STATISTICAL DATA MINING

ISM6137.901S17

PROJECT REPORT ON

Car Trends in www.carfax.com

AUTHORS:

Jijo Johny

Deepesh Kumar

Pradeep Raj Ooralath

Majed Alghamdi

Contents

Introduction

The aim of this project is to analyze the trends of car prices in www.carfax.com and to predict the variation of the prices based on their make, model, year of manufacture, mileage run by the car, type of engine, fuel type, exterior color and interior color. The data was obtained by scrapping the web site www.carfax.com. The scrapping was done by using a chrome extension "Web Scraper". The data obtained was then cleaned using R and then various models were created in order to analyze the trend of the price and also to predict the price on the most influencing variable. Many regression models were created on the data and the models were compared on R squared value, AIC, BIC and residual plots of each model. We have considered only sedans belonging to the segment. The data in Carfax related to the below cars were extracted from the website:

1. Honda Accord
2. Toyota Camry
3. Honda Civic
4. Hyundai Elantra
5. Ford Fusion
6. Kia Optima
7. Chevrolet Malibu
8. Nissan Altima
9. Ford Mustang

About www.carfax.com

Carfax, Inc. is a commercial web-based service that supplies vehicle history reports to individuals and businesses on used cars and light trucks for the American and Canadian consumers. CARFAX started with a vision - to be the leading source of vehicle history information for buyers and sellers of used cars. Today, CARFAX has the most comprehensive vehicle history database available in North America. Millions of consumers trust CARFAX to provide them with vehicle history information every year.

CARFAX receives information from more than 100,000 data sources including every U.S. and Canadian provincial motor vehicle agency plus many auto auctions, fire and police departments, collision repair facilities, fleet management and rental agencies, and more.

CARFAX Vehicle History Reports™ are available on all used cars and light trucks model year 1981 or later. Using the unique 17-character vehicle identification number (VIN), a CARFAX Report is instantly generated from their database of over 17 billion records.

Every CARFAX Report contains information that can impact a consumer's decision about a used vehicle. Some types of information that a CARFAX Report may include are:

Title information, including salvaged or junked titles

Flood damage history

Total loss accident history

Odometer readings

Lemon history

Number of owners

Accident indicators, such as airbag deployments

State emissions inspection results

Service records

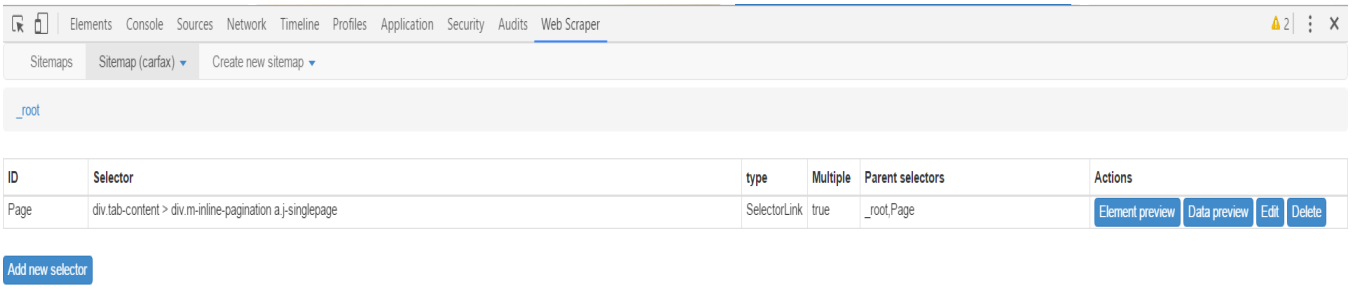
Vehicle use (taxi, rental, lease, etc.)

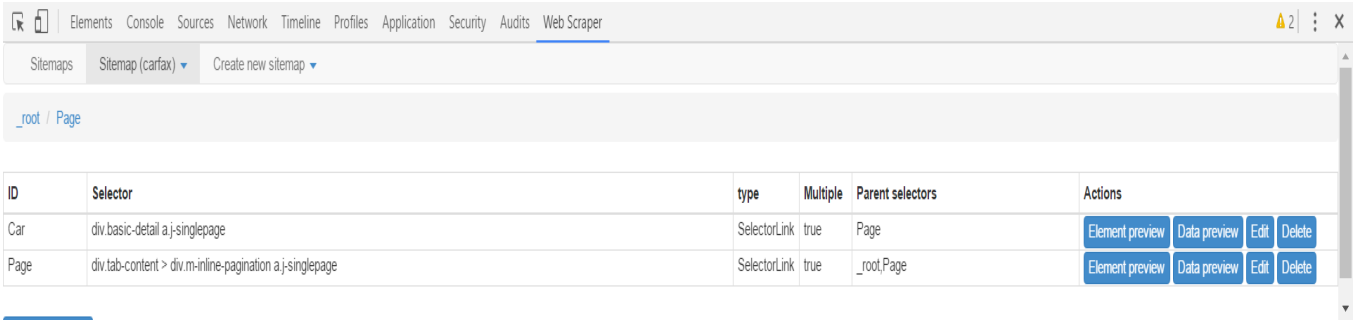
Web Scraping

Web scraping was done with the chrome extension “Web Scraper”. The following information were extracted from the website for each car:

1. Make
2. Model
3. Year
4. Mileage
5. Price
6. Engine
7. Transmission
8. Fuel type
9. MPG
10. Exterior color
11. Interior color

Web Scraping using “Web Scraper” Chrome extension:





Car Trends in www.carfax.com

Elements Console Sources Network Timeline Profiles Application Security Audits Web Scraper

Sitemaps | Sitemap (carfax) | Create new sitemap

_root / Page / Page

ID	Selector	type	Multiple	Parent selectors	Actions
Car	div.basic-detail a.j-singlepage	SelectorLink	true	Page	Element preview Data preview Edit Delete
Page	div.tab-content > div.m-inline-pagination a.j-singlepage	SelectorLink	true	_root,Page	Element preview Data preview Edit Delete

Elements Console Sources Network Timeline Profiles Application Security Audits Web Scraper

Sitemaps | Sitemap (carfax) | Create new sitemap

_root / Page / Car

ID	Selector	type	Multiple	Parent selectors	Actions
Item	div.vehicle-information	SelectorElement	false	Car	Element preview Data preview Edit Delete

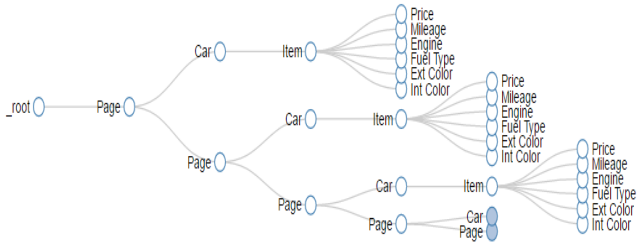
Add new selector

Elements Console Sources Network Timeline Profiles Application Security Audits Web Scraper

