Adam Hendel

DS710 – Assignment 3

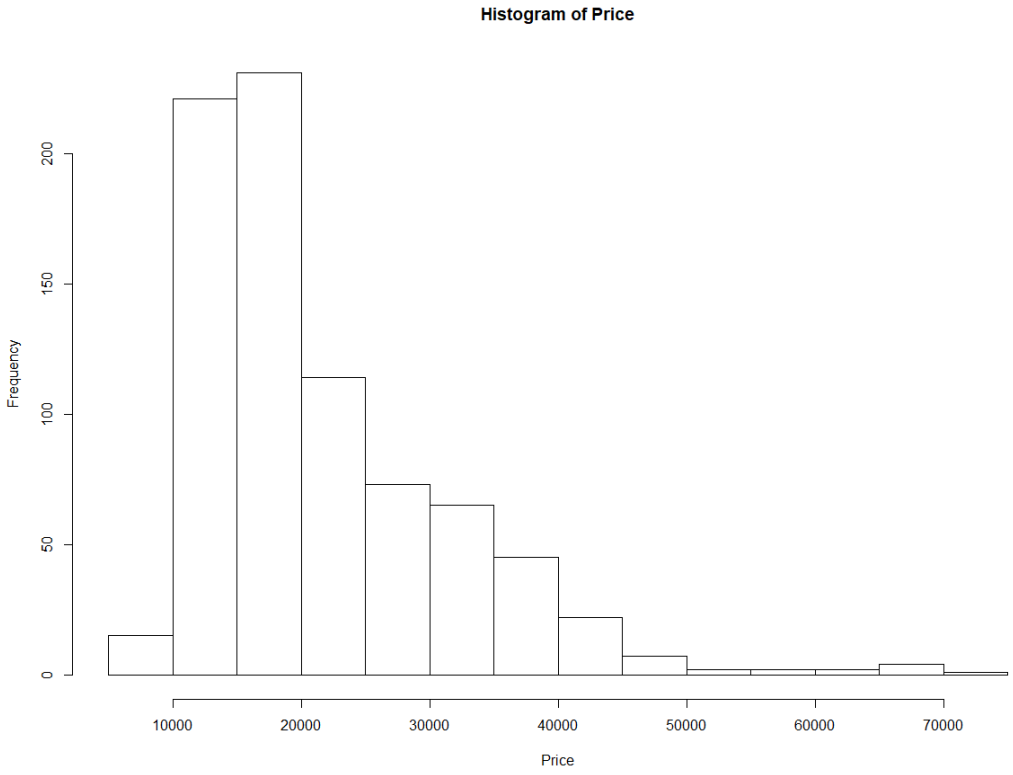
PART 1: ANALYZING USED CAR PRICES

1a. read on the data and attach

```
cars <- read.csv('Cars 2005.csv')
attach(cars)
```

 # 1b. make a histogram and describe the shape

the distribution is 'right skewed' since the tail is to the right and majority of observations to the left



 # 1c. what proportion of the cars cost between $10k and $20k?

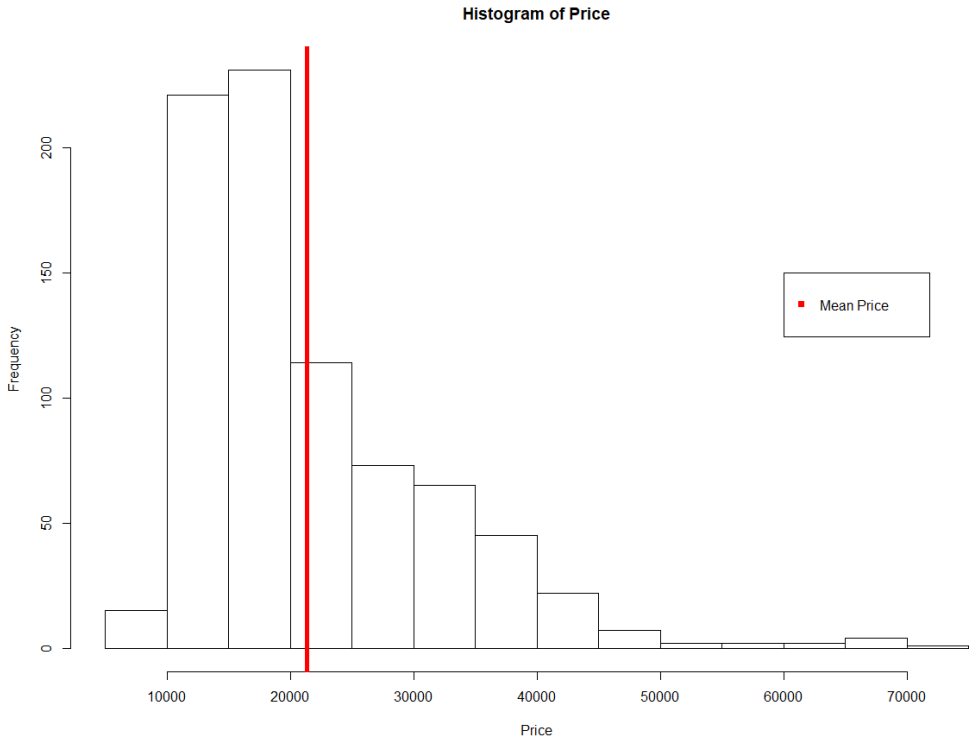
```
# total cars in set
tot.cars <- length(cars$Price)
# index of prices in the range
boolVec <- Price > 10000 & Price < 20000
# total cars in this range
tot.in.range <- sum(boolVec)
# proportion is ratio of the subset to the whole
proportion <- tot.in.range / tot.cars
print(proportion)
# 0.5621891
# About 56.2 % are within 10k and 20k
```

