Prediction of Used Car Price

Dhruvkumar Shah Student ID:- 300318529

24/04/2021

Introduction

In the following project I would like to predict selling price of used cars. For that I will be using this data set I found on Kaggle of selling price of USed cars in India.

I would also like to check which are the factors impacting the selling price of used car.

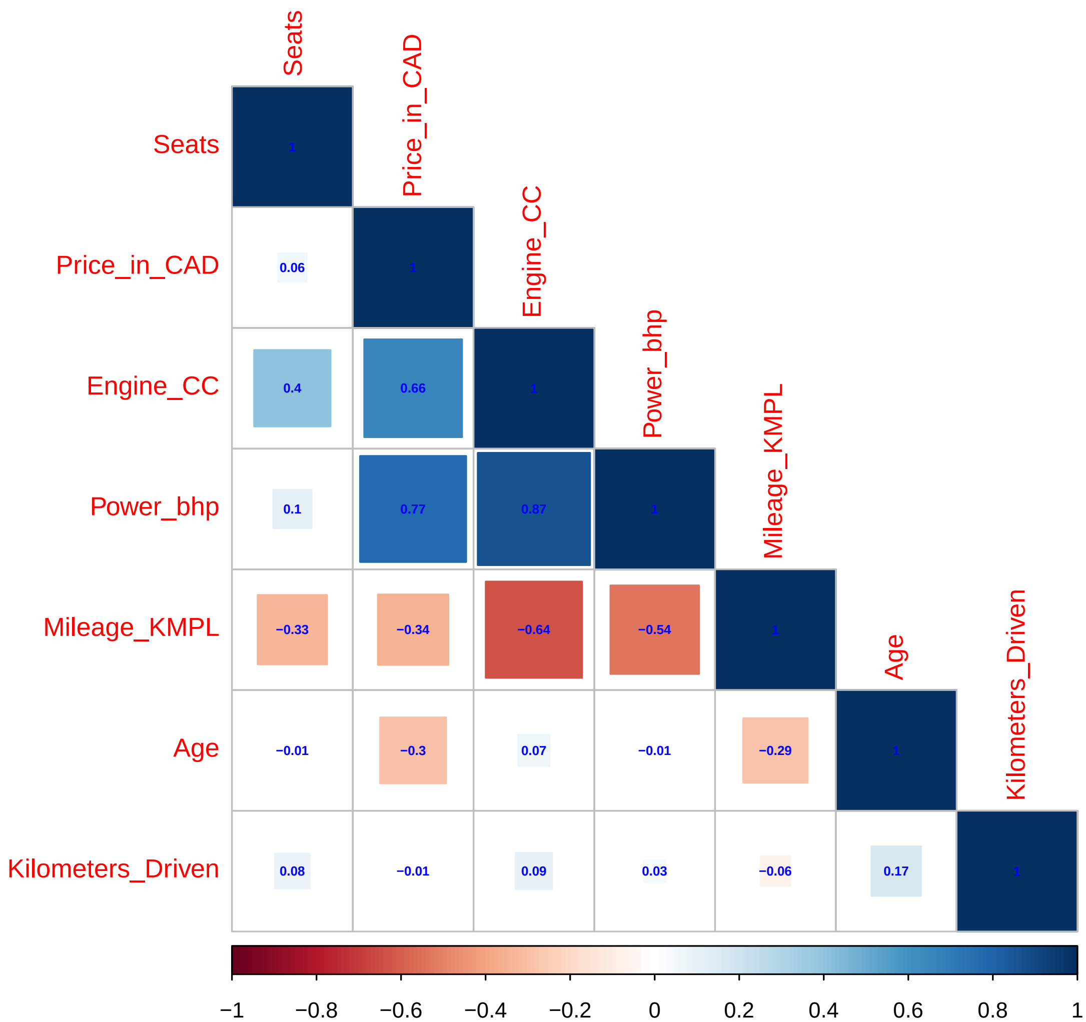
Data Set

```
##     Brand     Model Location Age Kilometers_Driven Fuel_Type Transmission
## 1  Maruti     Wagon   Mumbai  10             72000       CNG       Manual
## 2 Hyundai Creta 1.6     Pune   5             41000    Diesel       Manual
## 3   Honda      Jazz  Chennai   9             46000    Petrol       Manual
##   Owner_Type Mileage_KMPL Engine_CC Power_bhp Seats Price_in_CAD
## 1      First        26.60       998     58.16     5         2917
## 2      First        19.67      1582    126.20     5        20833
## 3      First        18.20      1199     88.70     5         7500
```

