INDIAN STATISTICAL INSTITUTE, NEW DELHI

MASTERS OF STATISTICS

Diamond Price Prediction

Data Analysis Project Work

Name: Shantanu Nayek
Roll Number: MD2218
Supervisor : Prof. Deepayan Sarkar

Contents

1 Introduction

Diamond is one of the most precious jewel in the earth. Due to its rarity , it is very expensive. Also ,more significantly different diamonds are of different costs. Depending on various characteristics, the cost varies. Here we are interested with a kind of data set which have particulars for different features may be categorical or continous along with their price. Due to its high cost, it will be very much interesting if we are succesful in obtaining an appropriate model for price prediction . Not only that but also, if we can establish some relationships between price and the factors based on the data, then it may be something worthwhile for the society.

2 Organization of the paper

2.1 The Problem

The data in hand is based on some attributes and characteristics of diamond .It has about 55000 data points for the analysis. The project focus on two main problems. We take the price of diamond as our reponse variable.First one is that is there really any relationship between the predictor variables and the response variable. Also is there any relationship within the predictors itself. Second one is that how we can provide an appropriate prediction procedure for the price given the data on the predictors. Also , whether it is really necessary to know all the information about the attribute for getting an idea about the price of diamond.

2.2 Objectives of the paper

The paper has two main objectives. First one is to study the relationships among the variables in the data. Initially , we wish to study whether there is any significant relationship among the predictor variables and the response variables.For that we used some graphical tools. Next , we intend to study the relationships, if any within the predictor variables. Also, to study the nature of the response variable was also a part.

Secondly, the intention was to give some idea regarding the price prediction of the diamond based on the predictor variables given in the data. We wished to get some idea is it really so , all the variables are significant in predicting the price of diamond. After getting some idea about the important variables , we wished to obtain a suitable model for the purpose of prediction.

3 The data

• **Name :** Diamonds

• **Source :** Kaggle

• **Source Link :** https://www.kaggle.com/datasets/shivam2503/diamonds

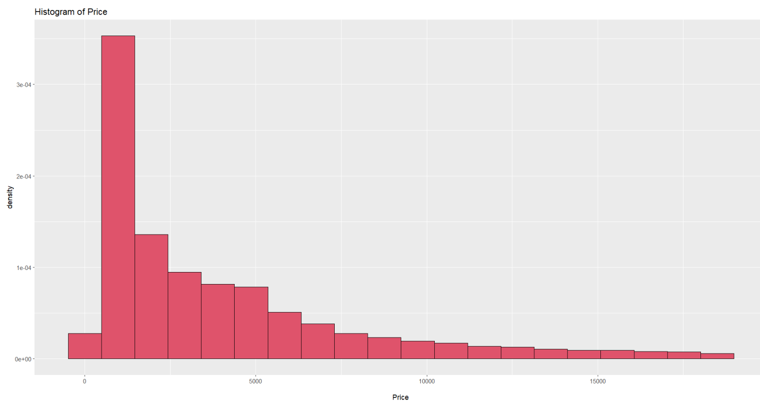
• **Description :** This classic dataset contains the prices and other attributes of almost 54,000 diamonds.

• **Reponse Variable :** Price : Price in US dollars ($326–$18,823)

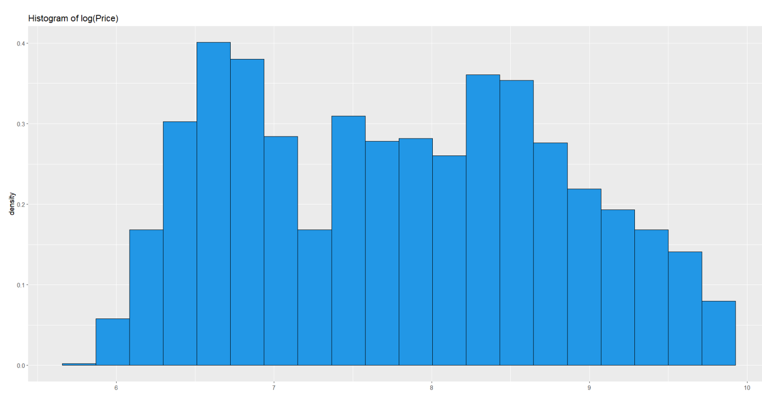
• **Predictors :**
1. carat : weight of the diamond (0.2–5.01)
2. cut : quality of the cut (Fair, Good, Very Good, Premium, Ideal)
3. color : diamond colour, from J (worst) to D (best)
4. clarity : a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
5. x : length in mm (0–10.74)
6. y : width in mm (0–58.9)
7. z : depth in mm (0–31.8)
8. depth : total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
9. table : width of top of diamond relative to widest point (43–95)

4 Exploratory Data Analysis

4.1 Nature of response variable



The response variable for the given data is Price (in U.S. dollars). The histogram of the variable price shows that it is positively skewed. So, if we use the data for the purpose of regression, we cannot do multiple linear regression using method of least squares as it violates the assumption of normality. So, we do a log transformation of the response.

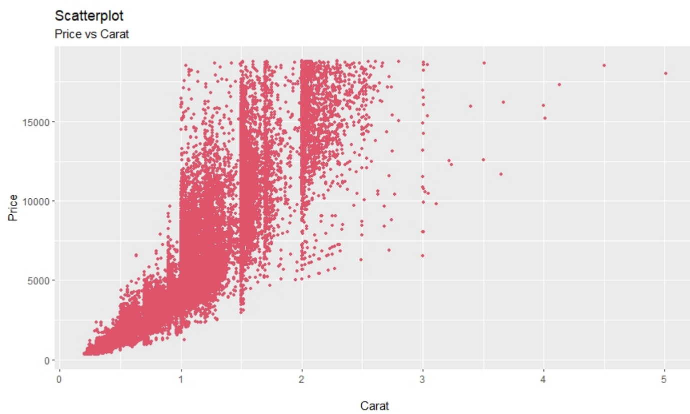


We observe that on log transformation of the price variable , the histogram is no more positively skewed. It had became almost symmetric roughly. So, applying method of least squares may not lead to much deviation in accuracy.

4.2 Relationships among response variables and continuous predictors

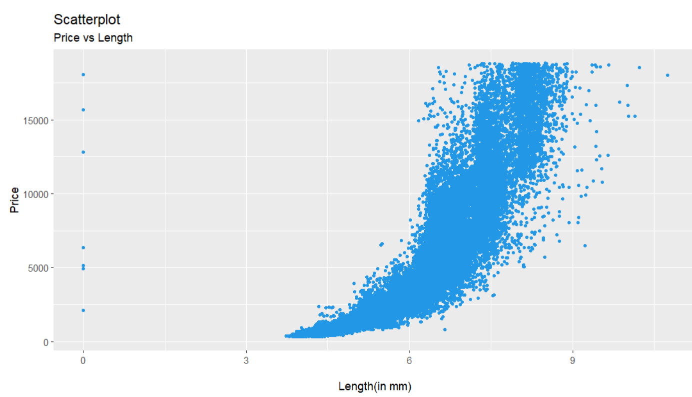
We wish to study whether there is any relationship between the predictors and response variable . Also , how the price of diamond gets influenced by the predictors.