ID	Selector	type	Multiple	Parent selectors	Actions
Price	li:nth-of-type(1)	SelectorText	false	Item	Element preview Data preview Edit Delete
Mileage	li:nth-of-type(2)	SelectorText	false	Item	Element preview Data preview Edit Delete
Engine	li:nth-of-type(4)	SelectorText	false	Item	Element preview Data preview Edit Delete
Fuel Type	li:nth-of-type(7)	SelectorText	false	Item	Element preview Data preview Edit Delete
Ext Color	li:nth-of-type(9)	SelectorText	false	Item	Element preview Data preview Edit Delete
Int Color	li:nth-of-type(10)	SelectorText	false	Item	Element preview Data preview Edit Delete

Elements Console Sources Network Timeline Profiles Application Security Audits Web Scraper

Sitemaps | Sitemap (carfax) | Create new sitemap



Dataset

Source of Data: www.carfax.com

Number of Variables: 12

Total number of instances: 2108

Variable	Type	Values
Make	Categorical	Honda, Toyota, Hyundai, Ford, etc
Model	Categorical	Accord, Civic, Camry, Mustang, etc
Year	Categorical	
Price	Numeric	
Mileage	Numeric	
Engine	Categorical	4,6 and 8 cylinder
Transmission	Categorical	Manual and automatic
Fuel type	Categorical	Gasoline, hybrid and flexible fuel type
MPG_CY	Numeric	MPG in cities
MPG_HWY	Numeric	MPG in highways
Ext Color	Categorical	Exterior color
Int Color	Categorical	Interior color

Raw data:

```
x<- read.csv("car_trends-master/Accord.csv")
head(x[1:6])
```

```
##                    Car          Price          Mileage         Engine
## 1  2012 HONDA ACCORD SE Price: $14,995  Mileage: 29,740 Engine:  4 Cyl
## 2  2008 HONDA ACCORD EX  Price: $9,500  Mileage: 85,912 Engine:  4 Cyl
## 3 2013 HONDA ACCORD EXL Price: $13,800  Mileage: 88,246 Engine:  4 Cyl
## 4  2005 HONDA ACCORD EX  Price: $4,995 Mileage: 146,455 Engine:  4 Cyl
## 5 2013 HONDA ACCORD EXL Price: $16,998  Mileage: 61,903 Engine:  6 Cyl
## 6 2012 HONDA ACCORD EXL Price: $14,987  Mileage: 82,202 Engine:  6 Cyl
##               Transmission            FuelType
## 1 Transmission:  Automatic Fuel Type: Gasoline
## 2 Transmission:  Automatic Fuel Type: Gasoline
## 3 Transmission:  Automatic Fuel Type: Gasoline
## 4 Transmission:  Automatic Fuel Type: Gasoline
## 5 Transmission:  Automatic Fuel Type: Gasoline
## 6 Transmission:  Automatic Fuel Type: Gasoline
```

```
#x<- read.csv("car_trends-master/Accord.csv")
head(x[7:9])
```

```
##                                          MPG
## 1 MPG City/Hwy:\n            \n            22/33
## 2 MPG City/Hwy:\n            \n            21/31
## 3 MPG City/Hwy:\n            \n            26/34
## 4 MPG City/Hwy:\n            \n            21/31
## 5 MPG City/Hwy:\n            \n            21/31
## 6 MPG City/Hwy:\n            \n            19/29
##                 ExtColor              IntColor
## 1  Exterior Color:  Black   Interior Color:  Tan
## 2    Exterior Color:  Red    Interior Color: N/A
## 3 Exterior Color:  Silver Interior Color:  Black
## 4  Exterior Color:  Black   Interior Color:  Tan
## 5   Exterior Color:  Gray  Interior Color:  Gray
## 6    Exterior Color:  Red Interior Color:  Black
```

Cleaned Data:

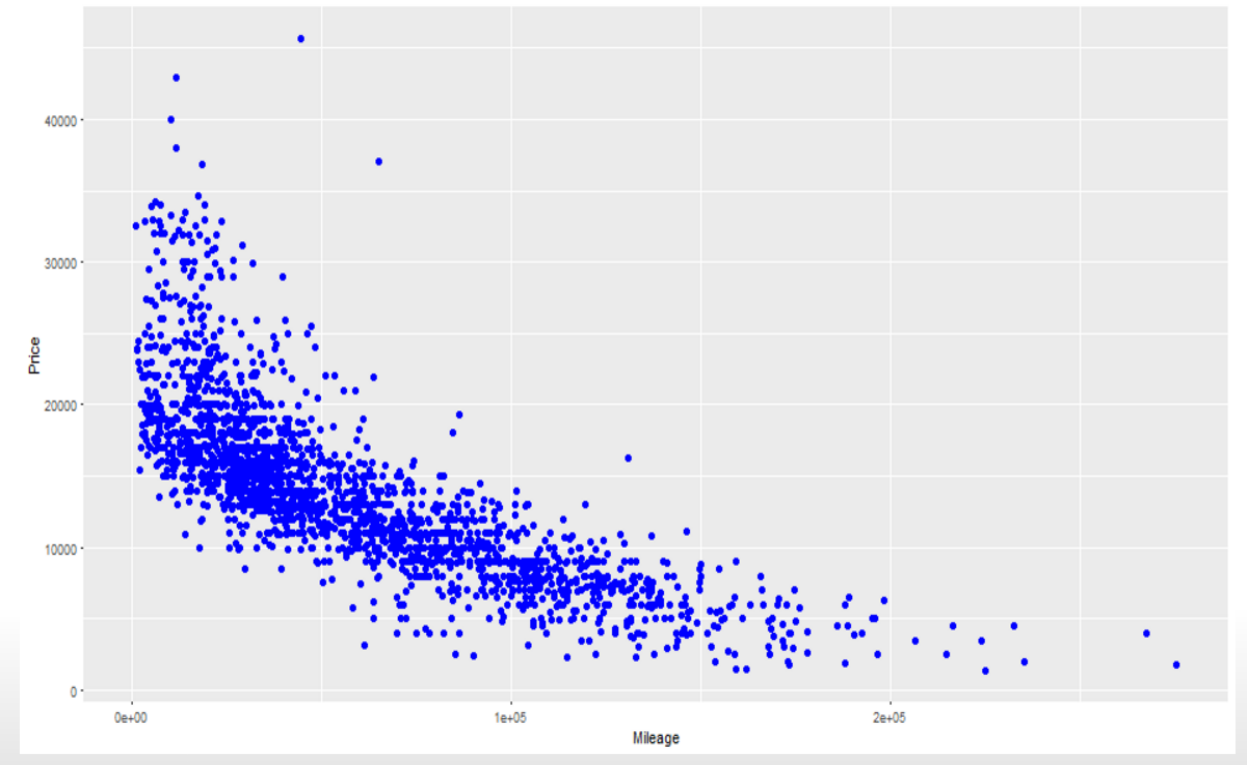
```
head(d)
```

```
##    Year  Make    Car Price Mileage Engine Transmission FuelType MPG_CY
## 1: 2012 HONDA ACCORD 14995   29740      4    Automatic Gasoline     22
## 2: 2008 HONDA ACCORD  9500   85912      4    Automatic Gasoline     21
## 3: 2013 HONDA ACCORD 13800   88246      4    Automatic Gasoline     26
## 4: 2005 HONDA ACCORD  4995  146455      4    Automatic Gasoline     21
## 5: 2013 HONDA ACCORD 16998   61903      6    Automatic Gasoline     21
## 6: 2012 HONDA ACCORD 14987   82202      6    Automatic Gasoline     19
##    MPG_HWY ExtColor IntColor
## 1:      33    Black      Tan
## 2:      31      Red      N/A
## 3:      34   Silver    Black
## 4:      31    Black      Tan
## 5:      31     Gray     Gray
## 6:      29      Red    Black
```