The data set consists of 13 Variables

```
## [1] 5872   13
```

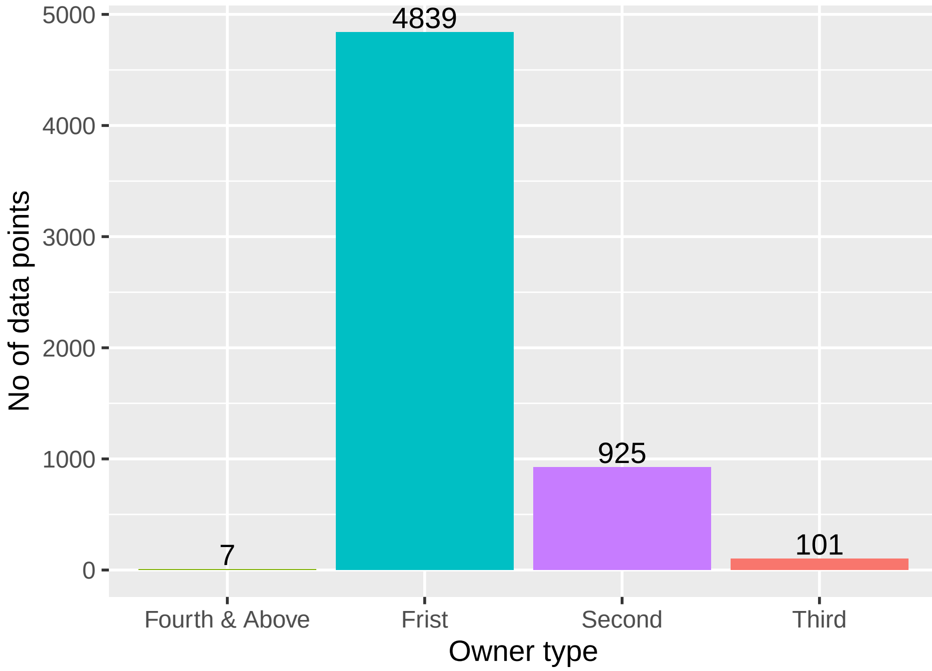
I would like to start the analysis by checking the correlation between different variables and impact of numerical variables on the selling price.

