**# AUTHOR: Brian Pankau**

**# DATE: 9 June 2017**

**# CLASS: UWLAX DS710 SUMMER 2017**

**# ASSIGNMENT: 3 R program**

**# CODE:**

**#**

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.)

ANSWER:

See code

CODE:

cars2005 = read.csv("C:/\_r-workspace/Cars 2005.csv")

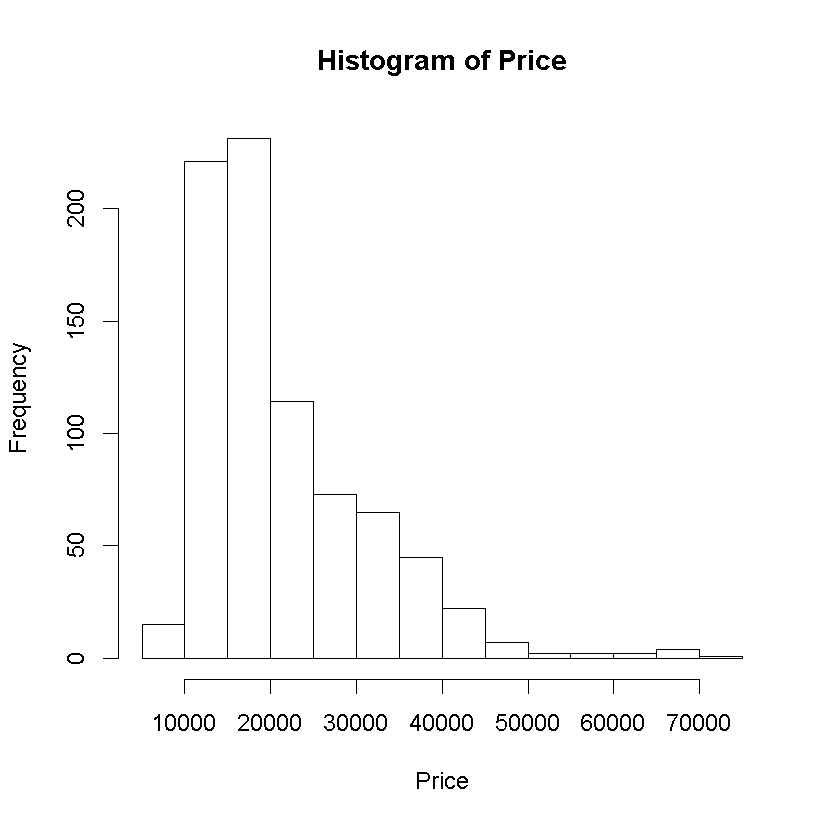
attach(cars2005)

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

ANSWER:  
Description: The histogram of Price is non-normal and is skewed to the right.

CODE & PLOTS:

with(cars2005, hist(Price))



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

ANSWER:

The proportion of cars in the data set that cost between $10K and $20K is 56.22 %.

CODE:

cars2005 = read.csv("C:/\_r-workspace/Cars 2005.csv")

attach(cars2005)

with(cars2005, hist(Price))

h <- hist(Price, breaks=5, plot=FALSE)

h$breaks

# [1] 0 10000 20000 30000 40000 50000 60000 70000 80000

h$counts

# [1] 15 452 187 110 29 4 6 1

h$counts[2]/sum(h$counts)\*100

# [1] 56.21891

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

ANSWER:

Mean = $21,343.14, Median = $18,025

The Mean is larger.

This makes sense since the mean is calculated by summing all the prices and then dividing by the number of observations, while the median is the middle value within the sorted list of prices.

CODE & PLOTS:

cars2005 = read.csv("C:/\_r-workspace/Cars 2005.csv")

attach(cars2005)

with(cars2005, hist(Price, plot=FALSE))

mean(Price)

# [1] 21343.14

median(Price)

# [1] 18025

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

ANSWER:

See code.

CODE & PLOTS:

pmean <- mean(Price)

pmedian <- median(Price)

hist(Price, breaks=5, col="green")

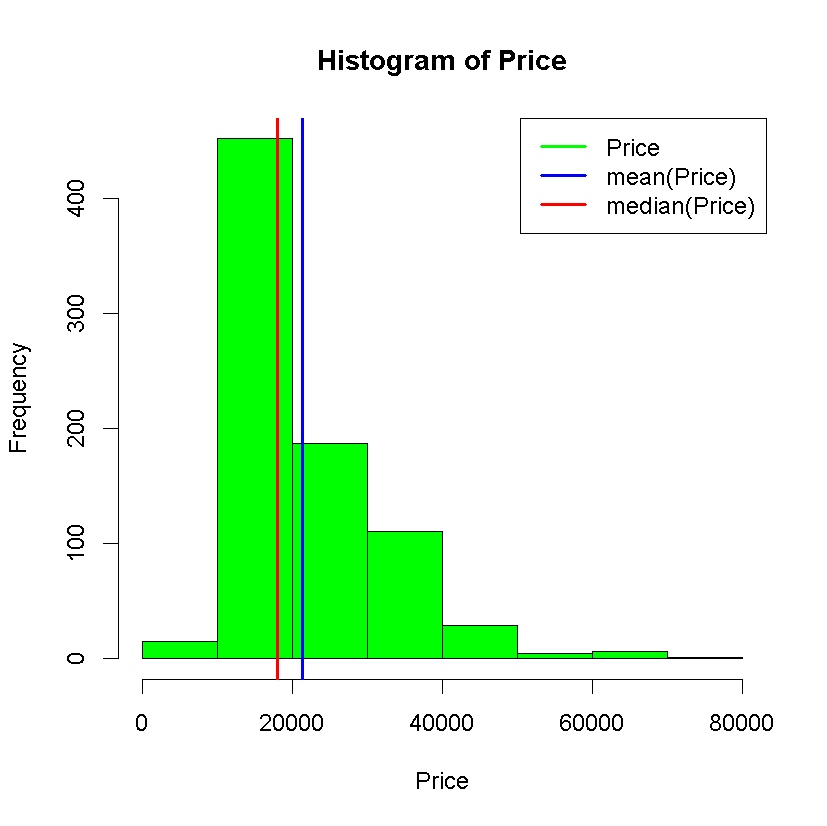
abline(v=pmean, col="blue", lwd=2)

abline(v=pmedian, col="red", lwd=2)

legend("topright",

c("Price", "mean(Price)", "median(Price)"),

col=c("green", "blue", "red"), lwd=2)



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?

ANSWER:

The transform function atan(Price) improves the normalization of the distribution by shifting the distribution to the right.

# skewness(normally distributed) -> 0.0

# skewness(Price) -> 1.56992

# skewness(atan(Price)) -> -0.2504575

CODE & PLOTS:

**1a. FIRST CHARACTERIZE THE NON TRANSFORMED VARIABLE**

# load package e1071 for skewness & kurtosis functions

install.packages('e1071')

library(e1071)

# Assess the skewness and kurtosis of the non-transformed Price variable

skewness(Price)

# [1] 1.56992

kurtosis(Price)

# [1] 3.247693

# Plot the non-transformed Price histogram probability Density function

hist(Price, breaks=5, prob=TRUE)

abline(v=pmedian, col="red", lwd=2)

abline(v=pmean, col="blue", lwd=2)

# calculate the mean and standard deviation for the non-transformed Price

m<-mean(Price)

m

# [1] 21343.14

std<-sqrt(var(Price))

