DSCI300 Mini Project 6 - Linear Regression

Nisi Mohan Kuniyil 300321388

21/11/2020

Problem Statement

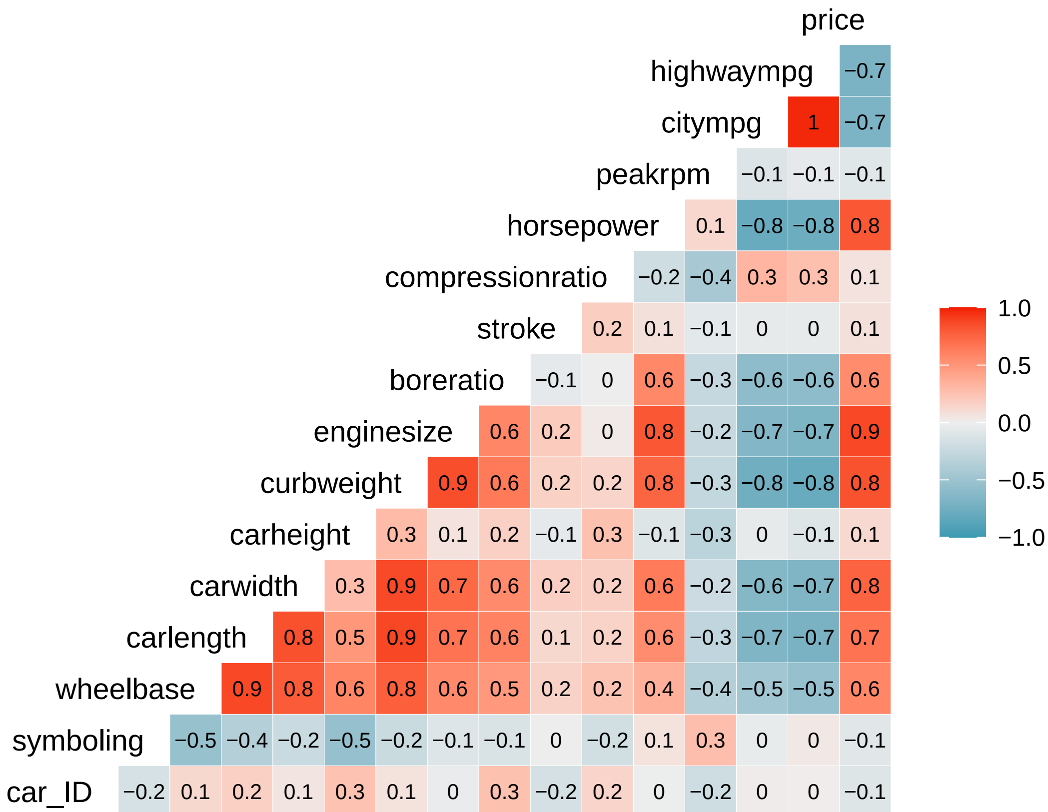
Regression Analysis of car prices in the US Automobile Market

An automobile company is planning to enter the US automotive market. As part of their rigorous market research, they would like to get a thorough understanding of the key factors that drive the price of cars in the US. They collected a decent amount of data about the cars that are currently on the market, recording a diverse set of attributes. Create a regression model that can predict the price of a car from a set of selected features, and also analyze how much these features can explain the variations of car prices in the US.

Solution

Exploratory Data Analysis

We start off by exploring the data variables and see if there is any correlation between them by plotting the correlation matrix. It shows that the features enginesize, curbweight, horsepower, carwidth, wheelbase, and carheight are positively correlated with price, whereas, highwaympg and citympg are negatively correlated with price.