- **Carat vs Price :**



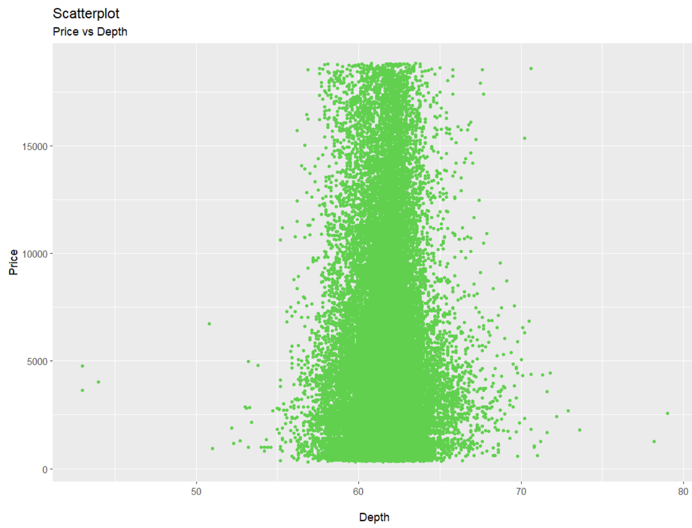
In the light of the given data , it seems that most of the observations are having weights in between 0 -2 carats. The price of diamond increases with increase in weight(in carat).Moreover , it seems that there is a clustering of observation for the different values of carat.

- **x vs Price :**



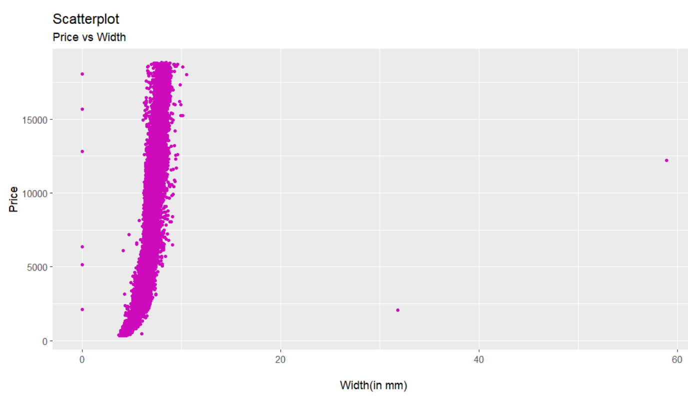
In the light of the given data , it seems that Price increases with the length (in mm)of diamond. There are some outliers for very low value of length.

- **Depth vs Price :**



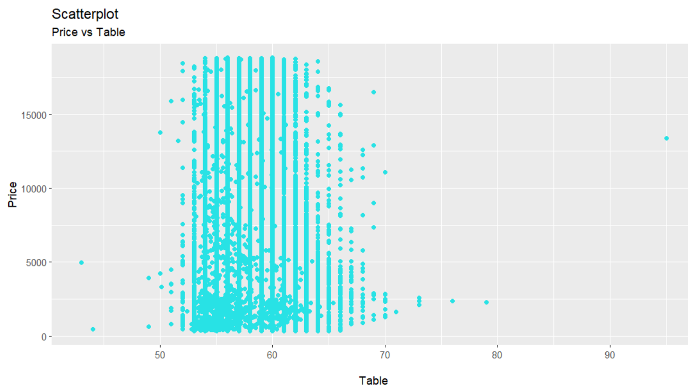
There are some observations which are outliers , some have very less depth and some has very high depth.Most of the observations have depth 55 mm -65 mm. The plot appears to be symmetric . It seems that the price of observations may be very high to low for different depths (in percentage).

- **y vs Price :**



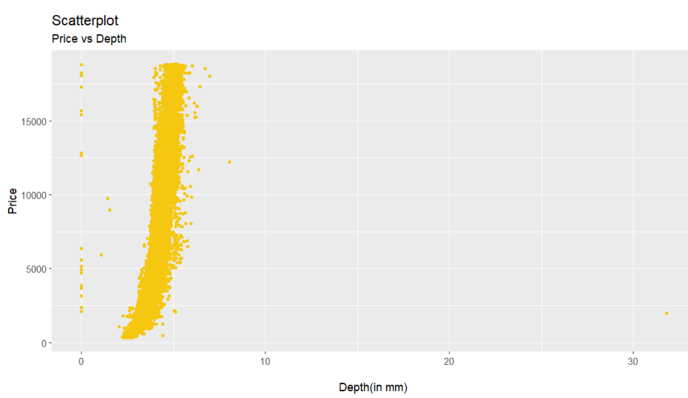
There are some outliers which made the plot less interprertable. If the plot would have been done except for the outliers , it seems that price increases with increase in width(in mm).

- **Table vs Price :**



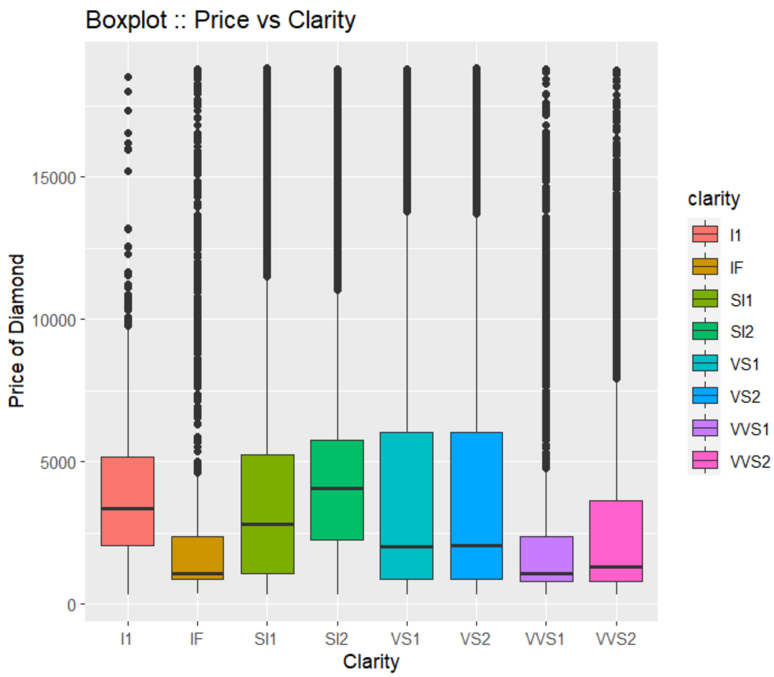
Here , table refers to the width of the top relative to the widest point . We observe there are some outliers . But most of the observations have values of table within 50 to 70.We observe that within this range for a fixed value of table , observations of very high to very low price are present.

- **z vs Price :**



There is an outlier in the observation which have depth more that 30 mm. So, the plot is tough to observe. If we observe ignoring the outliers then, it seems that price if diamond increases with increase in depth (in mm).

