**Preprocessing:**

library(rio)

data=import("6304 Module 5 Assignment Data.xlsx",sheet="Sheet 1")

colnames(data)=tolower(make.names(colnames(data)))

attach(data)

mydataset=subset(data,make=="cadillac" &(condition=="excellent"|condition=="good")&(cylinders=="6"|cylinders=="8")&(paint.color!="black" & paint.color!="green")&(year>="2006"&year<="2011"))

attach(mydataset)

set.seed(62067273)

primary=mydataset[sample(1:nrow(mydataset),70),]

**Analysis**

1.

str(primary)

**Output:**

> str(primary)

'data.frame': 70 obs. of 10 variables:

$ region : chr "augusta, GA" "daytona beach" "appleton-oshkosh-FDL" "visalia-tulare" ...

$ price : num 13495 9000 16995 4500 10495 ...

$ year : chr "2008" "2011" "2008" "2006" ...

$ make : chr "cadillac" "cadillac" "cadillac" "cadillac" ...

$ model : chr "escalade esv" "dts" "escalade esv" "dts performance pkg" ...

$ condition : chr "excellent" "excellent" "excellent" "excellent" ...

$ cylinders : num 8 8 8 8 8 6 6 6 8 8 ...

$ fuel : chr "gas" "gas" "gas" "gas" ...

$ odometer : num 168000 65383 139167 166280 173493 ...

$ paint.color: chr "brown" "white" "silver" "silver" ...

2.

primary$year = as.factor(primary$year)

primary$condition = as.factor(primary$condition)

primary$cylinders = as.factor(primary$cylinders)

primary.out = lm(price~year+condition+cylinders+paint.color, data=primary)

summary(primary.out)

**Output:**

> primary$year = as.factor(primary$year)

> primary$condition = as.factor(primary$condition)

> primary$cylinders = as.factor(primary$cylinders)

>

> primary.out = lm(price~year+condition+cylinders+odometer+paint.color, data=primary)

> summary(primary.out)

Call:

lm(formula = price ~ year + condition + cylinders + odometer +

paint.color, data = primary)

Residuals:

Min 1Q Median 3Q Max

-6966 -1298 0 1053 11186

Coefficients:

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

(Intercept) 4.027e+03 2.388e+03 1.687 0.097452 .

year2007 1.595e+03 1.386e+03 1.150 0.255033

year2008 4.484e+03 1.603e+03 2.798 0.007118 \*\*

year2009 3.721e+03 3.778e+03 0.985 0.329087

year2010 6.307e+03 1.439e+03 4.384 5.42e-05 \*\*\*

year2011 6.159e+03 1.521e+03 4.049 0.000166 \*\*\*

conditiongood 8.003e+02 9.578e+02 0.836 0.407064

cylinders8 4.435e+03 1.089e+03 4.072 0.000154 \*\*\*

odometer -9.146e-03 1.067e-02 -0.857 0.395226

paint.colorbrown -4.748e+01 2.323e+03 -0.020 0.983770

paint.colorcustom 2.761e+03 2.069e+03 1.335 0.187541

paint.colorgrey -1.543e+03 2.412e+03 -0.640 0.524866

paint.colorpurple 3.075e+03 4.907e+03 0.627 0.533529

paint.colorred 9.056e+02 1.936e+03 0.468 0.641788

paint.colorsilver 2.575e+02 1.694e+03 0.152 0.879719

paint.colorwhite 1.943e+03 1.849e+03 1.051 0.297987

---

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

Residual standard error: 3384 on 54 degrees of freedom

Multiple R-squared: 0.5403, Adjusted R-squared: 0.4126

F-statistic: 4.231 on 15 and 54 DF, p-value: 4.312e-05

3.

confint(primary.out)

**Output:**

> confint(primary.out)

