Chapter 2 - Statistical Learning

Thalles Quinaglia Liduares

01/03/2022

Applied Exercise 2.9

Upload packages

```
library(dplyr)
library(tibble)
```

Upload database

```
setwd("C:\\Program Files\\R\\Machine Learning")

data<-read.csv("auto-mpg.csv")

data1<-tibble::as_tibble(data) # Transforming the dataframe in tibble to facilitate the analy
sis.
```

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

```
glimpse(data1)
```

```
## Rows: 398
## Columns: 9
## $ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14, 2~
## $ cylinders    <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, 4, ~
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383, 34~
## $ horsepower   <chr> "130", "165", "150", "150", "140", "198", "220", "215", "~
## $ weight       <int> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, 385~
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8.5, ~
## $ model.year   <int> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7~
## $ origin       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, ~
## $ car.name     <chr> "chevrolet chevelle malibu", "buick skylark 320", "plymou~
```

The dataframe contains 9 variables.

7 variables are quantitative and 2 are qualitative.

(b) What is the range of each quantitative predictor? You can answer this using the range() function

```
num_var<- select_if(data1, is.numeric)

num_var # Note that only numerical variables were selectioned.
```

```
## # A tibble: 398 x 7
##      mpg cylinders displacement weight acceleration model.year origin
##    <dbl>     <int>        <dbl>  <int>        <dbl>      <int>  <int>
##  1    18         8          307   3504         12           70      1
##  2    15         8          350   3693         11.5         70      1
##  3    18         8          318   3436         11           70      1
##  4    16         8          304   3433         12           70      1
##  5    17         8          302   3449         10.5         70      1
##  6    15         8          429   4341         10           70      1
##  7    14         8          454   4354          9           70      1
##  8    14         8          440   4312          8.5         70      1
##  9    14         8          455   4425         10           70      1
## 10    15         8          390   3850          8.5         70      1
## # ... with 388 more rows
```

```
for (var in num_var) {

  print(range(var))

}
```

```
## [1]  9.0 46.6
## [1] 3 8
## [1]  68 455
## [1] 1613 5140
## [1]  8.0 24.8
## [1] 70 82
## [1] 1 3
```

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

The mean values are equal to

```
for (var in num_var) {
  print(round(mean(var),2))


}
```

```
## [1] 23.51
## [1] 5.45
## [1] 193.43
## [1] 2970.42
## [1] 15.57
## [1] 76.01
## [1] 1.57
```

And the `sd` for quantitative variables are equal to

```
for (var in num_var) {
  print(round(sd(var),2))
}
```

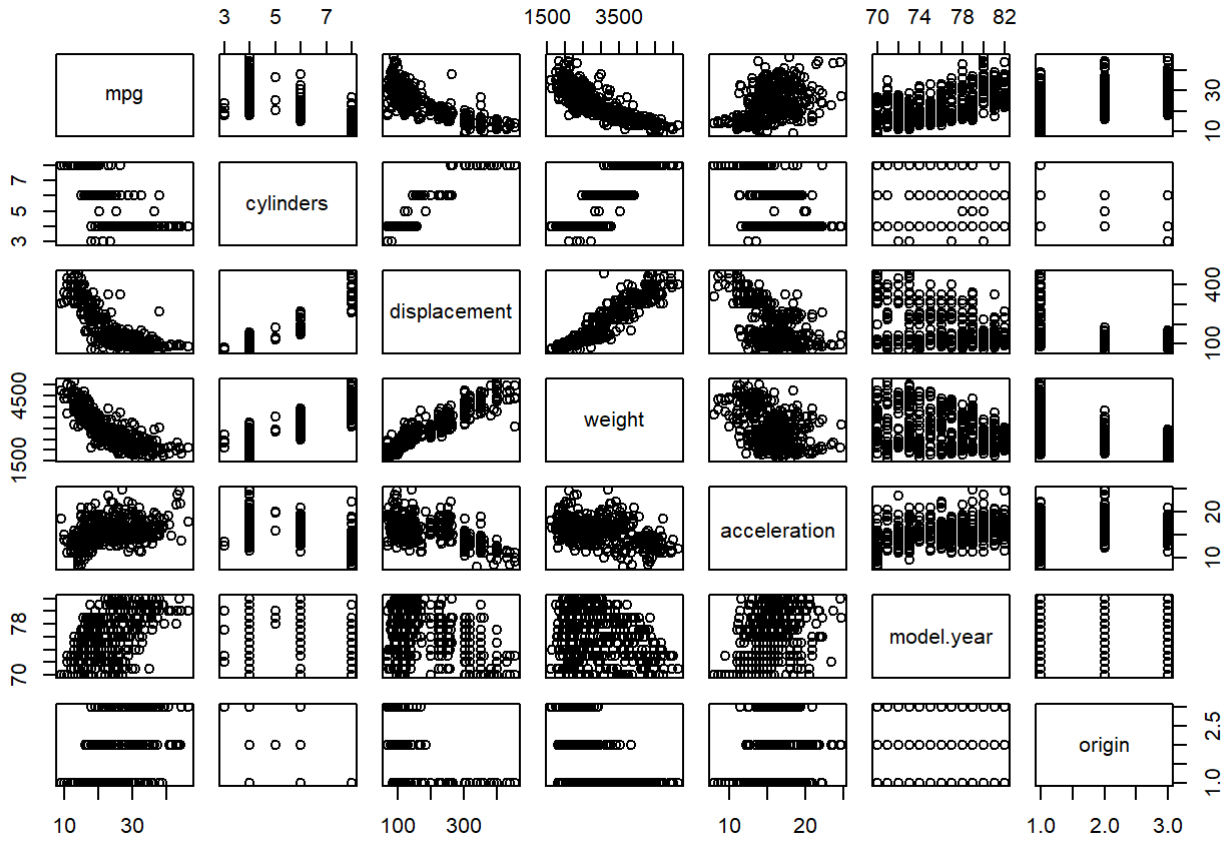
```
## [1] 7.82
## [1] 1.7
## [1] 104.27
## [1] 846.84
## [1] 2.76
## [1] 3.7
## [1] 0.8
```