- **Clarity and Price :**



• Clarity refers to the measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)).

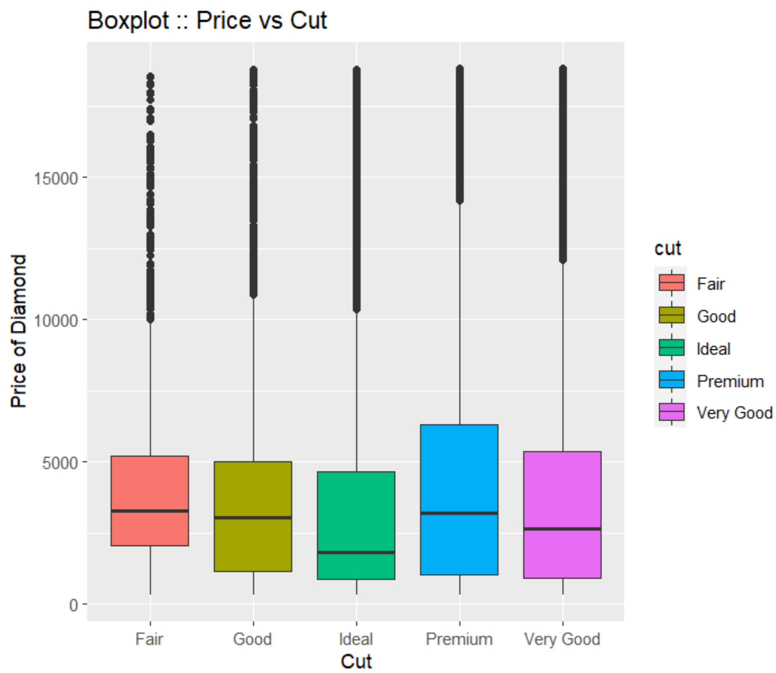
• The most significant thing is to observe is that there are significant outliers in each of the levels of clarity.

• Here , no as such pattern is observable for the price with better levels of clarity when other covariates are fixed.

• To be noticed , price is almost similar for the levels VS1 and VS2. This similar price factor is also approximately noticable in the levels IF and VVS1.

• It seems that the price is highest for the level SI2 and lowest for IF .

- **Cut and Price :**



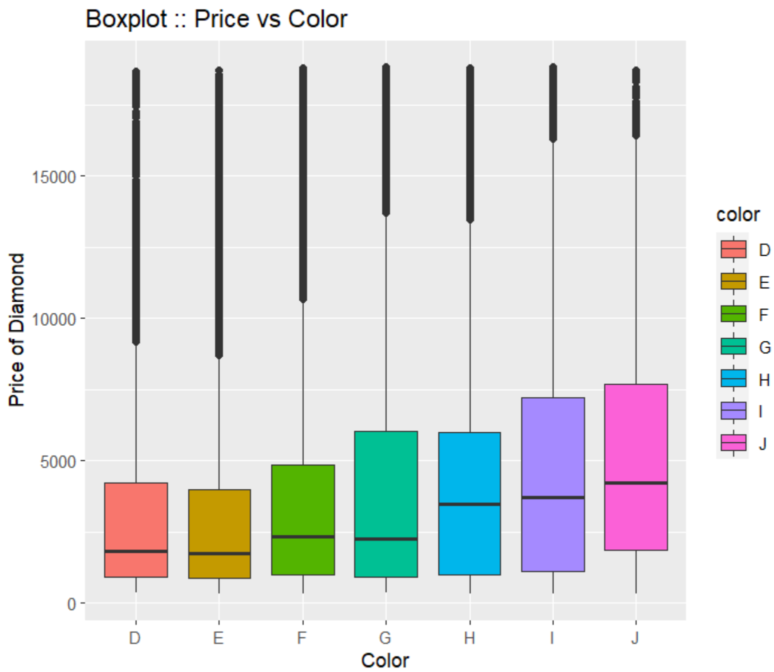
- Cut refers to the quality of the cut (Fair, Good, Very Good, Premium, Ideal).

- The most significant thing is to observe is that there are significant outliers in each of the levels of cut.

- Here , no as such pattern is observable for the price with better levels of cut when other covariates are fixed.

- It seems that the price is highest for the level Premium and lowest for Ideal .

• **Color and Price :**



• Color refers to the diamond colour, from J (worst) to D (best).

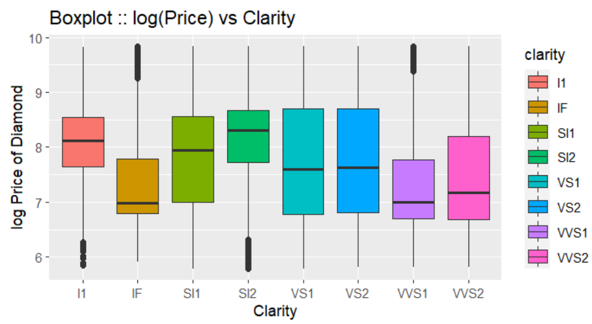
• The most significant thing is to observe is that there are significant outliers in each of the levels of cut.

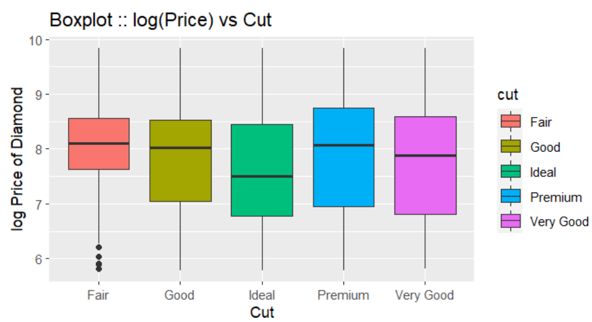
•The number of outliers in general decreases as the color of diamond become worse.

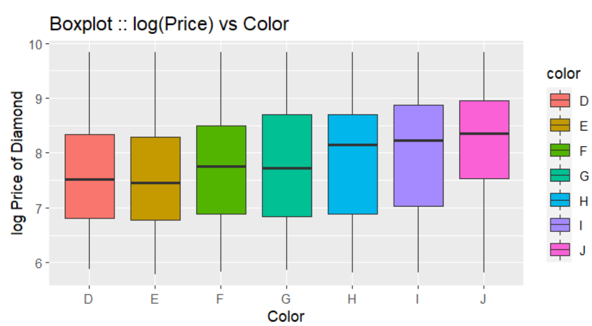
• Here , we can observe that price of diamond approximately increases on average as the color beomes worse (D to J)when other covariates are fixed.

• It seems that the price is highest for the level of color J and lowest for the level of color E .

