**Regression Model to predict Car Sales**

**Business Problem**

An Automotive industry group keeps track of the sales for a variety of personal cars. In an effort to be able to identify over- and underperforming Car models, I want to establish a relationship between vehicle sales and vehicle characteristics. Information concerning different manufacturer and models of cars is contained in car\_sales.csv

**What Drives Car Sales?**

**Null Hypothesis:** There is no significant difference in the Car Sales due to the Price in thousands Engine size, Length, Fuel efficiency, Vehicle type etc.

**Alternate Hypothesis:** There is significant difference in the Car Sales due to the Price\_in\_thousands Engine size, Length, Fuel efficiency, Vehicle type etc.

**Variables:**

Manufacturer

Model

Sales\_in\_thousands

year\_resale\_value

Vehicle\_type

Price\_in\_thousands

Engine\_size

Horsepower

Wheelbase

Width

Length

Curb\_weight

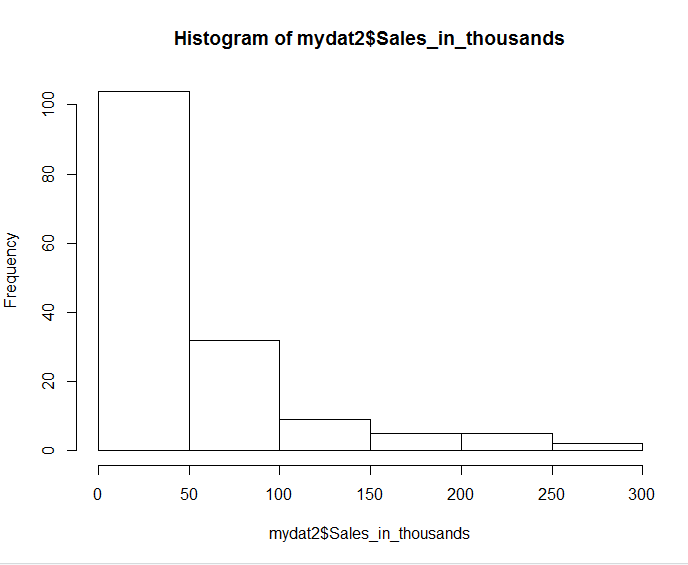
Fuel\_capacity

Fuel\_efficiency

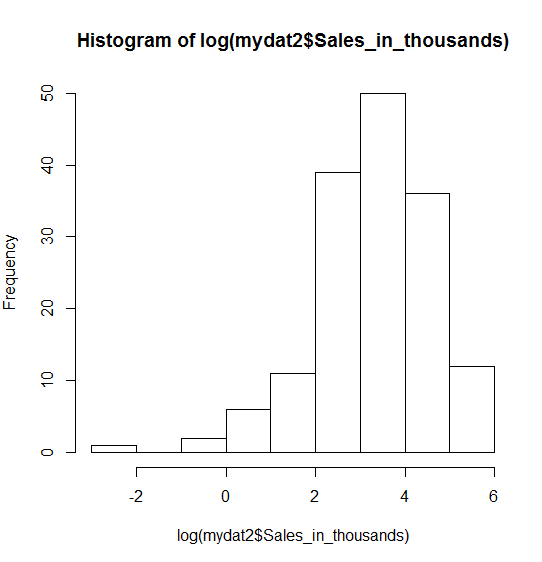
Latest\_Launch

Power\_perf\_factor

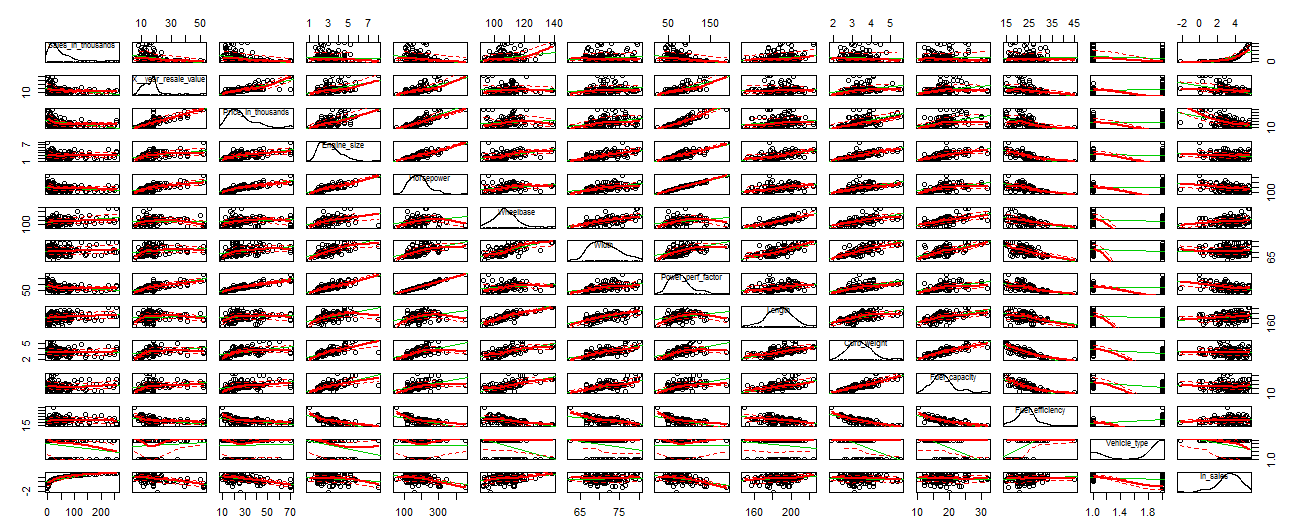
**Histogram of Dependent variable (Sales\_in\_thousands)**



Histogram of Dependent variable without transformation.(not normal)

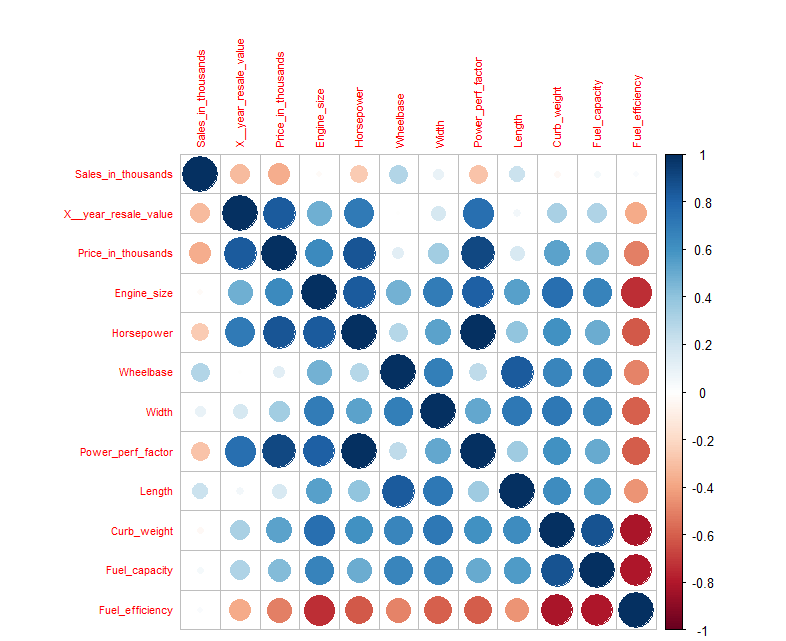


Histogram(Distribution) of Dependent variable with natural log transformation.(Behavior is more normal compared to actual distribution). Dependent variable into natural log transformed variable makes improvement in the model.