2.5 % 97.5 %

(Intercept) -7.600179e+02 8814.54125

year2007 -1.184491e+03 4374.13193

year2008 1.270632e+03 7696.72130

year2009 -3.853389e+03 11294.46649

year2010 3.422952e+03 9191.50033

year2011 3.109311e+03 9209.35401

conditiongood -1.119935e+03 2720.59806

cylinders8 2.251177e+03 6618.42557

odometer -3.054198e-02 0.01225

paint.colorbrown -4.705089e+03 4610.13512

paint.colorcustom -1.386153e+03 6908.65086

paint.colorgrey -6.378548e+03 3291.55050

paint.colorpurple -6.763005e+03 12912.98542

paint.colorred -2.975177e+03 4786.28592

paint.colorsilver -3.138005e+03 3653.01284

paint.colorwhite -1.763688e+03 5649.38959

**2 and 3 analysis:**

From the above analysis of the model, we can say that

the price of the car with 8 cylinders goes up by 2251 dollars compared to that of 6 cylinders

the price of the car with custom or white color goes down by 1763 dollars compared to that of other colors

As condition good coefficient doesn’t have a significant p-value, there wouldn't be much impact or contrast between colors which are excellent or good in condition.

Similarly, if it is a car bought in 2007 the price increases by 1595 dollars. In 2008 price increases by 4484.

P-value is significant for 2008, 2010,2011. The odometer coefficient doesn’t have a significant p-value and will not have much impact on the price of the car.

A car with any other color, excellent condition and 6 cylinders, ones bought in 2006 constitute to a range of 4027 dollars.

4.

Considering the p-values and R-square values we can interpret that the model is a better fit with explanation of the variance by 54.03% except that there are few beta coefficients like paint and condition good which are having high p-values and are non-significant, so we can eliminate them to find a better fit.

5.

**# Linearity**

plot(primary$price,primary.out$fitted.values,pch=19,

main="Actuals v. Fitteds")

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

cor(primary$price,primary.out$fitted.values)

**Output:**

> plot(primary$price,primary.out$fitted.values,pch=19,

+ main="Actuals v. Fitteds")

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

> cor(primary$price,primary.out$fitted.values)

[1] 0.7350259

Chart, scatter chart

Description automatically generated

From the above plot and correlation we can observe that the linearity is slightly deviated.

**#Normality**

qqnorm(primary.out$residuals,pch=19,

main="Residuals Normality Plot")

qqline(primary.out$residuals,lwd=3,col="red")

hist(primary.out$residuals,main="Histogram of Residuals",col="red")

plot(density(primary.out$residuals),lwd=3,main="Density Plot of Residuals")

skewness(primary.out$residuals)

kurtosis(primary.out$residuals)

**Output:**

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

> qqnorm(primary.out$residuals,pch=19,

+ main="Residuals Normality Plot")

> qqline(primary.out$residuals,lwd=3,col="red")

>

> hist(primary.out$residuals,main="Histogram of Residuals",col="red")

> plot(density(primary.out$residuals),lwd=3,main="Density Plot of Residuals")

>

> skewness(primary.out$residuals)

[1] 0.5507553

> kurtosis(primary.out$residuals)

[1] 4.916445

From above plots, we can say that the distribution is perfectly not normal.

For normal distribution, kurtosis should be 3 but the kurtosis value is 4.9.

Therefore, normality is deviated.

**#Equality of variance**

plot(primary$price,rstandard(primary.out),pch=19,

main="Equality of Variances Plot",ylim = c(-4,4))

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

**Output:**

Chart, scatter chart

Description automatically generated

From the above plot we can say that there are few deviations for equality of variance.

At lower values graph is over fitted and at higher values, it is under fitted.

6.

#Q6

newdata=data.frame(condition="excellent",odometer=183957,cylinders="8",year="2011",paint.color="red")

predict(primary.out,newdata,interval="predict")

Output:

> newdata=data.frame(condition="excellent",odometer=183957,cylinders="8",year="2011",paint.color="red")

> predict(primary.out,newdata,interval="predict")

fit lwr upr

1 13844.48 6262.526 21426.44

For the given description of the Cadillac, our model estimates an value of 13844 dollars and 48 cents. Since, all the above values mentioned are significant values we can say we can predict the value with accuracy.