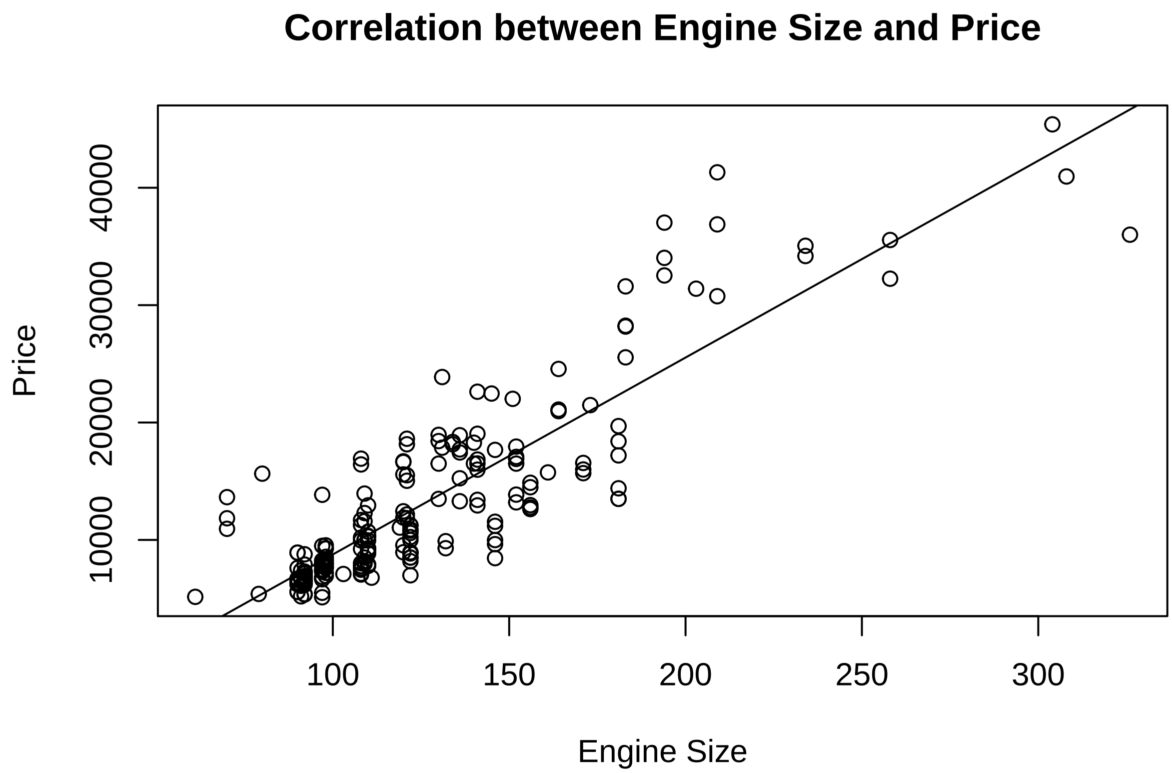
Modeling

The next step is to compute the individual R^2 values for the most promising features we identified from the correlation matrix. These values can be used to interpret the factors that have an influence on the price of a car. We omitted carwidth, carlength, and wheelbase because they all have a strong correlation with curbweight and it makes sense to get rid of the redundant variables.

	rSquared
enginesize	0.764129136
horsepower	0.653088356
peakrpm	0.007270487
citympg	0.470254895
highwaympg	0.486644493
curbweight	0.697734241

The best indicator of the price is the *enginesize* with an $R^2 \approx 0.764$.

R-squared measures the strength of the relationship between your model and the dependent variable on a 0-100% scale. Since there appeared to be a strong relationship between engine size and price of the car, a scatter plot has used to visualize this relationship. The scatter plot shows a positive correlation between engine size and price. Also, it can be seen that the data points are closer to the regression line.



Next, We develop a multiple regression equation using *enginesize*, *horsepower*, *peakrpm*, *citympg*, and *highwaympg* to estimate price, and check how well the regression model explains the variability in price compared to the individual R^2.

$$price = \ 97.75x_1 + 29.10x_2 + 2.10x_3 - 38.11x_4 + 120.3x_5 + 5.93x_6 - 30810$$

where x_1 represents engine size, x_2 represents *horsepower*, x_3 represents *peakrpm*, x_4 represents *citympg*, x_5 represents *highwaympg*, x_6 curb weight.

