**QMB Assignment 5**

**Name – Manish Sanjay Talekar**

**U54500765**

#Preprocessing

#Q1

**> master\_data <- read\_xlsx("6304 Module 5 Assignment Data.xlsx")**

**> colnames(master\_data)=tolower(make.names(colnames(master\_data)))**

**> names(master\_data)**

#Q2

**> cadillac <- subset(master\_data, make=="cadillac")**

**> year <- subset(cadillac, between(year, 2006, 2011))**

**> condition <- subset(year, ((condition=="good") | (condition== "excellent")))**

**> cylinders <- subset(condition, ((cylinders==6) | (cylinders==8)))**

**> color <- subset(cylinders, ((paint.color!="black") | (paint.color!="green")))**

**> set.seed(54500765)**

**> primary\_data <- sample\_n(color, 70)**

#Analysis

#q1

**> str(primary\_data)**

**tibble [70 x 10] (S3: tbl\_df/tbl/data.frame)**

**$ region : chr [1:70] "watertown, NY" "winchester, VA" "salem, OR" "tyler / east TX" ...**

**$ price : num [1:70] 5500 12994 4895 4499 2000 ...**

**$ year : chr [1:70] "2006" "2010" "2006" "2007" ...**

**$ make : chr [1:70] "cadillac" "cadillac" "cadillac" "cadillac" ...**

**$ model : chr [1:70] "dts" "cts awd" "dts" "cts" ...**

**$ condition : chr [1:70] "excellent" "excellent" "good" "excellent" ...**

**$ cylinders : num [1:70] 8 6 8 6 6 6 8 8 6 6 ...**

**$ fuel : chr [1:70] "gas" "gas" "gas" "gas" ...**

**$ odometer : num [1:70] 107494 52274 161000 195648 103331 ...**

**$ paint.color: chr [1:70] "grey" "custom" "silver" "blue" ...**

#q2

**> primary\_data$year = as.factor(primary\_data$year)**

**> primary\_data$condition = as.factor(primary\_data$condition)**

**> primary\_data$paint.color = as.factor(primary\_data$paint.color)**

**> primary\_data$cylinders = as.factor(primary\_data$cylinders)**

**> str(primary\_data)**

**tibble [70 x 10] (S3: tbl\_df/tbl/data.frame)**

**$ region : chr [1:70] "watertown, NY" "winchester, VA" "salem, OR" "tyler / east TX" ...**

**$ price : num [1:70] 5500 12994 4895 4499 2000 ...**

**$ year : Factor w/ 6 levels "2006","2007",..: 1 5 1 2 5 6 6 3 6 5 ...**

**$ make : chr [1:70] "cadillac" "cadillac" "cadillac" "cadillac" ...**

**$ model : chr [1:70] "dts" "cts awd" "dts" "cts" ...**

**$ condition : Factor w/ 2 levels "excellent","good": 1 1 2 1 1 1 1 1 1 1 ...**

**$ cylinders : Factor w/ 2 levels "6","8": 2 1 2 1 1 1 2 2 1 1 ...**

**$ fuel : chr [1:70] "gas" "gas" "gas" "gas" ...**

**$ odometer : num [1:70] 107494 52274 161000 195648 103331 ...**

**$ paint.color: Factor w/ 8 levels "blue","brown",..: 4 3 7 1 6 7 6 8 7 8 ...**

**> rm\_out <- lm(price~odometer+year+condition+paint.color+cylinders,**

**+ data = primary\_data)**

**> summary(rm\_out)**

**Call:**

**lm(formula = price ~ odometer + year + condition + paint.color +**

**cylinders, data = primary\_data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-8957.1 -998.2 22.3 1912.4 8480.2**