4.3 Plot of Logarithmic of Price vs Categorical Predictors



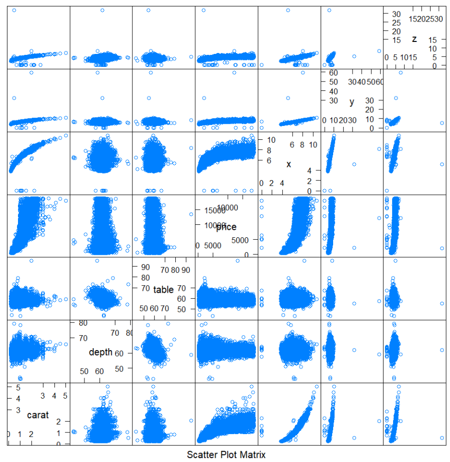




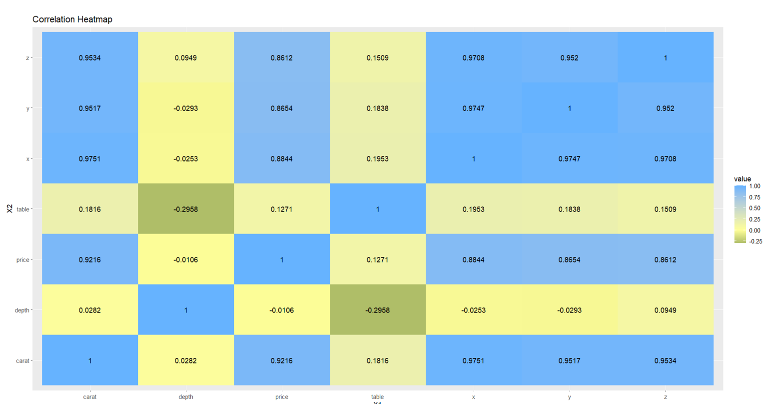
The above boxplots are based on taking log(price) as response. And we can observe that significantly the outliers are reduced and hence justified that the predictors are suitable for least squares taking this logarithimic price as response variable.

4.4 Multicollinearity and Correlation Matrix

Now , we intend to study whether there is any linear relationship within the predictor variables. For that we obtain a scatterplot among the predictors.



Since , the number of observations is very high and number of variables quite more , so the plot is hard to interpret. So, apparently we observe the correlation heat map to identify collinearity among the variables.



As it moves from yellow to blue via greenish-yellow , the magnitude of correlation increases. We observe that x, y, z , carat , these four variables have high correlation (more than0.95)within them. So, it seems that there is multicollinearity among the predictors.

4.5 Observations

• The response variable (Price in U.S. dollars) is positively skewed . Taking log transformation removes the skewness.

• The continuous predictors namely x, carat , y, z seems to be significant as with increase in the values of the predictors the value of the response increases.

•There are significant outliers in the values of the predictor y and z.

•It seems that table and depth (in percent) do not as such influence the price of diamond. That is in other words, for different values of table and depth(in percent) , almost diamond of prices low to high are significantly present.

•For the categorical predictors , Clarity, Color and Cut outliers are significantly present in each of the levels.

• For Cut and Clarity , there is no as such pattern noticable in the price of diamond with change of levels.

•It is apparent for the categorical predictor Color , that worse the color , higher the price of diamond on average , when effects of other covariates remain unchanged.

•From the correlation heat map , it seems that x, y, z, carat shows traces of linear relationships (high correlation).

5 Model for Price Prediction

5.1 Multiple Linear Regression

Based on the fact that some predictor variables have linear relationship between them, we eliminate some of the predictors with justifications (*). Also, we dummify the categorical variables for the purpose of linear regression.

***Justifications :**

• On regressing ,log price on x when other covariates are fixed, the proportion of variability explained by the regression equation of log price on x is approximately **0.7822**.

• On regressing log price on y when other covariates are fixed, the proportion of variability explained bythe regression equation of log price on y is approximately **0.7490**.

• On regressing log price on z when other covariates are fixed, the proportion of variability explained by the regression equation of log price on z is approximately **0.7418**.

• On regressing log price on carat when other covariates are fixed, the proportion of variability explained by the regression equation of log price on carat is approximately **0.8493**.

• So, we regress log price keeping carat and eliminate x, y and z in the model.

5.1.1 Test Set and Train Set

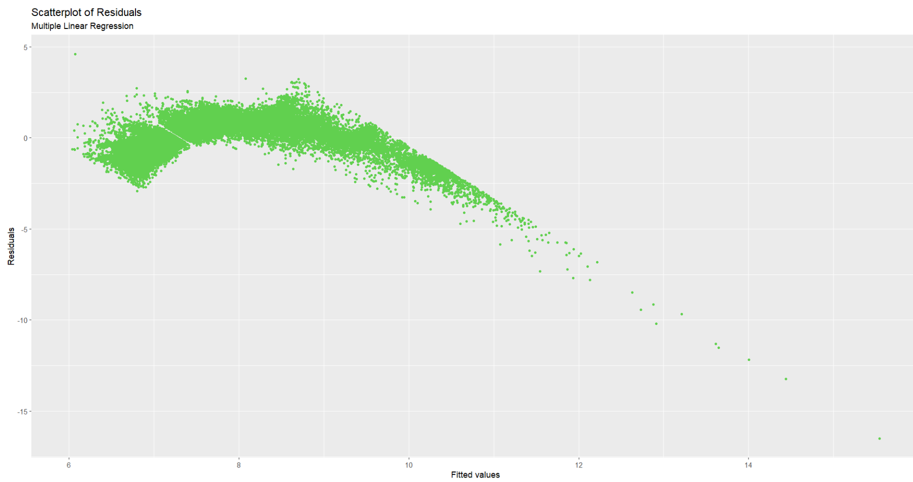
For multiple linear regression , we split the data into train set and test set. The train set contains 0.6 proportion of the total observations and train set contains 0.4 proportion of the total observations. We first regress price on other covariates in the train set and observe the residual plot. Then we use the obtained equation in the test set and observe how much deviation in the price of diamond occurs.