The R^2 value for this multiple regression model is 0.823. When a regression model accounts for more of the variance, the data points are closer to the regression line. When we used a single variable alone for the prediction the highest R^2 the value that we got was 0.764, whereas the multiple regression gives us better R^2.

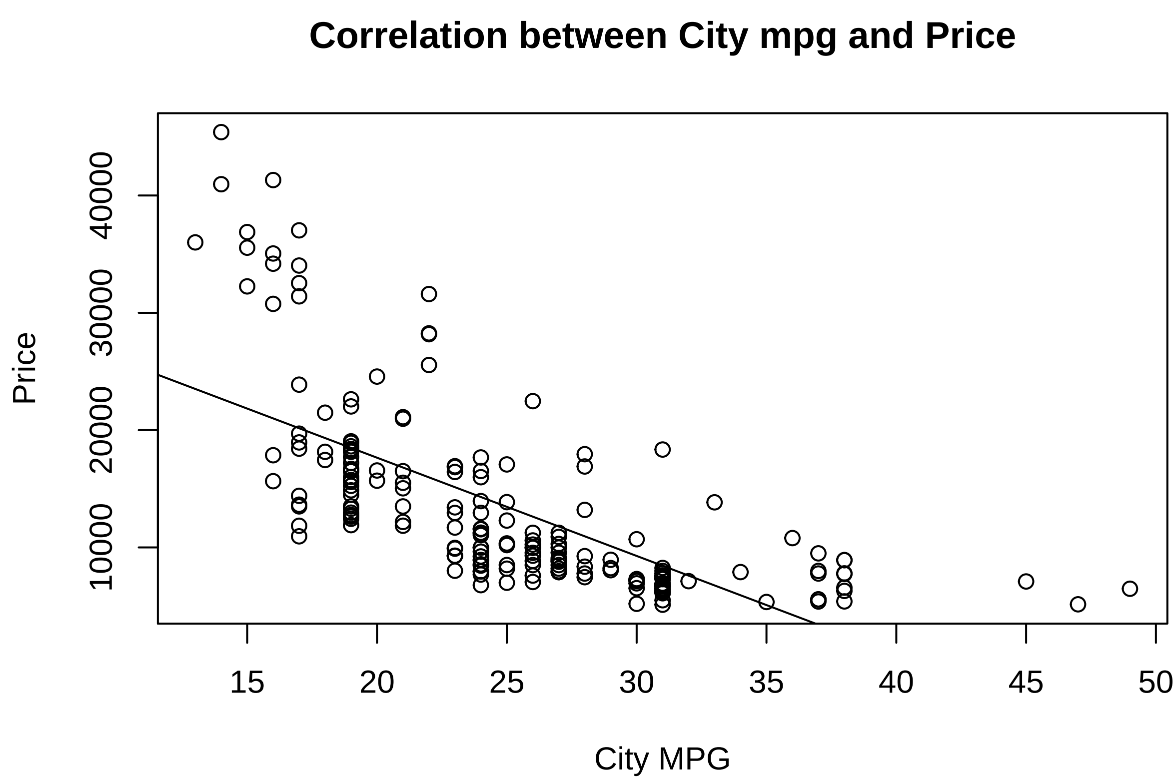
```
##
## Call:
## lm(formula = CarPrice_df$price ~ CarPrice_df$enginesize + CarPrice_df$horsepower +
##     CarPrice_df$peakrpm + CarPrice_df$citympg + CarPrice_df$highwaympg +
##     CarPrice_df$curbweight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9364.1 -1675.3    -7.9  1331.4 12985.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -3.081e+04  6.264e+03  -4.918 1.84e-06 ***
## CarPrice_df$enginesize  9.775e+01  1.410e+01   6.934 5.66e-11 ***
```

```
## CarPrice_df$horsepower  2.910e+01  1.532e+01   1.899  0.05896 .
## CarPrice_df$peakrpm     2.101e+00  6.786e-01   3.096  0.00225 **
## CarPrice_df$citympg    -3.811e+01  1.742e+02  -0.219  0.82701
## CarPrice_df$highwaympg  1.203e+02  1.649e+02   0.730  0.46643
## CarPrice_df$curbweight  5.926e+00  1.154e+00   5.135 6.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3414 on 198 degrees of freedom
## Multiple R-squared:  0.8228, Adjusted R-squared:  0.8174
## F-statistic: 153.2 on 6 and 198 DF,  p-value: < 2.2e-16
```

