Homework #1

Eric Pettengill

2.4.7 (a)-(d)

(a) Compute the Euclidian distance between each observation and test point $X_1 = X_2 = X_3 = 0$.

```
library(tidyverse)

x <- data.frame(x1 = c(0,2,0,0,-1,1),
                x2 = c(3,0,1,1,0,1),
                x3 = c(0,0,3,2,1,1),
                y = c('red', 'red', 'red', 'green', 'green', 'red'))

x %>% mutate(distance = sqrt(x1^2 + x2^2 +x3^2))
```

```
##   x1 x2 x3     y distance
## 1  0  3  0   red 3.000000
## 2  2  0  0   red 2.000000
## 3  0  1  3   red 3.162278
## 4  0  1  2 green 2.236068
## 5 -1  0  1 green 1.414214
## 6  1  1  1   red 1.732051
```

(b) What is our prediction with $K = 1$? Why?

Since the minimum distance occurs at the $5th$ observation with a value of 1.414214 our prediction for $K = 1$ is green.

(c) What is our prediction with $K = 3$? Why?

We take the three points in our test data closest the our test point of $X_1 = X_2 = X_3 = 0$ and observe that red occurs twice and green once, so our prediction would be green.

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

We would expect the best value for K be smaller since the larger the value for K, the decision boundary becomes more linear.

2.4.9 (a)-(f)

```
auto <- read.csv("Auto.csv", na.strings = "?")
auto <- na.omit(auto)
auto$horsepower <- as.numeric(auto$horsepower)
auto$weight <- as.numeric(auto$weight)
auto$year <- as.numeric(auto$year)
auto$cylinders <- as.factor(auto$cylinders)
auto$origin <- as.factor(auto$origin)
```

(a) Which of the variables are quantitative and qualitative?

Quantitative: MPG, DISPLACEMENT, HORSEPOWER, WEIGHT, ACCELERATION, YEAR

Qualitative: CYLINDERS, ORIGIN, NAME

(b) What is the range of each quantitative predictor?

```
autonew <- auto %>% select_if(is.numeric)
map(autonew, range)
```

```
## $mpg
## [1]  9.0 46.6
##
## $displacement
## [1]  68 455
##
## $horsepower
## [1]  46 230
##
## $weight
## [1] 1613 5140
##
## $acceleration
## [1]  8.0 24.8
##
## $year
## [1] 70 82
```

(c) What is the mean and standard deviation of each quantitative variable?

```
map_dbl(autonew, mean)
```

```
##          mpg displacement   horsepower       weight acceleration
##     23.44592    194.41199    104.46939   2977.58418     15.54133
##         year
##     75.97959
```

```
map_dbl(autonew, sd)
```

```
##          mpg displacement   horsepower       weight acceleration
##     7.805007   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

(d) Remove the 10th to 85th ovservations. What is the range, mean, and standard deviation of the remaining data?

```
autosubset <- autonew[-c(10:85), ]
map(autosubset, range)
```

```
## $mpg
## [1] 11.0 46.6
##
## $displacement
## [1]  68 455
##
## $horsepower
## [1]  46 230
##
## $weight
## [1] 1649 4997
##
## $acceleration
```

