R Notebook

(a) Specify which of the predictors are quantitative (measuring numeric properties such as size, or quantity), and which are qualitative (measuring non-numeric properties such as color, appearance, type etc.)? Keep in mind that a qualitative variable may be represented as a quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings.

Qualitative : Origin, Name, Horsepower Quantitative : mpg, cylinders ,displacement ,weight ,acceleration ,year

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(corrr)
library(ggcorrplot)
library(rworldmap)
```

```
## Loading required package: sp

## ### Welcome to rworldmap ###

## For a short introduction type :   vignette('rworldmap')
```

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(maptools)
```

```
## Checking rgeos availability: FALSE
##      Note: when rgeos is not available, polygon geometry     computations in maptools depend on gpcl
##      which has a restricted licence. It is disabled by default;
##      to enable gpclib, type gpclibPermit()
```

```
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
Auto <- read_csv("Auto.csv")
```

```
## Parsed with column specification:
## cols(
##   mpg = col_double(),
##   cylinders = col_double(),
##   displacement = col_double(),
##   horsepower = col_character(),
##   weight = col_double(),
##   acceleration = col_double(),
##   year = col_double(),
##   origin = col_double(),
##   name = col_character()
## )
```

```
Auto
```

```
## # A tibble: 397 x 9
##      mpg cylinders displacement horsepower weight acceleration  year origin
##    <dbl>     <dbl>        <dbl> <chr>       <dbl>        <dbl> <dbl>  <dbl>
##  1    18         8          307 130          3504         12      70      1
##  2    15         8          350 165          3693         11.5    70      1
##  3    18         8          318 150          3436         11      70      1
##  4    16         8          304 150          3433         12      70      1
##  5    17         8          302 140          3449         10.5    70      1
##  6    15         8          429 198          4341         10      70      1
##  7    14         8          454 220          4354          9      70      1
##  8    14         8          440 215          4312          8.5    70      1
##  9    14         8          455 225          4425         10      70      1
## 10    15         8          390 190          3850          8.5    70      1
## # ... with 387 more rows, and 1 more variable: name <chr>
```

(b) What is the range, mean and standard deviation of each quantitative predictor?

```
summary(Auto)
```

```
##       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
##
##
##
```

```
print("MPG")
```

```
## [1] "MPG"
```

```
summary(Auto$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   17.50   23.00   23.52   29.00   46.60
```

```
range1 <- max(Auto$mpg) - min(Auto$mpg)
range1
```

```
## [1] 37.6
```

```
sd(Auto$mpg)
```

```
## [1] 7.825804
```

```
print("cylinder")
```

```
## [1] "cylinder"
```

```
summary(Auto$cylinders)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   5.458   8.000   8.000
```

```r
range2 <- max(Auto$mpg) - min(Auto$mpg)
range2
```

```
## [1] 37.6
```

```r
sd(Auto$cylinders)
```

```
## [1] 1.701577
```

```r
print("displacement")
```

```
## [1] "displacement"
```

```r
summary(Auto$displacement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   104.0   146.0   193.5   262.0   455.0
```

```r
range3 <- max(Auto$displacement) - min(Auto$displacement)
range3
```

```
## [1] 387
```

```r
sd(Auto$displacement)
```

```
## [1] 104.3796
```

```r
print("weight")
```

```
## [1] "weight"
```

```r
summary(Auto$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1613    2223    2800    2970    3609    5140
```

```r
range5 <- max(Auto$weight) - min(Auto$weight)
range5
```

```
## [1] 3527
```

```r
sd(Auto$weight)
```

```
## [1] 847.9041
```

```r
print("acceleration")
```

```
## [1] "acceleration"
```

```r
summary(Auto$acceleration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   13.80   15.50   15.56   17.10   24.80
```

```r
range6 <- max(Auto$acceleration) - min(Auto$acceleration)
range6
```

```
## [1] 16.8
```

```r
sd(Auto$acceleration)
```

```
## [1] 2.749995
```

```r
print("year")
```

```
## [1] "year"
```

```r
summary(Auto$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.00   73.00   76.00   75.99   79.00   82.00
```

```r
range7 <- max(Auto$year) - min(Auto$year)
range7
```

```
## [1] 12
```

```r
sd(Auto$year)
```

```
## [1] 3.690005
```