 # 1d. Find the mean and median price. Which is larger, why does this make sense?

```
# mean price
mean(Price)
# 21343.14
# median price
median(Price)
# 18025
# the mean is higher. This make sense by visual inspection of the histogram because
# we see that the frequency of prices is skewed right, so the median will tend towards the higher
# density, which is a lower value (due to skewness).
```

 # 1e. Add a vertical line to the histogram at the mean price. Also add a legend

```
hist(Price)
abline(v=mean(Price), col = 'red', lwd = 5)
# text(x=mean(Price)+7000,y=150, labels='Mean Price ')
legend(x=60000, y=150, legend = 'Mean Price', col = 'red', pch = 15)
```



 # 1f. Transform price to reduce its skew, make a histogram of the transformed price.

fit a normal distribution to new price, graph the density curve on the same plot as histogram

how well does a normal distribution fit the transformed data?

transform price

log.Price <- log(Price)

plot histogram of transformed price

hist(log.Price)

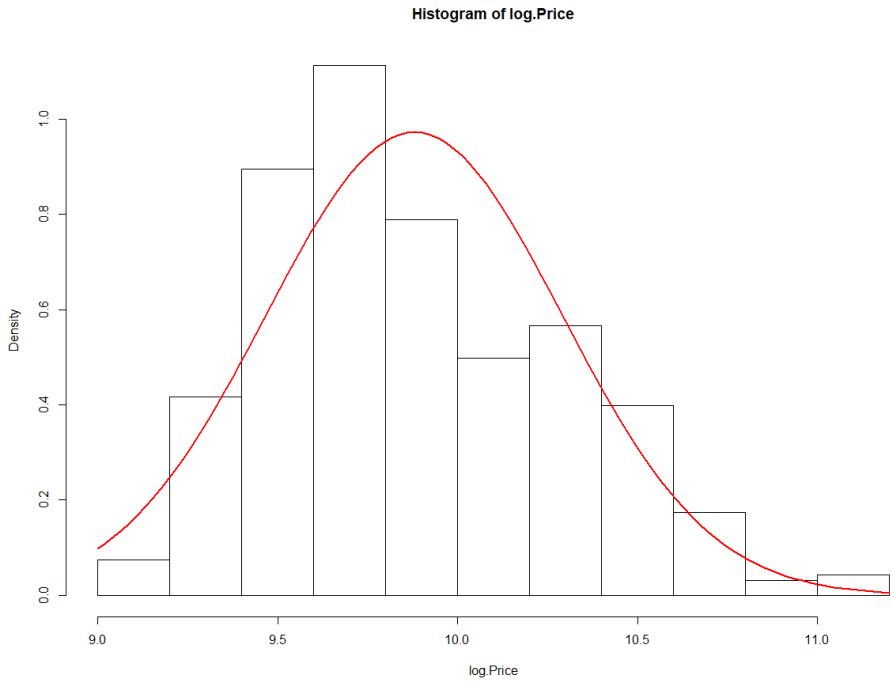
fit normal distribution to transformed price

curve(dnorm(x, mean = mean(log.Price), sd = sd(log.Price)),

add=T, col = 'red', lwd = 2)

it is a better fit to normal distribution now that we log transformed the Price by visual inspection

to normal distribution now that we log transformed the Price.

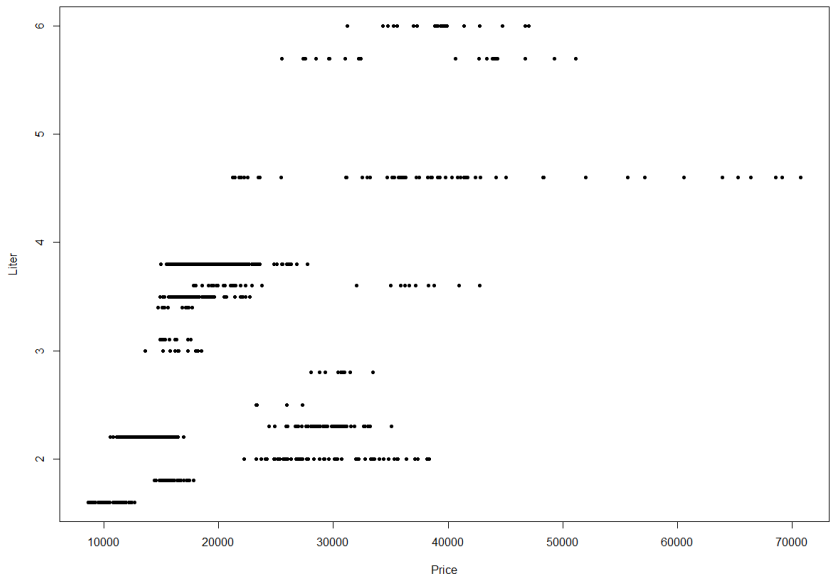


 # 1g Make a scatterplot of transformed price versus engine size, in liters. Describe relationship

between these two variables

plot(Price, Liter, pch = 16, cex = 0.7)