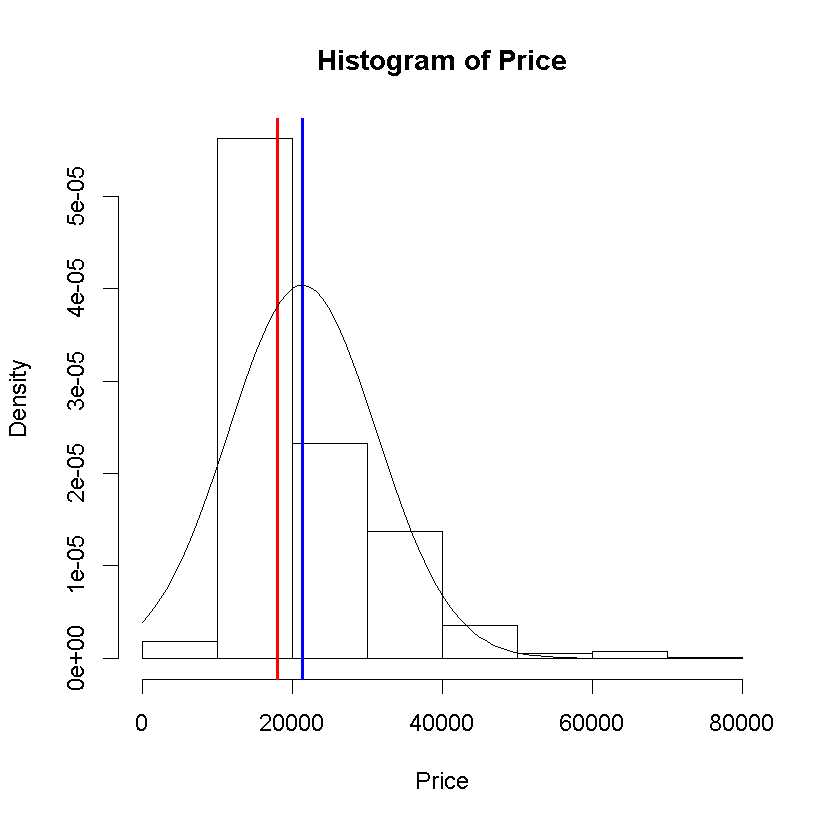
std

# [1] 9884.853

# plot on top of the histogram the density function

curve(dnorm(x, mean=m, sd=std), add=TRUE, main="atan(Price)")

**1b. PLOT OF THE NON-TRANFORMED PRICE VARIABLE**



**2a. TRANSFORM THE PRICE VARIABLE USING ATAN FUNCTION AND ASSESS GOODNESS OF FIT**

# ==> USE A ATAN FUNCTION TO TRANSFORM THE PRICE VARIABLE

# atan function shifts the distribution to the right

# skewness(normally distributed) -> 0.0

# skewness(Price) -> 1.56992

# skewness(atan(Price) improves the normalization of the distribution

skewness(atan(Price))

[1] -0.2504575

kurtosis(atan(Price))

[1] -0.6204877

patanmedian = median(atan(Price))

patanmedian

# [1] 1.570741

patanmean = mean(atan(Price))

patanmean

# [1] 1.570741

m<-mean(atan(Price))

m

# [1] 1.570741

std<-sqrt(var(atan(Price)))

std

# [1] 2.068974e-05

hist(atan(Price), breaks=5, prob=TRUE, main="Histogram of atan(Price)")

abline(v=patanmedian, col="red", lwd=2)

abline(v=patanmean, col="blue", lwd=2)

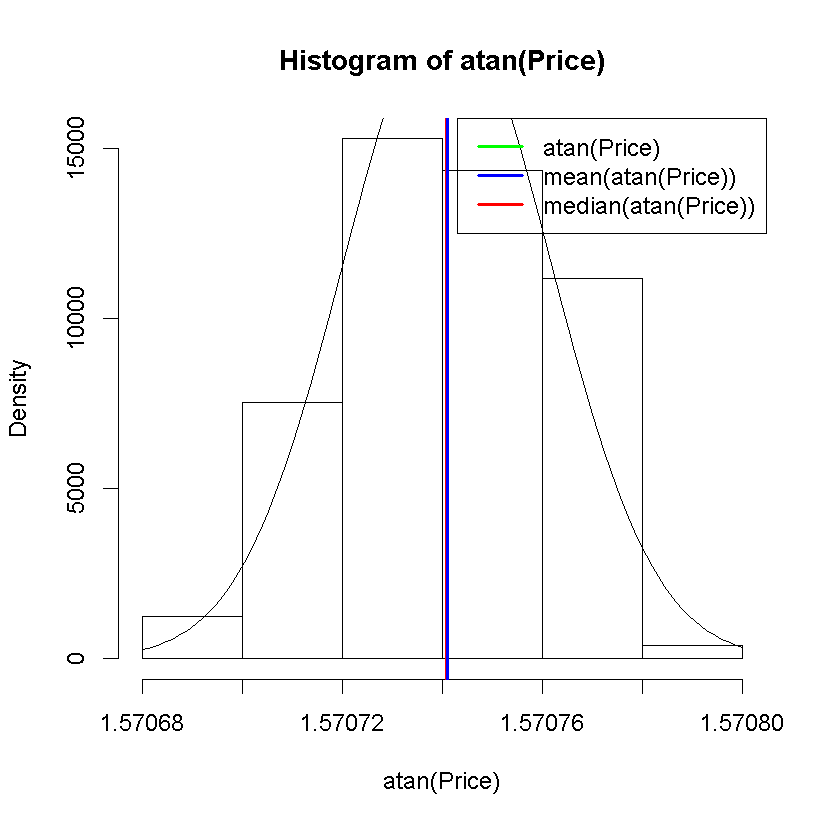
curve(dnorm(x, mean=m, sd=std), add=TRUE, main="atan(Price)")

legend("topright",

c("atan(Price)", "mean(atan(Price))", "median(atan(Price))"),

col=c("green", "blue", "red"), lwd=2)

**2b. PLOT OF THE TRANFORMED PRICE VARIABLE**



**ALSO IMPLEMENTED WITH QPLOT & GGPLOT**

# load ggplot2 library

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create dataframe

df <- data.frame(

+ price=myCars$Price,

+ price\_trans=atan(myCars$Price),

+ liter=myCars$Liter,

+ leather=myCars$Leather)

# PLOT THE PRICE

qplot(df$price,

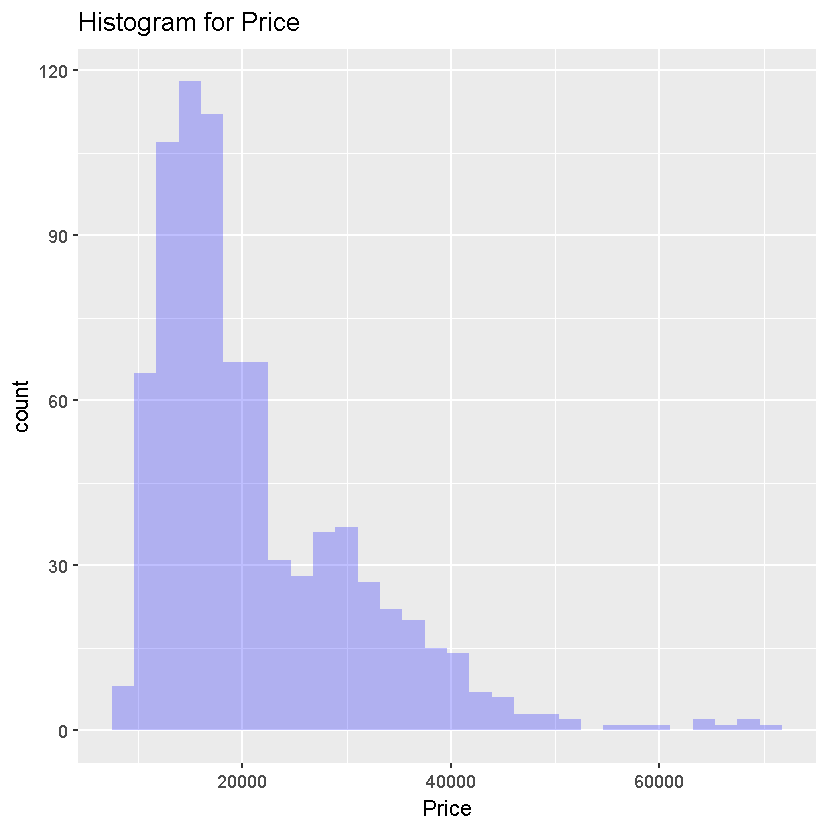
geom="histogram",

main="Histogram for Price",

xlab="Price",

fill=I("blue"),

alpha=I(.25))