**Scatter plot to check for Multicollinearity**

The variables which are showing completely random pattern are of no use for the model. The variables which are showing some kind of trend will be useful for our model. From this graph we can understand the histogram for each variable and choose the appropriate variable during refinement.

**Correlation Matrix Plotted graphically(**Before Refinement)



Deeper blue- correlation on positive side (Highly correlated).

Deeper red - correlation on negative side.

**StepAIC (Akaike Information Criteria)** for multicollinearity

**Final Model:**

fit3<-lm(ln\_sales ~ Price\_in\_thousands + Engine\_size + Length + Fuel\_efficiency + Vehicle\_type, data = mydat2)

summary(fit3)

Min 1Q Median 3Q Max

-4.8655 -0.4920 0.0585 0.6438 1.9195

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.665274 1.750235 -1.523 0.129898

Price\_in\_thousands -0.054009 0.008564 -6.307 2.99e-09 \*\*\*

Engine\_size 0.232789 0.136715 1.703 0.090675 .

Length 0.028122 0.007352 3.825 0.000191 \*\*\*

Fuel\_efficiency 0.090831 0.035328 2.571 0.011104 \*

Vehicle\_typePassenger -0.978133 0.246877 -3.962 0.000114 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9696 on 151 degrees of freedom

Multiple R-squared: 0.468, Adjusted R-squared: 0.4504

F-statistic: 26.56 on 5 and 151 DF, p-value: < 2.2e-16

Our P value for all the predictors seem to be significant. Therefore they influence in the change of Dependent

(Response)value. In real world the value of R-squared-46% is significant.