5.1.2 The Model

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.4997027  0.1378903  39.885  < 2e-16 ***
carat          2.1527046  0.0043929 490.040  < 2e-16 ***
depth         -0.0012027  0.0015616  -0.770   0.4412    
table          0.0064294  0.0011366   5.657 1.56e-08 ***
cut_Fair      -0.0659847  0.0127279  -5.184 2.18e-07 ***
cut_Good      -0.0127053  0.0076062  -1.670   0.0949 .  
cut_Ideal      0.0378040  0.0055340   6.831 8.56e-12 ***
cut_Premium   -0.0069067  0.0055717  -1.240   0.2151    
color_D        0.5873916  0.0101123  58.087  < 2e-16 ***
color_E        0.5362913  0.0096348  55.662  < 2e-16 ***
color_F        0.5407500  0.0095636  56.542  < 2e-16 ***
color_G        0.4602034  0.0093729  49.100  < 2e-16 ***
color_H        0.3236368  0.0095964  33.725  < 2e-16 ***
color_I        0.1584545  0.0101670  15.585  < 2e-16 ***
clarity_I1    -0.9487705  0.0182097 -52.102  < 2e-16 ***
clarity_IF     0.0870049  0.0121883   7.138 9.63e-13 ***
clarity_SI1   -0.2160349  0.0074397 -29.038  < 2e-16 ***
clarity_SI2   -0.4106912  0.0080353 -51.111  < 2e-16 ***
clarity_VS1   -0.0563066  0.0078948  -7.132 1.01e-12 ***
clarity_VS2   -0.1206586  0.0074125 -16.278  < 2e-16 ***
clarity_VVS1   0.0009962  0.0095956   0.104   0.9173    
---
```

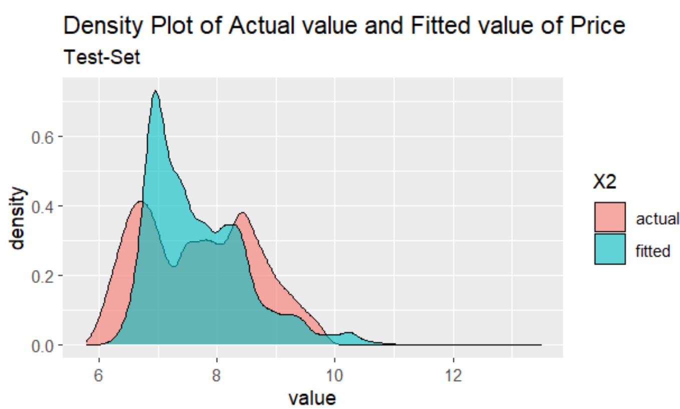
The above table gives the estimate of the regression coefficients based on the train set for the corresponding predictors.We observe that the predictors 'depth' , 'cut Premium' , 'clarity VVS1' , 'cut Good' are not that much significant in predicting the price. It is to be noted that we used the log of price as response (justification given in section 4.1). The proportion of variability explained by regression equation of log(price) on other covariates.

5.1.3 Residual Plot



The residual plot shows that there is a pattern in the residuals. There is no as such any clustering around the zero line.Hence, it shows that this regression equation is not worthwhile.

5.1.4 Fitting the model in test set



We observe that there is a significant deviation in the actual values of log price and fitted values of log price in the test set. So, overall the model so chosen, is not that much efficient for predicting price of diamond.

5.1.5 PRESS and Residual Sum of squares

PRESS=$\sum_{i=1}^{n}(e_i - e_{-i})^2$
RSS =$\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

The value of the PRESS statistic is **4103.543** and RSS is **1363.1**.

5.1.6 Observations

•The predictors 'depth' , 'cut Premium' , 'clarity VVS1' , 'cut Good' are insignificant in the light of the given data.

•The multiple r-square for this model is approximately 0.8876. That is it explains 0.8876 proportion of total variability.

•The residual plot shows pattern , hence signifies the lower efficacy of the model.

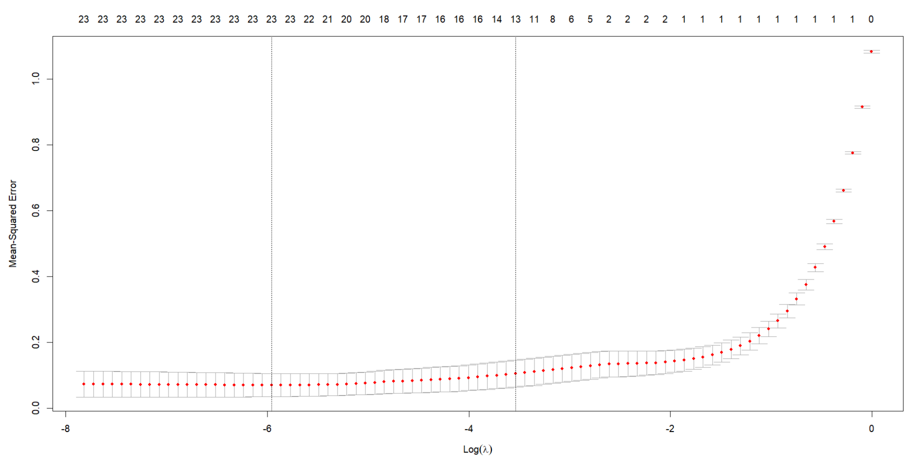
•The density plot of the fitted and actual values of the log (price) in the test set shows that there is significant deviation in the values, hence the model requires certain modifications.

5.2 Lasso Regression

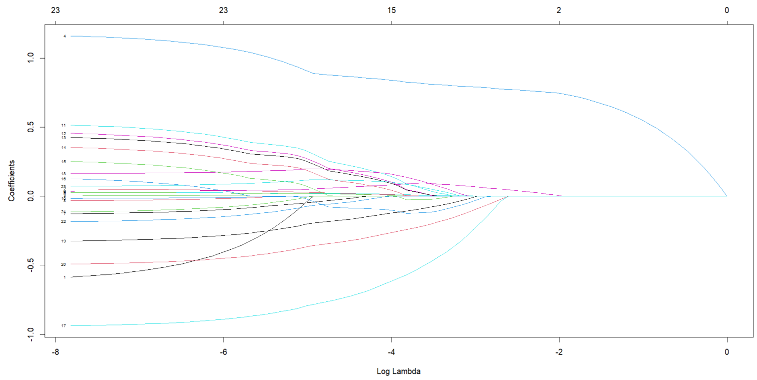
Lasso was introduced in order to improve the prediction accuracy and interpretability of regression models. It selects a reduced set of the known covariates for use in a model.Lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

5.2.1 Variable Selection

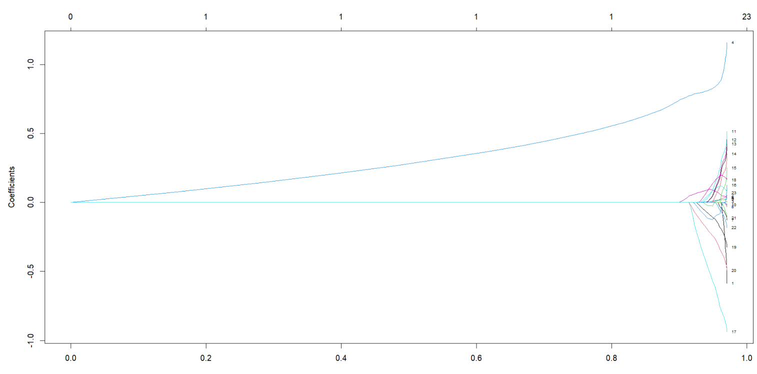
In our context , LASSO is mainly used to select the significant variables. The loss function have a variable lambda which is chosed based on minimum mean square error. Then based on that value of lambda we observe the number of non-zero predictors in the model. Then we observe the fraction deviance explained by those non-zero predictors. Then we collect those predictors and fit a linear model.



In the light of the given data, we observe that the mean square error is minimum when the value of log lambda is approximately **-5.8** and number of non - zero predictors 23 . To be noted that, considering one standard error the value of log lambda is approximately **-3.5** and the number of non-zero predictors is 13 . Thus the model becomes less complex for the second one. Hence we opt the value of lambda for which number of non-zero predictors is 13.



Here we clearly observe that for the value of log lambda -3.5 approximately , number of non-zero predictors is approximately 13.So, we wish to check how much proportionality of variation it explains. for that we observe the graph of fraction deviance.



Here it is observable that one non-zero predictor explains 0.8 proportion of total variability, a measure given by fraction deviance.So, significantly , 13 predictors explains the variability near about 0.9 proportion which is better than that of the multiple linear regression (in section 5.1.2)

	lambda.min	lambda.1se
(Intercept)	-1.234	2.833
carat	-0.395	.
depth	0.043	.
table	0.003	.
x	1.072	0.811
y	0.025	0.007
z	0.041	0.093
cut_Fair	-0.095	.
cut_Good	-0.016	.
cut_Ideal	0.022	.
cut_Premium	-0.009	.
color_D	0.423	0.039
color_E	0.366	0.008
color_F	0.338	0.008
color_G	0.269	.
color_H	0.163	-0.023
color_I	0.035	-0.115
clarity_I1	-0.888	-0.476
clarity_IF	0.175	0.090
clarity_SI1	-0.282	-0.079
clarity_SI2	-0.448	-0.202
clarity_VS1	-0.077	.
clarity_VS2	-0.146	.
clarity_VVS1	0.088	0.054

The above table gives the non -zero predictors corresponding to lambda minimum and lambda 1se . Now we collect the non-zero predictors and fit a linear model in order to obtain a better model .

5.2.2 Test Set and Train Set

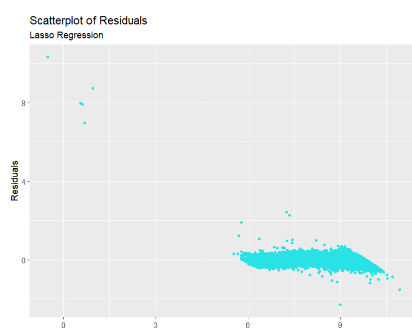
For multiple linear regression , we split the data into train set and test set. The train set contains 0.6 proportion of the total observations and train set contains 0.4 proportion of the total observations. We first regress price on other covariates in the train set and observe the residual plot. Then we use the obtained equation in the test set and observe how much deviation in the price of diamond occurs.

5.2.3 The Model