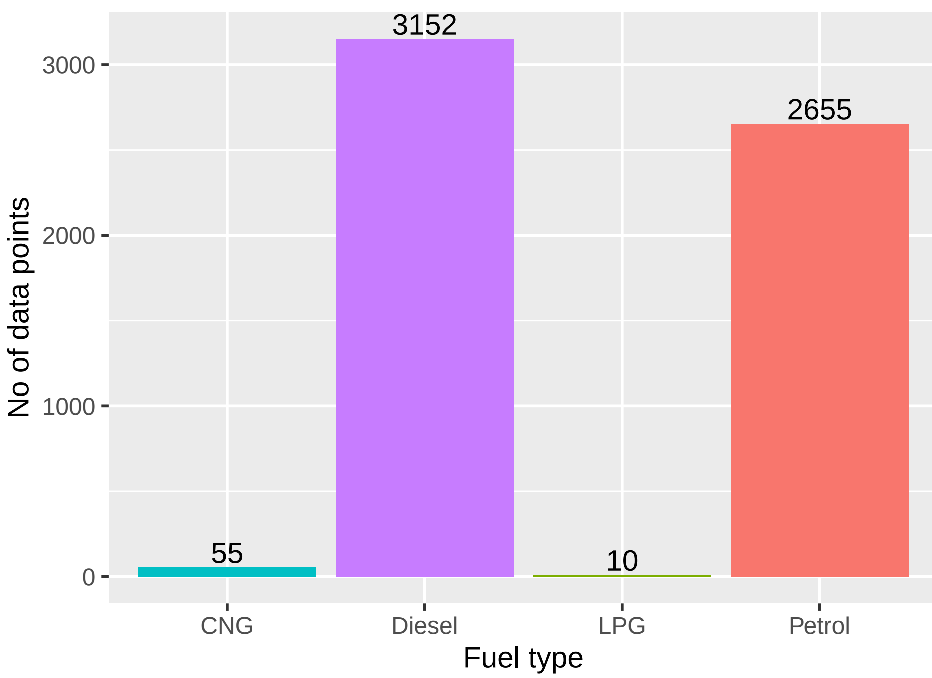


Correlation plot shows high correlation between Engine Cubic Capacity and Break Horse Power which makes sense because higher horse power cars will have higher power as well.

Will keep a track of the same as such high correlation might cause multicollnearity issues going forward.

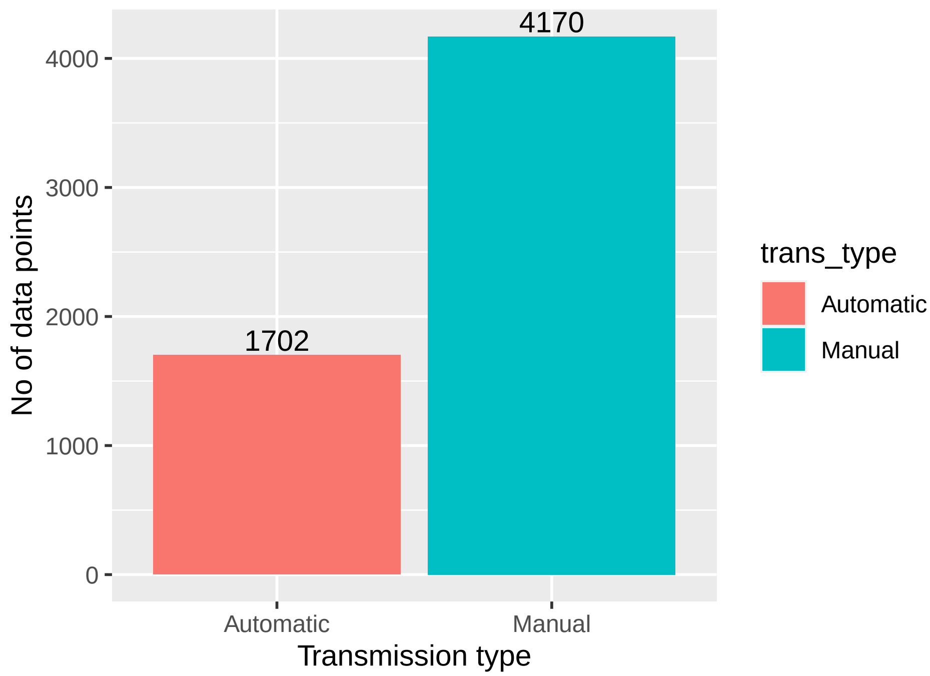
Next I would like to see the distribution of values in the categorical variables.

