Areka Raza Initial Code

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
data<-read.csv("bmw_pricing_challenge.csv", stringsAsFactors = FALSE,
sep=",", header = TRUE)

install.packages("corrplot")

## Installing package into 'C:/Users/areka/OneDrive/Documents/R/win-
library/4.1'
## (as 'lib' is unspecified)

## package 'corrplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\areka\AppData\Local\Temp\Rtmpqgqwhl\downloaded_packages

library(corrplot)

## corrplot 0.92 loaded

summary(data)

##   maker_key          model_key            mileage         engine_power
##  Length:4843        Length:4843        Min.   :    -64   Min.   :  0
##  Class :character   Class :character   1st Qu.: 102914   1st Qu.:100
##  Mode  :character   Mode  :character   Median : 141080   Median :120
##                                        Mean   : 140963   Mean   :129
##                                        3rd Qu.: 175196   3rd Qu.:135
##                                        Max.   :1000376   Max.   :423
##  registration_date      fuel           paint_color          car_type
##  Length:4843        Length:4843        Length:4843        Length:4843
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  feature_1       feature_2       feature_3       feature_4
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2181      FALSE:1004      FALSE:3865      FALSE:3881
##  TRUE :2662      TRUE :3839      TRUE :978       TRUE :962
##
##
##
##  feature_5       feature_6       feature_7       feature_8
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2613      FALSE:3674      FALSE:329       FALSE:2223
##  TRUE :2230      TRUE :1169      TRUE :4514      TRUE :2620  
##                                                                
##                                                                
##                                                                
##      price          sold_at         
##  Min.   :   100   Length:4843       
##  1st Qu.: 10800   Class :character  
##  Median : 14200   Mode  :character  
##  Mean   : 15828                     
##  3rd Qu.: 18600                     
##  Max.   :178500
```

```
#mileage cannot be -64, engine power cannot be 0.

sapply(data, class)
```

```
##         maker_key         model_key           mileage      engine_power 
##       "character"       "character"         "integer"         "integer" 
## registration_date              fuel       paint_color          car_type 
##       "character"       "character"       "character"       "character" 
##         feature_1         feature_2         feature_3         feature_4 
##         "logical"         "logical"         "logical"         "logical" 
##         feature_5         feature_6         feature_7         feature_8 
##         "logical"         "logical"         "logical"         "logical" 
##             price           sold_at 
##         "integer"       "character"
```

```
sum(is.na(data))
```

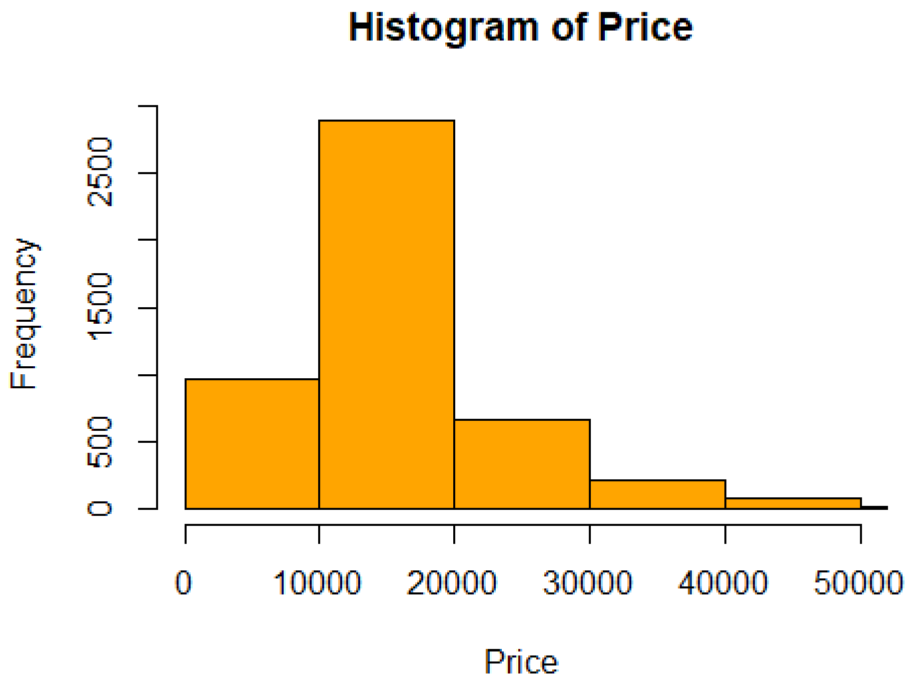
```
## [1] 0
```

