Module 5 Assignment

Rudy Fasano

11/30/2020

Exercise 2

2.2 Below, the work directory was set as well as required packages:

```
setwd("C:/Users/fasan/Desktop/500B/Module5/")
library(RMySQL)
library(ggplot2)
library(animation)
```

MySQL database, auto, was connected to in order to conduct statistical analyses:

cor.test() was conducted to find correlation coefficient amongst other statistical data. This resulted in the correlation coefficient being 0.865 which can be inferred that the two variables, weight and horsepower, have a highly positive correlation and relationship. A Linear regression model was also conducted. $Y = A + bX$ / Weight = 984.50 + 19.08 * horsepower

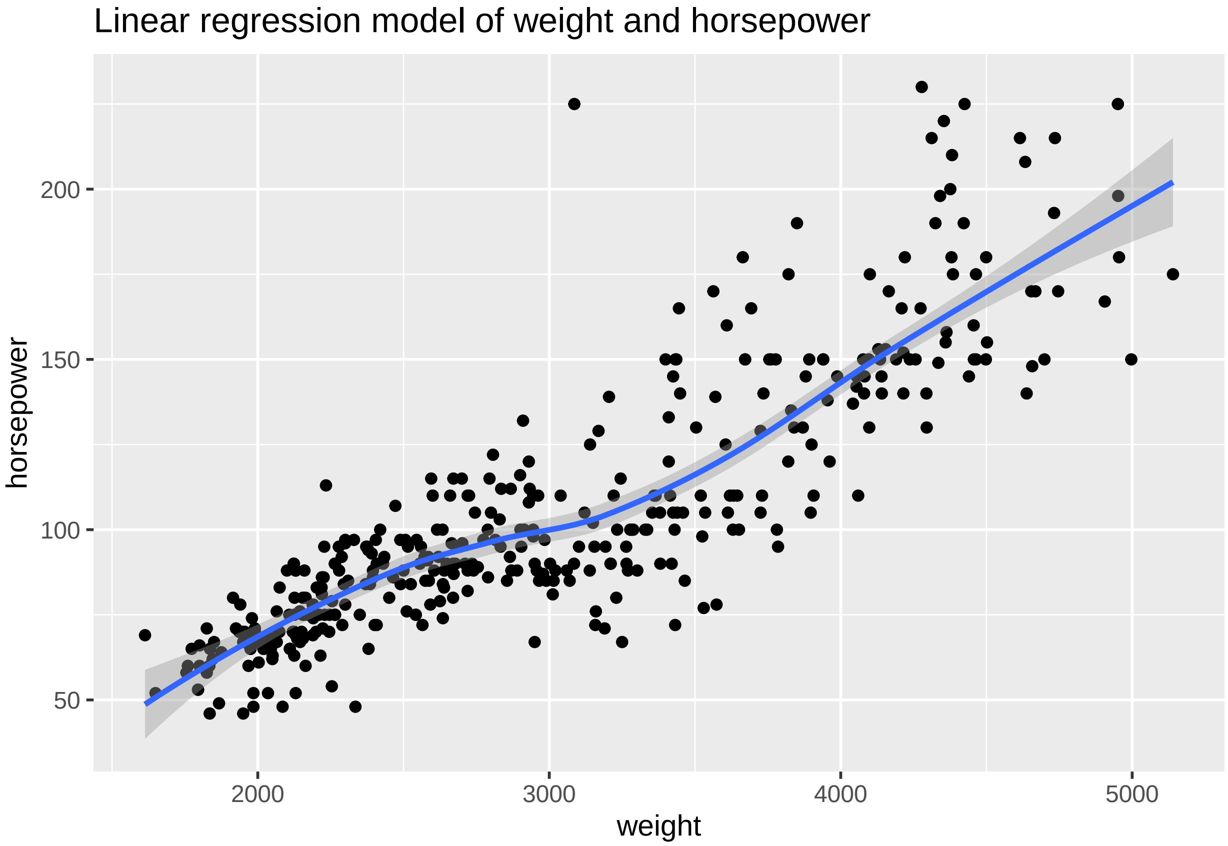
```
cor.test(df$weight, df$horsepower) ## correlation test
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$weight and df$horsepower
## t = 33.972, df = 390, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8371778 0.8875815
## sample estimates:
##       cor
## 0.8645377
```

```
##
linear_reg<-lm(weight~horsepower, data=df)
print(linear_reg)                            # Linear regression
```