As price increases so does the engine size, to a point. Around $50k the engine size levels off.



 # 1h Find correlation between transformed price and engine size in liters. Explain.

cor(log.Price, Liter)

0.5904097

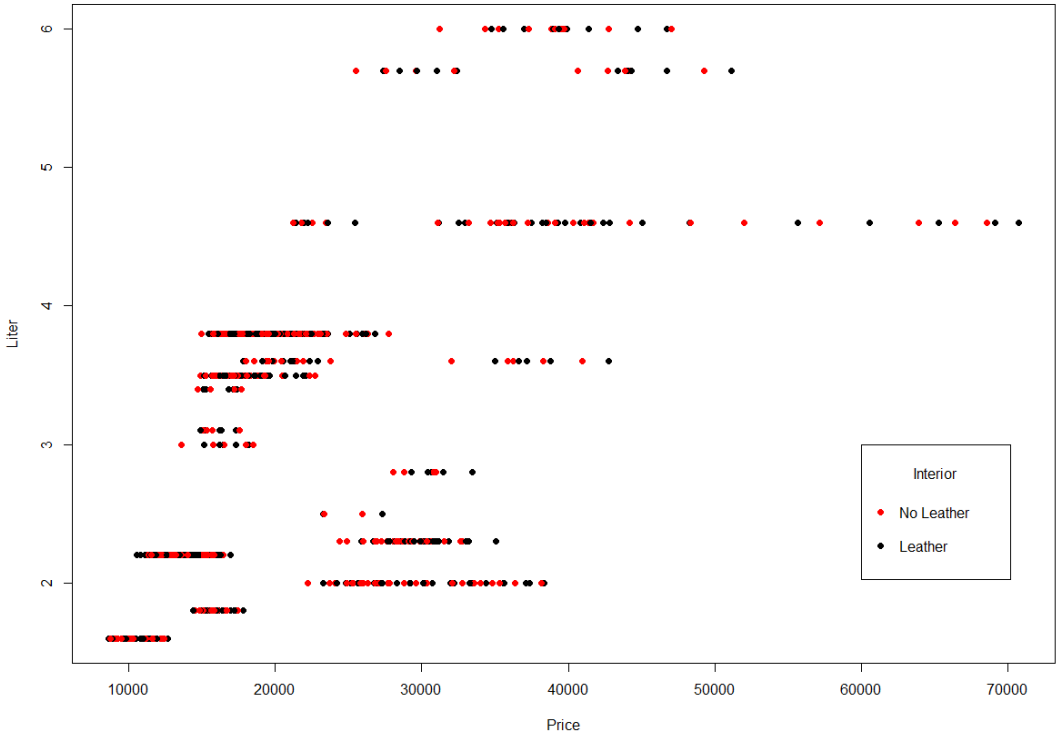
a correlation of 0.59, indicates there is a slight positive relationship between these two variables

As one variable increases, so does the other variable, but not perfectly, the relationship is not most correlated

 # 1i. Modify the scatterplot in g to use one color of plotting symbol for cars with leather

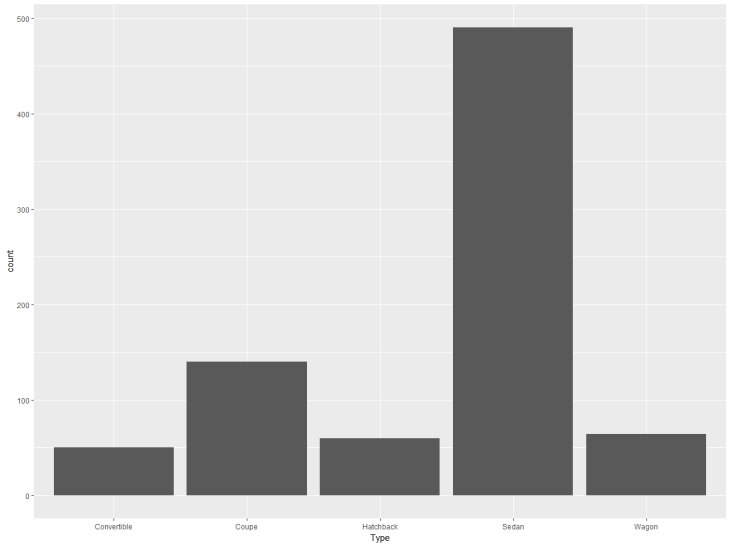
 # and a different color for cars without leather interiors, and add a legend.

```
labs <- levels(factor(Leather))
# inspect labs 0 no leather, 1 is leather
labs <- c('No Leather', 'Leather')
plot(Price, Liter, pch = 16, cex = 1, col = c('black', 'red'))
legend(x = 60000, y = 3, legend = labs, col = c('red', 'black'),
      pch = 16, title = 'Interior')
# cars[Price == max(Price),] check that the plot makes sense
```