Relationships between variables

Price vs Mileage

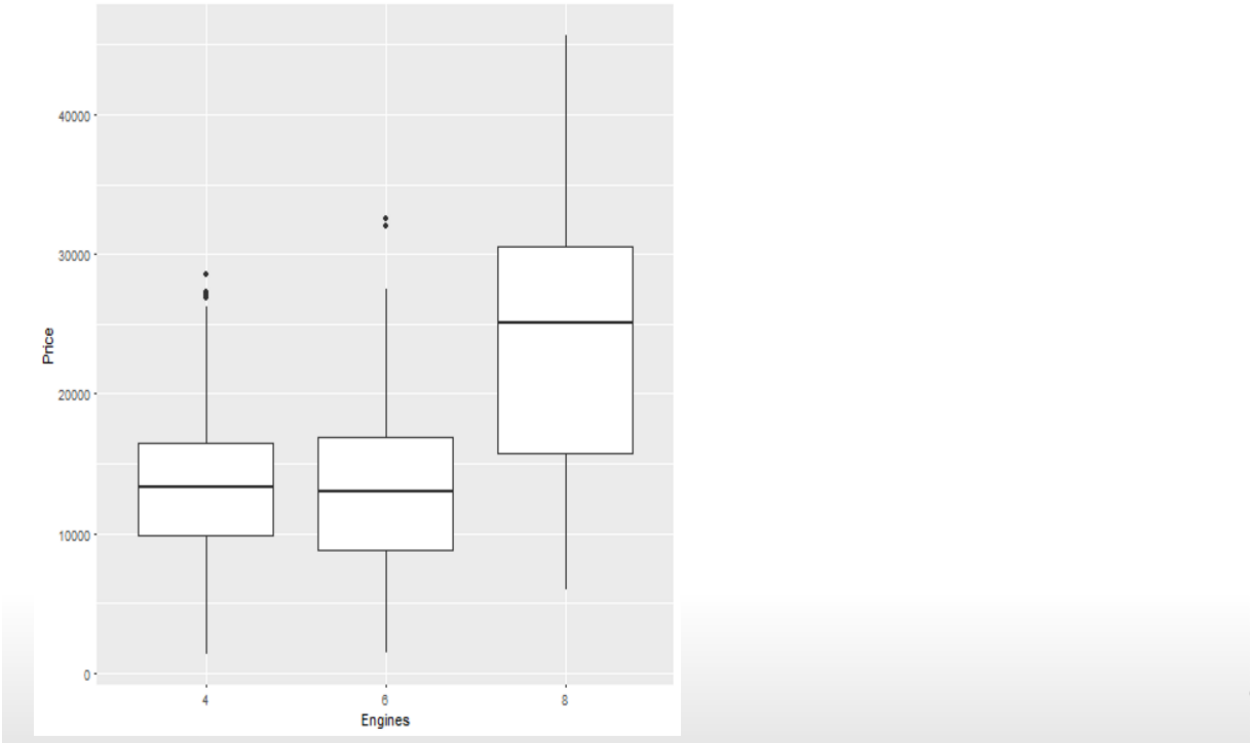
```
ggplot(d,aes(x=Mileage,y=Price))+ geom_point(size=2,color="blue",na.rm = TRUE)
```



Price and mileage have a negative relationship.

Price vs Engine

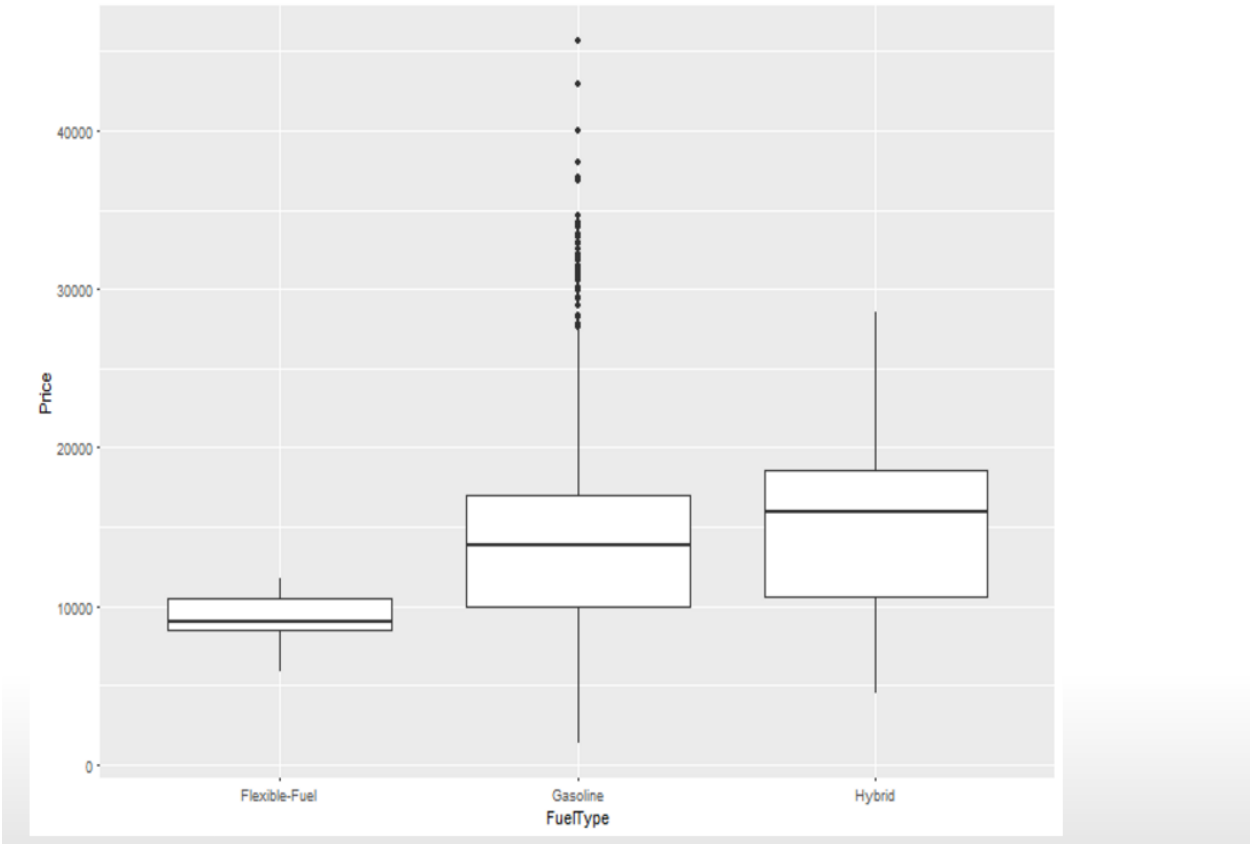
```
ggplot(d, aes(x=factor(Engine),y=Price))+ geom_boxplot(na.rm = TRUE) + xlab(label = "Engines")
```



