# Entrega 01 - MAE 0501

Vítor Garcia Comissoli - Nusp: 11810411

2024-09-03

## 9)

## a)

Primeiramente, para observar o tipo de variável de cada coluna, utilizou-se o cófigo abaixo:

```
class(data$mpg)
```

```
## [1] "numeric"
```

```
class(data$cylinders)
```

```
## [1] "integer"
```

```
class(data$displacement)
```

```
## [1] "numeric"
```

```
class(data$horsepower)
```

```
## [1] "character"
```

```
class(data$weight)
```

```
## [1] "integer"
```

```
class(data$acceleration)
```

```
## [1] "numeric"
```

```
class(data$year)
```

```
## [1] "integer"
```

```
class(data$origin)
```

```
## [1] "integer"
```

```
class(data$name)
```

```
## [1] "character"
```

```
summary(data)
```

```
##       mpg          cylinders      displacement    horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Length:397
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   Class :character
##  Median :23.00   Median :4.000   Median :146.0   Mode  :character
##  Mean   :23.52   Mean   :5.458   Mean   :193.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0
```

```
##  Max.   :46.60   Max.   :8.000   Max.   :455.0
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2223   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000
##  Median :2800   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2970   Mean   :15.56   Mean   :75.99   Mean   :1.574
##  3rd Qu.:3609   3rd Qu.:17.10   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##      name
##  Length:397
##  Class :character
##  Mode  :character
##
##
##
```

Como apresentado acima, obsertva-se que, tirando "name" e "origin" (que são categóricas), todas as outras variáveis são numéricas. Vale ressaltar que "horsepower" está como Character, o que não é o ideal, pois é uma variável numérica.

Transformou-se então as variáveis "horsepower" em numérica e "origin" e "name" em fatores, por meio do código abaixo:

```
data$horsepower<-as.numeric(data$horsepower)
```

```
## Warning: NAs introduced by coercion
```

```
data$origin<-as.factor(data$origin)
data$name<-as.factor(data$name)

data<-na.omit(data)
```

Então, ficamos com:

```
class(data$horsepower)
```

```
## [1] "numeric"
```

```
class(data$origin)
```

```
## [1] "factor"
```

```
class(data$name)
```

```
## [1] "factor"
```

```
summary(data)
```

```
##       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   1:245   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   2: 68   ford pinto        :  5
##  Median :15.50   Median :76.00   3: 79   toyota corolla    :  5
```

```
##  Mean   :15.54   Mean   :75.98          amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00          amc hornet        :  4
##  Max.   :24.80   Max.   :82.00          chevrolet chevette:  4
##                                         (Other)           :365
```

Assim, temos que "name" e "origin" são fatores, "horsepower" agora é numérica e as outras variáveis permanecem numéricas.

# b)

Temos que o intervalo de cada variável quantitativa é:

```
sapply(data[1:7],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
```

# c)

A média e o desvio padrão de cada variável quantitativa é:

```
print(sapply(data[1:7], mean))
```

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

```
print(sapply(data[1:7], sd))
```

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

# d)

O intervalo, a média e o desvio padrão de cada variável quantitativa, excluindo as linhas 10 a 85, é:

```
data2 <- data[-c(10:85),]

print(sapply(data2[,1:7], 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
```

```
print(sapply(data2[,1:7], mean))
```

