Exploratory Analysis

Cool Beans Programming

2023-04-23

Loading libraries and data

```
library(ISLR2)
data(Auto)
```

Exploring the data

```
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##
##   acceleration        year           origin                      name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                  (Other)           :365
```

All variables are quantitative except the name of the automobile.

Key metrics

```
apply(Auto[,1:7], 2, range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
```

The ranges of the variables are given above. The greatest range is among vehicle weight. Below, the mean and standard deviation of each quantitative predictor is given.

```
apply(Auto[,1:7], 2, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
apply(Auto[,1:7], 2, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

Removing Observations

```
Auto2 <-Auto[-c(10:85),]
```

```
apply(Auto2[,1:7],2,range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
```

```
apply(Auto2[,1:7],2,mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
##         year
##    77.145570
```

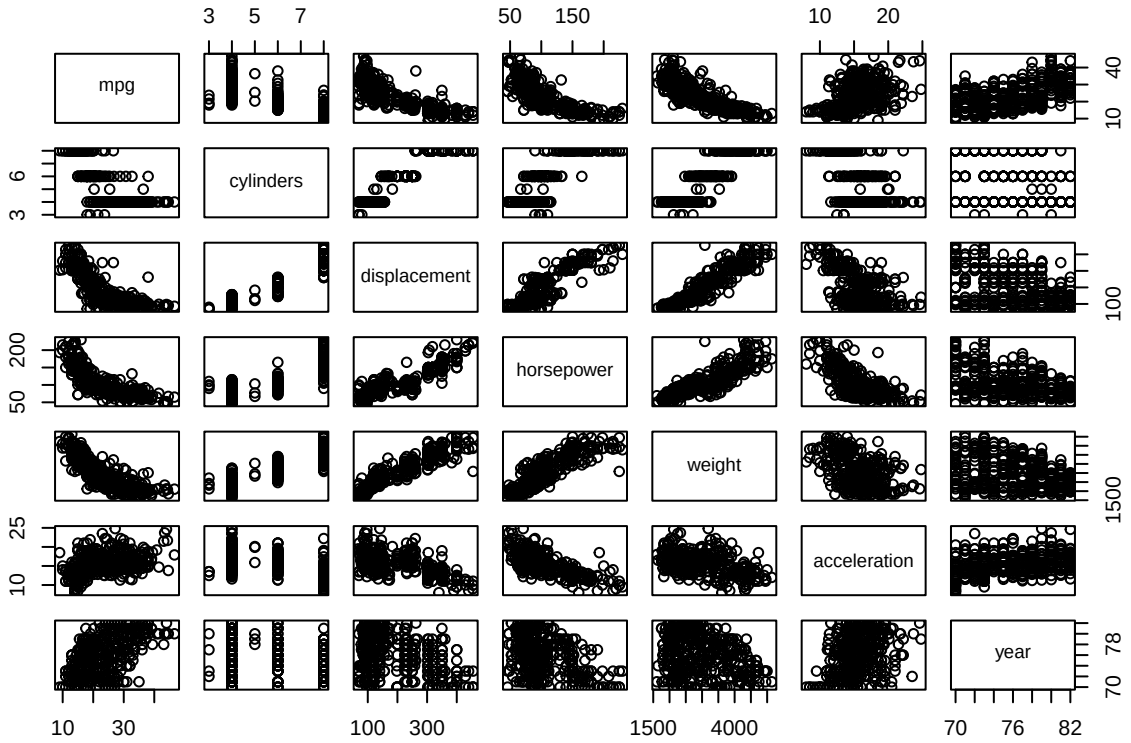
```
apply(Auto2[,1:7],2,sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

After removing observations 10 through 85, the range, mean, and standard deviation for all quantitative variables are relatively close.