1. 8 cylinder engines are generally more expensive.
2. 4 cylinder and 6 cylinder engines have similar price mostly.

Price vs Fuel-type

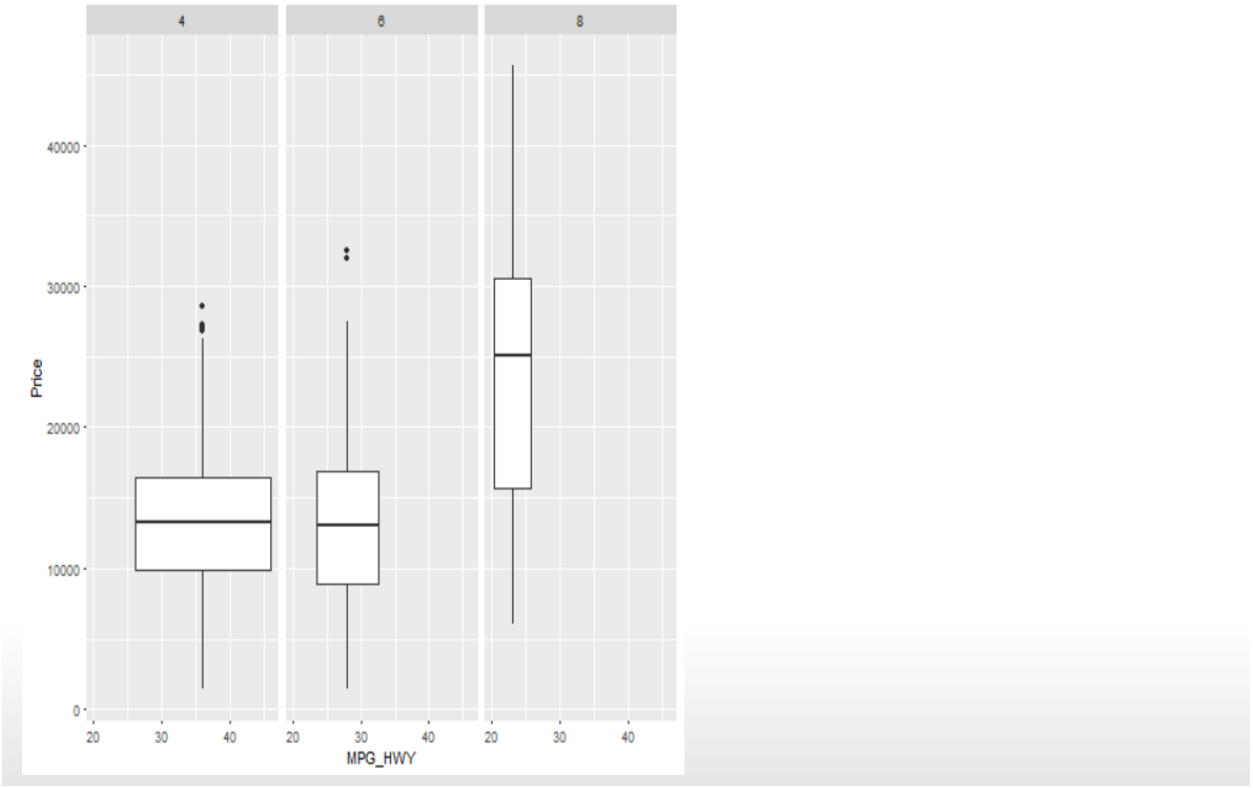
```
ggplot(d, aes(x=FuelType,y=Price))+ geom_boxplot(na.rm = TRUE)
```



Flexible fuel type cars are generally less expensive than hybrid and gasoline cars.