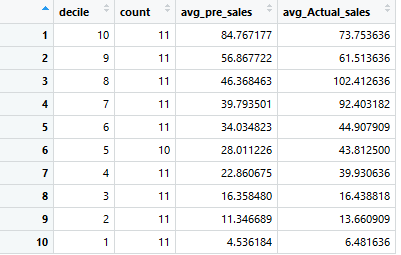
> vif(fit3)

Price\_in\_thousands Engine\_size Length Fuel\_efficiency Vehicle\_type

2.225829 3.362838 1.607816 3.725296 1.963830

**Decile based on predicted sales of training data:**

Expecting the Actual sales to follow rank order



Deviated from Rank Order because of unusual observation

coefficients(fit3)

(Intercept) Price\_in\_thousands Engine\_size Length Fuel\_efficiency

-2.66527362 -0.05400917 0.23278893 0.02812158 0.09083145

Vehicle\_typePassenger

-0.97813330

> confint(fit3, level=0.95)

2.5 % 97.5 %

(Intercept) -6.12338694 0.79283969

Price\_in\_thousands -0.07092900 -0.03708935

Engine\_size -0.03733183 0.50290970

Length 0.01359504 0.04264813

Fuel\_efficiency 0.02103082 0.16063208

Vehicle\_typePassenger -1.46591298 -0.49035362

> coefficients(fit3) # model coefficients

(Intercept) Price\_in\_thousands Engine\_size Length Fuel\_efficiency

-2.66527362 -0.05400917 0.23278893 0.02812158 0.09083145

Vehicle\_typePassenger

-0.97813330

> confint(fit3, level=0.95) # CIs for model parameters

2.5 % 97.5 %

(Intercept) -6.12338694 0.79283969

Price\_in\_thousands -0.07092900 -0.03708935

Engine\_size -0.03733183 0.50290970

Length 0.01359504 0.04264813

Fuel\_efficiency 0.02103082 0.16063208

Vehicle\_typePassenger -1.46591298 -0.49035362

**Equation**:

**Y= B0+ B1X1 +B2X2+B3X3+B4X4+B5X5**

Sales\_in\_thousands = (-2.665274 +

Price\_in\_thousands\* -0.054009+

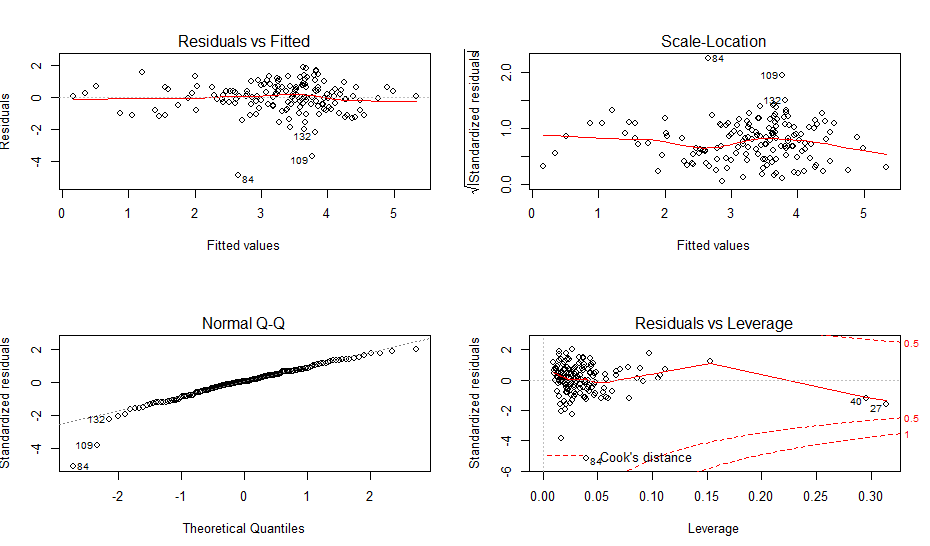
Engine\_size\* 0.232789+

Length\* 0.028122+

Fuel\_efficiency\*0.090831 +

Vehicle\_typePassenger\* -0.978133 )

**Diagnostic plots:**



**Summary:**From the Significant values we can conclude that the value of Car Sales (Sales\_in\_thousands) is influenced by Price\_in\_thousands, Engine size, Length, Fuel efficiency, Vehicle type. Our Fit3 model will be a good multivariable regression model to predict the future car sales data set.