**DS 710**

**R Programming Assignment**

**Assignment 3**

1. **Analyzing Used Car Prices**
2. Download Cars 2005.csv, load the data into R, and attach it.

(Dataset: “Car Data," submitted by Shonda Kuiper, Grinnell College. Dataset obtained from the Journal of Statistics Education (http://www.amstat.org/publications/jse). Accessed 3 June 2015. Used by permission of author.)

>cars = read.csv("C:/Users/pedbv9699/Desktop/DS 710/week\_3/Cars 2005.csv")

>attach(cars)

1. Make a histogram of the prices of cars in the data.  Describe the shape of the distribution.

>hist(Price)



1. What proportion of cars in the data set cost between $10,000 and $20,000?

for(price in Price){

+ if( (price > 10000) & (price < 20000) ){

+ count = count + 1

+ }

+ }

>print(count)

[1] 452

>length(Price)

[1] 804

452/804

[1] 0.5621891

1. Find the mean and median price.  Which is larger?  Why does this make sense?

>mean(Price)

[1] 21343.14

>median(Price)

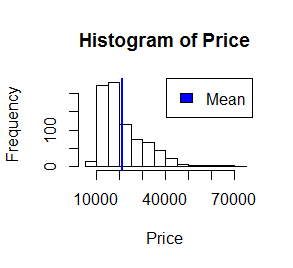
[1] 18025

**The mean is LARGER, This is common in right-skewed data, because mean is more sensitive to extreme values than the median. The mean tends to be pulled toward extreme values**

1. Add a vertical line to the histogram to denote the mean price.  Add a legend to the graph.

>abline(v = mean(Price), col = "blue", lwd = 2)

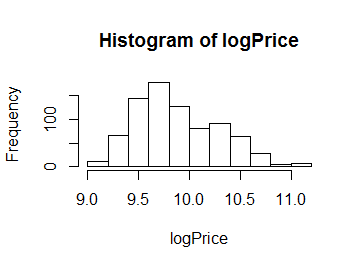
>legend("topright", legend = c("Mean"), fill = c("blue"))



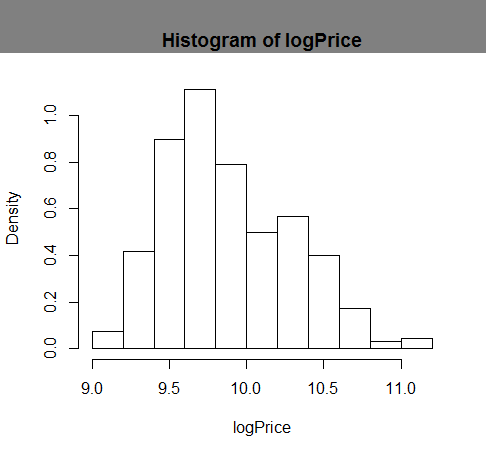
1. Transform the price to reduce its skew, and make a histogram of the transformed price.  Fit a normal distribution to the transformed price, and graph the normal density curve on the same plot as the histogram.  How well does a normal distribution fit the transformed data?

>logPrice=log(Price)

>hist(logPrice)



>hist(logPrice,prob = T)



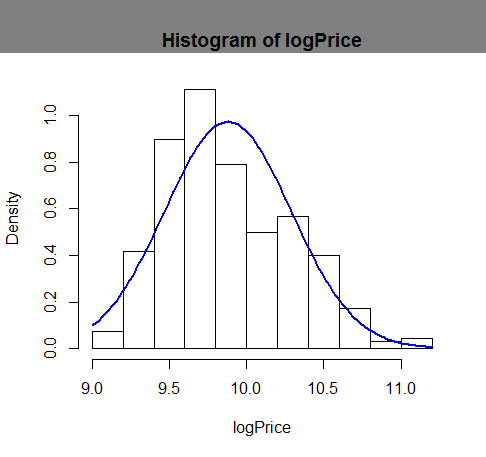
>mean(logPrice)

[1] 9.87905

>sd(logPrice)

[1] 0.4101115

>curve(dnorm(x, 9.87905, 0.4101115), add = T, col = "blue", lwplotd = 2 )