(c) Now remove the 40th through 80th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```r
new_auto <- Auto[-c(40:80),]
#Redefined Range , mean and SD
print("MPG")
```

```
## [1] "MPG"
```

```r
summary(new_auto$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   18.00   23.65   24.02   29.82   46.60
```

```r
range1 <- max(new_auto$mpg) - min(new_auto$mpg)
range1
```

```
## [1] 37.6
```

```r
sd(Auto$mpg)
```

```
## [1] 7.825804
```

```r
print("cylinder")
```

```
## [1] "cylinder"
```

```r
summary(new_auto$cylinders)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   5.399   6.000   8.000
```

```r
range2 <- max(new_auto$mpg) - min(new_auto$mpg)
range2
```

```
## [1] 37.6
```

```r
sd(new_auto$cylinders)
```

```
## [1] 1.659254
```

```r
print("displacement")
```

```
## [1] "displacement"
```

```r
summary(new_auto$displacement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   103.2   146.0   189.2   252.0   455.0
```

```r
range3 <- max(new_auto$displacement) - min(new_auto$displacement)
range3
```

```
## [1] 387
```

```r
sd(new_auto$displacement)
```

```
## [1] 100.8794
```

```r
print("weight")
```

```
## [1] "weight"
```

```r
summary(new_auto$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1649    2222    2782    2935    3508    4997
```

```r
range5 <- max(new_auto$weight) - min(new_auto$weight)
range5
```

```
## [1] 3348
```

```r
sd(new_auto$weight)
```

```
## [1] 810.8406
```

```r
print("acceleration")
```

```
## [1] "acceleration"
```

```r
summary(new_auto$acceleration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   14.00   15.50   15.61   17.02   24.80
```

```r
range6 <- max(new_auto$acceleration) - min(new_auto$acceleration)
range6
```

```
## [1] 16.8
```

```r
sd(new_auto$acceleration)
```

```
## [1] 2.712348
```

```r
print("year")
```

```
## [1] "year"
```

```
summary(new_auto$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.00   74.00   77.00   76.51   79.00   82.00
```