The coefficient t-value for engine size, horse power and curb weight are far away from zero as this would indicate we could easily reject the null hypothesis, that is we could declare a relationship between these variables and the price of the car. We do not have to consider the variables which have t-value closer to zero for predicting the price. Since these variables do not indicate any stronger relationship in predicting the price. From the above summary, it is clear that enginesize and curbweight are both significant at α 0.001 and 0.01 respectively.

It appears to be there is a negative correlation between city mpg and price from the correlation matrix plotted above. Furthermore, the linear model also shows that city mpg has a negative impact on the prediction of price.

To interpret this more easily, a scatter plot of price against city mpg is plotted. It appears to be there is a negative relationship between city mpg and price. Also, the observations are not fitted closer to the regression line.



The above model gives a clearer image of which factors affect the pricing of cars. From the model, we could say that some variables are not significant for the prediction of the car price. It is surprising to see that horsepower does not have much significance in the prediction of the car price. Generally, expensive supercars tend to have engines with high horsepower.

In order to understand the effect of citympg, highwaympg and horsepower, we developed another model removing those variables and observed how R^2 varies.

The second model regression equation is :

$$price = \ 11.56x_1 + 2.67x_2 + 5.606x_3 - 29400$$

where x_1 is the *enginesize*, x_2 is the *peakrpm*, and x_3 is the *curbweight.* All these three variables are significant at $\alpha = 0.001$ or 99.9% confidence level. Also, the R^2 value is 0.818 which means 81.8% of the variability in car price is explained by this model. The previous model has an R^2 of 0.823, which is not that significantly different from the second model R^2 0.818.

Since these two models have closer R-squared value, It is better to use the second model for the car price prediction because it is a simpler one.

Conclusion

From the analysis, we found out that the variables that have a major impact on the variances in car prices are enginesize, curbweight, and peakrpm. The R^2 of the model is high, with 81.8%. This model is significant at $\alpha = 0.001$ or 99.9% level.

Appendix 1: The Problem

Regression Analysis of car prices in the US Automobile Market

An automobile company is looking to enter the US automotive market. They are conducting market research to figure out

1. What are the key variables that affect the price of cars in the US.
2. How well those variables describe the price of a car.

Managerial Report

1. Since there are a lot of variables in the CarPrice dataset it is better to understand the relationship between each variable and car price. First, identify the most correlated features by using a correlation matrix.

2. From the correlation matrix take the variables which have a closer correlation with the price and use these variables to predict the car price. Here curb weight, engine size, horsepower, peak rpm, city mpg, highway mpg appear to have a relationship with price. Use only single variables and predict which of these five factors provides the best single predictor of the car price?

3. Develop an estimated regression equation that can be used to predict car price given the engine size(enginesize), horsepower, peak rpm(peakrpm), city miles per gallon(citympg), and highway miles per gallon(highwaympg). Test for individual significance and discuss your findings and conclusion.

4. From the multiple regression above if any variable does not have much significance remove those and predict car price using the rest of the variables. Based on the results of your analysis, what estimated regression equation would you recommend using to predict price? Provide an interpretation of the regression coefficients for this equation.

Appendix 2: Analysis

Read in the dataset.