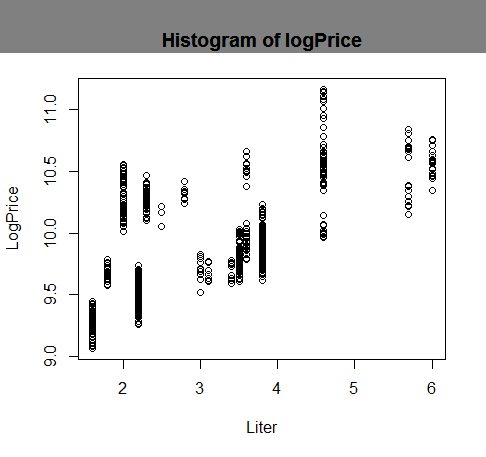
**Seems like it’s not perfect but it works and decent compared to what it would have been without the Log transformation**

1. Make a scatterplot of transformed price versus engine size, measured in liters.  Describe the relationship between these two variables.

>LogPrice =log(Price)

>plot(Liter,LogPrice)

>plot(Liter,LogPrice, col = c("red","black")[as.factor(Leather)])



**Seems to be showing that the size of the engine versus price is having a positive trend. But it is no way definite. 5 and 6 liter engines seem to be more expensive and right fully so as seen on plot(Liter, LogPrice) scatter plot**

1. Find the correlation between transformed price and engine size in liters.  Explain what it tells us.

>cor(LogPrice,Liter)

[1] 0.5904097

**This correlation means that the linear relationship between LogPrice and Liter is moderately Strong with .59 as the value of correlation**

1. Modify your scatterplot in part g to use one color of plotting symbol for cars with leather

interiors, and a different color for cars without leather interiors.  Add a legend to your plot.

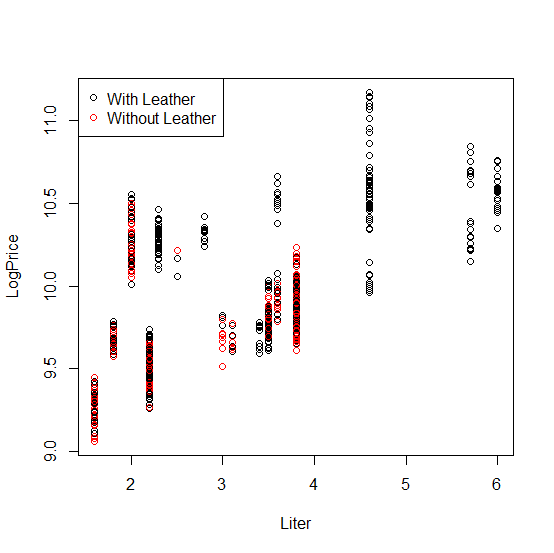
>toChange = which(Leather == 1)

>mycolors<- vector()

>mycolors[toChange ] = "red"

>plot(Liter,LogPrice, col = c("red","black")[as.factor(Leather)])

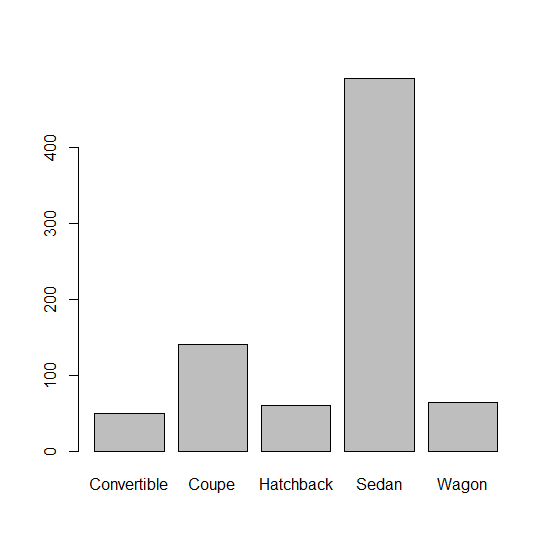
>legend( "topleft", legend = c("With Leather", "Without Leather"), col = c("black", "red"), pch = 21)



1. Make a barplot of the types (Sedan, Hatchback, etc.) of cars in the data.

>counts = table(Type)

