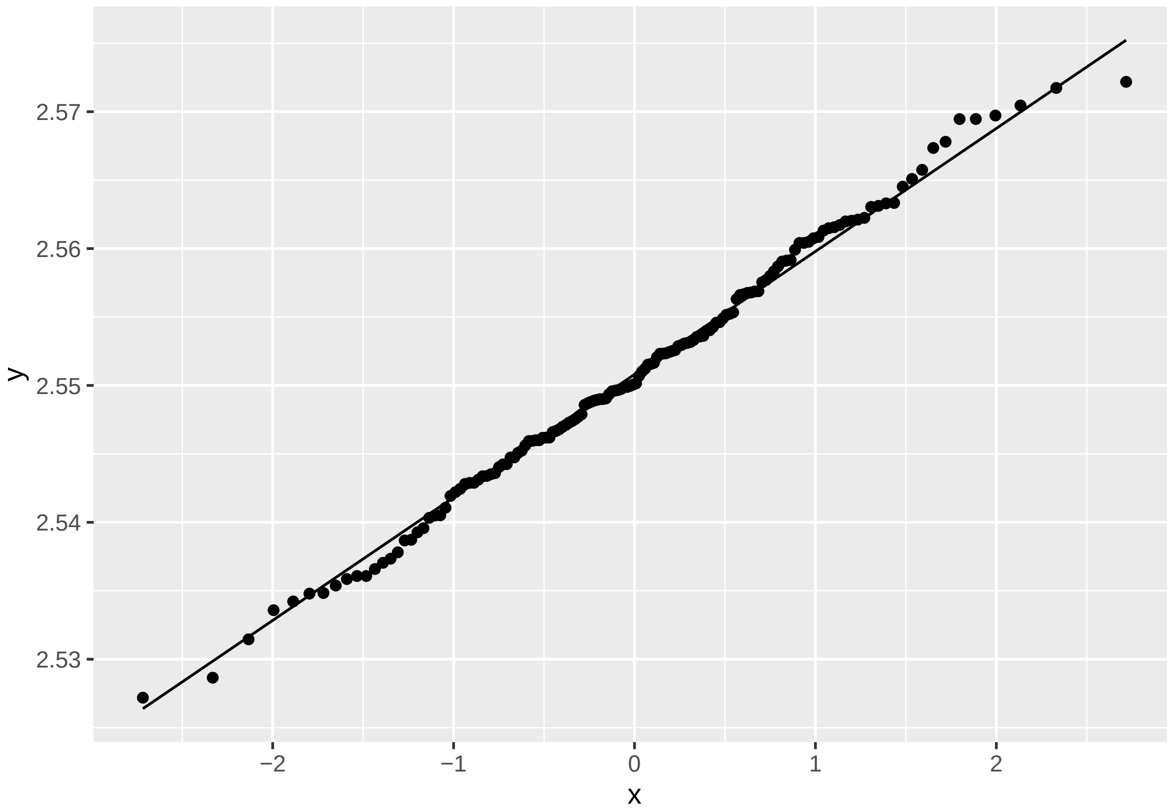
The histogram is unimodal, fairly symmetric, and fairly bell-shaped.

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.

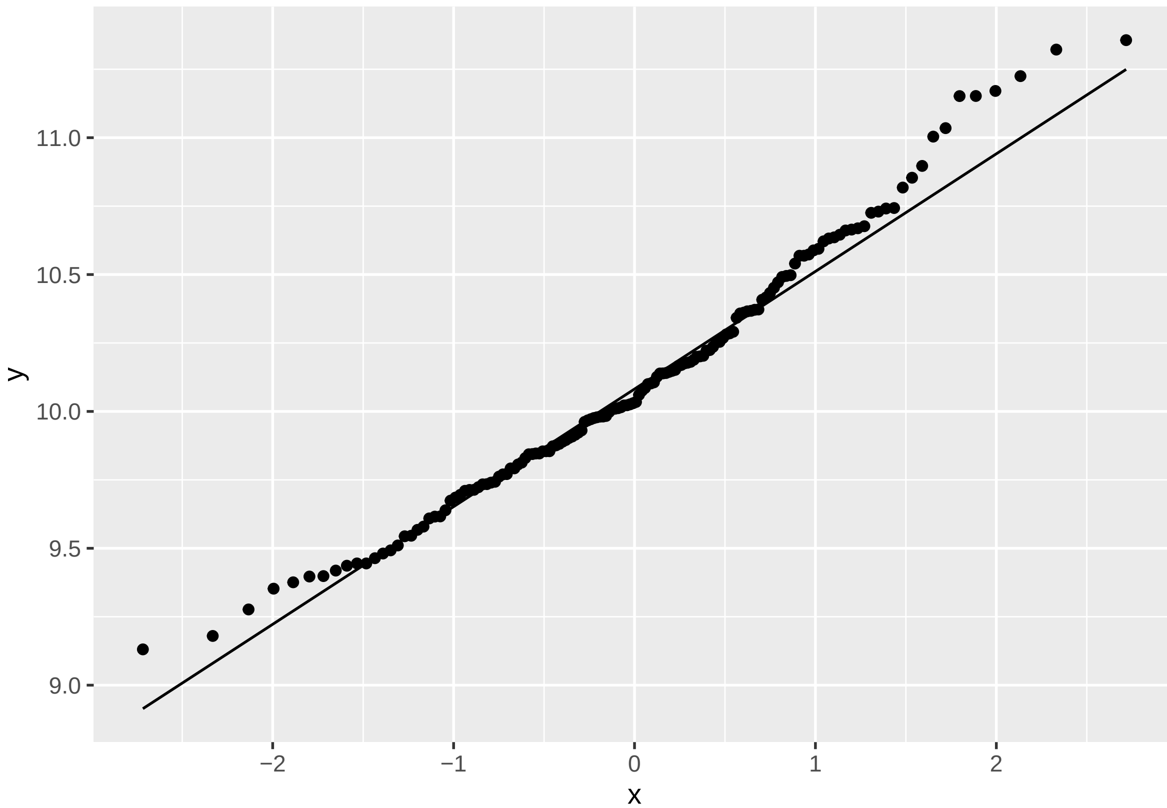
```r
p1 <- ggplot(car_sales, aes(sample = price))
p1 + stat_qq() + stat_qq_line()
```




```r
p2 <- ggplot(car_sales, aes(sample = bcprice))
p2 + stat_qq() + stat_qq_line()
```




```
p3 <- ggplot(car_sales, aes(sample = logprice))
p3 + stat_qq() + stat_qq_line()
```