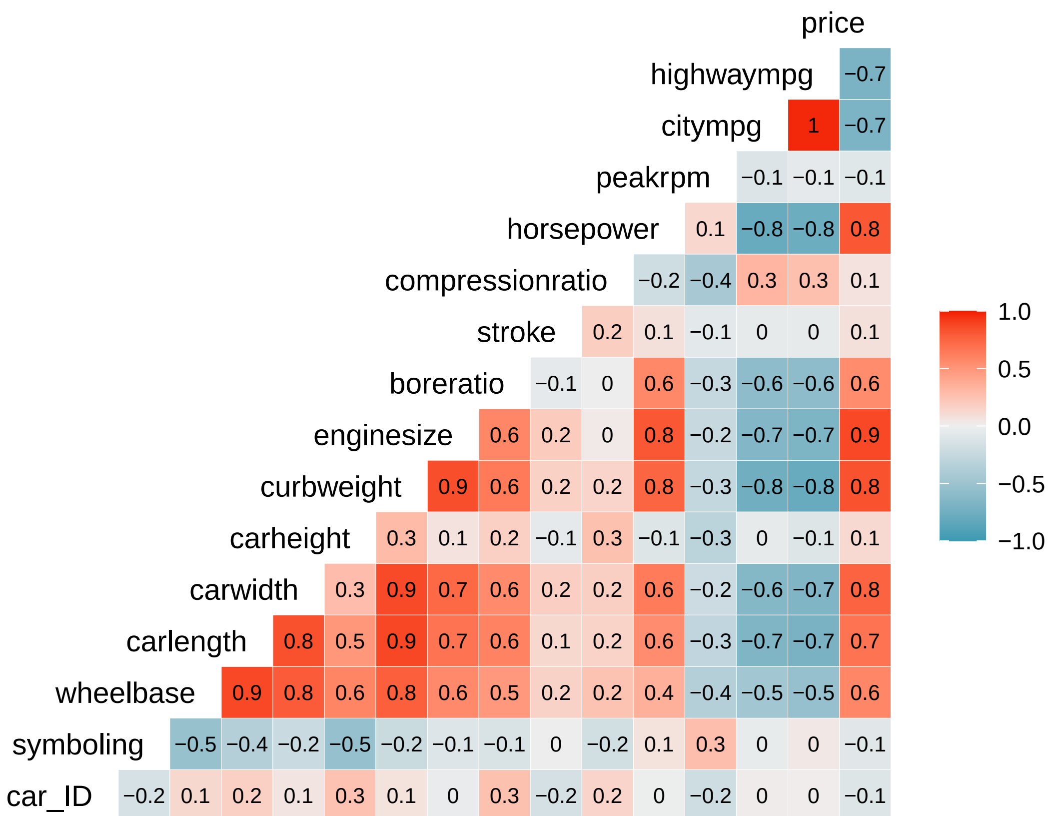
```
CarPrice <- read.csv("CarPrice.csv")
head(CarPrice)
```

```
##   car_ID symboling                  CarName fueltype aspiration doornumber
## 1      1         3       alfa-romero giulia      gas        std        two
## 2      2         3      alfa-romero stelvio      gas        std        two
## 3      3         1 alfa-romero Quadrifoglio      gas        std        two
## 4      4         2              audi 100 ls      gas        std       four
## 5      5         2               audi 100ls      gas        std       four
## 6      6         2                 audi fox      gas        std        two
##       carbody drivewheel enginelocation wheelbase carlength carwidth carheight
## 1 convertible        rwd          front      88.6     168.8     64.1      48.8
## 2 convertible        rwd          front      88.6     168.8     64.1      48.8
## 3   hatchback        rwd          front      94.5     171.2     65.5      52.4
## 4       sedan        fwd          front      99.8     176.6     66.2      54.3
## 5       sedan        4wd          front      99.4     176.6     66.4      54.3
## 6       sedan        fwd          front      99.8     177.3     66.3      53.1
##   curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke
## 1       2548       dohc           four        130       mpfi      3.47   2.68
## 2       2548       dohc           four        130       mpfi      3.47   2.68
## 3       2823       ohcv            six        152       mpfi      2.68   3.47
## 4       2337        ohc           four        109       mpfi      3.19   3.40
## 5       2824        ohc           five        136       mpfi      3.19   3.40
## 6       2507        ohc           five        136       mpfi      3.19   3.40
##   compressionratio horsepower peakrpm citympg highwaympg price
## 1              9.0        111    5000      21         27 13495
## 2              9.0        111    5000      21         27 16500
## 3              9.0        154    5000      19         26 16500
## 4             10.0        102    5500      24         30 13950
## 5              8.0        115    5500      18         22 17450
## 6              8.5        110    5500      19         25 15250
```