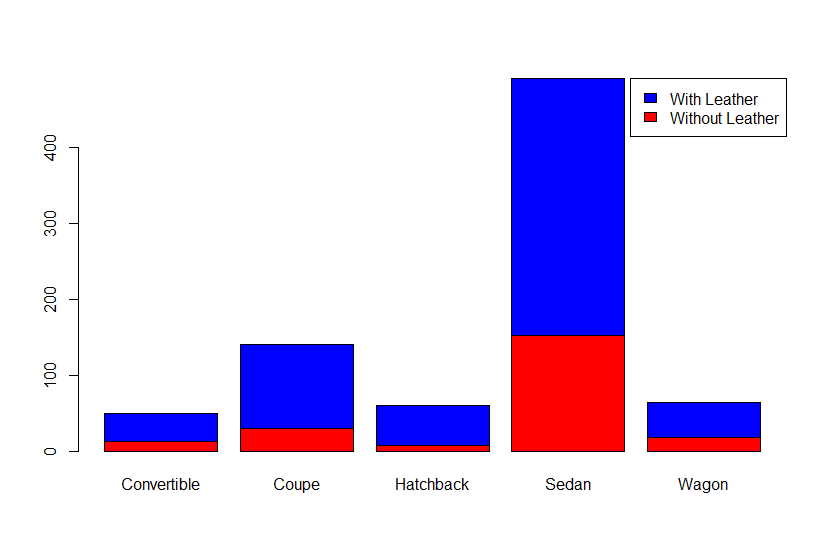
>barplot(counts)



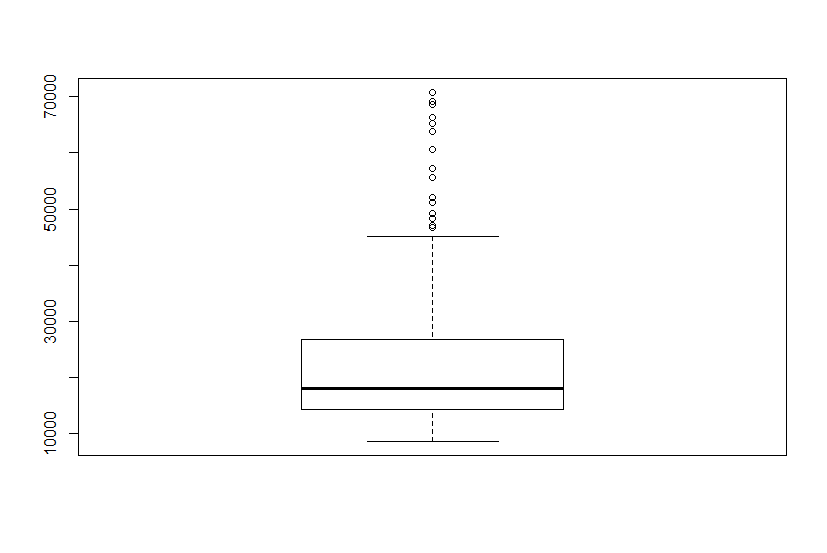
1. Make a barplot of the types of cars and whether they have leather interiors.  Add a legend to your plot.

>barplot( counts, col = c("red", "blue") )

>legend( "topright", legend = c("With Leather", "Without Leather"), fill = c("blue", "red") )



1. Make a boxplot of (untransformed) price by type of car.  In words, summarize what it shows.



**Its skewed right, and seems like the range us from below 10,000 to upward of 70,000. Q2 (median is set at 14,000. InterQuartile range is from 13,00 0 to 25,000 = q3 – q1**

1. Create two different histograms in a vertical stack that allow comparison of (untransformed) price according to whether the car has a leather interior.  Use the same horizontal axis for each to enable comparison, and use informative labels for each graph and the x-axis.

withLeather = which( Leather == 1 )

priceWithLeather = Price[ withLeather ]

noLeather = which( Leather == 0 )