The box-cox transformed data seems to be most normal. The log transformed data is relatively good, but we see slightly more flaring in the tails of the qq plot.

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:

```
mat <- matrix(NA, nrow = 9, ncol = 5)
mat[1, 2] <- mean(price) - 1 * sd(price)
mat[2, 2] <- mean(price) - 2 * sd(price)
mat[3, 2] <- mean(price) - 3 * sd(price)

mat[1, 3] <- mean(price) + 1 * sd(price)
mat[2, 3] <- mean(price) + 2 * sd(price)
mat[3, 3] <- mean(price) + 3 * sd(price)

mat[1, 5] <- sum(price >= mean(price) - 1 * sd(price) & price <=
    mean(price) + 1 * sd(price))/length(price) * 100
mat[2, 5] <- sum(price >= mean(price) - 2 * sd(price) & price <=
    mean(price) + 2 * sd(price))/length(price) * 100
mat[3, 5] <- sum(price >= mean(price) - 3 * sd(price) & price <=
    mean(price) + 3 * sd(price))/length(price) * 100
```

```r
mat[4, 2] <- mean(bcprice) - 1 * sd(bcprice)
mat[5, 2] <- mean(bcprice) - 2 * sd(bcprice)
mat[6, 2] <- mean(bcprice) - 3 * sd(bcprice)

mat[4, 3] <- mean(bcprice) + 1 * sd(bcprice)
mat[5, 3] <- mean(bcprice) + 2 * sd(bcprice)
mat[6, 3] <- mean(bcprice) + 3 * sd(bcprice)

mat[4, 5] <- sum(bcprice >= mean(bcprice) - 1 * sd(bcprice) &
    bcprice <= mean(bcprice) + 1 * sd(bcprice))/length(bcprice) *
    100
mat[5, 5] <- sum(bcprice >= mean(bcprice) - 2 * sd(bcprice) &
    bcprice <= mean(bcprice) + 2 * sd(bcprice))/length(bcprice) *
    100
mat[6, 5] <- sum(bcprice >= mean(bcprice) - 3 * sd(bcprice) &
    bcprice <= mean(bcprice) + 3 * sd(bcprice))/length(bcprice) *
    100


mat[7, 2] <- mean(logprice) - 1 * sd(logprice)
mat[8, 2] <- mean(logprice) - 2 * sd(logprice)
mat[9, 2] <- mean(logprice) - 3 * sd(logprice)

mat[7, 3] <- mean(logprice) + 1 * sd(logprice)
mat[8, 3] <- mean(logprice) + 2 * sd(logprice)
mat[9, 3] <- mean(logprice) + 3 * sd(logprice)

mat[7, 5] <- sum(logprice >= mean(logprice) - 1 * sd(logprice) &
    logprice <= mean(logprice) + 1 * sd(logprice))/length(logprice) *
    100
mat[8, 5] <- sum(logprice >= mean(logprice) - 2 * sd(logprice) &
    logprice <= mean(logprice) + 2 * sd(logprice))/length(logprice) *
    100
mat[9, 5] <- sum(logprice >= mean(logprice) - 3 * sd(logprice) &
    logprice <= mean(logprice) + 3 * sd(logprice))/length(logprice) *
    100
mat[, 1] <- c(1, 2, 3)
mat[, 4] <- c(68, 95, 99.7)
rownames(mat) <- c("Original", "", "", "Box-Cox", "", "", "Log",
    "", "")
colnames(mat) <- c("x", "xbar-k*s", "xbar+k*s", "Theoretical %",
    "Actual %")

library(knitr)
```

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

	x	xbar-k*s	xbar+k*s	Theoretical %	Actual %
Original	1	12913.15	41750.49	68.0	78.95
	2	-1505.52	56169.16	95.0	94.74
	3	-15924.18	70587.83	99.7	97.37
Box-Cox	1	2.54	2.56	68.0	66.45
	2	2.53	2.57	95.0	94.08
	3	2.52	2.58	99.7	100.00
Log	1	9.65	10.56	68.0	66.45
	2	9.19	11.02	95.0	94.08
	3	8.73	11.48	99.7	100.00

The original data was highly skewed, leading to the empirical rule not fitting well. By transforming, the distribution is getting slightly more spread out while also become much more symmetric. Because of this, we see improvements to the fit of the empirical rule to the transformed data, with both the log and box-cox transformations matching exactly.

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.

Car prices (and the prices of commodities more generally) often are very right-skewed because of luxury products. However, most "normal" cars have prices that follow a bell-shaped distribution. Additionally, the normal distribution has support from $-\infty$ to ∞ and car prices are typically nonnegative. It is useful to transform data into a form that resembles a normal distribution in order to apply statistical tests later on.

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?