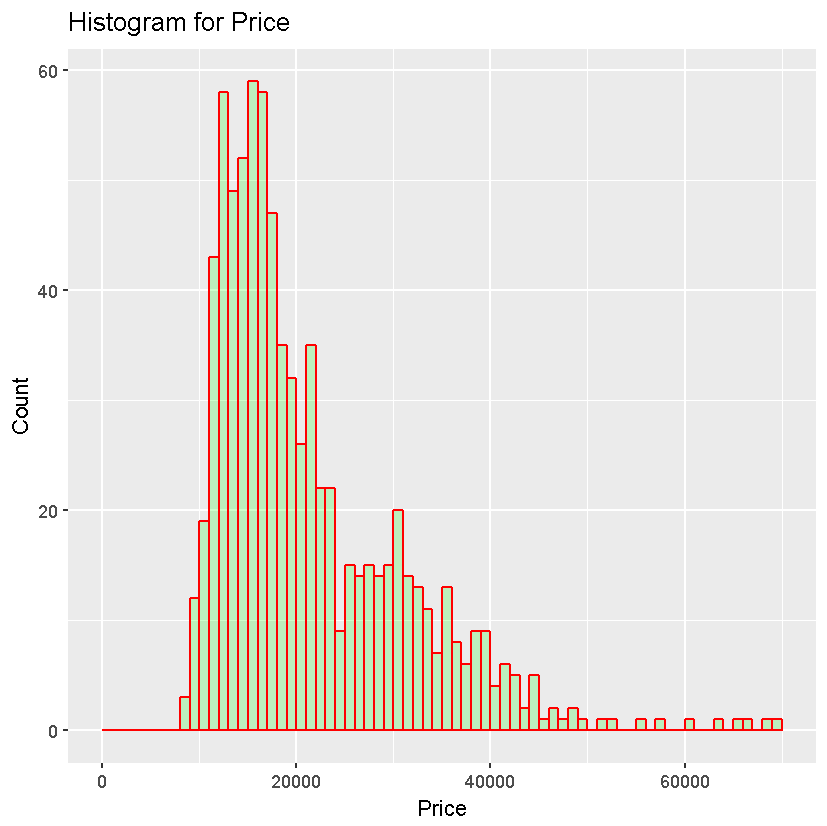


ggplot(data=df,aes(df$price))

+ geom\_histogram(breaks=seq(0,70000,by=1000),col="red",fill="green",alpha=.2)

+ labs(title="Histogram for Price")

+ labs(x="Price",y="Count")



# PLOT THE TRANSFORMED PRICE

qplot(df$price\_trans,

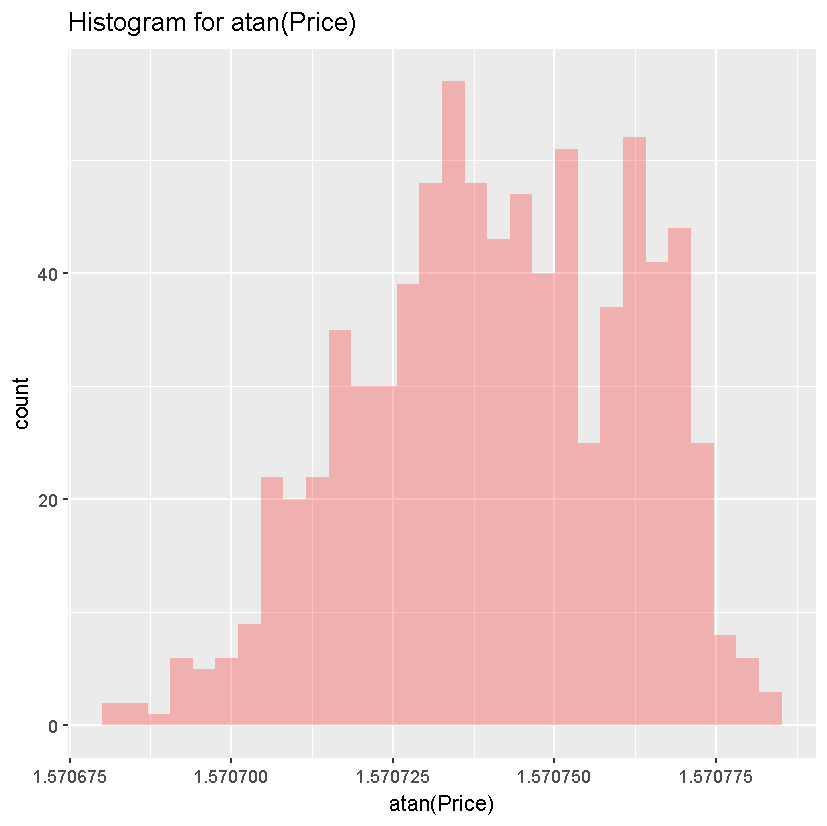
geom="histogram",

main="Histogram for atan(Price)",

xlab="atan(Price)",

fill=I("red"),

alpha=I(.25))

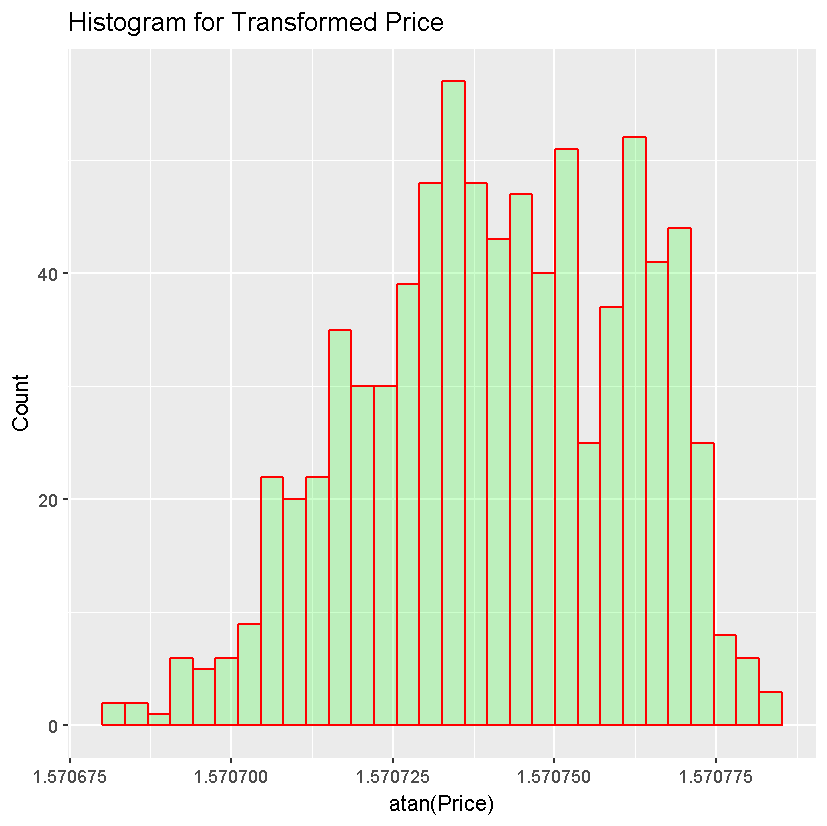


ggplot(data=df,aes(df$price\_trans))

+ geom\_histogram(col="red",fill="green",alpha=.2)

+ labs(title="Histogram for Transformed Price")

+ labs(x="atan(Price)",y="Count")



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

ANSWER:

As the size of the engine of the car increases, so does the price for the car positively increase.

CODE & PLOTS:

cars2005 = read.csv("C:/\_r-workspace/Cars 2005.csv")

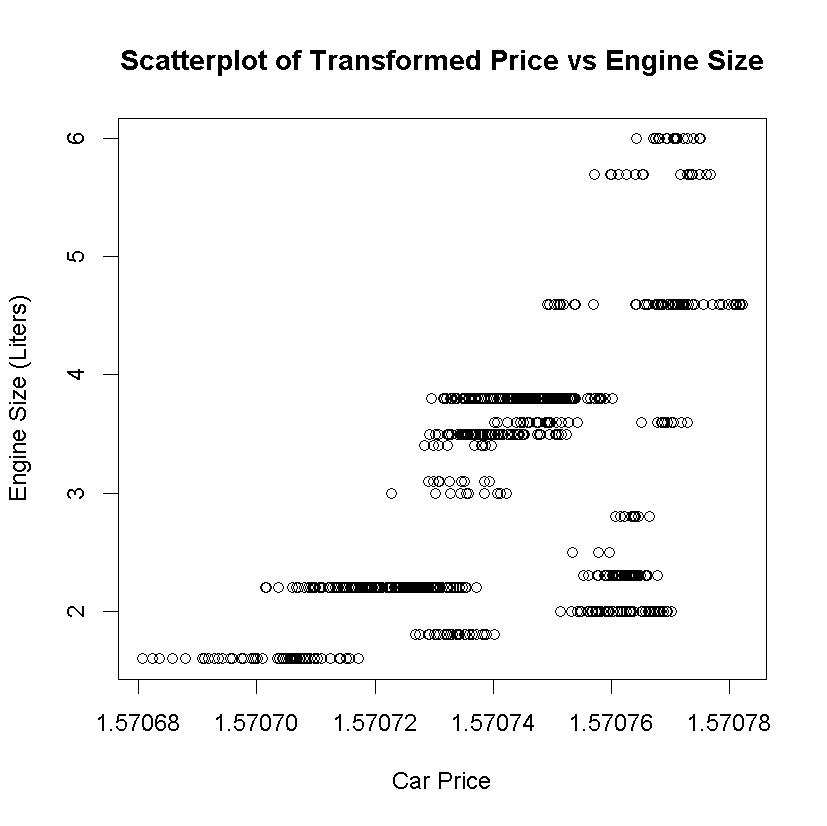
attach(cars2005)

plot(atan(Price),Liter,

main="Scatterplot of Transformed Price vs Engine Size",

xlab="Car Price",

ylab="Engine Size (Liters)")