```
# Creating a new column - Age --------------------------------------------

data$age<-data$age
data$sold_at<-as.Date(data$sold_at)
data$registration_date<-as.Date(data$registration_date)
data$age<-(data$sold_at-data$registration_date) / 365
data$age = as.numeric(data$age)
head(data$age)
```

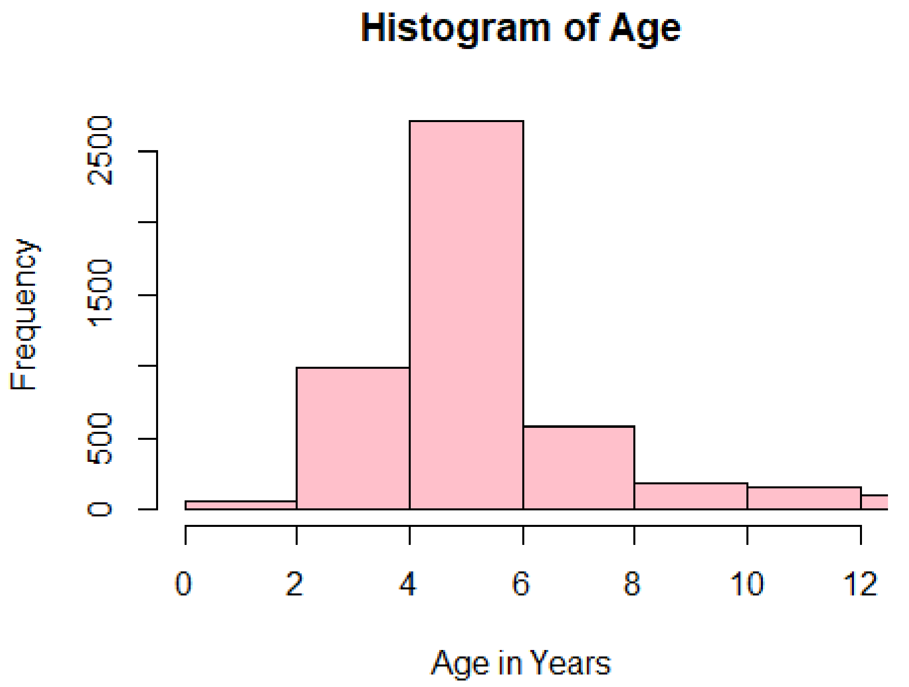
```
## [1] 5.920548 1.838356 5.841096 3.591781 3.334247 6.761644
```

```
# Renaming features ------------------------------------------------------


colnames(data)[9]<-"HasMoonRoof"
colnames(data)[10]<-"HasLeatherSeats"
colnames(data)[11]<-"HasHeatedSeats"
colnames(data)[12]<-"HasNavigationSystem"
colnames(data)[13]<-"HasBluetooth"
colnames(data)[14]<-"HasRemoteStart"
```

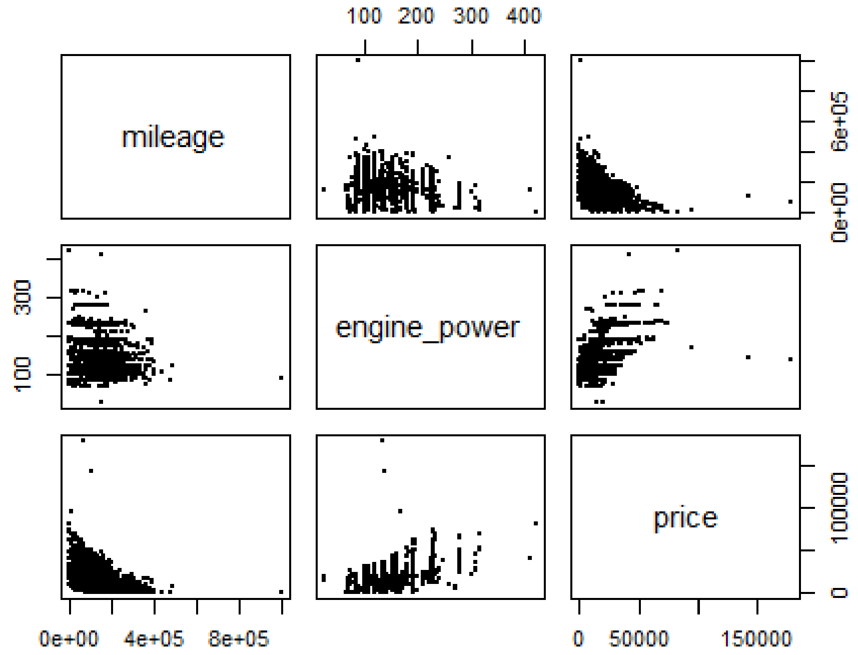
```r
colnames(data)[15]<-"HasBlindSpotMonitoring"
colnames(data)[16]<-"HasMSportPackage"