Question 1

Develop a correlation matrix to check the relationship exists between each variables and price.

```
ggcorr(CarPrice, label = TRUE, label_size = 2.9, hjust = 1, layout.exp = 2)
```



we will be taking only 7 columns from this to predict car price.

```
CarPrice_df <- CarPrice[,c("CarName","curbweight","enginesize","horsepower","peakrpm","citympg","highwa

head(CarPrice_df)
```

```
##                       CarName curbweight enginesize horsepower peakrpm citympg
## 1        alfa-romero giulia       2548        130        111    5000      21
## 2       alfa-romero stelvio       2548        130        111    5000      21
## 3 alfa-romero Quadrifoglio       2823        152        154    5000      19
## 4               audi 100 ls       2337        109        102    5500      24
## 5                audi 100ls       2824        136        115    5500      18
## 6                  audi fox       2507        136        110    5500      19
##   highwaympg price
## 1         27 13495
## 2         27 16500
## 3         26 16500
## 4         30 13950
## 5         22 17450
## 6         25 15250
```

Question 2

Running linear regression on single variables to see which variable is the best predictor of car price.

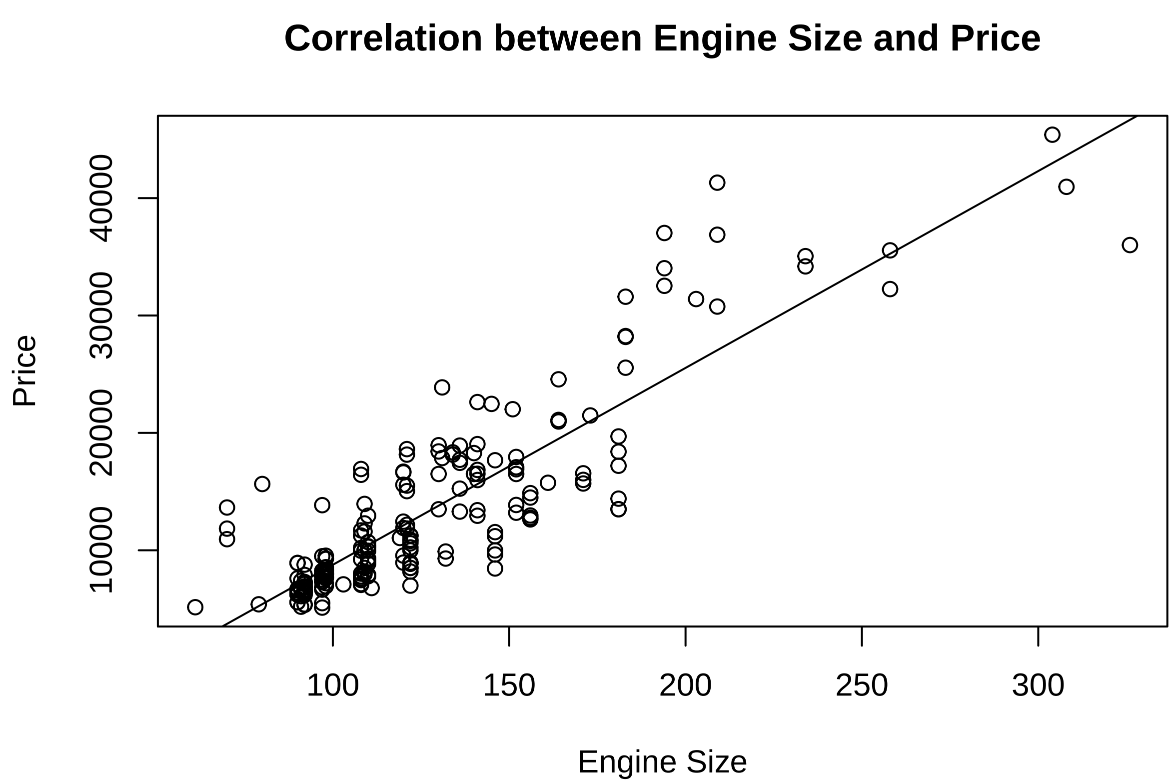
```
rSquared <- c(
enginesize = summary(lm(CarPrice_df$price ~ CarPrice_df$enginesize))$r.squared,
```

```r
horsepower = summary(lm(CarPrice_df$price ~ CarPrice_df$horsepower))$r.squared,
peakrpm = summary(lm(CarPrice_df$price ~ CarPrice_df$peakrpm))$r.squared,
citympg = summary(lm(CarPrice_df$price ~ CarPrice_df$citympg))$r.squared,
highwaympg = summary(lm(CarPrice_df$price ~ CarPrice_df$highwaympg))$r.squared,
curbweight = summary(lm(CarPrice_df$price ~ CarPrice_df$curbweight))$r.squared)
cbind(rSquared)
```

```
##               rSquared
## enginesize 0.764129136
## horsepower 0.653088356
## peakrpm    0.007270487
## citympg    0.470254895
## highwaympg 0.486644493
## curbweight 0.697734241
```