(d)Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
# In progress..
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
pairs(num_var)
```



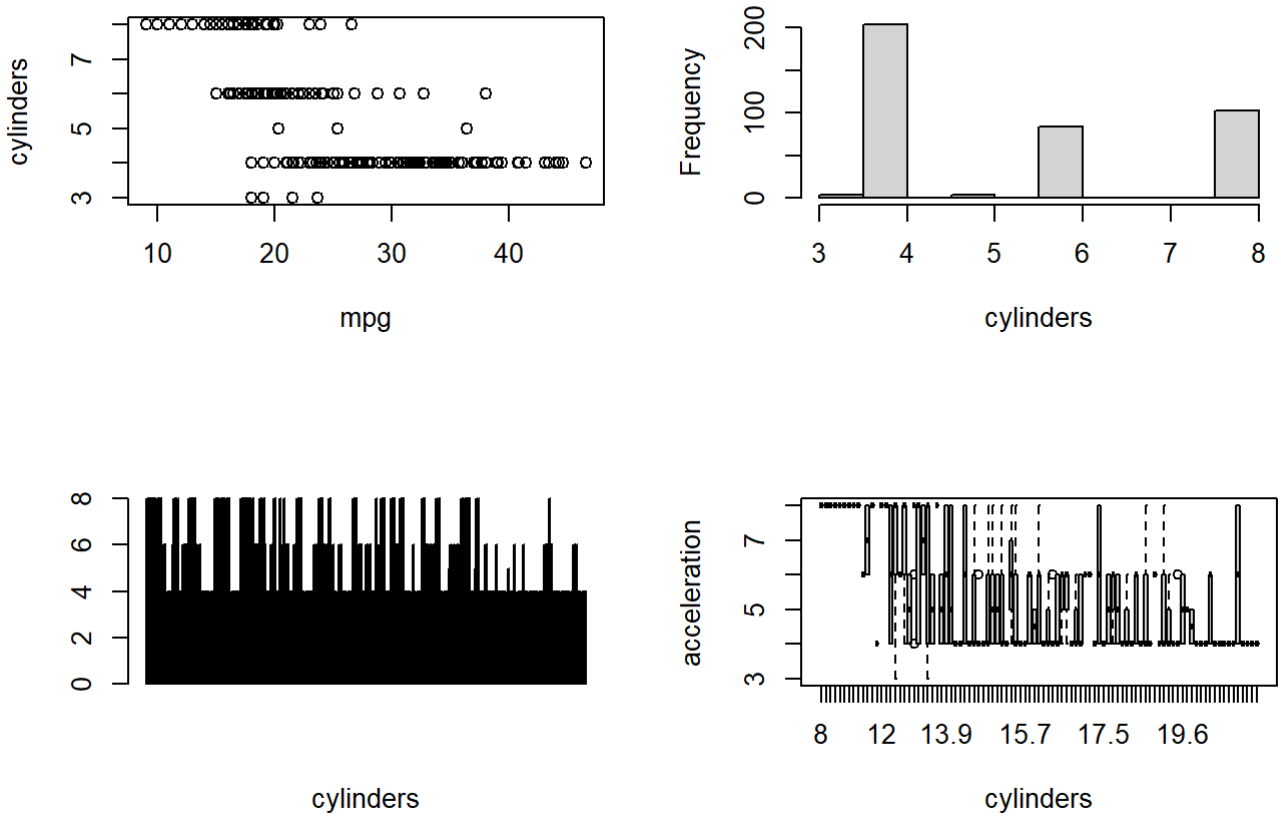
```
par(mfrow=c(2,2))

plot(data1$mpg,data1$cylinders, xlab="mpg", ylab="cylinders", main="")

hist(data1$cylinders, xlab="cylinders", main="")

barplot(data1$cylinders, xlab="cylinders", main="")

boxplot(data1$cylinders~data1$acceleration, xlab="cylinders",ylab="acceleration", main="")
```



(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

```
# In progress..
```