# Adjusting incorrect/logical values ----------------------------------------
------


data["mileage"][data["mileage"]==-64] <- 64
data["engine_power"][data["engine_power"]==0] <- mean(data$engine_power)

data$HasMoonRoof [data$HasMoonRoof == "true"] <- 1
data$HasMoonRoof [data$HasMoonRoof == "false"] <- 0

data$HasBluetooth[data$HasBluetooth == "true"] <- 1
data$HasBluetooth[data$HasBluetooth == "false"] <- 0

data$HasNavigationSystem [data$HasNavigationSystem == "true"] <- 1
data$HasNavigationSystem [data$HasNavigationSystem == "false"] <- 0

data$HasLeatherSeats [data$HasLeatherSeats == "true"] <- 1
data$HasLeatherSeats [data$HasLeatherSeats == "false"] <- 0

data$HasHeatedSeats [data$HasHeatedSeats == "true"] <- 1
data$HasHeatedSeats [data$HasHeatedSeats == "false"] <- 0

data$HasBlindSpotMonitoring [data$HasBlindSpotMonitoring == "true"] <- 1
data$HasBlindSpotMonitoring [data$HasBlindSpotMonitoring == "false"] <- 0

data$HasRemoteStart [data$HasRemoteStart == "true"] <- 1
data$HasRemoteStart [data$HasRemoteStart == "false"] <- 0

data$HasMSportPackage [data$HasMSportPackage == "true"] <- 1
data$HasMSportPackage [data$HasMSportPackage == "false"] <- 0

# Visualizing the data ----------------------------------------------------


hist(data$price, main="Histogram of Price", xlab="Price", col="orange",
xlim=(c(0,50000)))
```



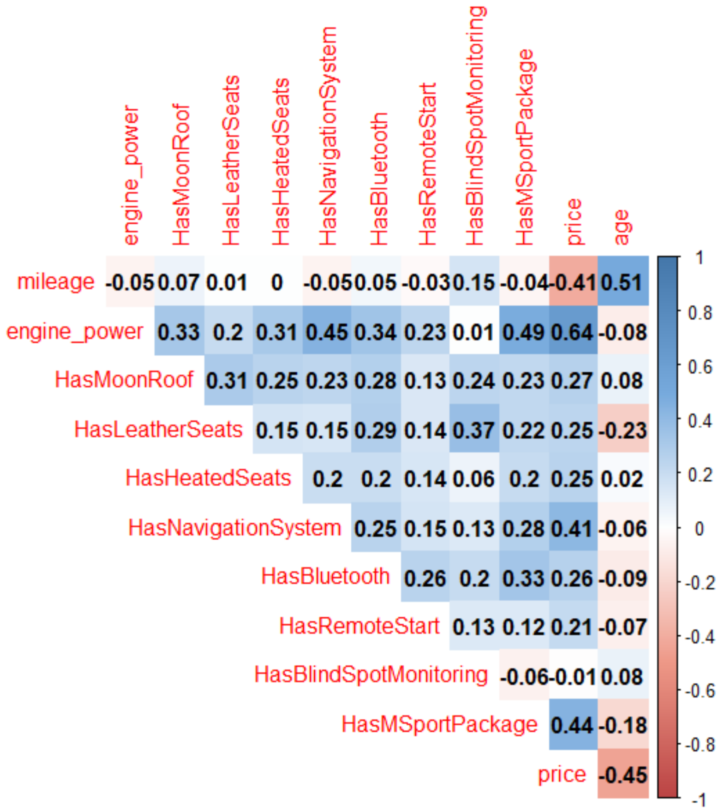
```
hist(data$age, main= "Histogram of Age", xlab="Age in Years", col="pink",
xlim=c(0,12))
```



```
pairs(data[c(3,4,17)],pch=19,cex=0.5)
```



```
cor<-cor(data[, unlist(lapply(data, is.numeric))])

library(corrplot)
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",
"#4477AA"))
corrplot(cor, method="color", col=col(200),
         diag=FALSE,
         type="upper",
         addCoef.col = "black"
)
```



```
# Normalizing the data ----------------------------------------------------


min_max_norm<-function(x){
  (x-min(x)) / (max(x)-min(x))
}
scaled_data<-
as.data.frame(lapply(data[c("age","price","mileage","engine_power"
,"HasMoonRoof", "HasLeatherSeats", "HasHeatedSeats", "HasNavigationSystem",
"HasBluetooth", "HasRemoteStart", "HasBlindSpotMonitoring",
"HasMSportPackage")], min_max_norm))
head(scaled_data)