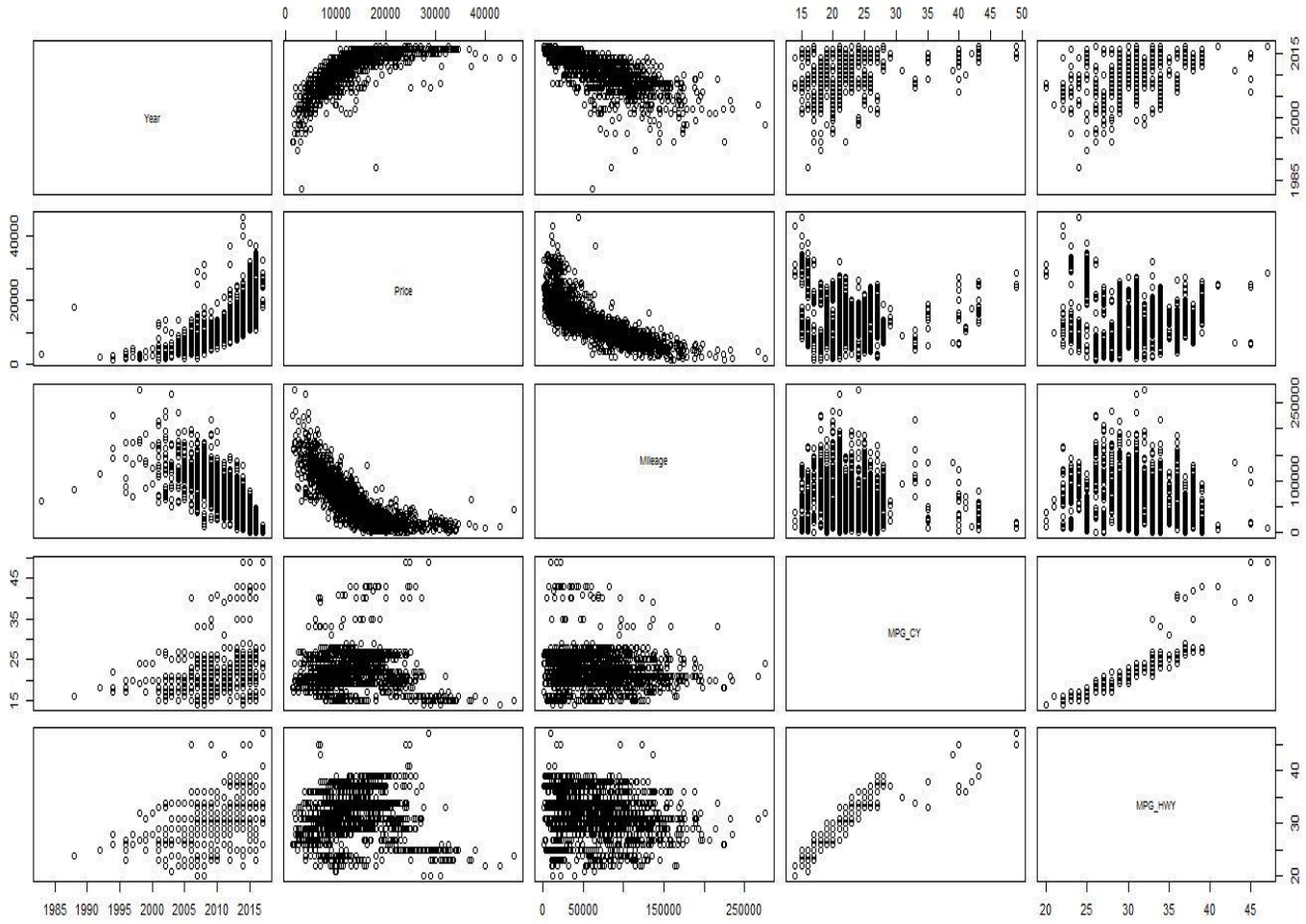
Price vs MPG_HWY conditioned on Engine

```
ggplot(d,aes(x=MPG_HWY,y=Price))+ geom_boxplot(na.rm = TRUE)+facet_grid(.~Engine)
```



MPG_HWY is lower for 8 cylinder cars but the price is generally higher because they are more powerful.

Correlation plot of all variables



Regression Models

Model 1

```
> mod1 <- lm(Price~Mileage,d)
> summary(mod1)

Call:
lm(formula = Price ~ Mileage, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-10431.0  -2294.2   -687.5   1461.3  30239.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.020e+04  1.401e+02  144.24   <2e-16 ***
Mileage     -1.082e-01  1.959e-03  -55.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3817 on 2102 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5918,    Adjusted R-squared:  0.5917
F-statistic:  3048 on 1 and 2102 DF,  p-value: < 2.2e-16
```

Model 2

```
> mod2 <- lm(Price~factor(Year)+Mileage,d)
> summary(mod2)

Call:
lm(formula = Price ~ factor(Year) + Mileage, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-9961.3 -1981.7   -516.1   1184.3 30359.2

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.111e+03  3.480e+03   2.044 0.041126 *
factor(Year)1988   1.634e+04  4.914e+03   3.326 0.000897 ***
factor(Year)1992   2.560e+03  4.916e+03   0.521 0.602556
factor(Year)1994   6.295e+03  4.029e+03   1.563 0.118304
factor(Year)1996   3.590e+03  3.810e+03   0.942 0.346171
factor(Year)1997   4.785e+03  4.017e+03   1.191 0.233766
factor(Year)1998   8.077e+03  3.907e+03   2.067 0.038826 *
factor(Year)1999   4.082e+03  4.016e+03   1.016 0.309560
factor(Year)2000   5.495e+03  4.018e+03   1.368 0.171505
factor(Year)2001   6.615e+03  3.595e+03   1.840 0.065936 .
factor(Year)2002   6.525e+03  3.567e+03   1.829 0.067525 .
factor(Year)2003   6.764e+03  3.587e+03   1.886 0.059489 .
factor(Year)2004   7.059e+03  3.563e+03   1.982 0.047667 *
factor(Year)2005   6.742e+03  3.517e+03   1.917 0.055386 .
factor(Year)2006   7.336e+03  3.512e+03   2.089 0.036865 *
factor(Year)2007   8.247e+03  3.498e+03   2.358 0.018488 *
factor(Year)2008   8.518e+03  3.494e+03   2.438 0.014843 *
factor(Year)2009   8.204e+03  3.496e+03   2.347 0.019025 *
factor(Year)2010   8.023e+03  3.491e+03   2.298 0.021659 *
factor(Year)2011   8.651e+03  3.495e+03   2.475 0.013385 *
factor(Year)2012   9.549e+03  3.485e+03   2.740 0.006196 **
factor(Year)2013   1.013e+04  3.482e+03   2.910 0.003658 **
factor(Year)2014   1.105e+04  3.480e+03   3.175 0.001522 **
factor(Year)2015   1.202e+04  3.481e+03   3.452 0.000567 ***
factor(Year)2016   1.464e+04  3.482e+03   4.205 2.72e-05 ***
factor(Year)2017   1.753e+04  3.620e+03   4.842 1.38e-06 ***
Mileage           -6.470e-02  3.191e-03 -20.275  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3474 on 2077 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.6658,    Adjusted R-squared:  0.6616
F-statistic: 159.2 on 26 and 2077 DF,  p-value: < 2.2e-16
```