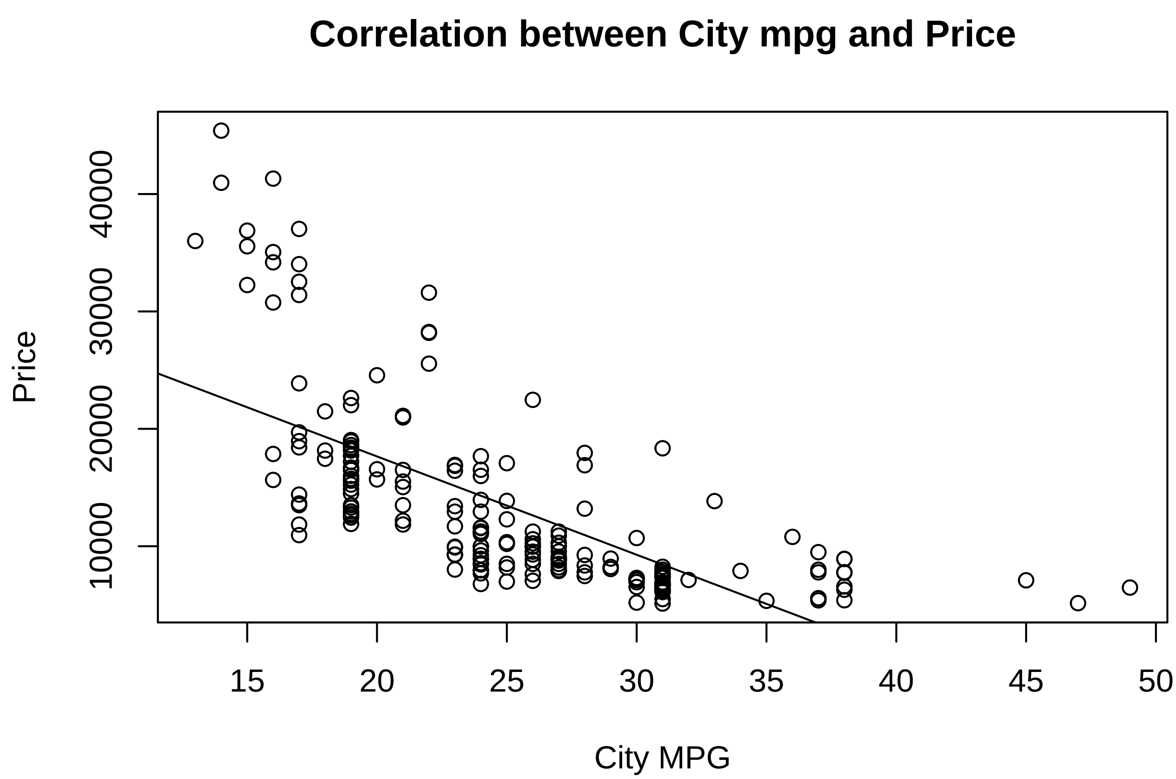
The best indicator of car price prediction is Engine size with R2 approx 0.764

```r
plot(CarPrice_df$enginesize, CarPrice_df$price,
     xlab = "Engine Size",
     ylab = "Price",
     main = "Correlation between Engine Size and Price"
     )

abline(lm(CarPrice_df$price ~ CarPrice_df$enginesize))
```