# GGPLOT IMPLEMENTATION

# load ggplot2 library

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create dataframe

df <- data.frame(

+ price=myCars$Price,

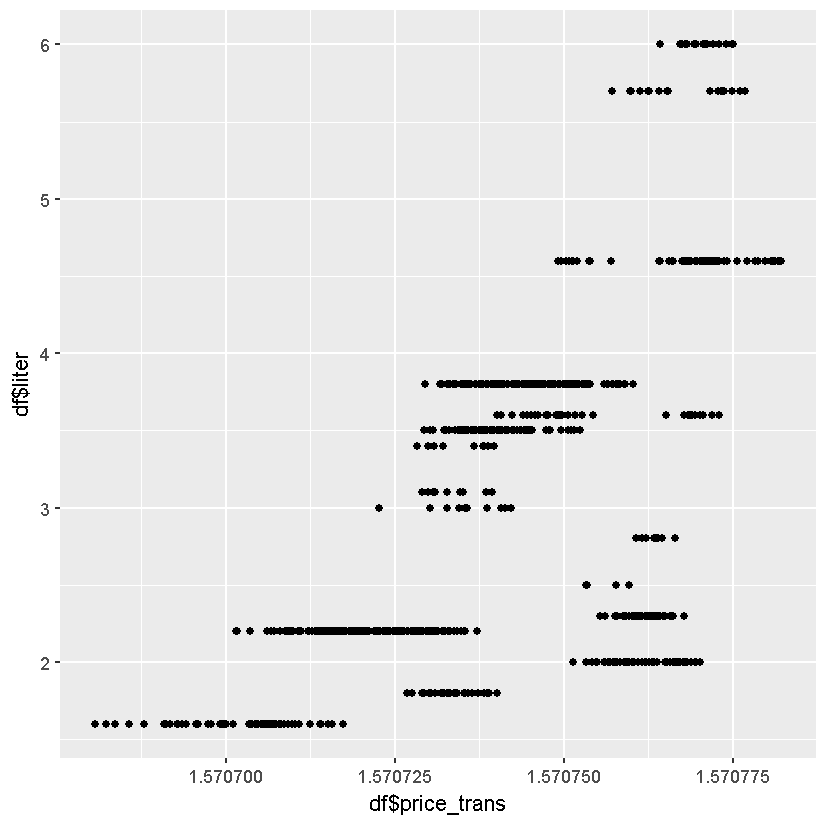
+ price\_trans=atan(myCars$Price),

+ liter=myCars$Liter,

+ leather=myCars$Leather)

# PLOT THE TRANSFORMED PRICE

ggplot(df, aes(x=df$price\_trans, y=df$liter)) + geom\_point()



# GGPLOT IMPLEMENTATION

# load ggplot2 library

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create dataframe

df <- data.frame(

+ price=myCars$Price,

+ price\_trans=atan(myCars$Price),

+ liter=myCars$Liter,

+ leather=myCars$Leather)

# PLOT THE PRICE

ggplot(data=df,aes(df$price))

+ geom\_histogram(breaks=seq(0,70000,by=1000),col="red",fill="green",alpha=.2)

+ labs(title="Histogram for Price")

+ labs(x="Price",y="Count")

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

ANSWER:

The correlation coefficient for the relationship between the non-transformed variable *price* and *liters* is 0.5581458 indicating that the relationship is positively linearly related and the relationship has an above moderate strength (above 0.5).

The correlation coefficient for the relationship between the transformed variable *price\_trans* and *liters* is 0.6017405 indicating that the relationship is positively linearly related and the relationship has an above moderate strength (above 0.5).

A comparison between the non-transformed *price* variable vs transformed *price\_trans* variable correlaton coficient results indicates that the application of the atan()transform function against the *price* variable (generating *price\_trans*) results in a stronger linear positive relationship with engine size in liters.

CODE & PLOTS:

setwd("c:/\_r-workspace")

dir()

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

summary(myCars)

names(myCars)

myCars$Make

attributes(myCars)

summary(myCars$Price)

# create a dataframe with Price, atan(Price), Liters

corr\_df <- data.frame(

price=myCars$Price,

price\_trans=atan(myCars$Price),

liter=myCars$Liter)

# Assess the new dataframe

summary(corr\_df$price\_trans)

hist(corr\_df$price\_trans, breaks=5)

# Calculate the correlation coefficient for price vs liters

> cor(corr\_df$price, corr\_df$liter)

# [1] 0.5581458

# Calculate the correlation coefficient for price\_trans vs liters

> cor(corr\_df$price\_trans, corr\_df$liter)

# [1] 0.6017405

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.

ANSWER:

See Code.

CODE & PLOTS:

# install graphics plotting library

install.packages('ggplot2')

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create dataframe

df <- data.frame(

price=myCars$Price,

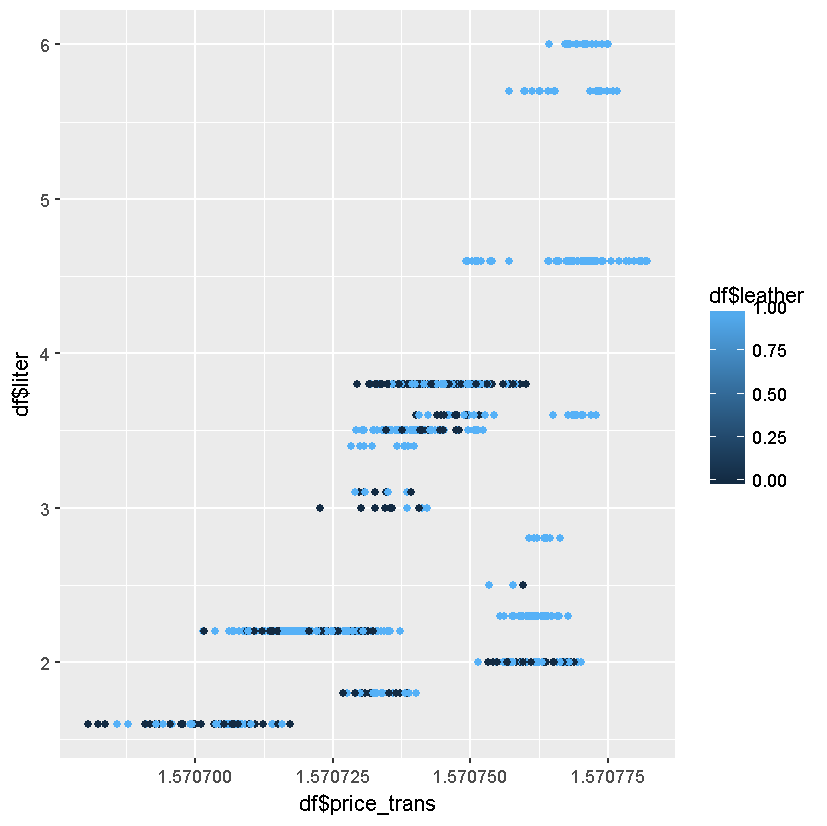
price\_trans=atan(myCars$Price),

liter=myCars$Liter,

leather=myCars$Leather)

# generate a scatterplot

ggplot(df, aes(x=df$price\_trans, y=df$liter, color=df$leather)) + geom\_point()



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

ANSWER:

See Code.