```
## [1]  8.5 24.8
##
## $year
## [1] 70 82
```

```
map_dbl(autosubset, mean)
```

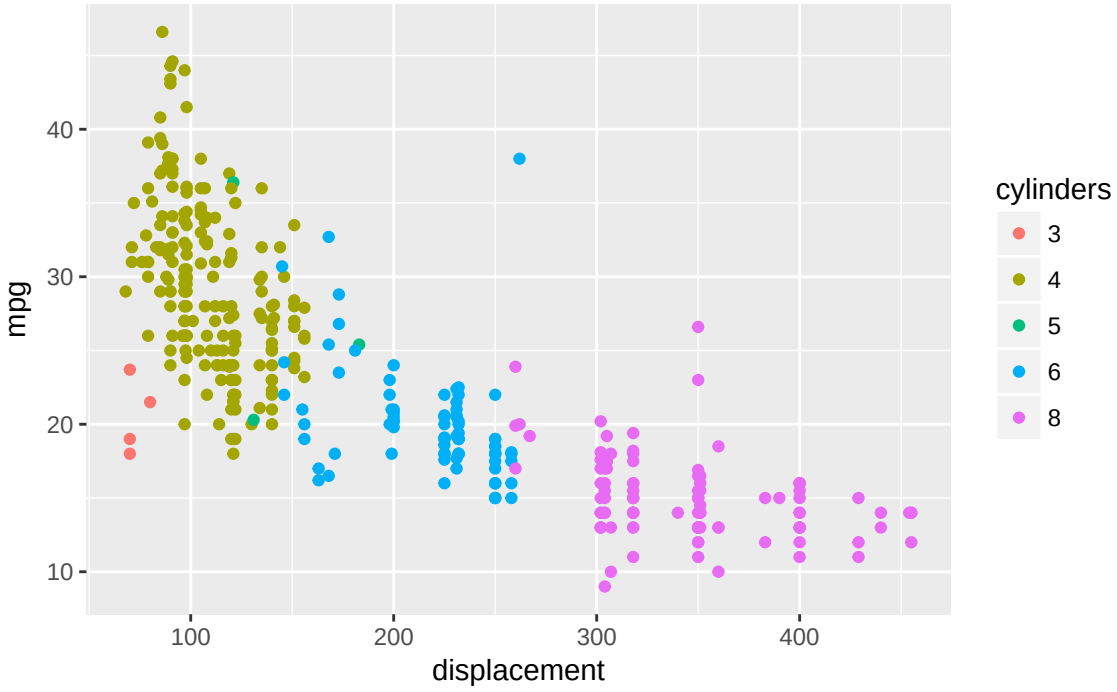
```
##          mpg displacement   horsepower       weight acceleration
##     24.40443    187.24051    100.72152   2935.97152     15.72690
##         year
##     77.14557
```

```
map_dbl(autosubset, sd)
```

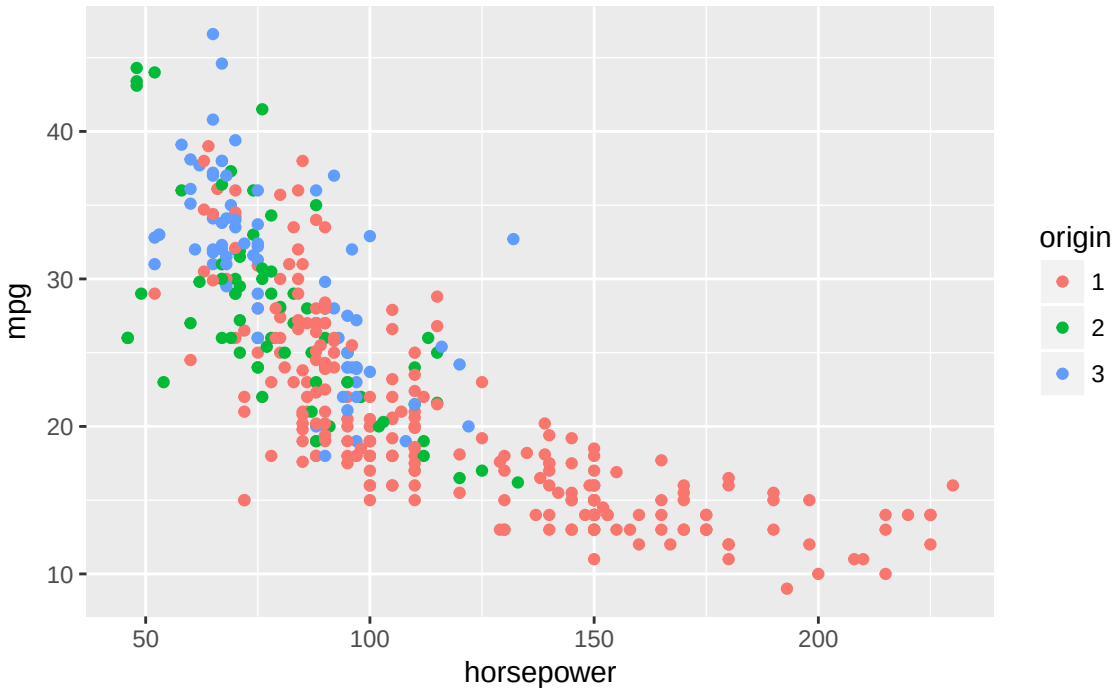
```
##          mpg displacement   horsepower       weight acceleration
##     7.867283    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