There may be a positive relationship between price and engine size.

```r
plot(CarPrice_df$citympg, CarPrice_df$price,
     xlab = "City MPG",
     ylab = "Price",
     main = "Correlation between City mpg and Price"
     )

abline(lm(CarPrice_df$price ~ CarPrice_df$citympg))
```



Question 3

Multiple linear regression using enginesize, horsepower, peakrpm, citympg, and highwaympg to estimate price.

```r
Price_lm <- lm(CarPrice_df$price ~ CarPrice_df$enginesize + CarPrice_df$horsepower +
CarPrice_df$peakrpm + CarPrice_df$citympg+ CarPrice_df$highwaympg + CarPrice_df$curbweight)
summary(Price_lm)
```

```
##
## Call:
## lm(formula = CarPrice_df$price ~ CarPrice_df$enginesize + CarPrice_df$horsepower +
##     CarPrice_df$peakrpm + CarPrice_df$citympg + CarPrice_df$highwaympg +
##     CarPrice_df$curbweight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9364.1 -1675.3    -7.9  1331.4 12985.9
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -3.081e+04  6.264e+03  -4.918 1.84e-06 ***
## CarPrice_df$enginesize   9.775e+01  1.410e+01   6.934 5.66e-11 ***
## CarPrice_df$horsepower   2.910e+01  1.532e+01   1.899  0.05896 .  
## CarPrice_df$peakrpm      2.101e+00  6.786e-01   3.096  0.00225 ** 
## CarPrice_df$citympg     -3.811e+01  1.742e+02  -0.219  0.82701    
## CarPrice_df$highwaympg   1.203e+02  1.649e+02   0.730  0.46643    
## CarPrice_df$curbweight   5.926e+00  1.154e+00   5.135 6.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3414 on 198 degrees of freedom
## Multiple R-squared:  0.8228, Adjusted R-squared:  0.8174 
## F-statistic: 153.2 on 6 and 198 DF,  p-value: < 2.2e-16
```

Question 4

After analyzing the first regression model, remove the variables that are not significant. Conduct another regression by using rest of the variables.

```r
Price_lmNew <- lm(CarPrice_df$price ~ CarPrice_df$enginesize  + CarPrice_df$peakrpm+ CarPrice_df$curbwei
summary(Price_lmNew)
```

```
##
## Call:
## lm(formula = CarPrice_df$price ~ CarPrice_df$enginesize + CarPrice_df$peakrpm + 
##     CarPrice_df$curbweight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9479.3 -1781.2    40.1  1324.6 13196.4 
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -2.940e+04  3.262e+03  -9.012  < 2e-16 ***
## CarPrice_df$enginesize   1.156e+02  1.098e+01  10.523  < 2e-16 ***
## CarPrice_df$peakrpm      2.670e+00  5.230e-01   5.104 7.67e-07 ***
## CarPrice_df$curbweight   5.606e+00  8.834e-01   6.346 1.44e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3432 on 201 degrees of freedom
## Multiple R-squared:  0.8182, Adjusted R-squared:  0.8154 
## F-statistic: 301.4 on 3 and 201 DF,  p-value: < 2.2e-16
```