CODE & PLOTS:

# install graphics plotting library

install.packages('ggplot2')

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create master dataframe

df <- data.frame(

price=myCars$Price,

mileage=(myCars$Mileage),

model=myCars$Model,

trim=myCars$Trim,

type=myCars$Type,

cylinder=myCars$Cylinder,

liter=myCars$Liter,

doors=myCars$Doors,

cruise=myCars$Cruise,

sound=myCars$Sound,

leather=myCars$Leather,

price\_trans=atan(myCars$Price))

# determine the categorical counts for type

summary(df$type)

# Convertible Coupe Hatchback Sedan Wagon

# 50 140 60 490 64

# create a dataframe for the boxplot

df\_type <- data.frame(car\_type=c("Convertible", "Coupe", "Hatchback", "Sedan", "Wagon"),+

len=(c(50,140,60,490,64)))

# verify the dataframe

head(df\_type)

# car\_type len

# 1 Convertible 50

# 2 Coupe 140

# 3 Hatchback 60

# 4 Sedan 490

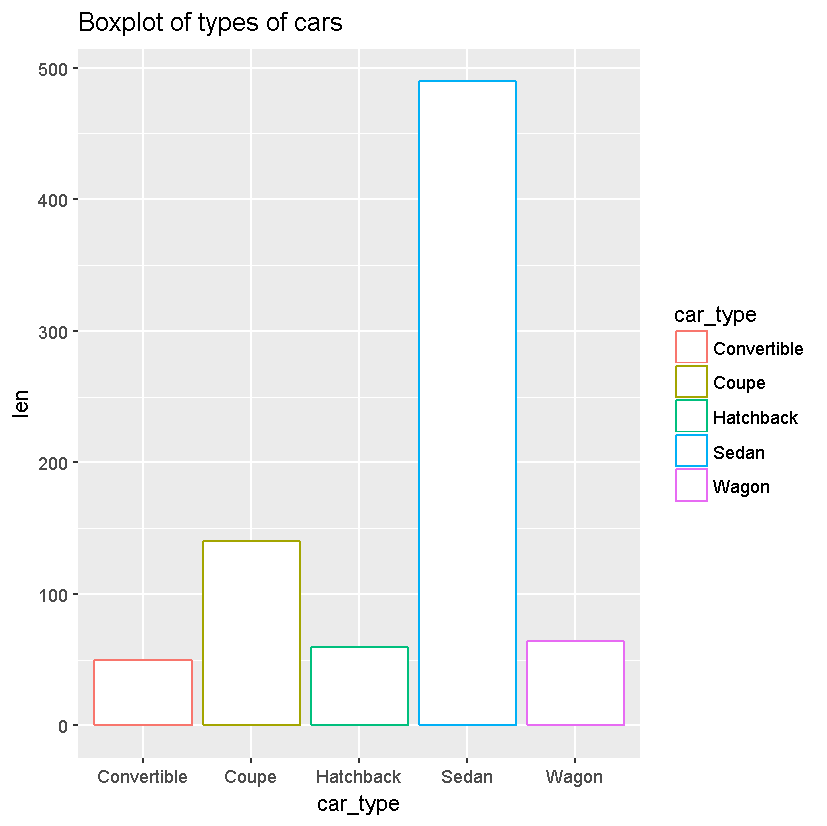
# 5 Wagon 64

# generate the boxplot

ggplot(data=df\_type, aes(x=car\_type, y=len, color=car\_type))+

geom\_bar(stat="identity", fill="white")+

labs(title="Boxplot of types of cars")



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

# map out counts

#car\_type: Convertible Coupe Hatchback Sedan Wagon

#car\_count: 50 140 60 490 64

#leather\_yes: 37 110 52 337 46

#leather\_no: 13 30 8 153 18

# create a dataframe

df\_cars\_interior <- data.frame(

interior = c("leather","leather","leather","leather","leather","fabric","fabric","fabric","fabric","fabric"),

car\_type = c("Convertible","Coupe","Hatchback","Sedan","Wagon","Convertible","Coupe","Hatchback","Sedan","Wagon"),

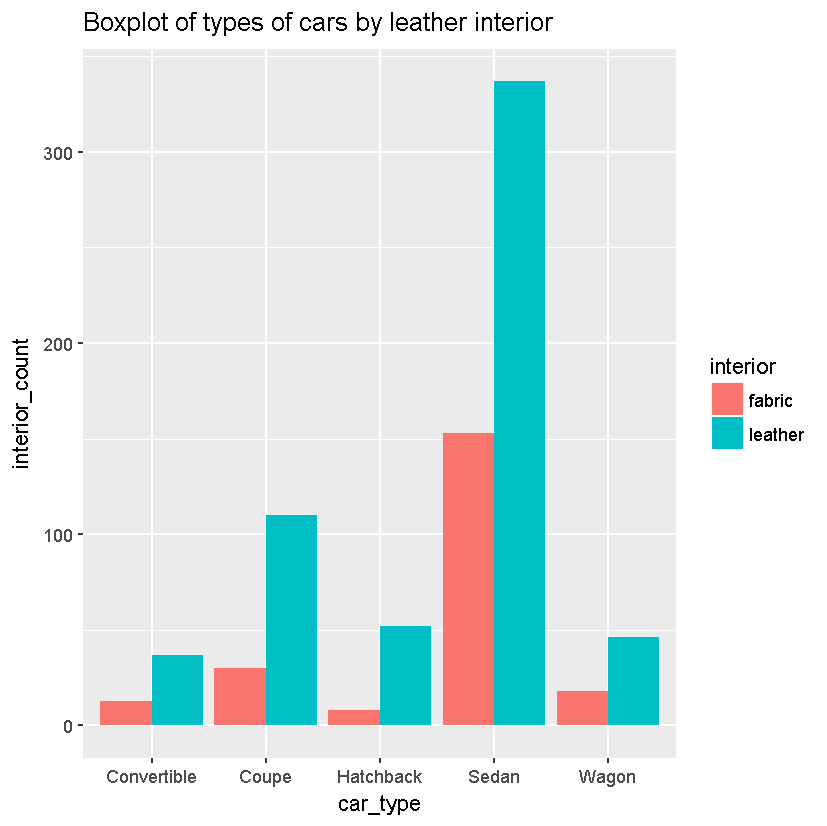
interior\_count = c(37,110,52,337,46,13,30,8,153,18))

# generate the boxplot

ggplot(data=df\_cars\_interior, aes(x=car\_type, y=interior\_count, fill=interior))+

geom\_bar(stat="identity", position=position\_dodge())+

labs(title="Boxplot of types of cars by leather interior")



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

ANSWER:

On average, convertibles are the most expensive, hatchbacks are the least expensive; sedans are the most popular, while the convertibles are the least popular.

CODE & PLOTS:

# install graphics plotting library

install.packages('ggplot2')

library(ggplot2)

# read data file

setwd("c:/\_r-workspace")

myCars <-read.csv(file="myCars2005.csv", head=TRUE, sep=",")

# create master dataframe

df <- data.frame(

price=myCars$Price,

mileage=(myCars$Mileage),

model=myCars$Model,

trim=myCars$Trim,

type=myCars$Type,

cylinder=myCars$Cylinder,

liter=myCars$Liter,

doors=myCars$Doors,

cruise=myCars$Cruise,

sound=myCars$Sound,

leather=myCars$Leather,

price\_trans=atan(myCars$Price))

# Determine the average price per type of car -- map out counts

car\_type: Convertible,Coupe,Hatchback,Sedan,Wagon

car\_count: 50,140,60,490,64

avg\_price: 40831.71,17726.93,14170.93,21067.93,22859.25

# create a dataframe

df\_cars\_avg\_price <- data.frame(

car\_type = c("Convertible","Coupe","Hatchback","Sedan","Wagon"),

car\_count = c(50,140,60,490,64),

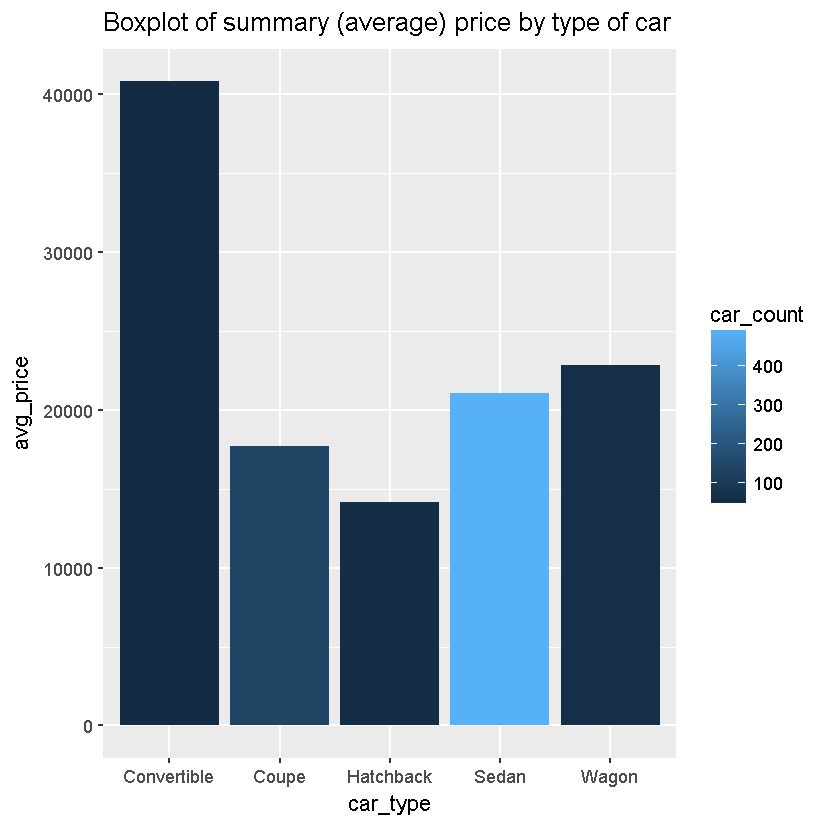
avg\_price = c(40831.71,17726.93,14170.93,21067.93,22859.25))