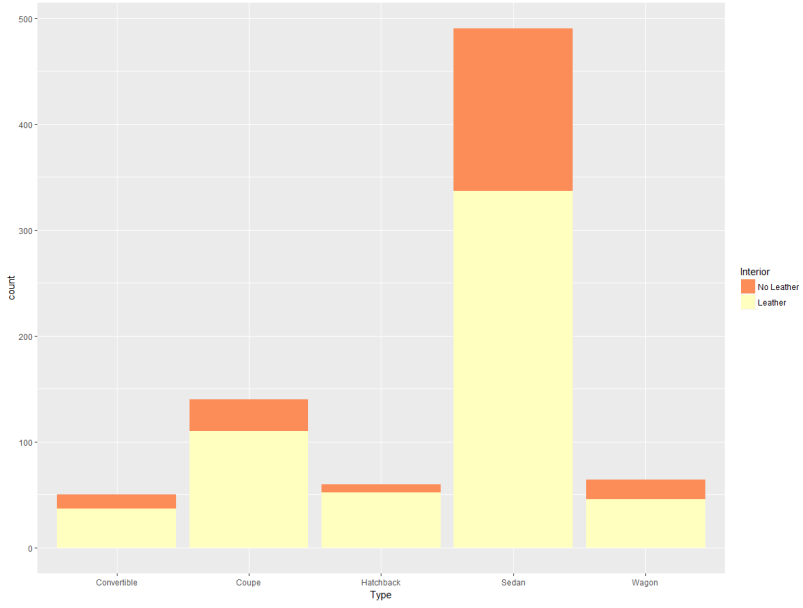
 # 1j. Make a barplot of the types

```
# ggplot

library(ggplot2)

base <- ggplot(cars, aes(x=Type))

base + geom_bar(stat='count')
```



 # 1k. Make a barplot of the types of cars and whether they have interior leather. Add a legend

```
base +

  geom_bar(stat='count', aes(fill = factor(Leather))) +

  scale_fill_brewer(palette = 'Spectral', name = 'Interior', labels = c('0'='No Leather', '1' = 'Leather'))
```



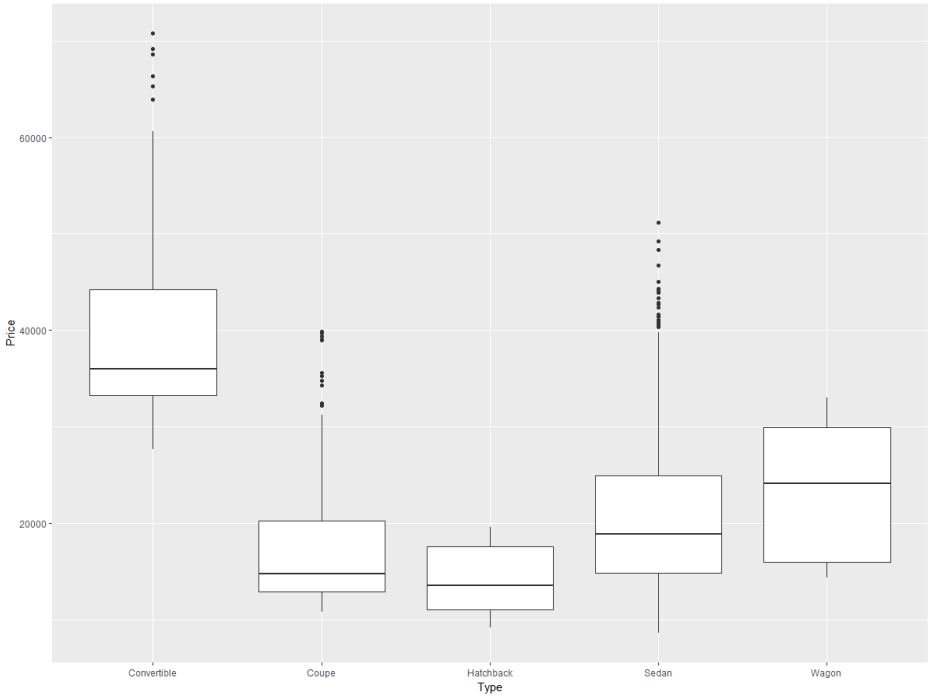
 # 1l. Make a boxplot of (untransformed) price by type of car. In words, summarize what it shows.

```
boxBase <- ggplot(cars)

boxBase +

  geom_boxplot(aes(Type, Price))
```

The box plots show the distribution of price across the types of cars. We see that hatchbacks have the most closely bunched distribution of prices by the size of the box and length of whiskers while sedans and convertibles are both right skewed, with outliers at the high end of the price range.