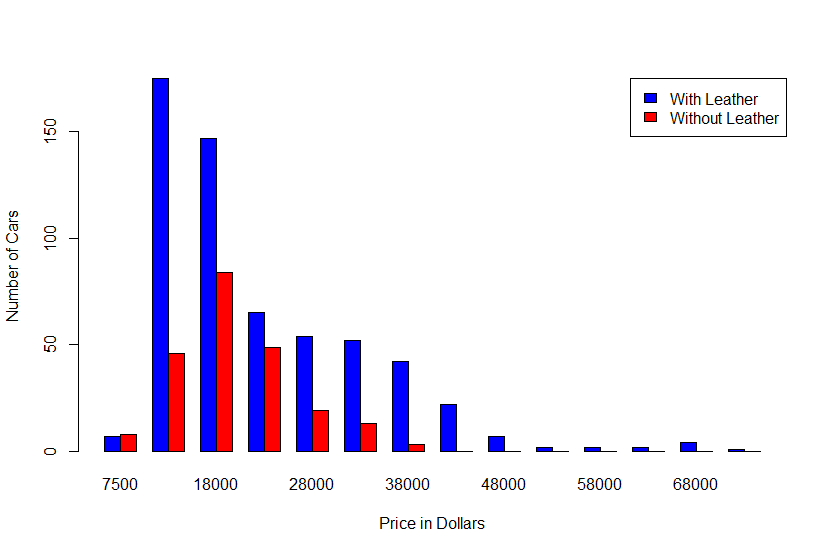
priceNoLeather = SalePrice[ noLeather ]

>hist(priceNoLeather, xlab = "Price in Dollars", ylab ="Number of Cars", main = "Price without Leather")

>hist(priceWithLeather, xlab = "Price in Dollars", ylab ="Number of Cars", main = "Price with Leather")



1. Create a side-by-side histogram to allow the same comparison as in part l.  Add a legend to your plot.



>multhist( list(priceWithLeather, priceNoLeather), col=c("blue", "red") ,xlab = "Price in Dollars", ylab ="Number of Cars")

>legend( "topright", legend = c("With Leather", "Without Leather"), fill = c("blue", "red") )

Submit a .doc, .docx, or .pdf document containing your R code, output and graphs, and interpretations (where requested) to parts a-n.  It is not necessary to include R code where you were testing code or where you made a mistake--just submit the final version.

1. **Analyzing the running speed of mammals**
2. In R, type

install.packages("quantreg")

data(Mammals, package="quantreg")

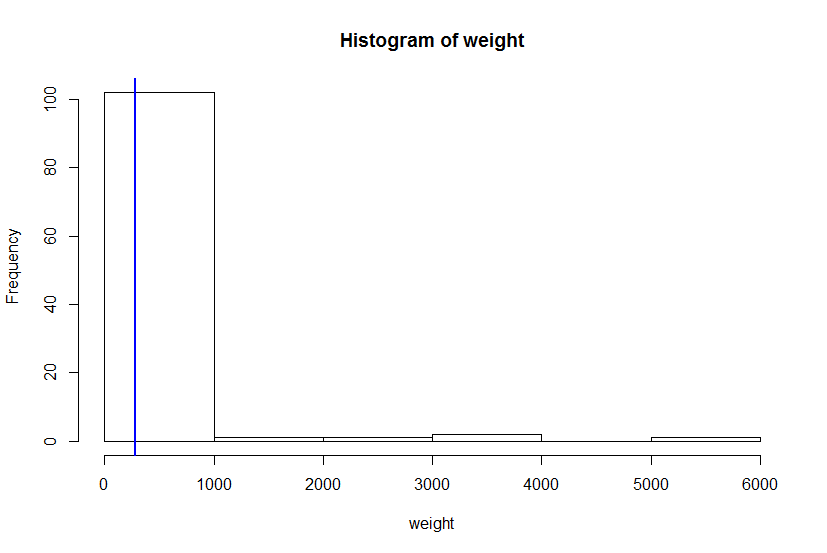
This will load a data set called Mammals, on the maximum land speed of various species of mammal.  Attach the data and look at the first few lines.

(Source:  Garland, T. (1983) The relation between maximal running speed and body mass in terrestrial mammals, *J. Zoology*, 199, 1557-1570.