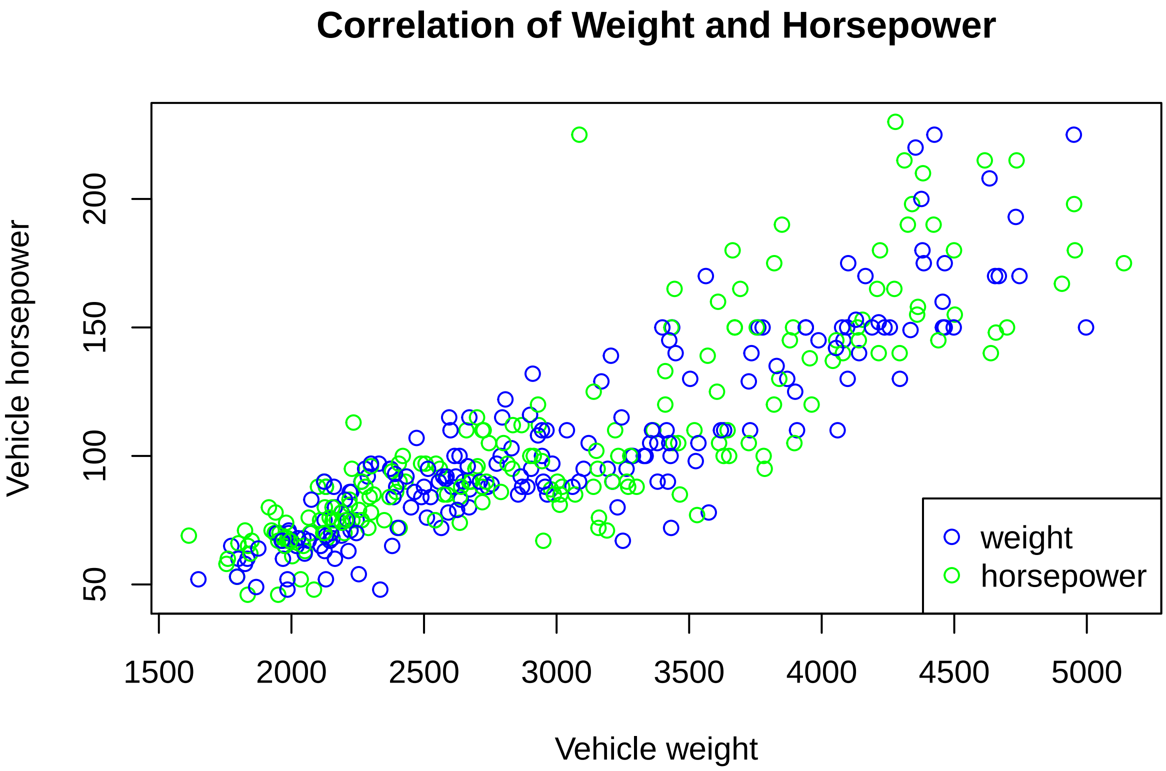
```
##
## Call:
## lm(formula = weight ~ horsepower, data = df)
##
## Coefficients:
## (Intercept)   horsepower
##      984.50        19.08
```

```
ggplot(df, aes(x=weight, y=horsepower))+ ## regression model
  geom_point() + stat_smooth() + ggtitle('Linear regression model of weight and horsepower')
```