Median price of convertibles are higher than all other car types, however, there are outliers in coupes and sedans that are higher price.

We see that hatchbacks price are consistently below 20k, while convertibles are generally above $30k, but there are several observations of convertibles well over $60k.



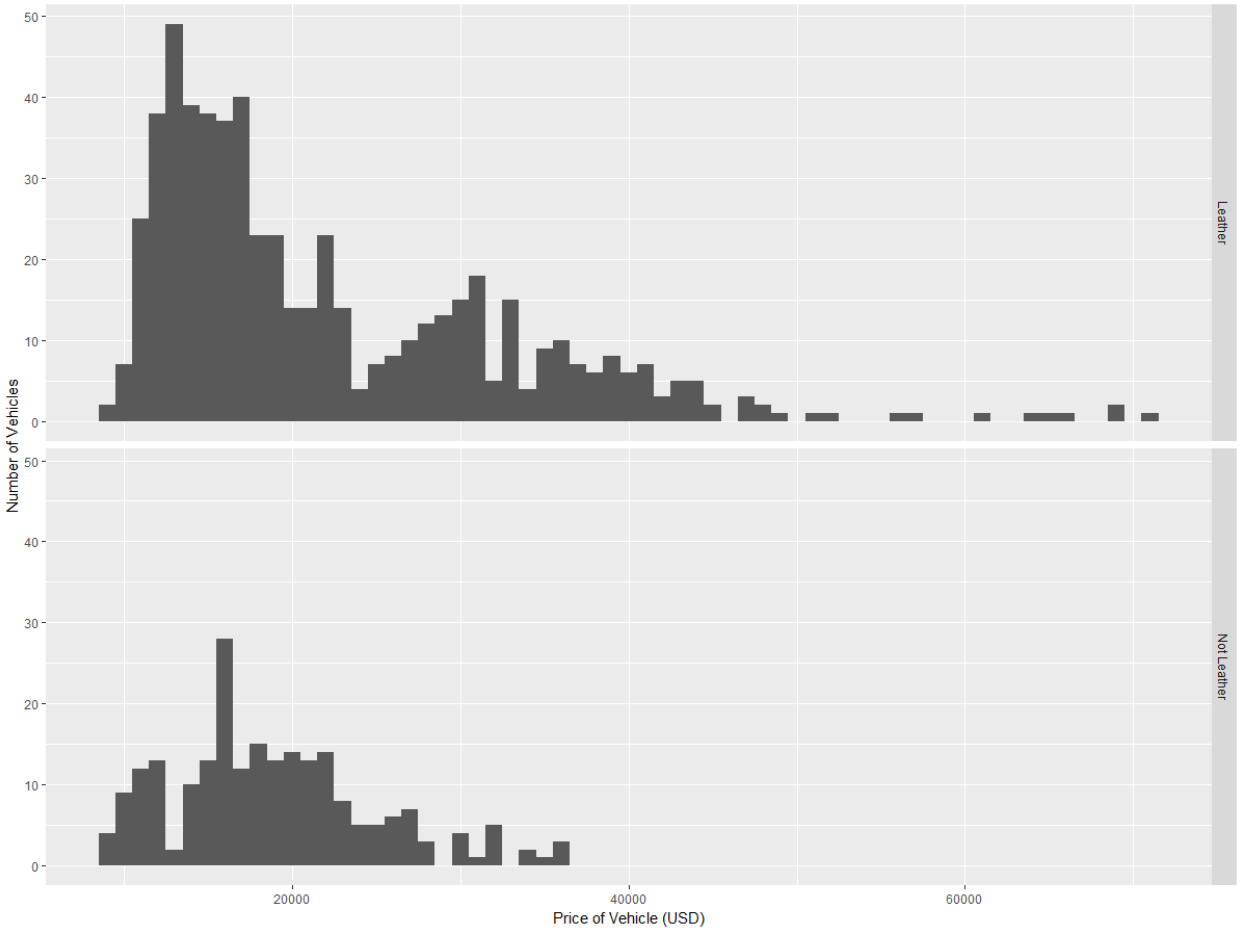
 # 1m. Create two different histograms in a vertical stack that allow comparison of (untransformed)

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.