Model 3

```
>
> mod3 <- lm(Price~factor(Year)+Mileage+FuelType,d)
> summary(mod3)

Call:
lm(formula = Price ~ factor(Year) + Mileage + FuelType, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-9941.3 -1971.3  -517.5  1211.4 30390.3

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.952e+03  3.581e+03   1.662 0.096656 .
factor(Year)1988    1.634e+04  4.905e+03   3.331 0.000879 ***
factor(Year)1992    2.555e+03  4.907e+03   0.521 0.602609
factor(Year)1994    6.284e+03  4.021e+03   1.563 0.118273
factor(Year)1996    3.585e+03  3.803e+03   0.943 0.346013
factor(Year)1997    4.778e+03  4.010e+03   1.192 0.233508
factor(Year)1998    8.065e+03  3.900e+03   2.068 0.038773 *
factor(Year)1999    4.076e+03  4.009e+03   1.017 0.309325
factor(Year)2000    5.489e+03  4.010e+03   1.369 0.171219
factor(Year)2001    6.608e+03  3.589e+03   1.841 0.065715 .
factor(Year)2002    6.518e+03  3.561e+03   1.831 0.067309 .
factor(Year)2003    6.758e+03  3.581e+03   1.887 0.059259 .
factor(Year)2004    7.053e+03  3.556e+03   1.983 0.047463 *
factor(Year)2005    6.736e+03  3.511e+03   1.919 0.055150 .
factor(Year)2006    7.278e+03  3.506e+03   2.076 0.038039 *
factor(Year)2007    8.163e+03  3.492e+03   2.338 0.019502 *
factor(Year)2008    8.501e+03  3.487e+03   2.438 0.014861 *
factor(Year)2009    8.155e+03  3.489e+03   2.337 0.019523 *
factor(Year)2010    8.035e+03  3.485e+03   2.306 0.021234 *
factor(Year)2011    8.658e+03  3.488e+03   2.482 0.013142 *
factor(Year)2012    9.589e+03  3.479e+03   2.756 0.005896 **
factor(Year)2013    1.006e+04  3.476e+03   2.895 0.003836 **
factor(Year)2014    1.102e+04  3.473e+03   3.172 0.001536 **
factor(Year)2015    1.200e+04  3.475e+03   3.452 0.000568 ***
factor(Year)2016    1.463e+04  3.476e+03   4.208 2.68e-05 ***
factor(Year)2017    1.720e+04  3.616e+03   4.756 2.11e-06 ***
Mileage            -6.460e-02  3.188e-03 -20.262  < 2e-16 ***
FuelTypeGasoline    1.153e+03  8.634e+02   1.336 0.181799
FuelTypeHybrid      2.494e+03  9.835e+02   2.536 0.011289 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3468 on 2075 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6629
F-statistic: 148.7 on 28 and 2075 DF,  p-value: < 2.2e-16
```

Models 1, 2, 3 Comparison

```
> cbind(AIC(mod1,mod2,mod3),BIC(mod1,mod2,mod3))
     df       AIC df       BIC
mod1  3 40678.79  3 40695.74
mod2 28 40308.17 28 40466.41
mod3 30 40302.47 30 40472.02
> anova(mod1,mod2,mod3)
Analysis of Variance Table

Model 1: Price ~ Mileage
Model 2: Price ~ factor(Year) + Mileage
Model 3: Price ~ factor(Year) + Mileage + FuelType
  Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
1   2102 3.0619e+10
2   2077 2.5071e+10 25 5548167177 18.4526 < 2.2e-16 ***
3   2075 2.4956e+10  2  115311649  4.7939  0.008372 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model 4

```
> mod4 <- lm(Price~factor(Year)+Mileage+factor(Engine)+factor(FuelType),d)
> summary(mod4)

Call:
lm(formula = Price ~ factor(Year) + Mileage + factor(Engine) +
    factor(FuelType), data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-8473.6 -1537.9   -17.4  1252.7 21345.2

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.088e+03  2.605e+03   0.802 0.422853
factor(Year)1988         7.870e+03  3.560e+03   2.211 0.027155 *
factor(Year)1992         3.237e+03  3.555e+03   0.911 0.362557
factor(Year)1994         5.098e+03  2.912e+03   1.751 0.080160 .
factor(Year)1996         3.129e+03  2.753e+03   1.137 0.255871
factor(Year)1997         5.236e+03  2.906e+03   1.802 0.071731 .
factor(Year)1998         7.141e+03  2.826e+03   2.526 0.011600 *
factor(Year)1999         3.483e+03  2.901e+03   1.200 0.230103
factor(Year)2000         5.889e+03  2.906e+03   2.027 0.042835 *
factor(Year)2001         4.636e+03  2.601e+03   1.782 0.074816 .
factor(Year)2002         4.920e+03  2.580e+03   1.907 0.056636 .
factor(Year)2003         5.302e+03  2.593e+03   2.044 0.041036 *
factor(Year)2004         6.093e+03  2.576e+03   2.365 0.018113 *
factor(Year)2005         6.244e+03  2.544e+03   2.455 0.014177 *
factor(Year)2006         5.706e+03  2.541e+03   2.246 0.024832 *
factor(Year)2007         7.353e+03  2.531e+03   2.906 0.003705 **
factor(Year)2008         8.316e+03  2.527e+03   3.291 0.001015 **
factor(Year)2009         8.604e+03  2.528e+03   3.403 0.000678 ***
factor(Year)2010         8.809e+03  2.525e+03   3.489 0.000496 ***
factor(Year)2011         9.583e+03  2.528e+03   3.791 0.000154 ***
factor(Year)2012         1.050e+04  2.521e+03   4.164 3.26e-05 ***
factor(Year)2013         1.151e+04  2.520e+03   4.568 5.20e-06 ***
factor(Year)2014         1.241e+04  2.517e+03   4.932 8.78e-07 ***
factor(Year)2015         1.377e+04  2.519e+03   5.466 5.15e-08 ***
factor(Year)2016         1.591e+04  2.520e+03   6.312 3.36e-10 ***
factor(Year)2017         1.957e+04  2.620e+03   7.467 1.20e-13 ***
Mileage                 -4.408e-02  2.355e-03 -18.717  < 2e-16 ***
factor(Engine)6          1.780e+03  1.622e+02  10.973  < 2e-16 ***
factor(Engine)8          9.775e+03  2.270e+02  43.063  < 2e-16 ***
factor(FuelType)Gasoline 1.981e+03  6.309e+02   3.141 0.001709 **
factor(FuelType)Hybrid   4.120e+03  7.200e+02   5.722 1.21e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2510 on 2073 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.826,     Adjusted R-squared:  0.8234
F-statistic: 327.9 on 30 and 2073 DF,  p-value: < 2.2e-16
```

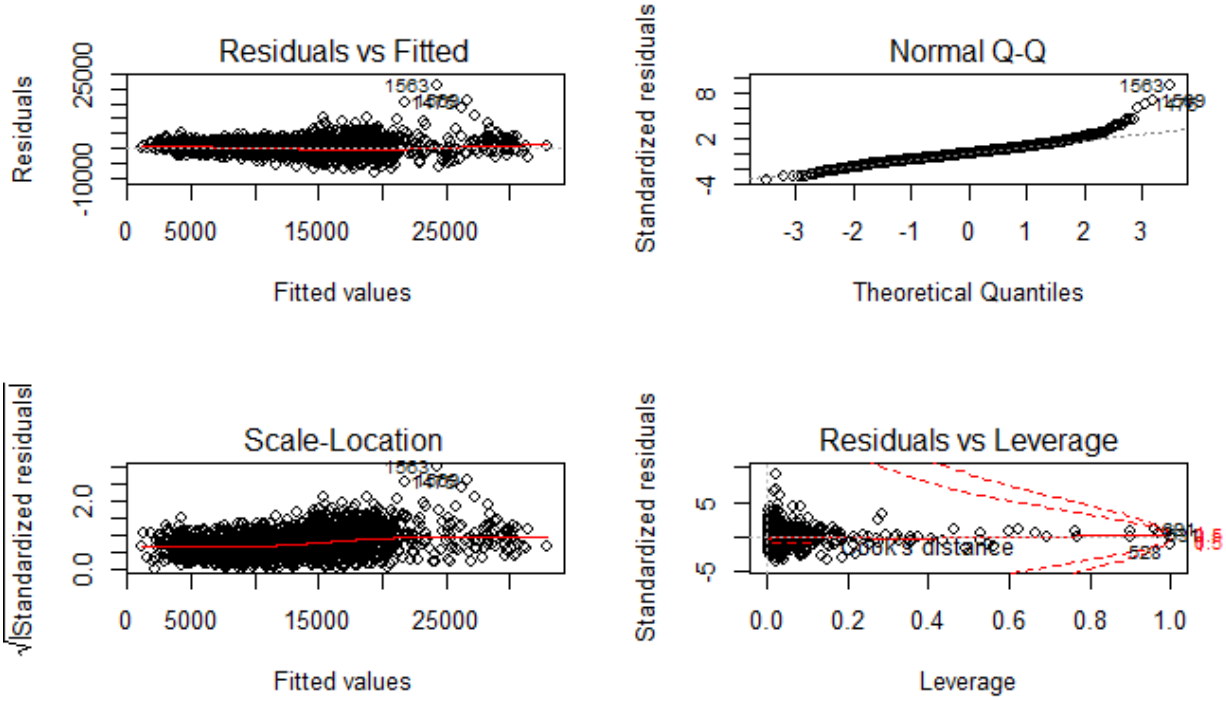
Model 5