# generate the boxplot

ggplot(data=df\_cars\_avg\_price, aes(x=car\_type, y=avg\_price, fill=car\_count))+

geom\_bar(stat="identity", position=position\_dodge())+

labs(title="Boxplot of summary (average) price by type of car")



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.

ANSWER:

See Code.

CODE & PLOTS:

# map out counts – leather

#car\_type: Convertible,Coupe,Hatchback,Sedan,Wagon

#car\_count: 37,110,52,337,46

#car\_price: 43675.39,17756.37,14723.66,22211.93,25209.96

# map out counts – fabric

#car\_type: Convertible,Coupe,Hatchback,Sedan,Wagon

#car\_count: 13,30,8,153,18

#car\_price: 32738.13,17619.0,10578.25,18548.12,16851.86

# create a dataframe - leather

df\_cars\_leather <- data.frame(

car\_type = c("Convertible","Coupe","Hatchback","Sedan","Wagon"),

car\_count = c(37,110,52,337,46),

avg\_price = c(43675.39,17756.37,14723.66,22211.93,25209.96))

# create a dataframe - fabric

df\_cars\_fabric <- data.frame(

car\_type = c("Convertible","Coupe","Hatchback","Sedan","Wagon"),

car\_count = c(13,30,8,153,18),

avg\_price = c(43675.39,17756.37,14723.66,22211.93,25209.96))

# generate the boxplot - leather

l<-ggplot(data=df\_cars\_leather, aes(x=car\_type, y=avg\_price, fill=car\_count))+

geom\_bar(stat="identity", position=position\_dodge())+

labs(title="Boxplot average price by car type with leather")

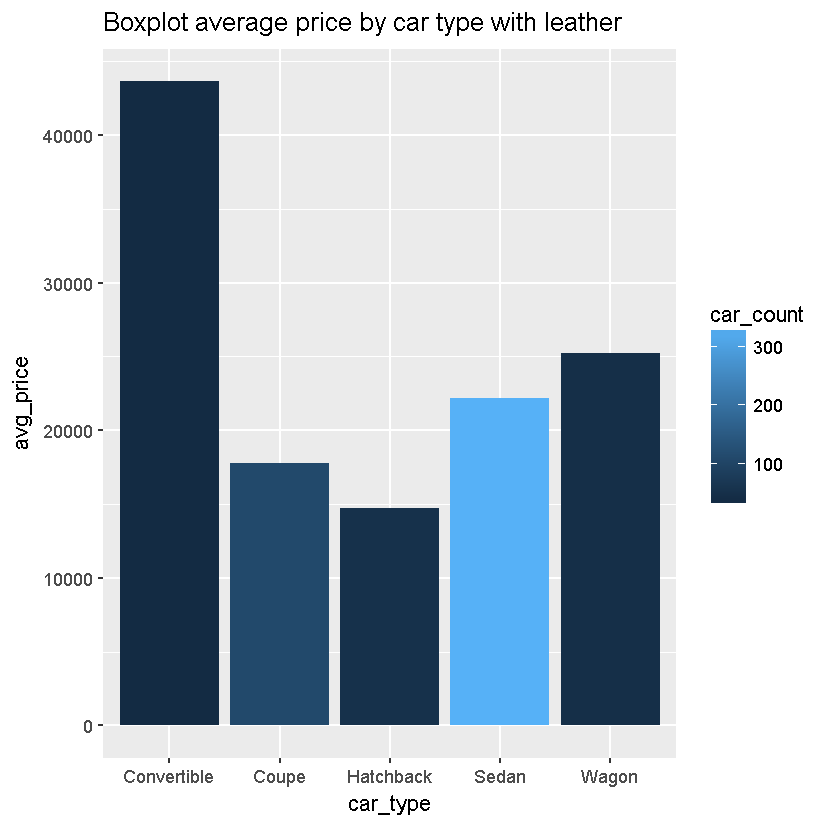
f<-ggplot(data=df\_cars\_fabric, aes(x=car\_type, y=avg\_price, fill=car\_count))+

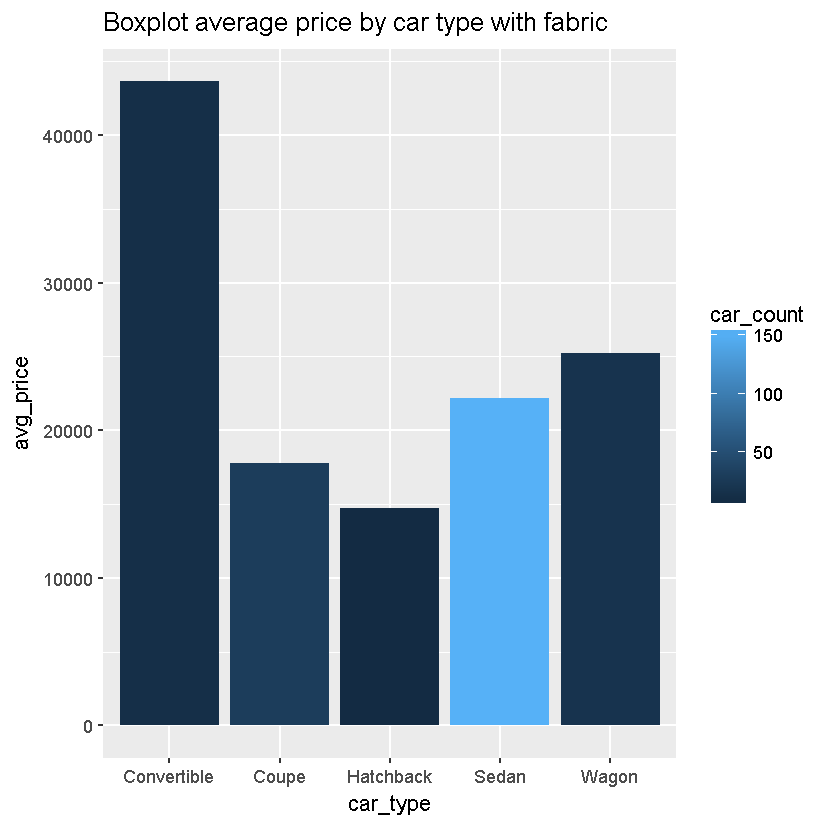
geom\_bar(stat="identity", position=position\_dodge())+

labs(title="Boxplot average price by car type with fabric")

#output the barcharts

l





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

# map out counts

car\_type:Convertible,Coupe,Hatchback,Sedan,Wagon,Convertible,Coupe,Hatchback,Sedan,Wagon

average\_price:43675.39,17756.37,14723.65,22211.93,25209.96,32738.13,17619.0,10578.25,18548.12,16851.86

interior: "leather","leather","leather","leather","leather","fabric","fabric","fabric","fabric","fabric"

interior\_count:37,110,52,337,46,13,30,8,153,18

# create a dataframe

df\_cars\_other <- data.frame(

car\_type = c("Convertible","Coupe","Hatchback","Sedan","Wagon","Convertible","Coupe","Hatchback","Sedan","Wagon"),

avg\_price = c(43675.39,17756.37,14723.65,22211.93,25209.96,32738.13,17619.0,10578.25,18548.12,16851.86),

interior = c("leather","leather","leather","leather","leather","fabric","fabric","fabric","fabric","fabric"),

car\_count = c(37,110,52,337,46,13,30,8,153,18))