```
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2.386274   0.007173  332.674  < 2e-16 ***
x             0.849862   0.005067  167.719  < 2e-16 ***
y             0.014448   0.003742    3.861 0.000113 ***
z             0.135251   0.006076   22.259  < 2e-16 ***
color_D       0.229842   0.004188   54.880  < 2e-16 ***
color_E       0.169452   0.003728   45.457  < 2e-16 ***
color_F       0.147092   0.003690   39.862  < 2e-16 ***
color_H      -0.030066   0.003840   -7.831    5e-15 ***
color_I      -0.160587   0.004388  -36.596  < 2e-16 ***
clarity_I1   -0.827601   0.010560  -78.372  < 2e-16 ***
clarity_IF    0.281774   0.006843   41.179  < 2e-16 ***
clarity_SI1  -0.190907   0.003024  -63.137  < 2e-16 ***
clarity_SI2  -0.363547   0.003505 -103.717  < 2e-16 ***
clarity_VVS1  0.182870   0.005007   36.520  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2187 on 33631 degrees of freedom
Multiple R-squared:  0.9558,    Adjusted R-squared:  0.9558
F-statistic: 5.599e+04 on 13 and 33631 DF,  p-value: < 2.2e-16
```

On fitting a linear model using the selected predictors, we obtain the regression coefficients from the above table and all of the predictors seems to be significant in the light of the data. Moreover, the r-square observed as approximately **0.95** which seems to be quiet high.

5.2.4 Residual Plot



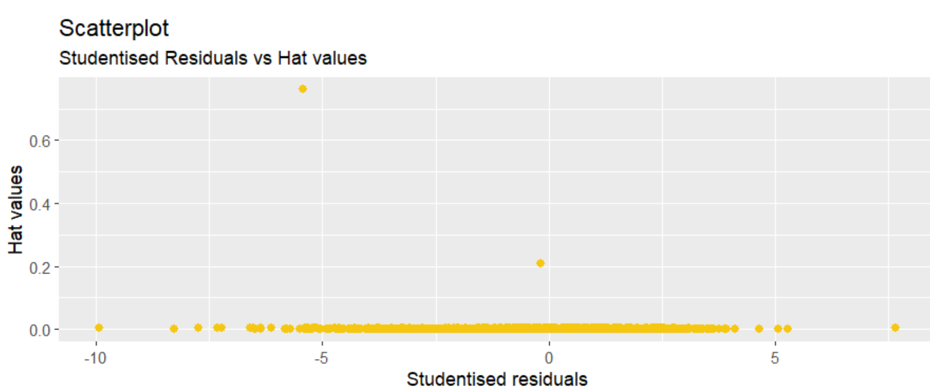
We observe that there are some outliers in the residual plot. So, we wish to remove the observations corresponding to the outliers and fit a linear model using the same variables.

5.2.5 Modifications

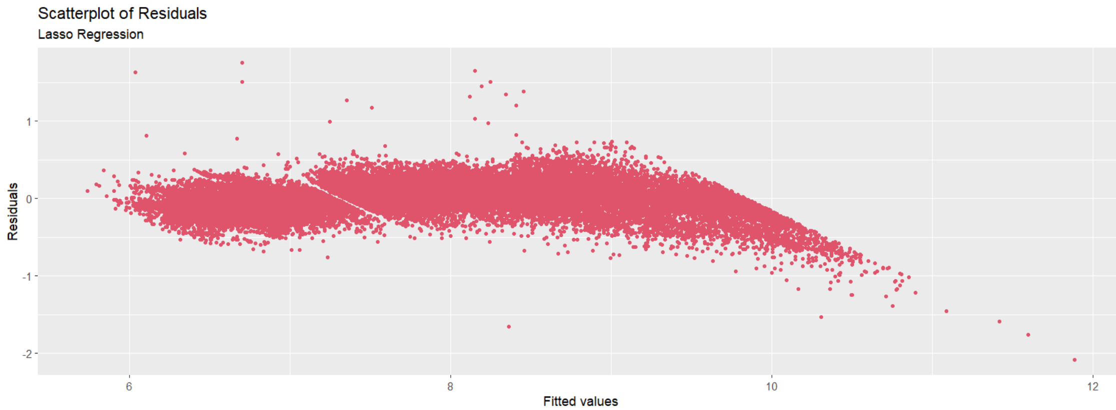
```
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2.343296   0.006602  354.931  < 2e-16 ***
x             0.742226   0.006097  121.735  < 2e-16 ***
y            -0.013906   0.003493   -3.981 6.86e-05 ***
z             0.367770   0.009199   39.980  < 2e-16 ***
color_D       0.233205   0.003841   60.716  < 2e-16 ***
color_E       0.175474   0.003420   51.315  < 2e-16 ***
color_F       0.148263   0.003384   43.813  < 2e-16 ***
color_H      -0.032810   0.003521   -9.318  < 2e-16 ***
color_I      -0.163157   0.004024  -40.549  < 2e-16 ***
clarity_I1   -0.847424   0.009693  -87.431  < 2e-16 ***
clarity_IF    0.288385   0.006275   45.959  < 2e-16 ***
clarity_SI1  -0.194895   0.002774  -70.263  < 2e-16 ***
clarity_SI2  -0.373236   0.003217 -116.029  < 2e-16 ***
clarity_VVS1  0.185217   0.004592   40.330  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2006 on 33625 degrees of freedom
Multiple R-squared:  0.9629,    Adjusted R-squared:  0.9629
F-statistic: 6.707e+04 on 13 and 33625 DF,  p-value: < 2.2e-16
```

On removing the observations corresponding to the outliers in the residual plot and fitting a linear model with the variables selected using LASSO , we obtained the estimates of the regression coefficients. We observe that all the predictors are significant and moreover the r-square increases by 0.01 unit than the previous model.

Checking for Influential Observations

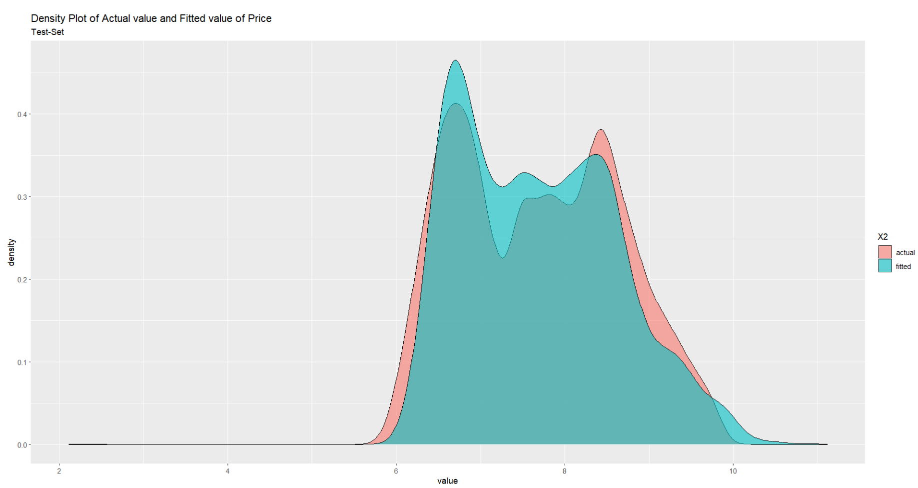


From the graph , there is no as such any values which have both studentised residual as well as hat value high. So, in the light of the given data it seems that there is no as such any influential observation.



From the residual plot we observe that all the residuals lie within -2 ands 2 . Moreover, it is well spread around the zero line and there is not as such systematic pattern.So, the model apparently seems to be a suitable model. Just for confirmation , we wish to check how the model behaves on the test set.

5.2.6 Fitting the model in the test set



The above plot suggests that there are some deviations in the actual and fitted values of the response. But , it shows a significant improvement than that of the previous case where we performed the linear regression by eliminating variables using correlation heat map.

5.2.7 PRESS and Residual Sum of squares

PRESS=$\sum_{i=1}^{n}(e_i - e_{-i})^2$
$RSS = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

The value of the PRESS statistic is **2124.267** and RSS is **832.9**.

5.2.8 Observations

• LASSO enabled to choose 13 most significant predictors which explains approximately 95 percent of the total variability. But the residual plot showed some outliers.

• On removing the outliers , the proportion of explained variability increases by 1 percent.

• From the estimates of the regression coefficients , on performing t - test gives all the selected regressors to be significant.

• Moreover, the plot of hat values and studentised residuals gives that there is not as such any influential observations.

• The residual plot do not shows any systematic pattern and well spread near the zero line, hence justifies the efficacy of the model.

• The density plot of the actual values and fitted values on the test set shows significant improvement than that of in the previous case (Section 5.1.4).

6 Conclusion

In the light of the given data , it seems that price of diamond significantly depends on 'carat' , 'x' , 'y' , 'z' and 'depth(in percent)' , 'table' do not as such affect price of diamond. Moreover , there are some outliers corresponding to each covariates. For , the categorical predictors 'clarity' , 'cut' , 'color' the outliers are significantly visible in each levels. This is due to the high positively skewed nature of the price of diamond. The outliers are reduced on observing log price for each levels of the categorical variables.

While finding a suitable model , the model obtained by ordinary least square based on the predictors chosen by observing multicollinearity by correlation heat map is less efficient than that of the model obtained by least squares based on the variables chosen by LASSO. The proportion of variability explained in the first model is around 0.88 using 20 predictors .Whereas the proportion of variability explained by the second model is approximately 0.96 , after some modifications considering 13 predictors. Also the value of PRESS decreases significantly for the second case .To be observed , the value of residual sum of squares in the test set also significantly decreases for the second case.

7 Appendix

Code 1: R Code for the Paper