```
Call:
lm(formula = Price ~ factor(Year) + Mileage + factor(Engine) +
    factor(FuelType) + factor(Year) * Mileage, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-8775.6 -1497.8  -106.4  1251.0 21496.9

Coefficients: (3 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.067e+04  9.265e+03   1.151 0.249786
factor(Year)1988         1.112e+04  4.801e+03   2.316 0.020677 *
factor(Year)1992         1.078e+04  8.501e+03   1.268 0.205056
factor(Year)1994        -9.950e+03  1.170e+04  -0.850 0.395270
factor(Year)1996        -1.041e+04  9.962e+03  -1.045 0.296130
factor(Year)1997        -7.350e+03  1.023e+04  -0.718 0.472635
factor(Year)1998        -7.721e+03  1.044e+04  -0.739 0.459822
factor(Year)1999        -1.024e+04  1.175e+04  -0.872 0.383457
factor(Year)2000        -4.755e+03  1.105e+04  -0.430 0.667019
factor(Year)2001        -9.005e+03  9.416e+03  -0.956 0.339009
factor(Year)2002        -7.953e+03  9.359e+03  -0.850 0.395537
factor(Year)2003        -1.014e+04  9.382e+03  -1.081 0.279802
factor(Year)2004        -8.068e+03  9.389e+03  -0.859 0.390226
factor(Year)2005        -4.984e+03  9.359e+03  -0.533 0.594431
factor(Year)2006        -6.916e+03  9.338e+03  -0.741 0.458996
factor(Year)2007        -2.749e+03  9.280e+03  -0.296 0.767055
factor(Year)2008        -5.223e+02  9.276e+03  -0.056 0.955105
factor(Year)2009        -8.304e+02  9.270e+03  -0.090 0.928629
factor(Year)2010        -3.522e+02  9.268e+03  -0.038 0.969687
factor(Year)2011         1.156e+03  9.273e+03   0.125 0.900818
factor(Year)2012         2.096e+03  9.256e+03   0.226 0.820901
factor(Year)2013         3.961e+03  9.249e+03   0.428 0.668466
factor(Year)2014         4.680e+03  9.244e+03   0.506 0.612723
factor(Year)2015         7.288e+03  9.246e+03   0.788 0.430648
factor(Year)2016         8.662e+03  9.244e+03   0.937 0.348836
factor(Year)2017         1.214e+04  8.105e+03   1.498 0.134306
Mileage                 -1.850e-01  1.456e-01  -1.271 0.204035
factor(Engine)6          1.781e+03  1.566e+02  11.377  < 2e-16 ***
factor(Engine)8          9.791e+03  2.208e+02  44.345  < 2e-16 ***
factor(FuelType)Gasoline 2.027e+03  6.105e+02   3.320 0.000915 ***
factor(FuelType)Hybrid   4.076e+03  6.973e+02   5.845 5.87e-09 ***
```

```
Residual standard error: 2414 on 2051 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8407,    Adjusted R-squared:  0.8366
F-statistic: 208.1 on 52 and 2051 DF,  p-value: < 2.2e-16
```

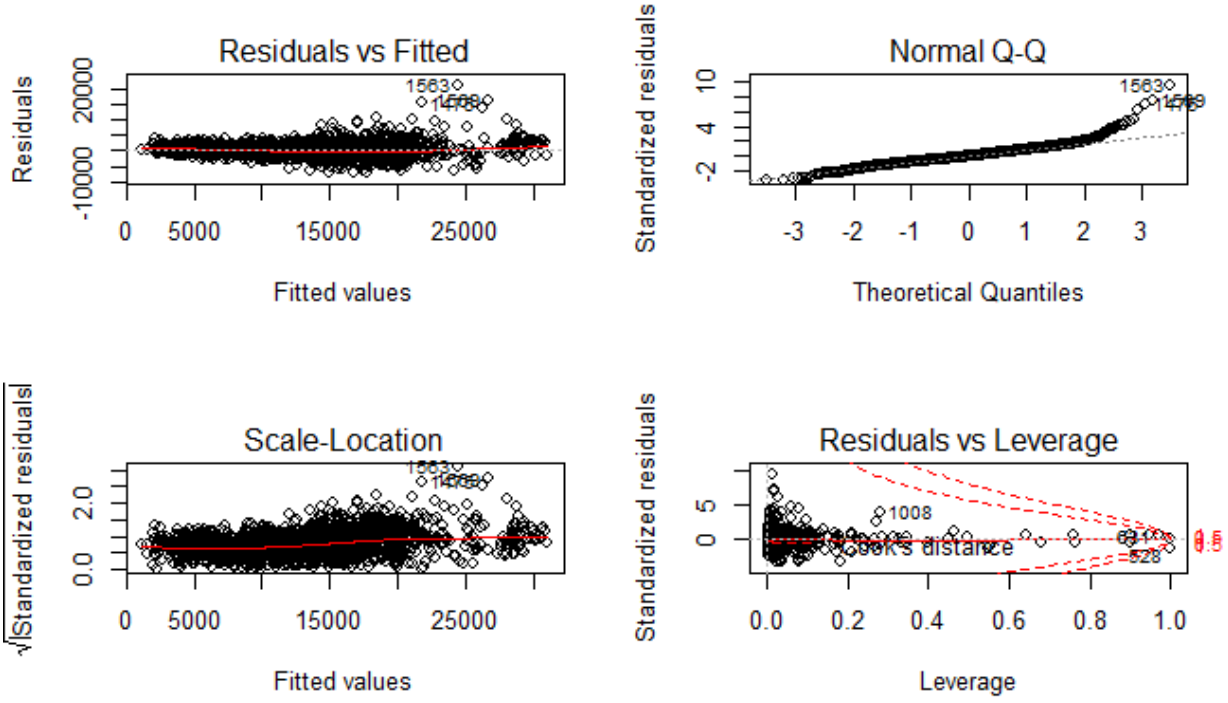
Model 6