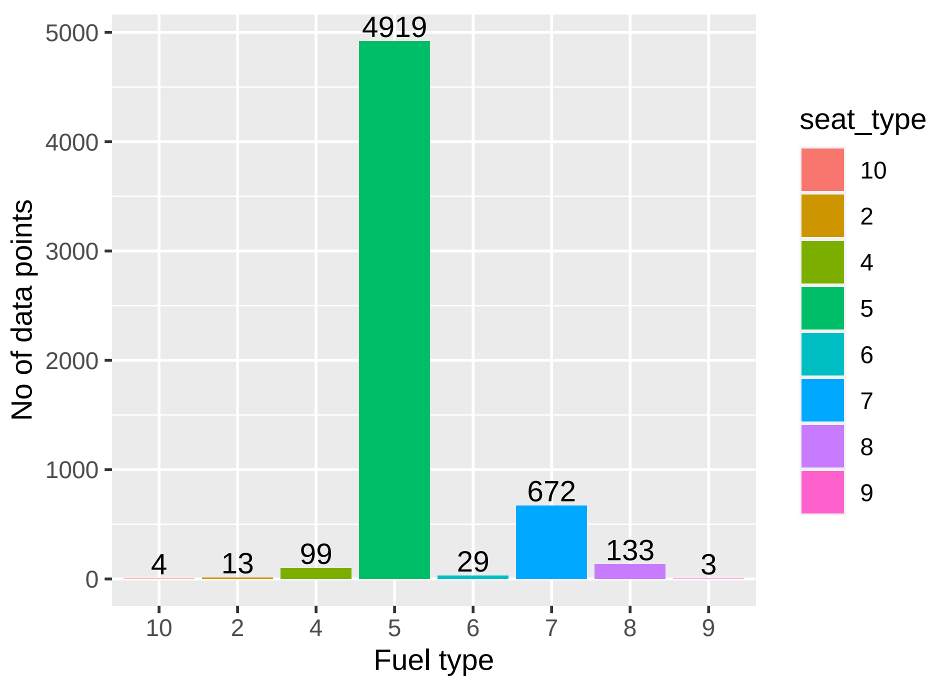




There looks to be high number of First hand car owners compared to others therefore I would like add a single encoded variable so that I can have total of First hand owners as one group and others in another group.

From the above bar plot I would like to make a column to accommodate Fuel Type as Diesel or Others.