```
#Part 1
#How do the predictors influence the response?

#Libraries
library(lattice)
library(fastDummies)
library(caTools)
library(glmnet)
library(dplyr)
library(ggplot2)
library(GGally)
library(reshape)
library(car)

#Reading the data
setwd("D:/")
data=read.csv("Diamonds.csv")
attach(data)
str(data)
View(data)


#Scatterplot of Continuous Predictors
#vs Response

#carat
ggplot(NULL,aes(x=carat,y=price))+
  geom_point(size=1,col=2)+
  labs(title="Scatterplot",
       subtitle="Price vs Carat",
       x="\nCarat",
       y="Price",
       col="Index")

#depth
ggplot(NULL,aes(x=depth,y=price))+
  geom_point(size=1,col=3)+
```

```
  labs(title="Scatterplot",
       subtitle="Price vs Depth",
       x="\nDepth",
       y="Price",
       col="Index")

#table
ggplot(NULL,aes(x=table,y=price))+
  geom_point(size=1.5,col=5)+
  labs(title="Scatterplot",
       subtitle="Price vs Table",
       x="\nTable",
       y="Price",
       col="Index")

#x
ggplot(NULL,aes(x=x,y=price))+
  geom_point(size=1,col=4)+
  labs(title="Scatterplot",
       subtitle="Price vs Length",
       x="\nLength(in mm)",
       y="Price",
       col="Index")

#y
ggplot(NULL,aes(x=y,y=price))+
  geom_point(size=1,col=6)+
  labs(title="Scatterplot",
       subtitle="Price vs Width",
       x="\nWidth(in mm)",
       y="Price",
       col="Index")

#z
ggplot(NULL,aes(x=z,y=price))+
  geom_point(size=1,col=7)+
  labs(title="Scatterplot",
       subtitle="Price vs Depth",
       x="\nDepth(in mm)",
       y="Price",
       col="Index")
```

```
#Histogram of continuous predictors

#carat
ggplot(NULL,aes(x=log(carat)))+
  geom_histogram(fill=4,col=1,bins=25,
                 aes(y=..density..))+
  labs(title="Histogram of log(Carat)",
       x="\nCarat",
       col="Index")

#table
ggplot(NULL,aes(x=table))+
  geom_histogram(fill=3,col=1,bins=8,
                 aes(y=..density..))+
  labs(title="Histogram of table",
       x="\ntable",
       col="Index")