```
Call:
lm(formula = Price ~ factor(Year) + Mileage + factor(Engine) +
    factor(FuelType) + factor(Year) * Mileage + factor(Transmission) +
    factor(ExtColor) + factor(IntColor), data = d)

Residual standard error: 2380 on 2025 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8412
F-statistic: 143.8 on 78 and 2025 DF,  p-value: < 2.2e-16
```



Model 7

```
Call:
lm(formula = Price ~ factor(Year) + Mileage + factor(Engine) +
    factor(FuelType) + factor(Year) * Mileage + factor(Transmission) +
    factor(Car), data = d)

   -

Residual standard error: 2221 on 2041 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8658,    Adjusted R-squared:  0.8617
F-statistic: 212.4 on 62 and 2041 DF,  p-value: < 2.2e-16
```



Prediction

- smp_size <- floor(0.40*nrow(d))
- set.seed(123)
- train_ind <- sample(seq_len(nrow(d)),size=smp_size)
- train <- d[train_ind,]
- test <- d[-train_ind,]
- pred<- predict.lm(mod9,test,type="response")
- err <- pred-test$Price
- pred.prob <- err <2000
- table(pred.prob)

Model	Wrongly Predicted	Correctly Predictioed	Accuracy
mod 5	211	1054	83.3202
mod 6	210	1055	83.3992
mod 7	183	1082	85.5336

Hypothesis Testing

1. Carfax claims that the average price of the Ford Mustang cars they sold in the past year is $19000. KBB, the main competitor of Carfax, believed that this value is lower than the actual average because the average price of the Ford Mustang cars that KBB sold in the previous year was more than $20000. So, they decided to prove that Carfax is lying by taking a sample. (Significance level : 5%)

   H_a : Average Price > $19000

   H_0: Average Price <= $19000

```
> M <- mean(mus$Price); M;
[1] 19381.24

> n <- length(mus$Price); n;
[1] 352

> s <- sd(mus$Price); s;
[1] 7927.969

> se <- s/sqrt(n); se;
[1] 422.5621

> Z <- (M - 19000)/se; Z;
[1] 0.9022073

> P <- pnorm(Z); P;
[1] 0.8165266
```

   Since, P > Significance level, we don't have enough evidence to reject H_0. Hence, we cannot conclude that the average price is greater than $18000.

2. Carfax claims that the Ford Mustang cars they sell has an average MPG of 28 in highways. KBB wants to prove that this claim is false (Significance level = 0.05).

H_a : Average MPG < 28

H_0: Average MPG >=28

```
> M <- mean(mus$MPG_HWY); M;
[1] 26.52841

> n <- length(mus$MPG_HWY); n;
[1] 352

> s <- sd(mus$MPG_HWY); s;
[1] 2.915581

> se <- s/sqrt(n); se;
[1] 0.155401

> Z <- (M - 28)/se; Z;
[1] -9.469635

> P <- pnorm(Z); P;
[1] 1.404113e-21
```

Since, P < Significance level, we have enough evidence to reject H_0. Hence, we can conclude that the average MPG in highways is less than 28.

Confidence Interval

1. Constructing a 95% **confidence interval** for the average price of Ford Mustang. Based on this confidence interval, what is the **minimum price** that a buyer can expect (with 95% confidence)?

```
> M <- mean(mus$Price); M;
[1] 19381.24

> n <- length(mus$Price); n;
[1] 352

> s <- sd(mus$Price); s;
[1] 7927.969

> se <- s/sqrt(n); se;
[1] 422.5621

> moe <- qnorm(0.975) * se; moe;
[1] 828.2066

> ci <- M + c(-moe, moe); ci;
[1] 18553.03 20209.45
```

Minimum price that a buyer can expect is $18553

2. Constructing a 95% **confidence interval** for the average mileage of Ford Mustang. Based on this confidence interval, what is the **minimum mileage** that a buyer can expect (with 95% confidence)?

```
> M <- mean(mus$Mileage); M;
[1] 43469.36

> n <- length(mus$Mileage); n;
[1] 352

> s <- sd(mus$Mileage); s;
[1] 33745.19

> se <- s/sqrt(n); se;
[1] 1798.625

> moe <- qnorm(0.975) * se; moe;
[1] 3525.24

> ci <- M + c(-moe, moe); ci;
[1] 39944.12 46994.60
```

Minimum mileage that a buyer can expect is around 40000 miles.

3. In an annual sales report that Carfax published, they showed that the average price of the Nissan Altima cars they sold is greater than that of Toyota Camry. Let's see whether this is correct by analyzing the sample we collected.

Nissan Altima:

```
> M <- mean(alt$Price); M;
[1] 13528.43

> n <- length(alt$Price); n;
[1] 366

> s <- sd(alt$Price); s;
[1] 4621.014

> se <- s/sqrt(n); se;
[1] 241.5443

> moe <- qnorm(0.975) * se; moe;
[1] 473.4181

> ci <- M + c(-moe, moe); ci;
[1] 13055.01 14001.85
```

95% confidence interval: [13055 14002]

Toyota Camry:

```
> M <- mean(cam$Price); M;
[1] 12268.13

> n <- length(cam$Price); n;
[1] 332

> s <- sd(cam$Price); s;
[1] 4726.381

> se <- s/sqrt(n); se;
[1] 259.3939

> moe <- qnorm(0.975) * se; moe;
[1] 508.4026

> ci <- M + c(-moe, moe); ci;
[1] 11759.73 12776.53
```

95% confidence interval: [11760 12777]

The confidence intervals of the average prices of Altima and Camry does not overlap. There, we can conclude that the average price of Altima is greater than that of Camry, with a **95% confidence**.

Conclusion:

Major factors that influence the price of a car are:

- Year: A car that is built recently have a higher price compared to other cars. 2014 built cars have a higher price over others.
- Mileage: A car that ran more fetch a lower price.
- The engine model has a major take in influencing the price.
- Exterior color influences a buyer while purchasing a car.