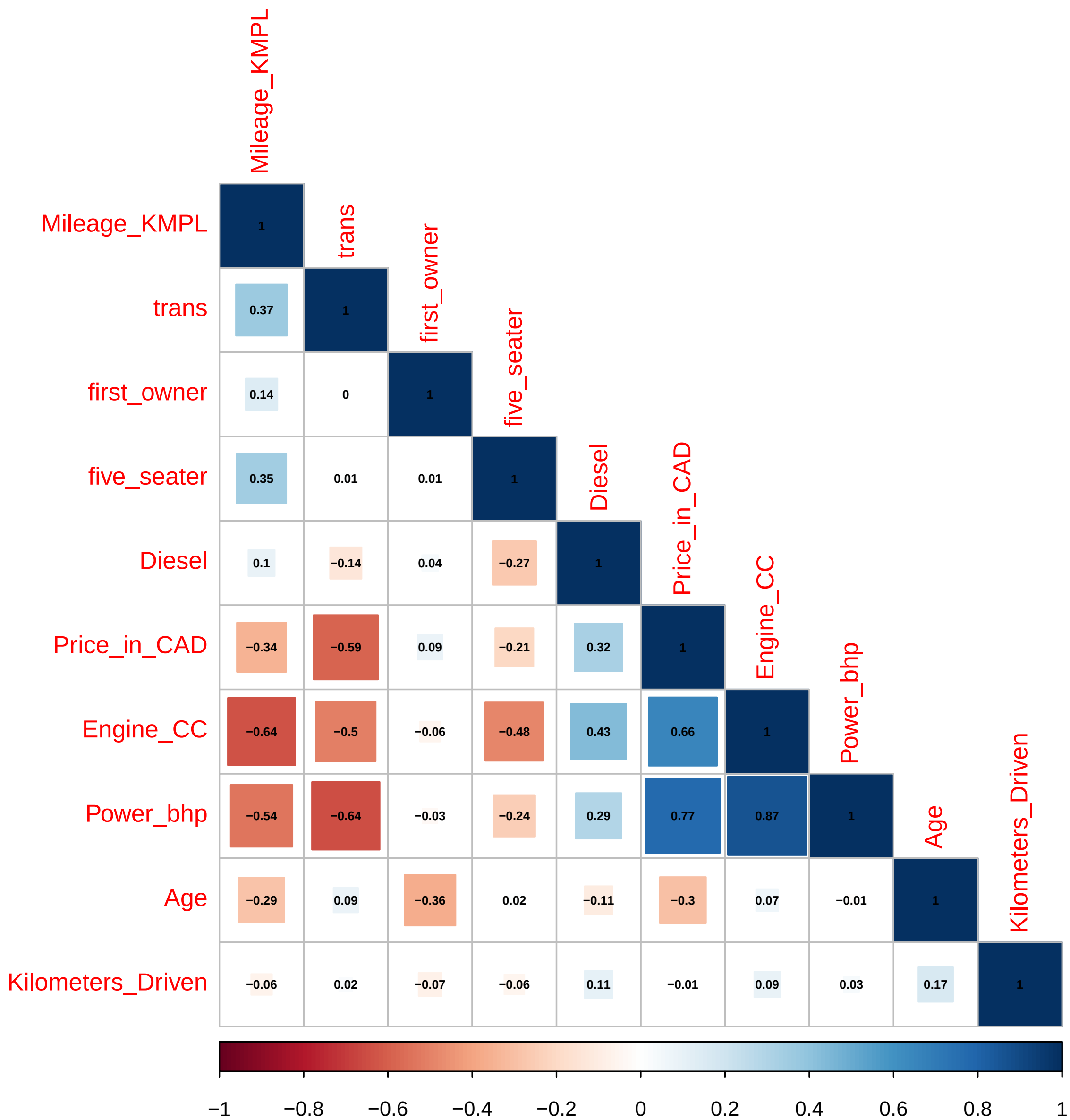




Converting Transmission into categorical variables where Encoding Manual as 1 and Automatic as 0.

Majority of the cars are 5 seater therefore we can bifurcate the data in 5 seater and others.

Plotting correlation matrix again with all the numerical and categorical variables



We can observe negative correlation tranmission and Price and this this relationship would be interesting to observe in the Regression model.

Regression model 1

```
##
## Call:
## lm(formula = Price_in_CAD ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76740  -5003   -985   3397 205705
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.137e+04  1.965e+03   5.787 7.70e-09 ***
## Age               -1.529e+03  6.776e+01 -22.567  < 2e-16 ***
## Kilometers_Driven -4.339e-02  5.882e-03  -7.377 1.94e-13 ***
## Mileage_KMPL      -2.665e+02  6.258e+01  -4.259 2.10e-05 ***
## Engine_CC         -2.050e+00  8.320e-01  -2.465   0.0138 *
## Power_bhp          2.407e+02  7.683e+00  31.323  < 2e-16 ***
## first_owner        2.278e+02  4.752e+02   0.479   0.6316
## Diesel             4.830e+03  4.619e+02  10.457  < 2e-16 ***
## trans             -4.741e+03  4.965e+02  -9.548  < 2e-16 ***
## five_seater       -1.473e+03  5.722e+02  -2.575   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10940 on 4252 degrees of freedom
## Multiple R-squared:  0.6994, Adjusted R-squared:  0.6988
## F-statistic:  1099 on 9 and 4252 DF,  p-value: < 2.2e-16
```

VIF values.

```
##               Age Kilometers_Driven      Mileage_KMPL         Engine_CC
##          1.640284          1.547594          2.670157          9.511531
##         Power_bhp       first_owner            Diesel             trans
##          6.623844          1.156510          1.885733          1.863804
##       five_seater
##          1.633872
```

Running our first regression model gives us some useful insights about the variables.

Negative coefficient of Engine_CC variable is opposite to the relation we observed in the correaltion matrix.

High Vif values of Engine_CC and Power_bhp further assures that this is caused due to multicollinearity.

Threfore we can drop Engine_CC variable from our model.

Variable first owner also has a high p value which suggests that it isnt a significant variable in the prediction. More over its p value suggests thats it is not a significant variable therefore we can drop it from our model.