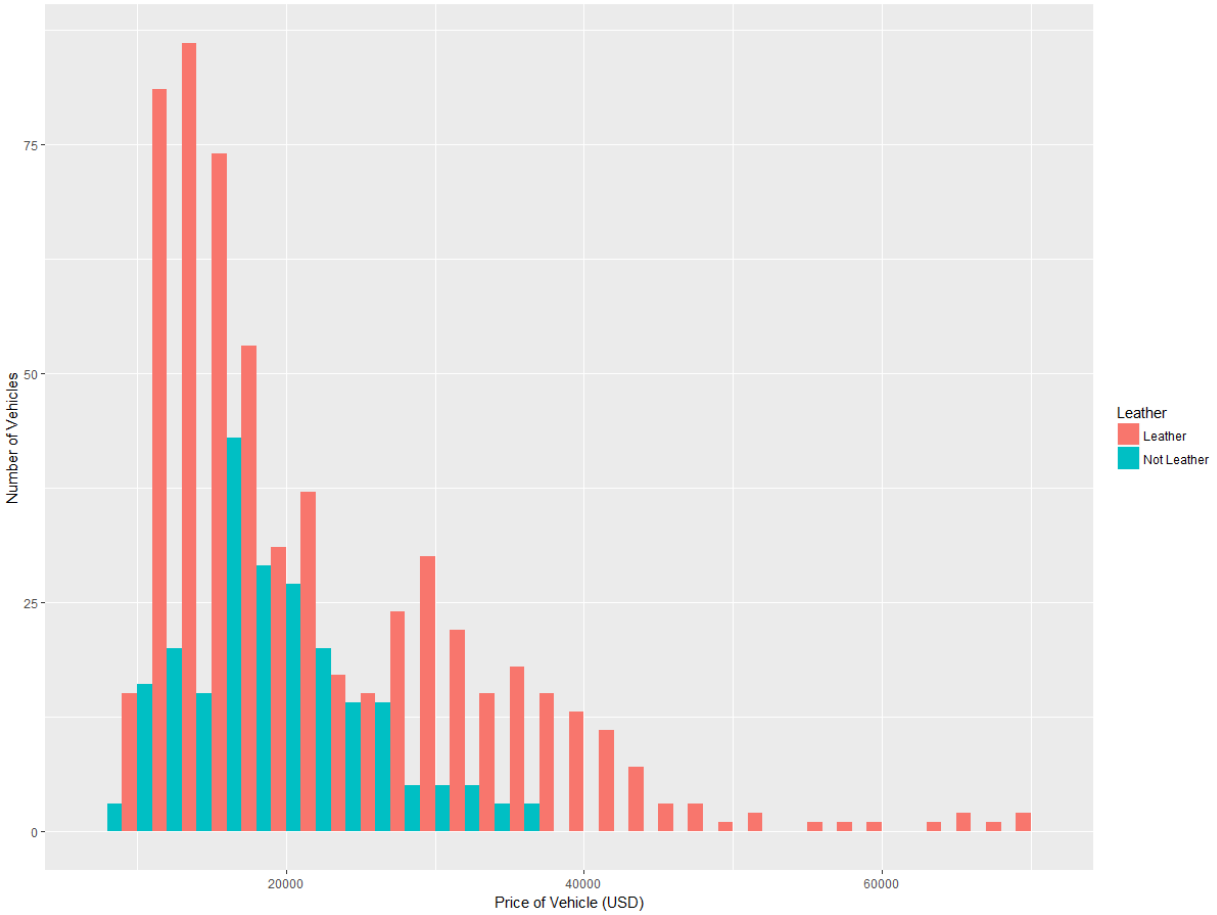
to modify the Leather column for ease of use

```
cars.hw <- cars
cars.hw$Leather[cars.hw$Leather == 0] <- 'Not Leather'
cars.hw$Leather[cars.hw$Leather == 1] <- 'Leather'
ggplot(cars.hw) +
 geom_histogram(aes(x=Price), binwidth = 1000) +
 facet_grid(Leather ~ .) +
 scale_x_continuous(name = 'Price of Vehicle (USD)') +
 scale_y_continuous(name = 'Number of Vehicles')
```



 # 1n. Create a single histogram with side-by-side bars to allow the same comparison in part m.

```
# add a legend
ggplot(cars.hw) +
  geom_histogram(aes(x=Price, fill = Leather),
                 binwidth = 2000, position = 'dodge') +
  scale_x_continuous(name = 'Price of Vehicle (USD)') +
  scale_y_continuous(name = 'Number of Vehicles')
```