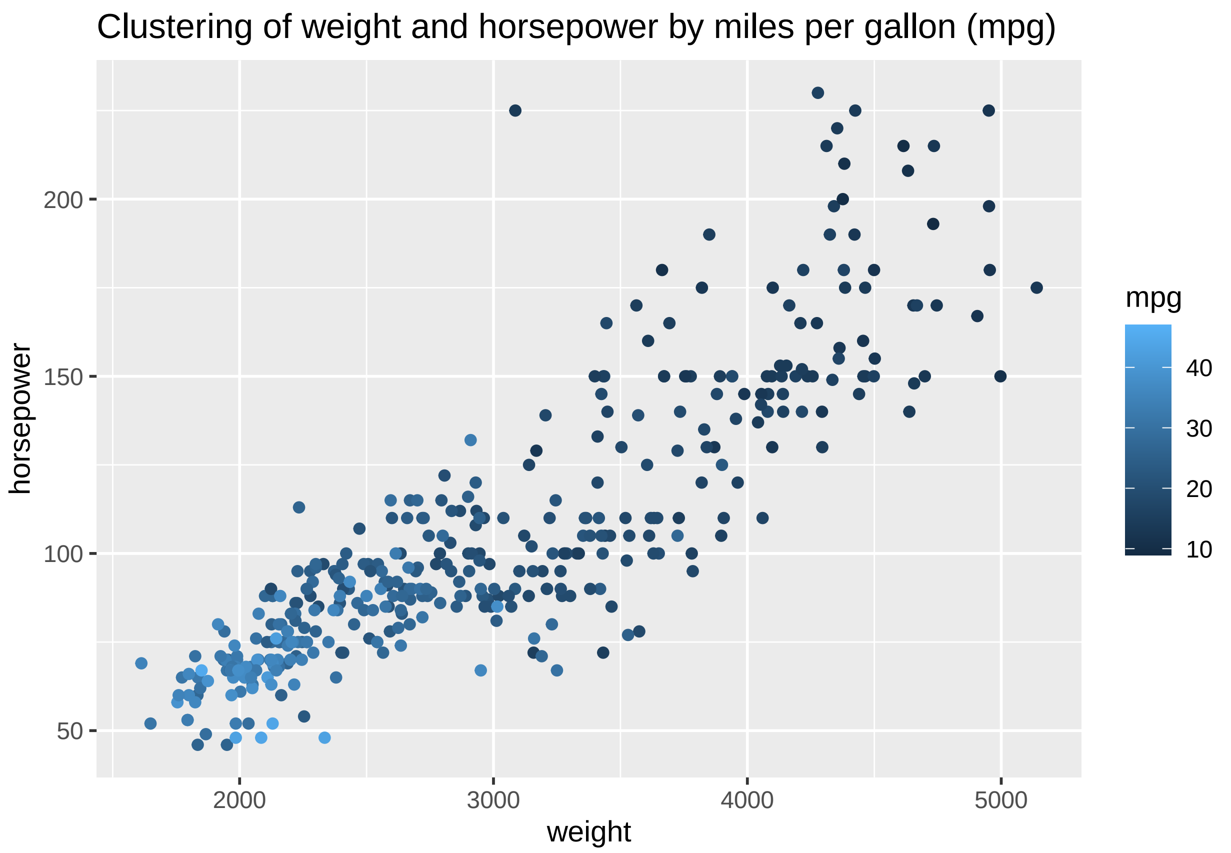
The following scatterplot was created to visualize how each variable relates as weight increases:

```
plot(df$weight,df$horsepower, main='Correlation of Weight and Horsepower', ### visualization of both va
     col=c('blue','green'),
           xlab='Vehicle weight',
           ylab='Vehicle horsepower')
legend("bottomright", legend =c('weight', 'horsepower'), col =c('blue','green'),
       pch = c(1, 1))
```



Clustering by miles per gallon was performed in order to identify any trends between vehicle weights and horsepower.

```
ggplot(df, aes(weight, horsepower, color=mpg))+geom_point()+ ### scatterplot
  ggtitle('Clustering of weight and horsepower by miles per gallon (mpg)')
```