Regression model 2

```
##
## Call:
## lm(formula = Price_in_CAD ~ Age + Kilometers_Driven + Mileage_KMPL +
##     Power_bhp + Diesel + trans, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76567  -5049  -1027   3379 205031
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.357e+03  1.466e+03   5.702 1.27e-08 ***
## Age               -1.560e+03  6.427e+01 -24.277  < 2e-16 ***
## Kilometers_Driven -4.433e-02  5.819e-03  -7.618 3.16e-14 ***
## Mileage_KMPL      -2.304e+02  5.117e+01  -4.503 6.88e-06 ***
## Power_bhp          2.260e+02  4.643e+00  48.674  < 2e-16 ***
## Diesel             4.508e+03  3.955e+02  11.396  < 2e-16 ***
## trans             -4.649e+03  4.825e+02  -9.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10950 on 4255 degrees of freedom
## Multiple R-squared:  0.6988, Adjusted R-squared:  0.6983
## F-statistic:  1645 on 6 and 4255 DF,  p-value: < 2.2e-16
```

VIF values

```
##               Age Kilometers_Driven      Mileage_KMPL         Power_bhp
##          1.473456          1.512469          1.782854          2.415512
##            Diesel             trans
##          1.380709          1.757353
```

The second regression model seems to be working well but I would like drop Kilometers driven variable and try again since its has very low coefficient.