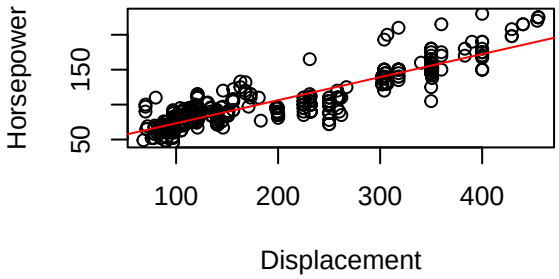
Investigate Predictors

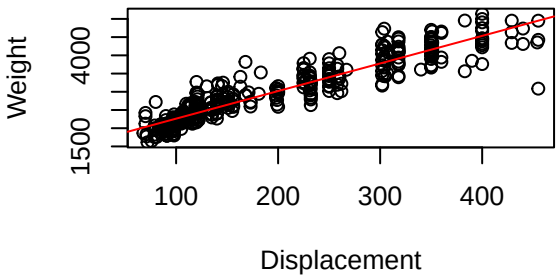
```
pairs(Auto[,1:7])
```

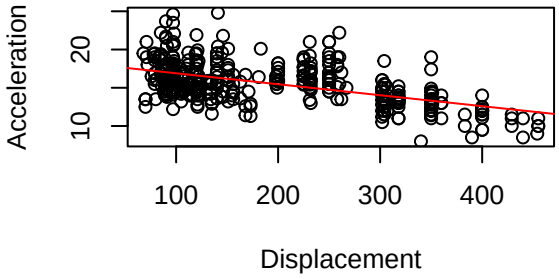


Many of the variables appear to have some sort of relationship though not necessarily a linear one. MPG has a non-linear relationship with most of the other predictors while displacement has a linear relationship with horsepower, weight, and acceleration.

```
par(mfrow=c(2,2))

plot(Auto$displacement, Auto$horsepower, xlab="Displacement", ylab="Horsepower")
abline(lm(Auto$horsepower ~ Auto$displacement), col="red")

plot(Auto$displacement, Auto$weight, xlab="Displacement", ylab="Weight")
abline(lm(Auto$weight ~ Auto$displacement), col="red")

plot(Auto$displacement, Auto$acceleration, xlab="Displacement", ylab="Acceleration")
abline(lm(Auto$acceleration ~ Auto$displacement), col="red")
```



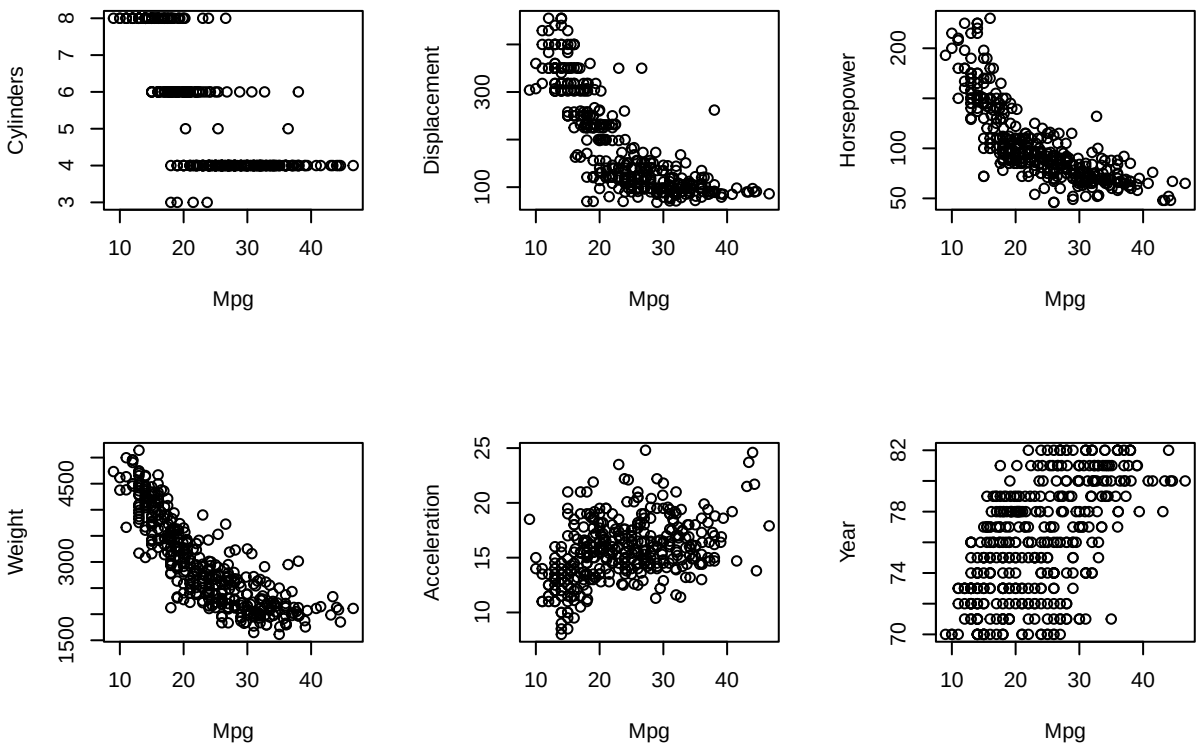




Making Predictions

To predict gas millage on the basis of the other variables, there are a few variables that may be useful in predicting mpg.

```
par(mfrow=c(2,3))

plot(Auto$mpg, Auto$cylinders, xlab="Mpg", ylab="Cylinders")
plot(Auto$mpg, Auto$displacement,xlab="Mpg", ylab="Displacement")
plot(Auto$mpg, Auto$horsepower,xlab="Mpg", ylab="Horsepower")
plot(Auto$mpg, Auto$weight,xlab="Mpg", ylab="Weight")
plot(Auto$mpg, Auto$acceleration,xlab="Mpg", ylab="Acceleration")
plot(Auto$mpg, Auto$year,xlab="Mpg", ylab="Year")
```



The set of plots above clearly indicate that there is a non-linear relationship between millage and weight, displacement, and horsepower. There may be some type of relationship between millage and acceleration as well.