```
range7 <- max(new_auto$year) - min(new_auto$year)
range7
```

```
## [1] 12
```

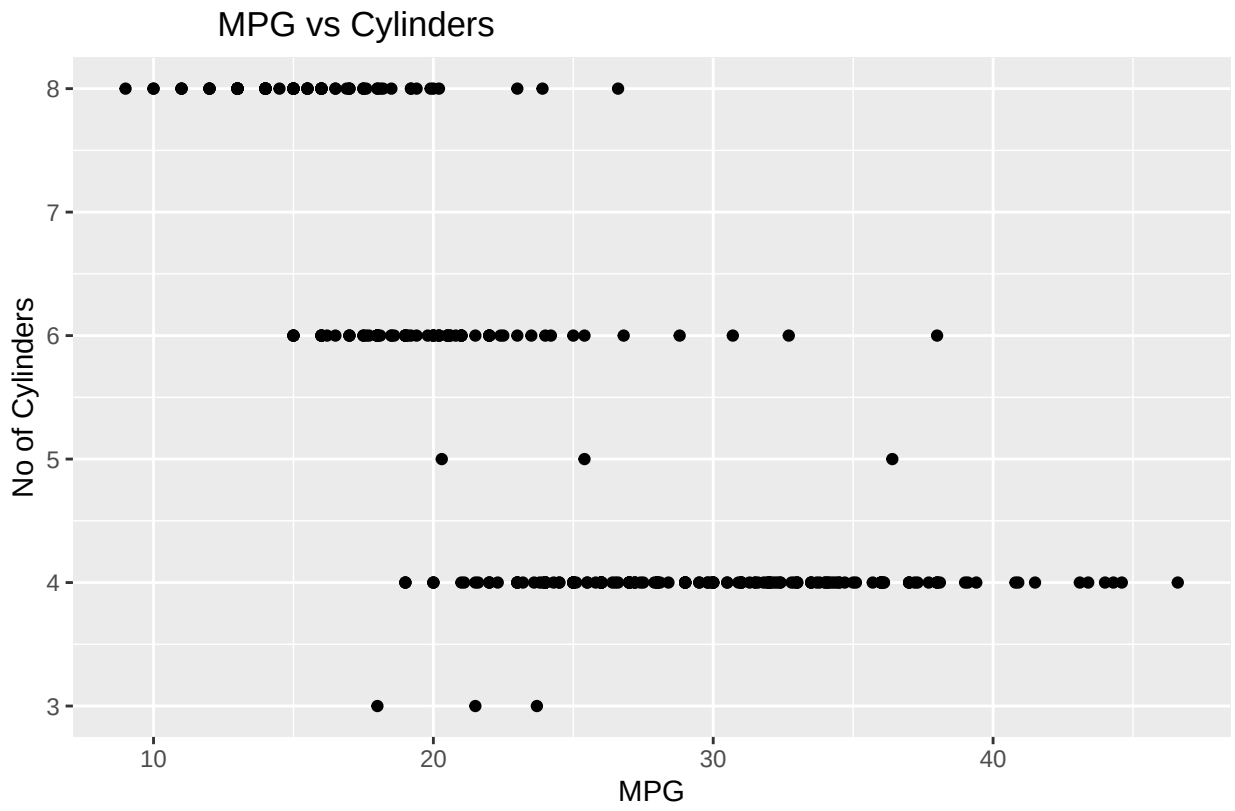
```
sd(new_auto$year)
```

```
## [1] 3.553609
```

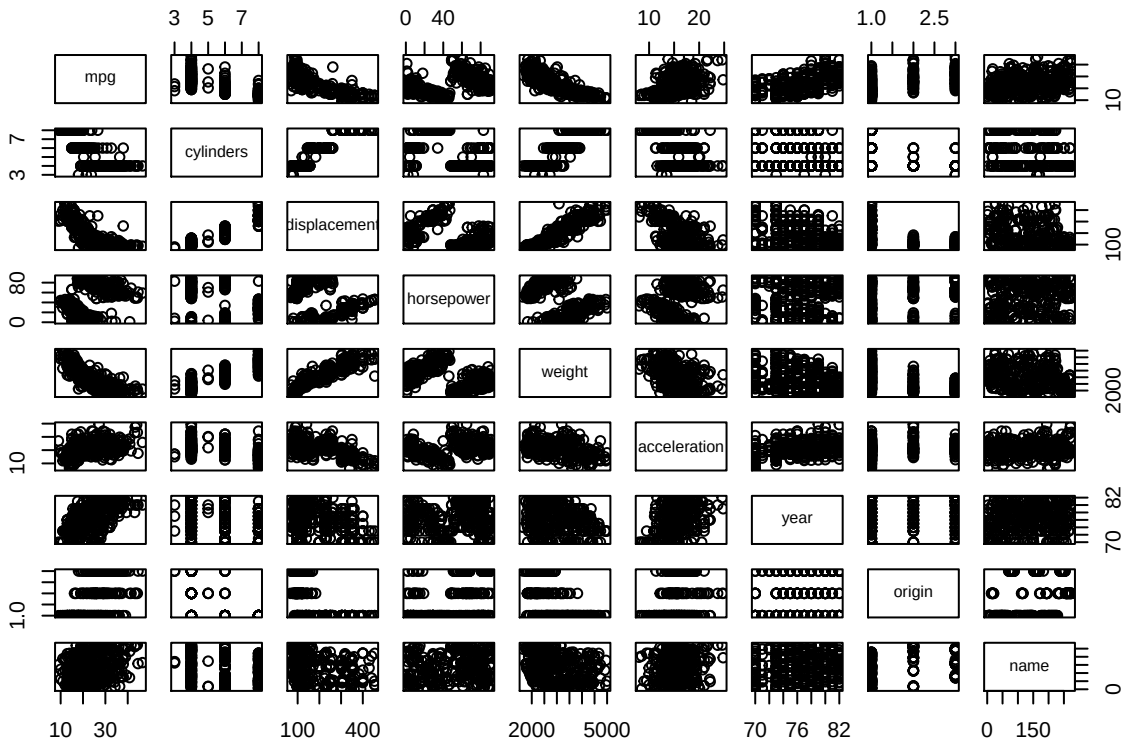
(d) Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create a correlation matrix for the relevant variables.