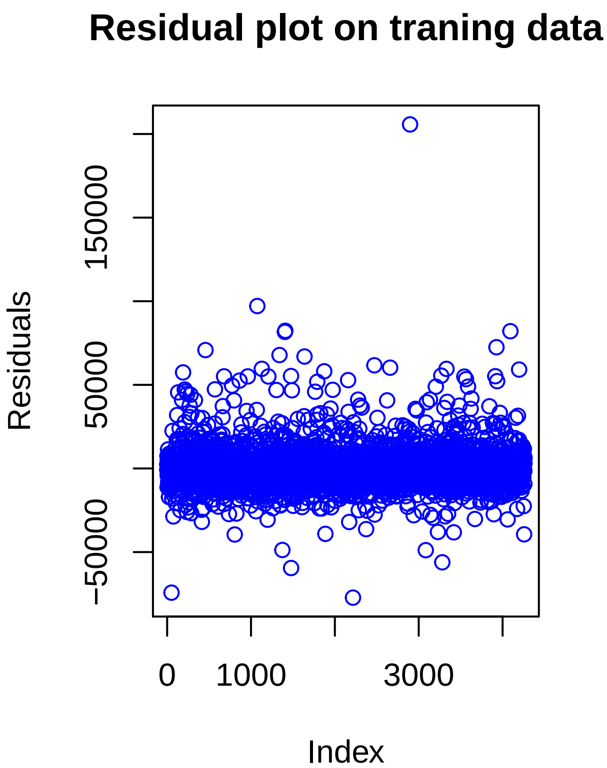
Regression model 3

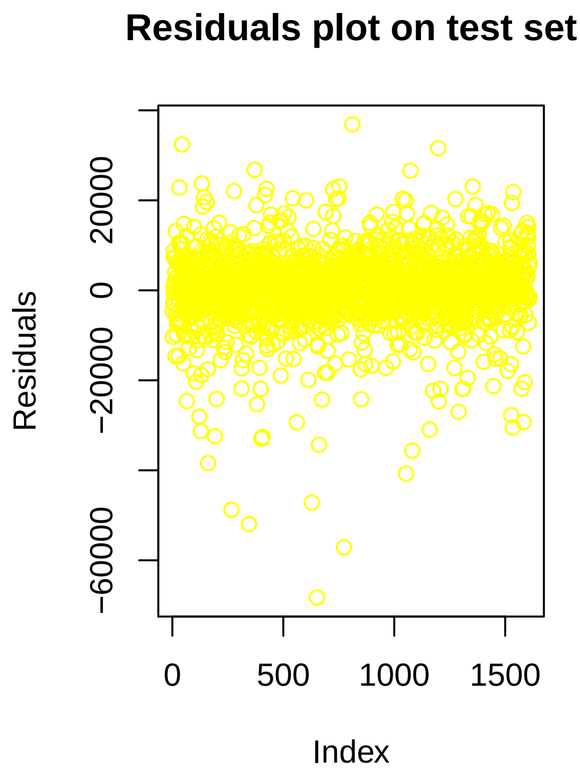
```
##
## Call:
## lm(formula = Price_in_CAD ~ Age + Mileage_KMPL + Power_bhp +
##     Diesel + trans, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77230  -5076   -904   3373 205720
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6819.679   1461.478   4.666 3.16e-06 ***
## Age          -1789.344     57.169 -31.299  < 2e-16 ***
## Mileage_KMPL  -175.762     51.007  -3.446 0.000575 ***
## Power_bhp      229.145      4.656  49.216  < 2e-16 ***
## Diesel        3406.441    370.628   9.191  < 2e-16 ***
## trans        -5025.354    483.120 -10.402  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11020 on 4256 degrees of freedom
## Multiple R-squared:  0.6947, Adjusted R-squared:  0.6943
## F-statistic:  1936 on 5 and 4256 DF,  p-value: < 2.2e-16
```

```
##          Age Mileage_KMPL    Power_bhp       Diesel        trans
##     1.150604     1.747796     2.396506     1.196289     1.738914
```

Regression model 3 seems doing a good job on train data but I would like plot the residuals plot and check to confirm the same.

It explains almost 70% variability in the data.





The residual plot suggests that there isnt any pattern in the residuals and the residuals are spread around zero therefore there arent many improvement we can make on the model.

Final equation

The final equation of price is given as

$y = 5830.232 - 1748.434 * x_1 - 151.227 * x_2 + 232.998 * x_3 + 3207.500 * x_4 - 5053.908 * x_5$

x_1 = Age x_2 = Mileage_KMPL x_3 = Power_bhp x_4 = Diesel x_5 = trans

Conclusions

From our analysis we can conclude the following

- Age, Diesel and transmission type seems to be having maximum impact on the resale value of the car.
- Manual transmission type hinders the resale value of car while Automatic transmission increases the resale value.
- Diesel cars also help the resale value as compared to petrol cars.

Code Appendix

```
library(regclass)
library(readr)
library(dplyr)
library(ggplot2)
library(caTools)

df<-read.csv("usedcar_clean.csv")

head(df,3)

dim(df)

df_numerical<-df %>% select(-c(Brand,Model,Location,Fuel_Type,Transmission,Owner_Type))

library(corrplot)
corrplot(cor(df_numerical), method="square",type = "lower", order="hclust", addCoef.col = "blue",number
```

```
owners_num<-c(nrow(df[df$Owner_Type=="First",]),
     nrow(df[df$Owner_Type=="Second",]),
     nrow(df[df$Owner_Type=="Third",]),
     nrow(df[df$Owner_Type=="Fourth & Above",]))
owners_type <- c("Frist","Second","Third","Fourth & Above")

owners <- cbind(owners_type,owners_num)



fuel_num<-c(nrow(df[df$Fuel_Type=="CNG",]),
     nrow(df[df$Fuel_Type=="Diesel",]),
     nrow(df[df$Fuel_Type=="LPG",]),
     nrow(df[df$Fuel_Type=="Petrol",]))
fuel_type <- levels(as.factor(df$Fuel_Type))

fuel <- cbind(fuel_type,as.numeric(fuel_num))

trans_num<-c(nrow(df[df$Transmission=="Automatic",]),
     nrow(df[df$Transmission=="Manual",]))
trans_type <- levels(as.factor(df$Transmission))

transmission <- cbind(trans_type,as.numeric(trans_num))

seat_num<-c(nrow(df[df$Seats==2,]),
     nrow(df[df$Seats==4,]),
     nrow(df[df$Seats==5,]),
     nrow(df[df$Seats==6,]),
     nrow(df[df$Seats==7,]),
     nrow(df[df$Seats==8,]),
     nrow(df[df$Seats==9,]),
     nrow(df[df$Seats==10,]))
seat_type <- levels(as.factor(df$Seats))

seats <- cbind(as.character(seat_type),as.numeric(seat_num))
```