(e) Using the full dataset, investigate the predictors graphically.

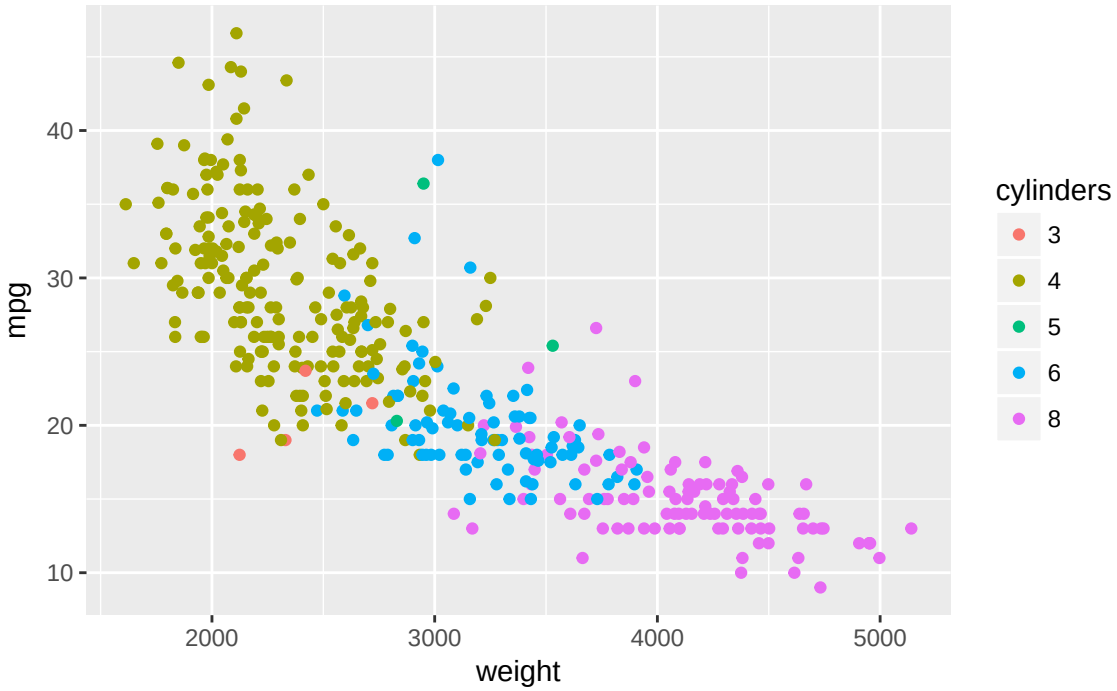
```
ggplot(auto) +
  geom_point(aes(x = displacement, y = mpg, color = cylinders))
```




```
ggplot(auto) +
  geom_point(aes(x = horsepower, y = mpg, color = origin))
```




```
ggplot(auto) + geom_point(aes(x= weight, y=mpg, color=cylinders))
```



(f) Suppose we wish to predict MPG. Do your plots suggest that any of the variables might be useful in predicting MPG.

Yes, there seems to be a relationship between the miles per gallon a car gets based on its displacement, horsepower, weight, and the number of cylinders. More precisely, the less a car weighs, the lower its displacement and horsepower, and the lower number of cylinders it has, the more miles per gallon the car will get.