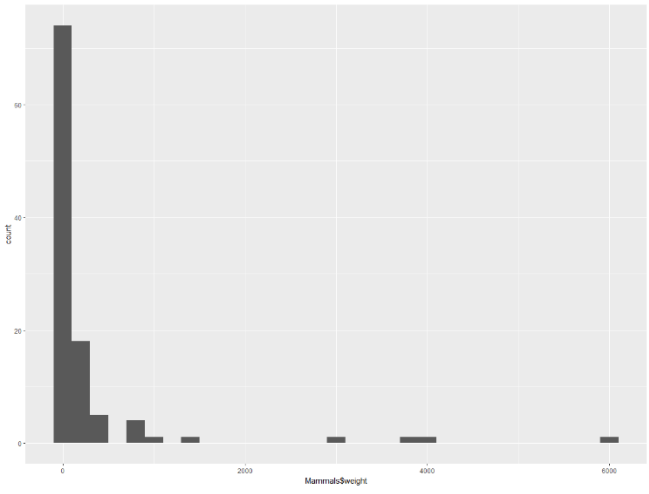


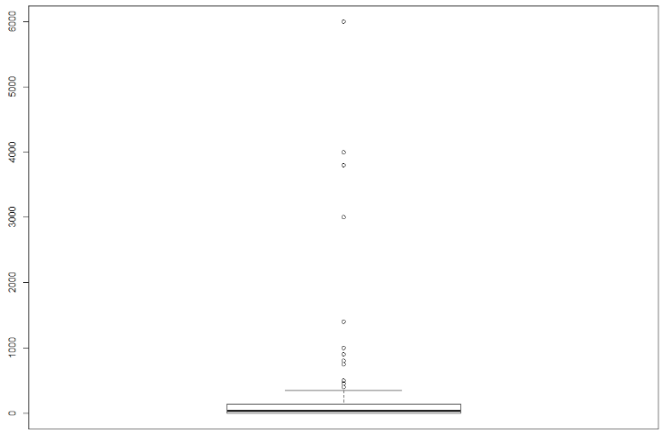
PART 2: ANALYSING RUNNING SPEED OF MAMMALS

```
# 2a load data

install.packages("quantreg")

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

 # 2b Decide whether either of the quantitative variables should be transformed.

```
# justify the decision using plots and descriptive statistics

# inspect data

str(Mammals)

# 'data.frame':   107 obs. of  4 variables:

#   $ weight  : num  6000 4000 3000 1400 400 350 300 260 250 3800 ...

# $ speed   : num  35 26 25 45 70 70 64 70 40 25 ...

# $ hoppers : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...

# $ specials: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...


# plots for weight

qplot(Mammals$weight, binwidth = 200)

boxplot(Mammals$weight)

summary(Mammals$weight)
```





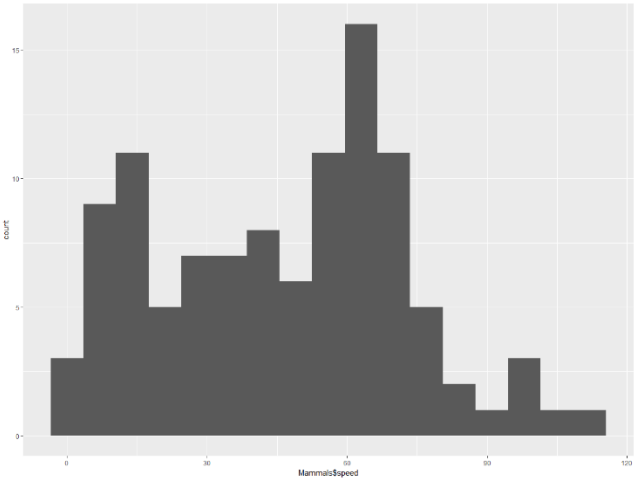

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.016    1.700   34.000  278.688  142.500 6000.000
```

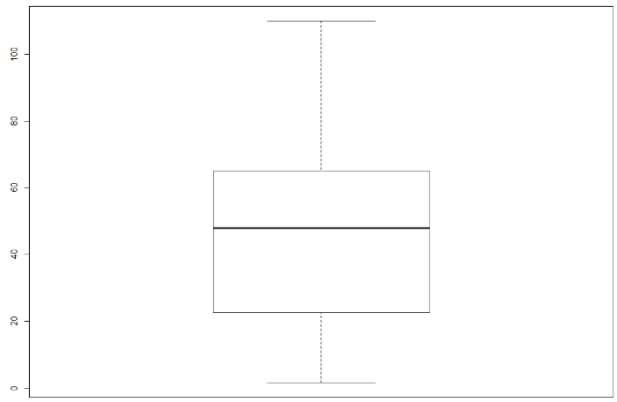
plots for speed

qplot(Mammals$speed, binwidth = 2)

boxplot(Mammals$speed)

summary(Mammals$weight)



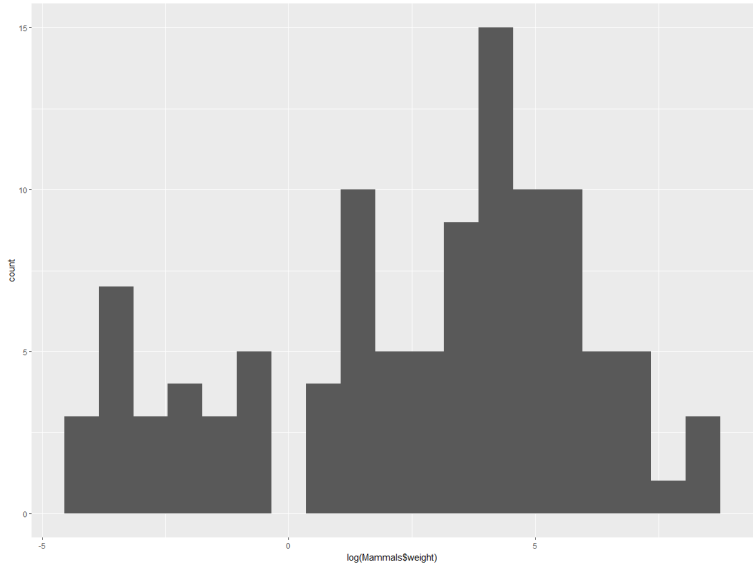


```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.60   22.50   48.00   46.21   65.00  110.00 
```

weight is heavily skewed right. We can see this from both the histogram with most values <1 and several values >> 100, and will likely benefit from a log transform. Speed is much closer to a normal distribution and probably does not need a transform.