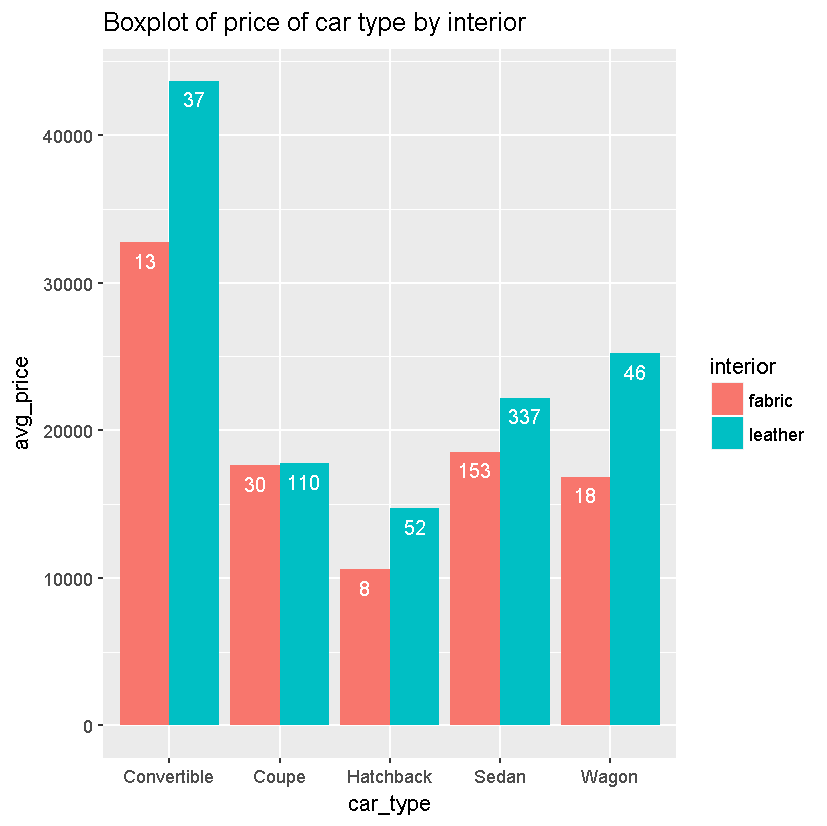
# generate the boxplot

ggplot(data=df\_cars\_other, aes(x=car\_type, y=avg\_price, fill=interior))+

geom\_bar(stat="identity", position=position\_dodge())+

geom\_text(aes(label=car\_count), vjust=1.6, color="white", position = position\_dodge(0.9), size=3.5)+

labs(title="Boxplot of price of car type by interior")



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.)

data("Mammals", package="quantreg")

head(Mammals)

# weight speed hoppers specials

# 1 6000 35 FALSE FALSE

# 2 4000 26 FALSE FALSE

# 3 3000 25 FALSE FALSE

# 4 1400 45 FALSE FALSE

# 5 400 70 FALSE FALSE

# 6 350 70 FALSE FALSE

summary(Mammals)

# weight speed hoppers specials

# Min. : 0.016 Min. : 1.60 Mode :logical Mode :logical

# 1st Qu.: 1.700 1st Qu.: 22.50 FALSE:96 FALSE:97

# Median : 34.000 Median : 48.00 TRUE :11 TRUE :10

# Mean : 278.688 Mean : 46.21

# 3rd Qu.: 142.500 3rd Qu.: 65.00

# Max. :6000.000 Max. :110.00

length(Mammals$weight)

[1] 107

length(Mammals$speed)

[1] 107

length(Mammals$hoppers)

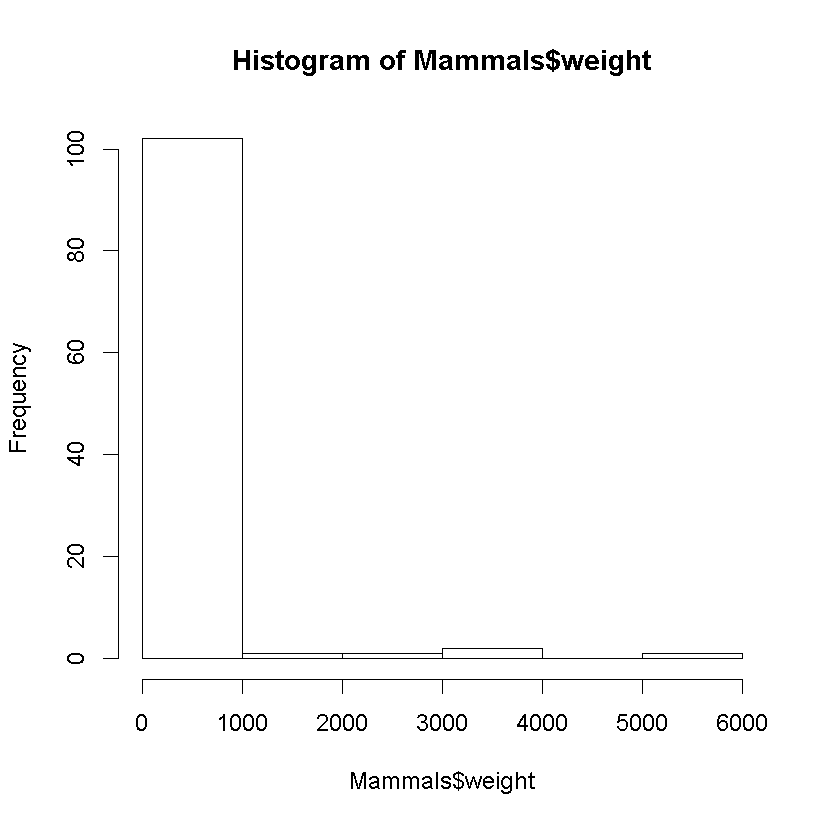
[1] 107

length(Mammals$specials)

[1] 107

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

hist(Mammals$weight)



OBSERVATION: WEIGHT IS SKEWED TO THE RIGHT. REQUIRES TRANSFORMATION – USE WEIGHT ^ 0.0625

hist((Mammals$weight)^.0625)

m=mean(Mammals$weight^.0625)

m

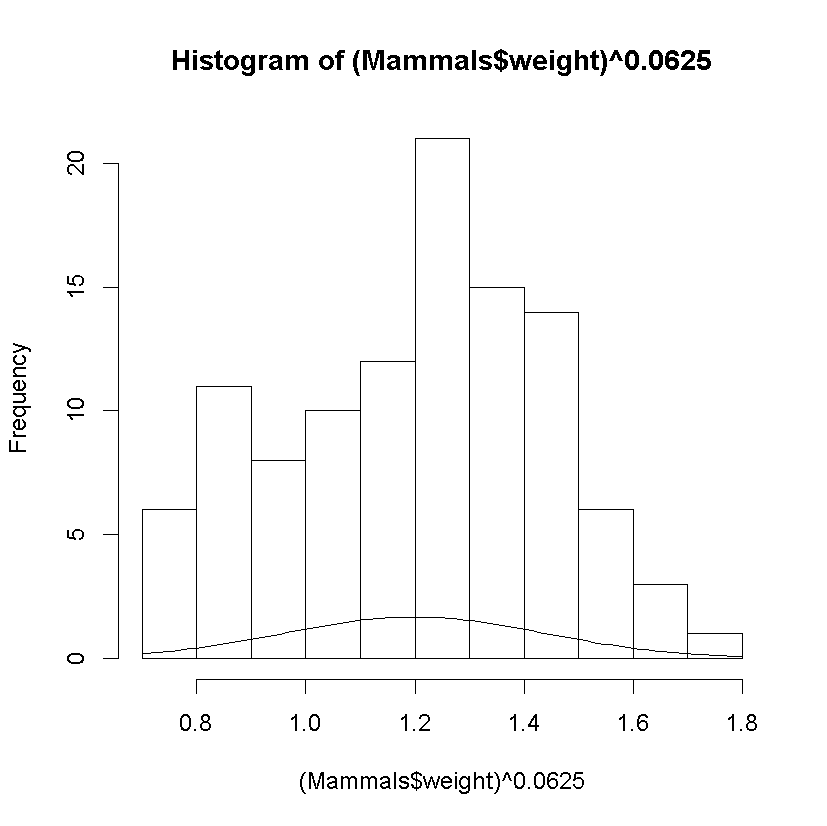
# [1] 1.199333

std=sqrt(var(Mammals$weight^.0625))

std

# [1] 0.2388167

curve(dnorm(x, mean=m, sd=std), add=TRUE, main="(Mammals$weight)^0.0625")