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

color = c("#cabcd2","#dfabc1","#c123ea","#c453ea")
ggplot(as.data.frame(owners), aes(x = as.factor(owners_type), y = as.numeric(owners_num),fill =color ))
  geom_col()+
  xlab("Owner type")+
  ylab("No of data points")+
  guides(fill = FALSE)+
  geom_text(aes(label = owners_num), vjust = -0.2)
```

```r
color = c("#cfb5d2","#dgcde1","#baeeea","#4adeea")
ggplot(as.data.frame(fuel), aes(x = as.factor(fuel_type), y = as.numeric(fuel_num),fill =color )) +
  geom_col()+
  xlab("Fuel type")+
  ylab("No of data points")+
  guides(fill = FALSE)+
  geom_text(aes(label = fuel_num), vjust = -0.2)
```

```r
color = c("#1fb5d2","#dccde1")
ggplot(as.data.frame(transmission), aes(x = as.factor(trans_type), y = as.numeric(trans_num),fill =trans
  geom_col()+
  xlab("Transmission type")+
  ylab("No of data points")+
  geom_text(aes(label = trans_num), vjust = -0.2)
```

```r
color = c("#cfb556","#dg3de1","#b13eea","#4de13a")
ggplot(as.data.frame(seats), aes(x = as.factor(seat_type), y = as.numeric(seat_num),fill = seat_type))
  geom_col()+
  xlab("Fuel type")+
  ylab("No of data points")+
    geom_text(aes(label = seat_num), vjust = -0.2)
```

```r
    # Encoding categorical variables


df$first_owner <- NA
for (a in 1:nrow(df)){

  if (df$Owner_Type[a] == "First"){
    df$first_owner[a] <- 1
  }else{
    df$first_owner[a] <-0
  }
}

df$Diesel <- NA
for (a in 1:nrow(df)){

  if (df$Fuel_Type[a] == "Diesel"){
    df$Diesel[a] <- 1
  }else{
    df$Diesel[a] <-0
```

```
  }
}

df$trans <- NA
for (a in 1:nrow(df)){

  if (df$Transmission[a] == "Manual"){
    df$trans[a] <- 1
  }else{
    df$trans[a] <-0
  }
}

df$five_seater <- NA
for (a in 1:nrow(df)){
  if (df$Seats[a] == 5){
    df$five_seater[a] <- 1
  }else{
    df$five_seater[a] <-0
  }
}


# Corelation plot of cleaned data

df_cleaned  <-df %>% select(-c(Brand,Model,Location,Fuel_Type,Transmission,Owner_Type,Seats))
corrplot(cor(df_cleaned), method="square",type = "lower", order="hclust", addCoef.col = "black",number.


# Regression 1

split <-  sample.split(df_cleaned$Price_in_CAD, SplitRatio = 0.7)
train <- subset(df_cleaned, split == TRUE)
test <- subset(df_cleaned, split == FALSE)

reg1 <- lm(formula = Price_in_CAD ~ .   , data = train )
summary(reg1)

VIF(reg1)

# Regression model 2

reg2 <- lm(formula = Price_in_CAD ~ Age + Kilometers_Driven + Mileage_KMPL + Power_bhp + Diesel + trans
summary(reg2)
VIF(reg2)


# Regression model 3
reg3 <- lm(formula = Price_in_CAD ~ Age  + Mileage_KMPL + Power_bhp + Diesel + trans, data = train )
summary(reg3)

VIF(reg3)
# residuals plot of prediction on test value and train values of p
```

```r
par(mfrow=c(1,2))
```

```r
plot(reg3$residuals,ylab="Residuals",col = "blue",main= "Residual plot on traning data")

pred <- predict(reg3, newdata = test)


plot((pred-test$Price_in_CAD),main = "Residuals plot on test set",col = "yellow",ylab="Residuals")
```

Data downloaded from the following link

https://www.kaggle.com/gothamv/usedcarprices