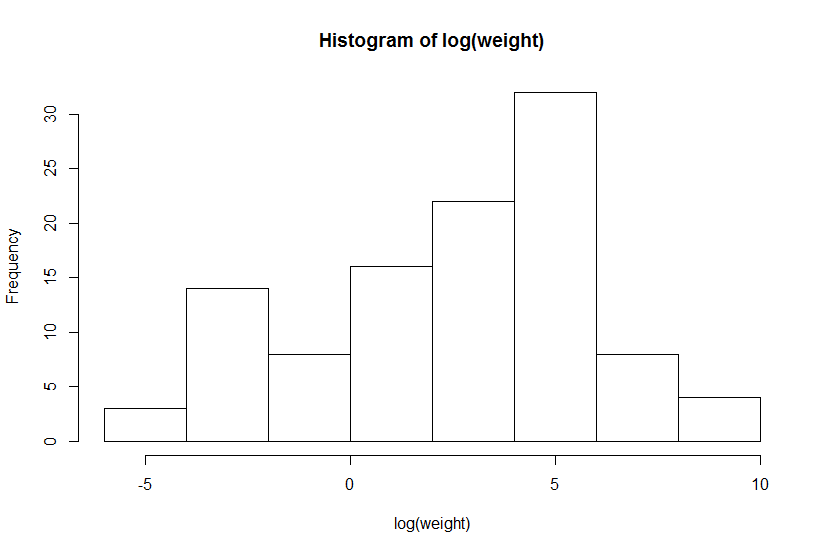
Metadata:  <http://vincentarelbundock.github.io/Rdatasets/doc/quantreg/Mammals.html>, accessed 7 June 2015.)

b. Decide whether either of the quantitative variables should be transformed.  Justify your decision using appropriate plots and/or descriptive statistics.

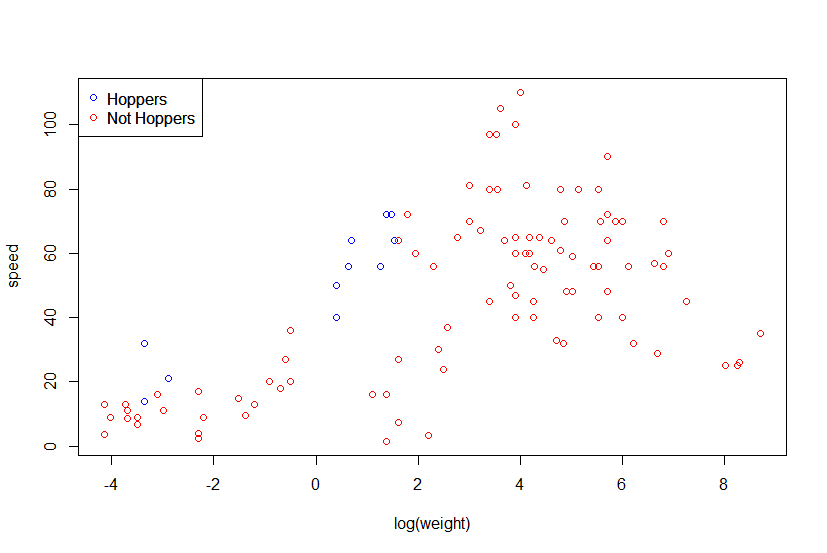
**Heavily skewed Right**



**I think it would be reasonable to use log transform to reduce the skew for Weight Histogram. Below is the Log transformation, as you can see its not more Symmetric. This would help make it fit normal distribution if we transform and go ahead use Log(weight) As we can see below after the log transform we achieve a more normal distribution.**



c.  Use appropriate graphs and/or descriptive statistics to describe the relationship between maximum land speed and body weight.  Does it matter whether the animal is a “hopper” (such as a kangaroo)?  Explain why you chose the graphs and/or statistics that you chose.



>plot(log(weight),speed, col = c("red","blue")[as.factor(hoppers)])

>legend( "topleft", legend = c("Hoppers", "Not Hoppers"), col = c("blue", "red"), pch = 21)

I chose this graph to compare directly hoppers vs non hoppers and also selected their speed against their weigh in transformed Log to have a clear comparison. We can see that as weigh increases the speed also seems to follow path, a moderate trend. I wouldn’t say linear but definitely some weak positive correlation can be observed. I do not however think Hoppers Vs Non Hoppers speed differentiates too much as it seems non hoppers looks pretty consistent with the rest of the non hoppers

>cor(log(weight),speed)

[1] 0.5751193

Shows the correlation between the two as weak as expected

Submit your graphs and written analysis and justifications as a .doc, .docx., or .pdf document.