```
SP <- ggplot(new_auto, aes(x=new_auto$mpg ,y = new_auto$cylinders)) +geom_point()
SP <- SP +labs(title = "
               MPG vs Cylinders",x="MPG", y="No of Cylinders")
show(SP)
```

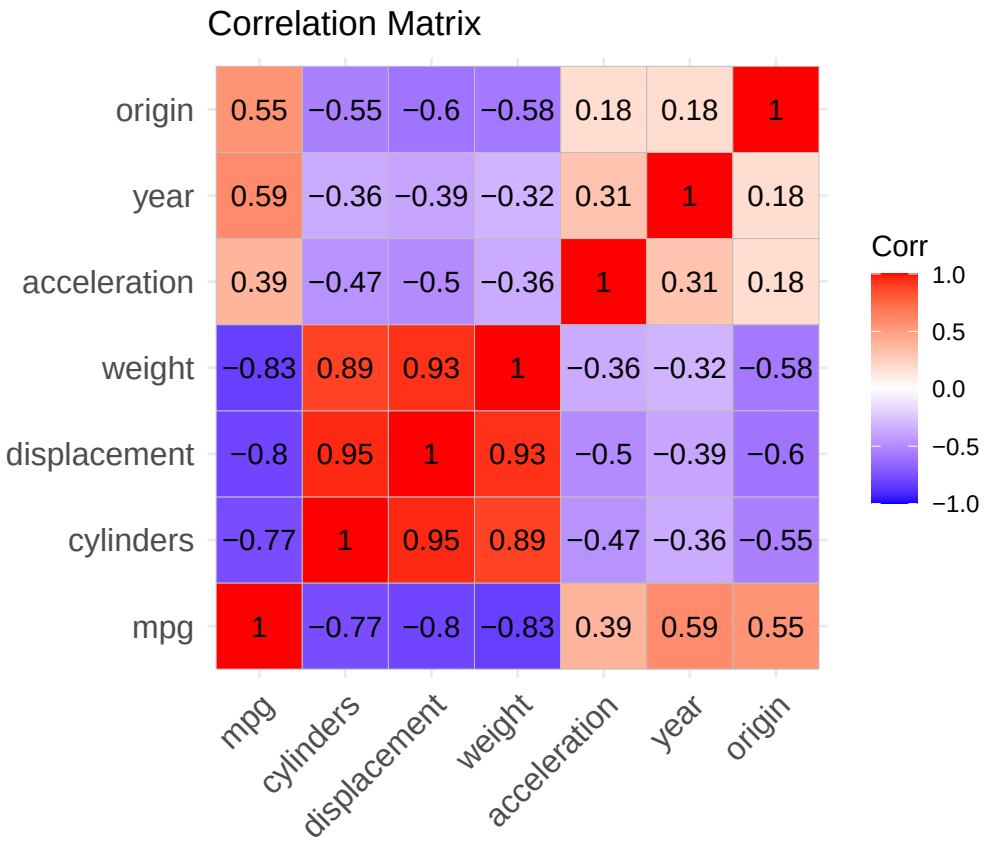
```
## Warning: Use of ‘new_auto$mpg‘ is discouraged. Use ‘mpg‘ instead.