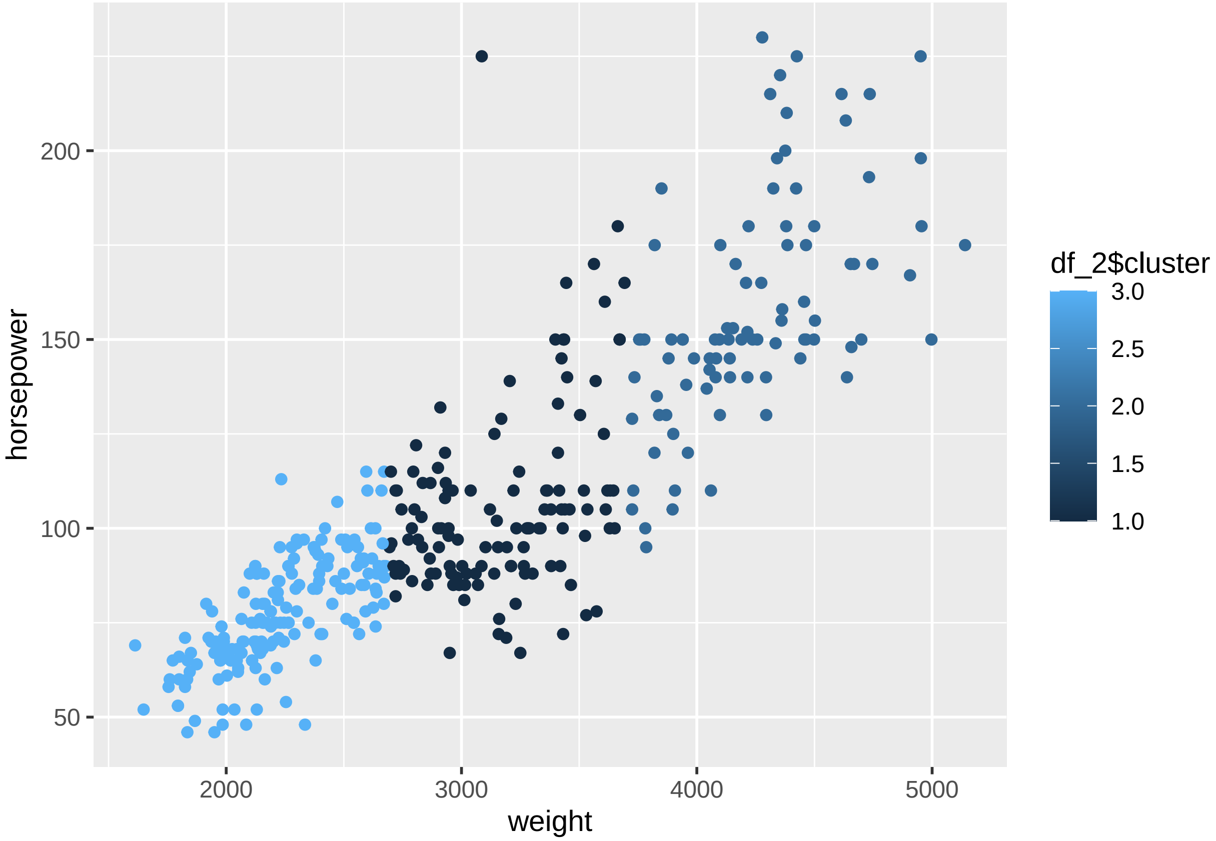


Further kmeans clustering was performed to find any viable groupings. Using 3 groups (K = 3) resulted in having 88.3% of well grouped data. We can infer from the plot that the relationship between weight and horsepower is strongest and most relevant at lower weights.

```
set.seed(20)
df_2<-kmeans(df[,4:5], 3, nstart=20) ### kmeans clustering
df_2
```

```
## K-means clustering with 3 clusters of sizes 124, 89, 179
##
## Cluster means:
##   horsepower   weight
## 1  105.79839 3163.081
## 2  157.73034 4242.427
## 3   77.06704 2220.196
##
## Clustering vector:
##   [1] 1 1 1 1 1 2 2 2 2 2 1 1 2 1 3 1 1 3 3 3 3 3 3 3 3 2 2 2 2 3 3 3 3 1 1 1 1
##  [38] 2 2 2 2 2 2 2 1 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 1 2 2 2 2 3 2 2 2
##  [75] 2 1 3 1 3 3 3 3 3 3 2 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 3 2 2 2 2 1 3 3 3 3
## [112] 3 3 3 2 2 3 3 3 1 1 3 1 1 1 1 1 3 3 3 3 2 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3 3
## [149] 3 3 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 3 3 1 3 1 3 3 1 3 1 1 1 1 3 3 3 3 3 3 3
## [186] 2 2 2 2 1 1 1 1 3 3 3 3 1 1 1 1 3 3 3 3 1 2 1 1 2 2 2 2 2 3 3 3 3 3 2 2 2
## [223] 2 1 1 1 1 2 2 2 2 3 1 3 1 3 3 3 3 1 3 1 3 3 3 3 3 1 2 1 1 1 1 1 1 1 1 1 1
## [260] 1 1 1 1 2 3 3 3 3 3 1 1 3 1 1 1 1 3 3 1 1 1 1 1 2 2 2 2 2 2 1 2 3 3 3 3 1
## [297] 2 1 1 3 3 3 3 3 3 1 3 3 3 3 3 3 1 1 1 3 1 3 3 3 3 1 3 3 3 1 1 3 3 3 1 3 3
## [334] 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 2 1 1 3 3 3 3 3 1 1 3
## [371] 3 3 3 3 3 3 3 3 3 3 1 1 3 1 3 3 1 1 3 3 3 1
##
## Within cluster sum of squares by cluster:
## [1] 10657321 10479237 11823699
##  (between_SS / total_SS =  88.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
ggplot(df, aes(weight, horsepower, color=df_2$cluster))+geom_point() ## plotting kmeans clusters
```



After analyzing the data as well as the clustering and kmeans clustering, it can be concluded that there is definitely a high, positively trending relationship between vehicle weights and horsepower. With the clustering, we notice that cars at lower weights tend to have less horse power at a much higher capacity than their heavier counterparts. Utilizing the kmeans clustering techniques, we can visibly see the three groups. At lower weight and at a much higher concentration, vehicles tend to confidently have low horsepower. The medium-sized family economic vehicles, which would be the second, middle trending group, begins to spread more as horse power fluctuates sporadically and is shown to have the most variety in spread of horsepower as weight increases.. However, it can also be confidently noted that the heavier the vehicle, the more horsepower it would typically have. The third and the last group depicts the heavier, muscle car styled vehicle group. In conclusion, as weight increases, the vehicles tend to have higher horsepower. In addition to this analysis, it is prudent to understand the function and purpose of the vehicle in order to properly assess the data; such as the potential outliers that resulted in being high-performing vehicles that ultimately remained within the analysis. “‘