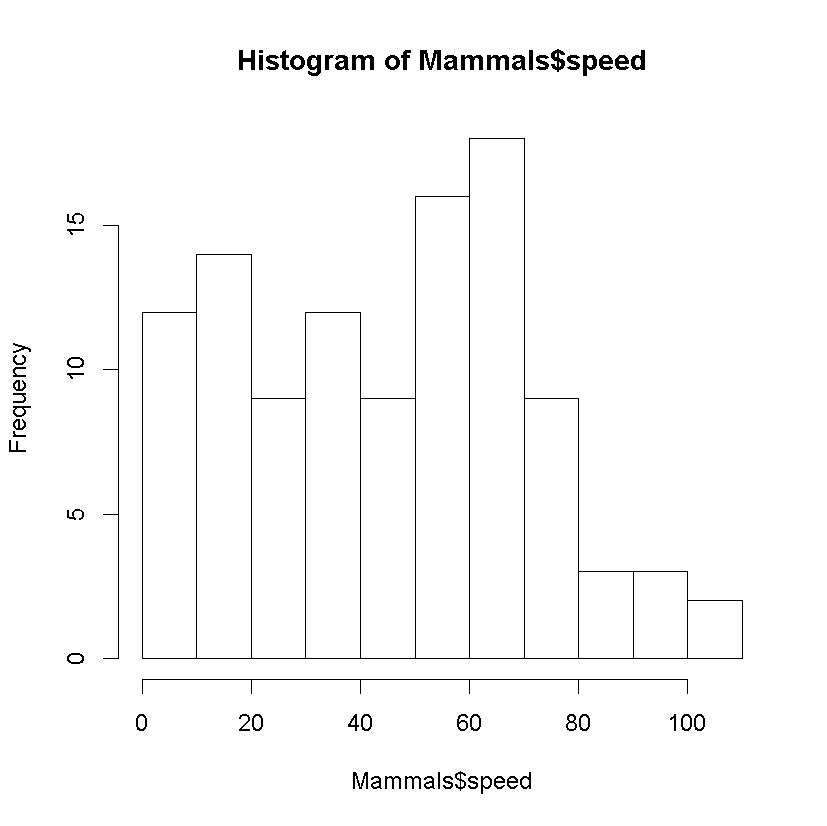
skewness(Mammals$weight)

# [1] 4.797331

skewness(Mammals$weight^0.0625)

# [1] -0.1531968

hist(Mammals$speed)



OBSERVATION: SPEED IS NON-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.

# create a dataframe with weight\_trans, , Liters

corr\_df <- data.frame(

weight=Mammals$weight,

weight\_trans=Mammals$weight^0.0265,

speed=Mammals$speed,

hop=Mammals$hoppers,

spec=Mammals$specials)

# Assess the new dataframe

summary(corr\_df$weight\_trans)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 0.8962 1.0140 1.0980 1.0747 1.1404 1.2593

hist(corr\_df$weight\_trans, breaks=5)

# Calculate the correlation coefficient for weight vs speed

cor(corr\_df$weight, corr\_df$speed)

# [1] -0.06653467

# Calculate the correlation coefficient for weight\_trans vs speed

cor(corr\_df$weight\_trans, corr\_df$speed)

# [1] 0.5573166

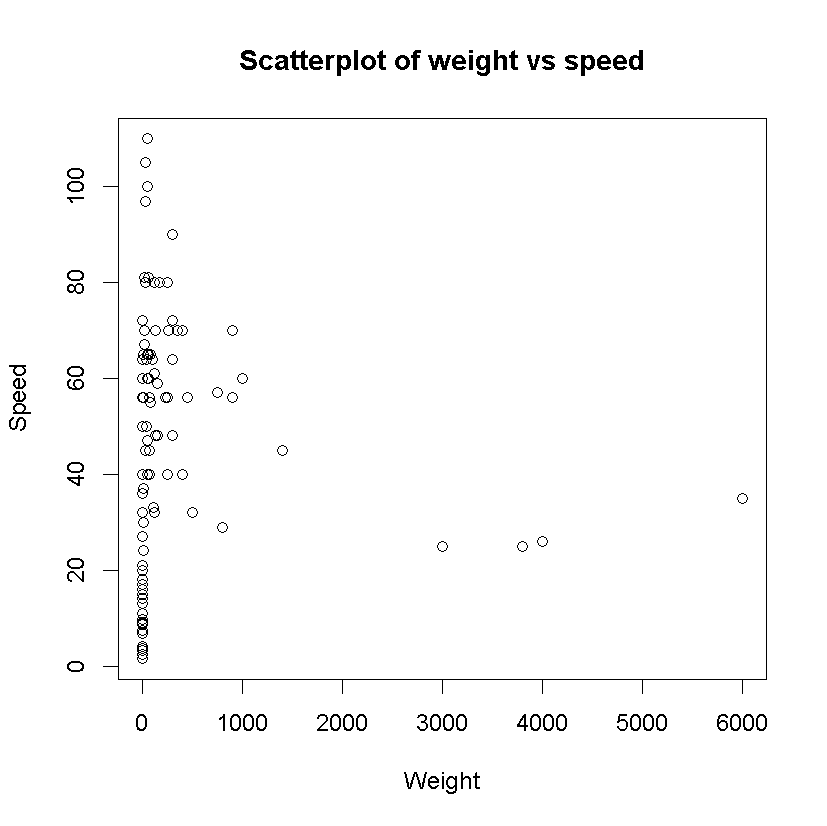
# generate a scatterplot

plot(Mammals$weight,Mammals$speed,

main="Scatterplot of weight vs speed",

xlab="Weight",

ylab="Speed")



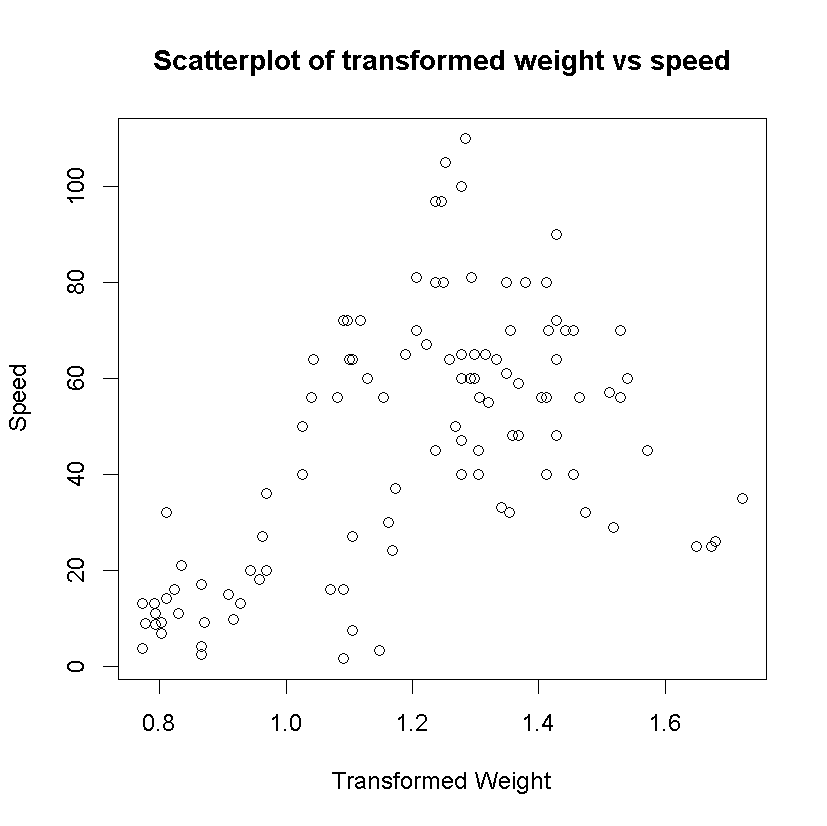
# generate a scatterplot

plot(Mammals$weight^0.0625,Mammals$speed,

main="Scatterplot of transformed weight vs speed",

xlab="Transformed Weight",

ylab="Speed")



Submit a **single** .docx or .pdf document containing your R code, output and graphs, and interpretations (where requested).  To facilitate a quick turnaround on grading, please keep all parts of a problem together, rather than putting code at the end of the file.

* Please include your full name at the top of your document.
* It is not necessary to include R code where you were testing code or where you made a mistake--just submit the final version.