**Coefficients:**

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

**(Intercept) 6.924e+03 2.607e+03 2.656 0.010386 \***

**odometer -2.297e-02 1.172e-02 -1.960 0.055221 .**

**year2007 1.904e+03 1.607e+03 1.185 0.241348**

**year2008 3.307e+03 1.868e+03 1.771 0.082288 .**

**year2009 3.898e+03 2.597e+03 1.501 0.139242**

**year2010 6.755e+03 1.686e+03 4.006 0.000190 \*\*\***

**year2011 7.106e+03 1.879e+03 3.781 0.000392 \*\*\***

**conditiongood 7.898e+02 1.110e+03 0.712 0.479708**

**paint.colorbrown -2.655e+03 3.077e+03 -0.863 0.391977**

**paint.colorcustom 3.693e+02 2.569e+03 0.144 0.886252**

**paint.colorgrey -4.097e+03 2.423e+03 -1.691 0.096663 .**

**paint.colorpurple 7.933e+02 4.994e+03 0.159 0.874386**

**paint.colorred -3.478e+02 2.916e+03 -0.119 0.905502**

**paint.colorsilver -1.599e+03 1.974e+03 -0.810 0.421331**

**paint.colorwhite -2.541e+02 2.231e+03 -0.114 0.909713**

**cylinders8 5.630e+03 1.266e+03 4.445 4.4e-05 \*\*\***

**---**

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

**Residual standard error: 3830 on 54 degrees of freedom**

**Multiple R-squared: 0.5413, Adjusted R-squared: 0.4139**

**F-statistic: 4.248 on 15 and 54 DF, p-value: 4.102e-05**

#q3

**> confint(rm\_out)**

**2.5 % 97.5 %**

**(Intercept) 1.696437e+03 1.215123e+04**

**odometer -4.648121e-02 5.317203e-04**

**year2007 -1.318053e+03 5.125282e+03**

**year2008 -4.377505e+02 7.051402e+03**

**year2009 -1.309247e+03 9.104514e+03**

**year2010 3.374684e+03 1.013549e+04**

**year2011 3.338276e+03 1.087377e+04**

**conditiongood -1.434989e+03 3.014510e+03**

**paint.colorbrown -8.824580e+03 3.513785e+03**

**paint.colorcustom -4.781668e+03 5.520186e+03**

**paint.colorgrey -8.954935e+03 7.613515e+02**

**paint.colorpurple -9.219585e+03 1.080618e+04**

**paint.colorred -6.193492e+03 5.497955e+03**

**paint.colorsilver -5.556064e+03 2.357664e+03**

**paint.colorwhite -4.726029e+03 4.217770e+03**

**cylinders8 3.090724e+03 8.168862e+03**

INTERPRETATION –

* From the above stated beta coefficients, their p values, and their confidence interval. We can conclude that year (values – 2006, 2010, 2011), cylinder (value – 6,8) are the significant variables in our model which are affecting the price variable. We have reached this conclusion because the p values corresponding to these variables is less than 0.05 and therefore, we reject the null hypothesis.
* The cars which were made in the year of 2010 will increase the price of the car by $6755.
* The cars which were made in the year of 2011 will increase the price of the car by $7106.
* The car which has 8 cylinders will increase the price of the car by $5630.
* The B0 coefficient states that car, which was made in 2006, has 6 cylinders, has blue color and in excellent condition will have the price of $6924.
* Confidence interval for all variables used is 95%.

#q4

INTERPRETATION – based on the adjusted R-squared value (0.4139 i.e., which is very low) and the residual standard error (3830 i.e., very high value), it is safe to say the model is not a good fit for the primary data set.