##          age      price    mileage engine_power HasMoonRoof
HasLeatherSeats
## 1 0.19376680 0.06278027 0.14030323    0.1884422           1
1
## 2 0.04540476 0.39013453 0.01386068    0.7336683           1
1
## 3 0.19087922 0.05661435 0.18317585    0.2386935           0
0
## 4 0.10913074 0.14013453 0.12793109    0.2763819           1
1
## 5 0.09977098 0.18665919 0.09700274    0.3391960           1
1
## 6 0.22433536 0.09529148 0.15224050    0.5025126           1
1
##   HasHeatedSeats HasNavigationSystem HasBluetooth HasRemoteStart
## 1              0                   0            1              1
## 2              0                   0            0              1
## 3              0                   0            1              0
## 4              0                   0            1              1
## 5              0                   0            0              1
## 6              0                   0            1              1
##   HasBlindSpotMonitoring HasMSportPackage
## 1                      1                0
## 2                      1                1
## 3                      1                0
## 4                      1                1
## 5                      1                1
## 6                      1                1
```

```
# Training the dataset ----------------------------------------------------

train_index<-sample(1:nrow(data),0.7*nrow(data))
train.set<-scaled_data[train_index,]
test.set<-scaled_data[-train_index,]


# Multiple Linear Regression ----------------------------------------------


fit=lm(price~ age + mileage + engine_power + HasMoonRoof + HasLeatherSeats +
HasHeatedSeats + HasNavigationSystem + HasBluetooth + HasRemoteStart +
HasBlindSpotMonitoring +HasMSportPackage , data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = price ~ age + mileage + engine_power + HasMoonRoof +
##     HasLeatherSeats + HasHeatedSeats + HasNavigationSystem +
##     HasBluetooth + HasRemoteStart + HasBlindSpotMonitoring +
##     HasMSportPackage, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24207  -2337   -220   1813 159882
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9106.22909  476.18917  19.123  < 2e-16 ***
## age                    -932.11315   38.84680 -23.995  < 2e-16 ***
## mileage                  -0.03854    0.00156 -24.698  < 2e-16 ***
## engine_power            107.27249    2.65950  40.336  < 2e-16 ***
## HasMoonRoof            1609.00099  182.68195   8.808  < 2e-16 ***
## HasLeatherSeats         491.47920  233.86467   2.102 0.035644 *
## HasHeatedSeats         1030.41476  212.45635   4.850 1.27e-06 ***
## HasNavigationSystem    2828.75985  226.45749  12.491  < 2e-16 ***
## HasBluetooth           -320.68982  183.64069  -1.746 0.080824 .
## HasRemoteStart          668.08897  195.47755   3.418 0.000637 ***
## HasBlindSpotMonitoring  346.53828  361.13218   0.960 0.337310
## HasMSportPackage       1848.06370  192.16750   9.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5507 on 4831 degrees of freedom
## Multiple R-squared:  0.644,  Adjusted R-squared:  0.6432
## F-statistic: 794.6 on 11 and 4831 DF,  p-value: < 2.2e-16
```

```
price = 9106.22909 + (-932.11315 * 20) + (-0.03854 * 1000) + (107.27249 *
145) + (1609.00099 * 1) + (491.47920 * 0) + (1030.41476 * 1) + ( 2828.75985 *
1) + (-320.68982 * 0)
+ (668.08897 * 0) + (346.53828 * 0) + (1848.06370 * 1)
```

```
## [1] 1848.064
```

```
price
```

```
## [1] 11448.11
```

```
#The price of a vehicle that is 20 years old, has mileage of up to 1000km and
engine power of 145 with the following features: Moon roof, heated seats,
navigation system and MSport Package is $11448.11


confint(fit)
```

```
##                                2.5 %        97.5 %
## (Intercept)             8.172682e+03  1.003978e+04
## age                    -1.008271e+03 -8.559557e+02
## mileage                -4.159638e-02 -3.547832e-02
## engine_power            1.020587e+02  1.124863e+02
## HasMoonRoof             1.250861e+03  1.967141e+03
## HasLeatherSeats         3.299800e+01  9.499604e+02
## HasHeatedSeats          6.139036e+02  1.446926e+03
## HasNavigationSystem     2.384800e+03  3.272720e+03
## HasBluetooth           -6.807091e+02  3.932951e+01
## HasRemoteStart          2.848640e+02  1.051314e+03
## HasBlindSpotMonitoring -3.614452e+02  1.054522e+03
## HasMSportPackage        1.471328e+03  2.224799e+03
```