## Warning: Use of ‘new_auto$cylinders‘ is discouraged. Use ‘cylinders‘ instead.
```



```r
#Make a matrix now
plot(new_auto)
```




```r
df <- subset(new_auto,select = -c(4,9))
corr_val <- cor(df)
ggcorrplot(corr_val, lab = TRUE,title = "Correlation Matrix" )
```

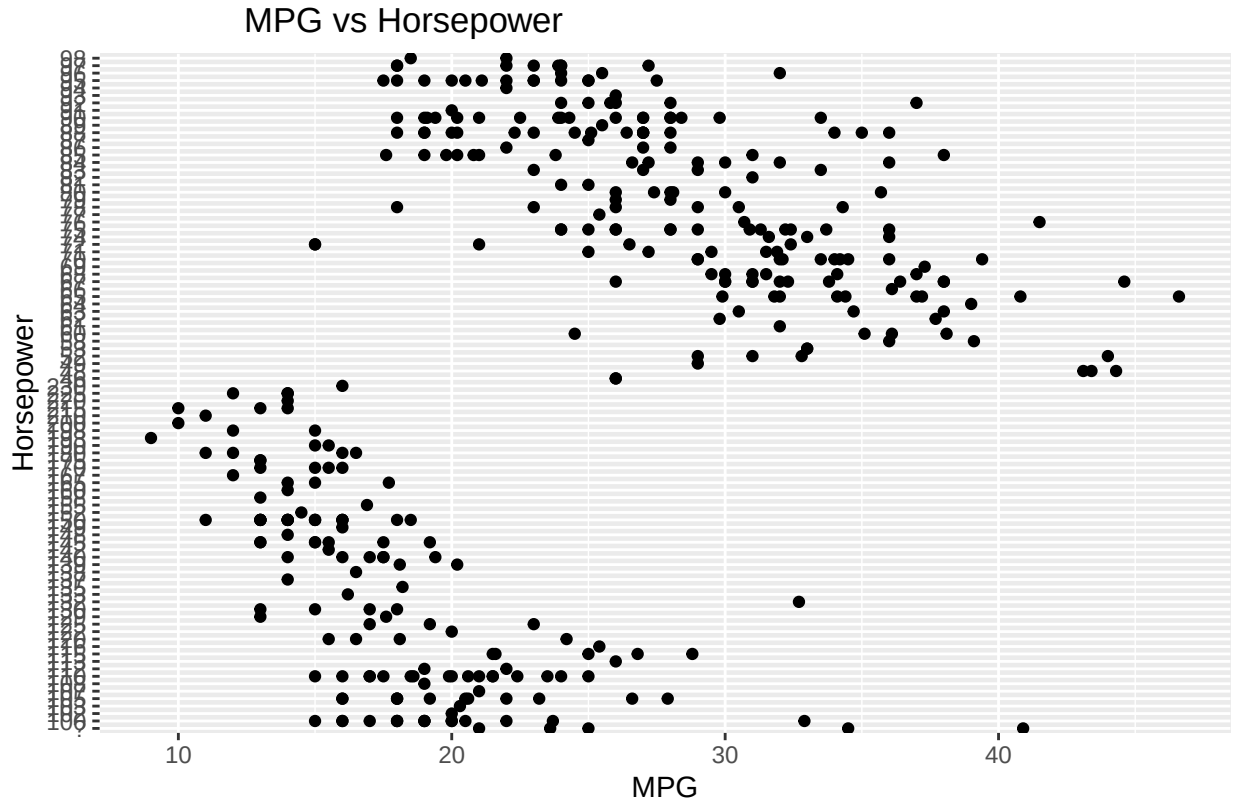


(e) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Which, if any, of the other variables might be useful in predicting mpg? Justify your answer based on the prior correlations.

```
SP <- ggplot(new_auto, aes(x=mpg ,y = horsepower)) +geom_point()
SP <- SP +labs(title = "
               MPG vs Horsepower",x="MPG", y="Horsepower")
print("From the dataset provided above, we can concur that the Miles per gallon type does not have an ol
```

```
## [1] "From the dataset provided above, we can concur that the Miles per gallon type does not have an
```

```
show(SP)
```


MPG vs Horsepower
Horsepower
MPG
10
20
30
40