#q5

**> par(mfrow=c(2,2))**

**> plot(primary\_data$price,rm\_out$fitted.values,pch=19,**

**+ main=paste("Actual v. Fitted, Cars, r=",**

**+ round(cor(primary\_data$price,rm\_out$fitted.values),3)))**

**> abline(0,1,col="red",lwd=3)**

**> qqnorm(rm\_out$residuals,pch=19,**

**+ main="Residuals QQ Plot, Cars")**

**> qqline(rm\_out$residuals,col="red",lwd=3)**

**> hist(rm\_out$residuals,col="red",**

**+ main="Residuals, Cars",freq=FALSE)**

**> curve(dnorm(x,mean(rm\_out$residuals),**

**+ sd(rm\_out$residuals)),**

**+ from=min(rm\_out$residuals),**

**+ to=max(rm\_out$residuals),**

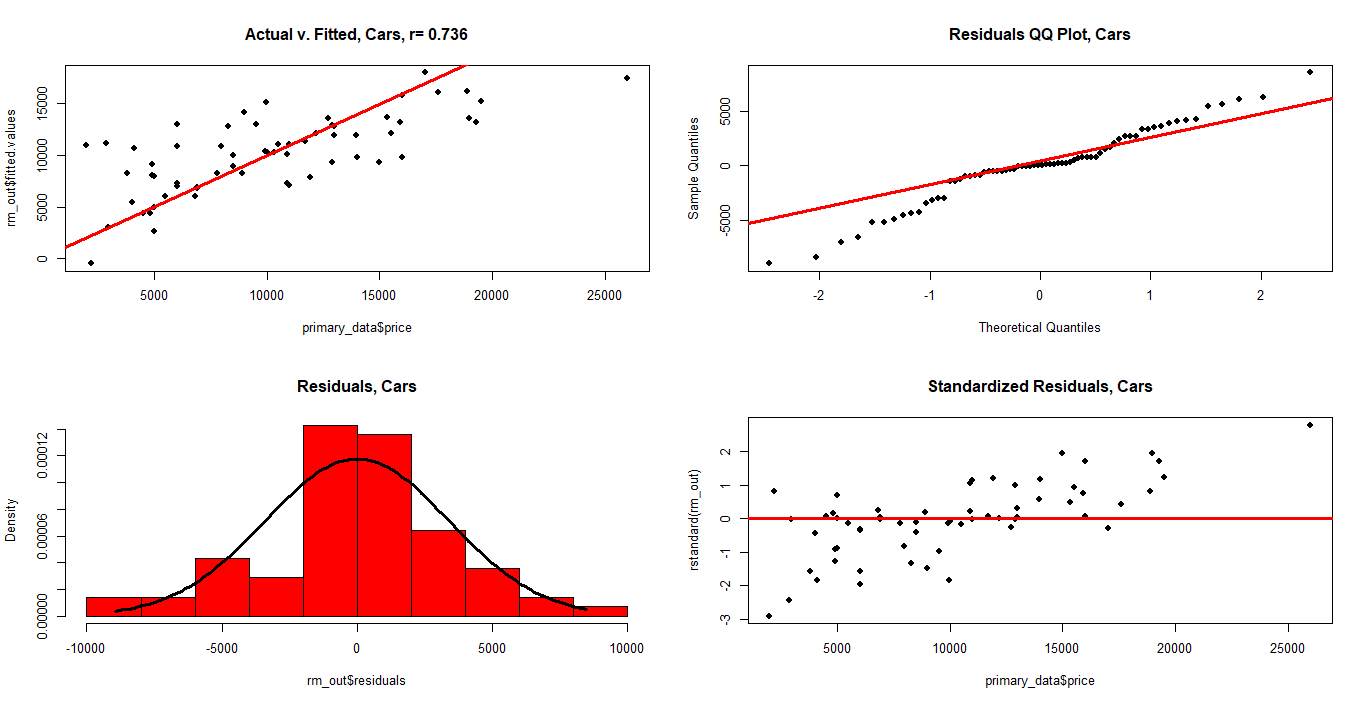
**+ lwd=3,add=TRUE)**

**> plot(primary\_data$price,rstandard(rm\_out),pch=19,**

**+ main="Standardized Residuals, Cars")**

**> abline(0,0,col="red",lwd=3)**

**> par(mfrow=c(1,1))**



**> skewness(rm\_out$residuals)**

**[1] -0.2973444**

**> kurtosis(rm\_out$residuals)**

**[1] 3.44982**

INTERPRETATIONS – Based on the above graphs and the skewness, kurtosis.

* The skewness which should be around 0 seems to be satisfied with little deviation.
* Kurtosis which should be around 3 seems to be satisfied with little deviation.
* Graph 1 shows the data to linear along the line as there is no pattern.
* Graph 2 and 3 shows the residuals to be normally distributed.
* Graph 4 does not show any pattern among the standard residuals.

Therefore, I can conclude that the model conforms with the LINE assumptions of regressions.

#q6

**> test=data.frame(odometer=183957,condition="excellent", year="2011",**

**+ cylinders="8", paint.color="red")**

**> predict(rm\_out,test,interval="predict")**

**fit lwr upr**

**1 15085.51 5817.364 24353.66**

**> predict(rm\_out,test,interval="confidence")**

**fit lwr upr**

**1 15085.51 9896.18 20274.85**

INTERPRETATION – The data for which are predicting the pricing is well within the bounds of the model. But since the model is not a good fit and the confidence interval (5817.364, 24353.66) is too large it is safe to say that the pricing is not accurate and usable.