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

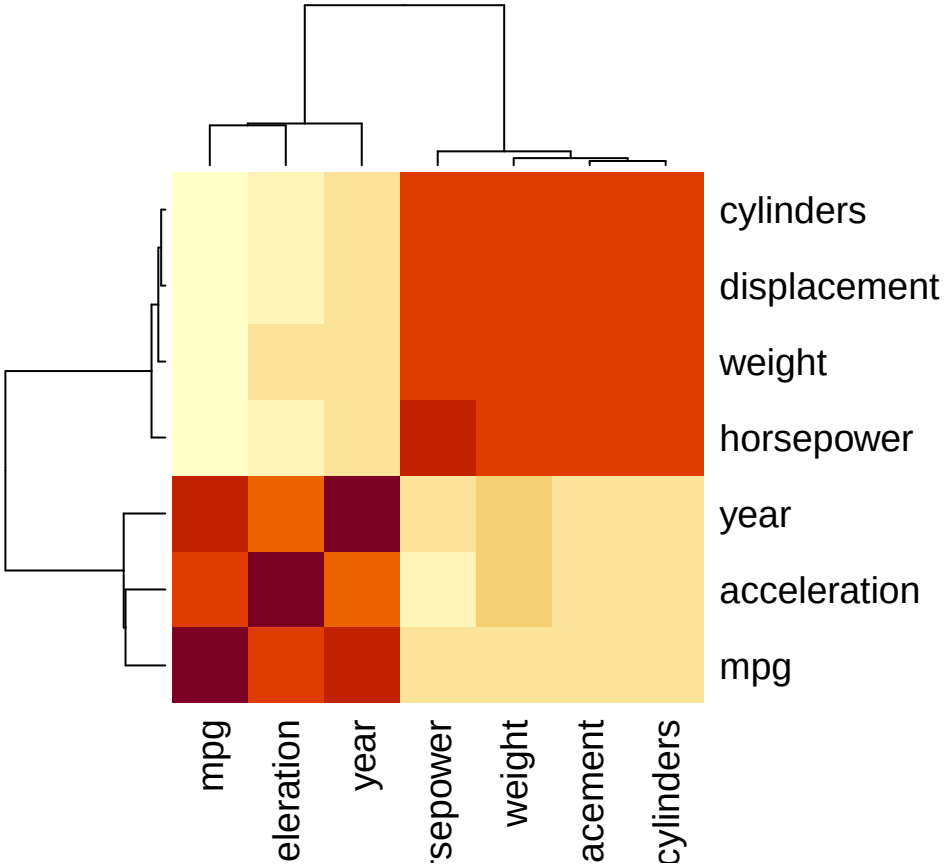
```
print(sapply(data2[,1:7], sd))
```

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

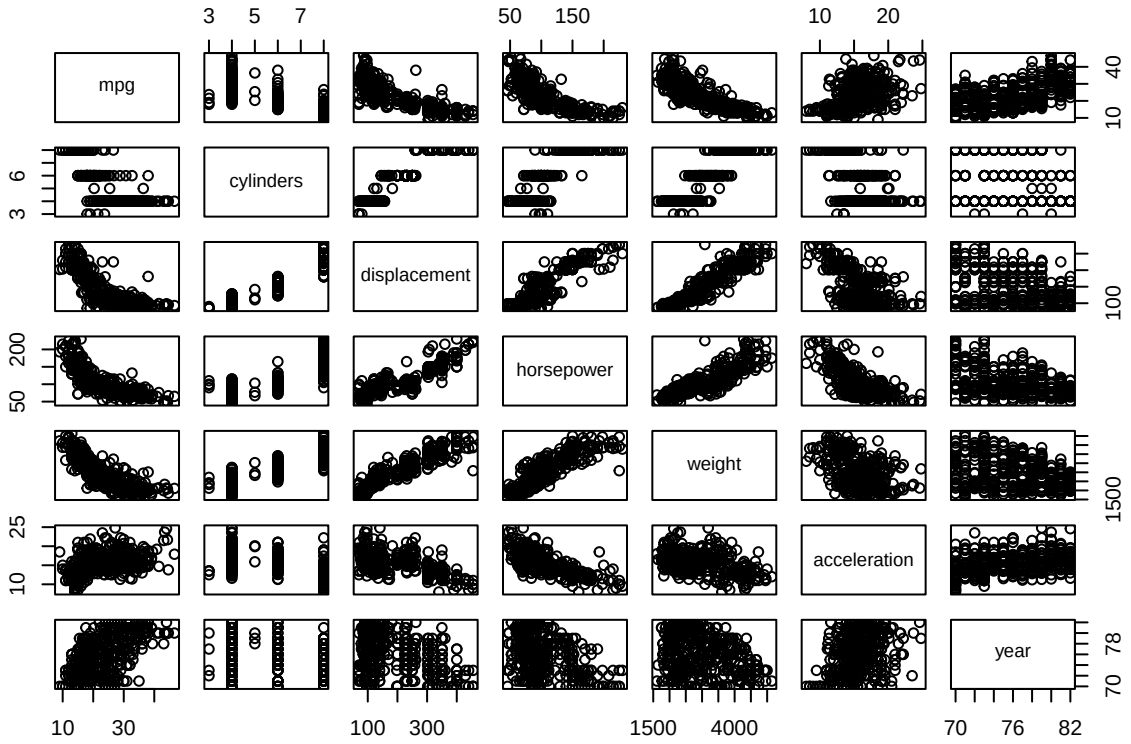
## e)

Para analisar a correlação entre as variáveis, utilizou-se o heatmap e o pairs (múltiplos scatterplots entre cada combinação de variáveis), como mostrado abaixo:

```
heatmap(cor(as.matrix(data[1:7])))
```




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




Realizou-se também a correlação de pearson entre as variáveis, como mostrado abaixo:

```
cor(data[1:7])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

Após a análise do heatmap e do pairs, observa-se que as variáveis que aparentam ser mais correlacionadas são “displacement” e “weight” (correlação de pearson de aproximadamente 0.93) e “displacement” e “cylinders” (correlação de pearson de aproximadamente 0.95).

## f)

De acordo com os resultados obtidos no item anterior, as variáveis que seriam mais úteis na predição de “mpg” seriam “cylinders”, “displacement”, “horsepower” e “weight”, pois são as que apresentam maiores correlações (em módulo) com “mpg” (correlações acima de 0.75, em módulo).