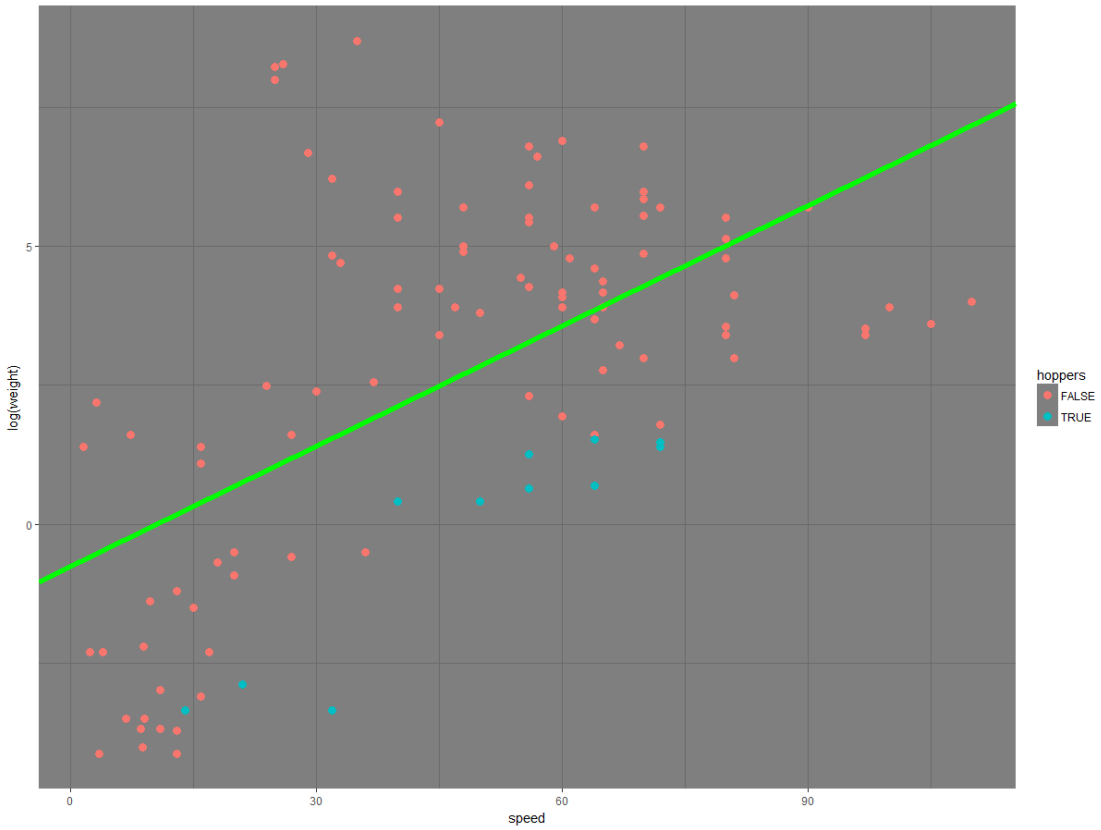
plot log transform weight

qplot(log(Mammals$weight), binwidth = 1) # somewhat log-normal



#2c 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.

```
# correlation between weight and speed

cor(Mammals$weight, Mammals$speed)

# 0.06653467 --- close to zero correlation


# correlation between log(weight) and speed

cor(log(Mammals$weight), log(Mammals$speed))

# 0.5751193 -- positive correlation indicates moderate linear relationship


# are any hopper and special? -- no

Mammals$hoppers & Mammals$specials

# [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE

# [20] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE

# [39] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE

# [58] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE

# [77] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE

# [96] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

library(ggplot2)

model <- lm(log(Mammals$weight) ~ Mammals$speed)

ggplot(Mammals, aes(x = speed, y = log(weight))) +
  geom_point(aes(col = hoppers), size = 3) +
  theme_dark() +
  geom_abline(slope = model$coefficients[2], intercept = model$coefficients[1],
              col = 'green', size = 2)
```



summary(model)

```
Call:
lm(formula = log(Mammals$weight) ~ Mammals$speed)

Residuals:
   Min     1Q Median     3Q    Max
-4.898 -2.290  0.122  1.776  7.203

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.76607    0.53469  -1.433    0.155
Mammals$speed  0.07225    0.01003   7.204 9.24e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.759 on 105 degrees of freedom
Multiple R-squared:  0.3308,   Adjusted R-squared:  0.3244
F-statistic: 51.89 on 1 and 105 DF,  p-value: 9.24e-11
```

there is a positive linear relationship between weight and land speed in these mammals. Log(weight) increases as peed increases. Whether the mammal is a ‘hopper’ does not seem to have an impact speed relative to weight.