#depth
ggplot(NULL,aes(x=log(depth)))+
  geom_histogram(fill=2,col=1,bins=8,
                 aes(y=..density..))+
  labs(title="Histogram of Depth",
       x="\nDepth",
       col="Index")

#x
ggplot(NULL,aes(x=x))+
  geom_histogram(fill=5,col=1,bins=10,
                 aes(y=..density..))+
  labs(title="Histogram of Length(in mm)",
       x="\nLength(in mm)",
       col="Index")

#y
ggplot(NULL,aes(x=log(y)))+
  geom_histogram(fill=4,col=1,bins=20,
                 aes(y=..density..))+
  labs(title="Histogram of Width(in mm)",
       x="\nWidth(in mm)",
```

```
       col="Index")

#z
ggplot(NULL,aes(x=log(z)))+
  geom_histogram(fill=3,col=1,bins=20,
                 aes(y=..density..))+
  labs(title="Histogram of Depth(in mm)",
       x="\nDepth(in mm)",
       col="Index")

#Boxplot of Categorical Variables

#Clarity
ggplot(data=NULL,aes(x=as.factor(clarity),y=log(price)
                     ,fill=clarity))+
  geom_boxplot()+
  labs(title = 'Boxplot :: Price vs Clarity',
       x='Clarity',
       y='Price of Diamond')

#Cut
ggplot(data=NULL,aes(x=as.factor(cut),y=price
                     ,fill=cut))+
  geom_boxplot()+
  labs(title = 'Boxplot :: Price vs Cut',
       x='Cut',
       y='Price of Diamond')

#Color
ggplot(data=NULL,aes(x=as.factor(color),y=price
                     ,fill=color))+
  geom_boxplot()+
  labs(title = 'Boxplot :: Price vs Color',
       x='Color',
       y='Price of Diamond')

#Pair-Pair plot of continuous variables
#splom(data[,c(2,6,7,8,9,10,11)])

#Correlation Heat-map
data_con=data[,c(2,6,7,8,9,10,11)]
```

```
cor(data_con)
corr = data.matrix(cor(data_con[sapply(data_con,
                                        is.numeric)]))
mel = melt(corr)
mel
ggplot(mel, aes(X1,X2))+geom_tile(aes(fill=value)) +
  geom_text(aes(label = round(value, 4)))+
  scale_fill_gradient2(low='#003300',mid = '#ffff99' ,high='
     #66b3ff') +
  labs(title = 'Correlation Heatmap')

#Observing the response
ggplot(NULL,aes(x=price))+
  geom_histogram(fill=2,col=1,bins=20,
                 aes(y=..density..))+
  labs(title="Histogram of Price",
       x="\nPrice",
       col="Index")

#Note : Positively skewed , so log transformation done

ggplot(NULL,aes(x=log(price)))+
  geom_histogram(fill=4,col=1,bins=20,
                 aes(y=..density..))+
  labs(title="Histogram of log(Price)",
       x="\nlog(Price)",
       col="Index")





#Part 2
#How to find a simple model for prediction of price of Diamond
   ?

#Dummy variable creation for the Categorical Variables
data=data[,-1]
data1=dummy_cols(data,select_columns = c('cut','color','
   clarity'))
#Working data
```

```
data2=data1[,-c(2,3,4,7,15,22,30)]

#Note : From the correlation matrix carat, x , y, z seem to
   have multicollinearity
summary(lm(log(data$price)~data2$x))
summary(lm(data$price~data2$y)) #0.749
summary(lm(data$price~data2$z)) #0.7418
summary(lm(data$price~data2$carat)) #0.8493

#Ordinary Multiple Linear Regression
data3=data2[,-c(4,5,6)]
View(data3)
View(data1)
#Train Set and Test Set
y=data1$price
y
set.seed(seed=4567)
train=which(sample.split(y,0.6)==TRUE)
train
train_data=cbind(y_p=log(y[train]),data1[train,-c
   (2,3,4,7,8,9,10,15,22,30)])
test_data=cbind(y_p=log(y[-train]),data1[-train,-c
   (2,3,4,7,8,9,10,15,22,30)])
head(train_data)
View(train_data)
y1=train_data$y_p
model1=lm(y_p~.,train_data)
model1
summary(model1)$coefficients #0.8876
ei1=residuals(model1)
hii1=hatvalues(model1)
sum((ei1/(1-hii1))^2)  #4103.543

val=cbind(fitted=predict(model1,test_data),actual=test_data
   [,1])
res=rstandard(model1)
fit_tr=predict(model1,train_data)
#Plot of Residuals
ggplot(NULL,aes(x=fit_tr,y=res))+
  geom_point(size=1,col=3)+
  labs(title="Scatterplot of Residuals",
```

```
       subtitle="Multiple Linear Regression",
       x="Fitted values",
       y="Residuals",
       col="Index")



rSqr=sum((val[,1]-val[,2])^2);#2124.267

#actual and fitted value in test set
ggplot(NULL,aes())+
  geom_histogram(col=1,
                 aes(val[,1],fill=4,y=..density..))+
  geom_histogram(col=2,
                 aes(val[,2],fill=3,y=..density..))+
  labs(title="Histogram of Actual value and Fitted value of
     Price",
       x="\nPrice",
       col="Index") #sky=fitted ,blue=actual


meltdata=melt(val)
p1 = ggplot(data=meltdata,aes(value,fill=X2))+
  geom_density(alpha=.6)+
  labs(title="Density Plot of Actual value and Fitted value of
      Price",
       subtitle = "Test-Set",
       col="Index")
p1

#Lasso Regression
set.seed(seed=1234)
train=which(sample.split(y,0.6)==TRUE)
train
train_data=cbind(y_p=log(y[train]),data1[train,-c
   (2,3,4,7,15,22,30)])
test_data=cbind(y_p=log(y[-train]),data1[-train,-c
   (2,3,4,7,15,22,30)])
head(train_data)
View(data1)
X= model.matrix( ~ . - y_p - 1,train_data)
```

```
fm.lasso= glmnet(X, train_data$y_p, alpha = 1)
plot(fm.lasso, xvar = "lambda", label = TRUE)
plot(fm.lasso, xvar = "dev", label = TRUE)
cv.lasso <- cv.glmnet(X, train_data$y_p, alpha = 1, nfolds = 50)
plot(cv.lasso) #21 non- zero predictor ;log lambda= -5.5

s.cv <- c(lambda.min = cv.lasso$lambda.min, lambda.1se = cv.lasso$lambda.1se)
round(coef(cv.lasso, s = s.cv), 3) # corresponding coefficients

View(test_data)
fit_lasso=predict(cv.lasso,s="lambda.1se",newx=data.matrix(test_data[,-1]))

View(train_data)
#Least square using Lasso Model
data_new=train_data[,-c(2,3,4,8,9,10,11,15,22,23)]
model_lasso=lm(y_p~.,data_new)
summary(model_lasso)
res1=residuals(model_lasso)
hii2=hatvalues(model_lasso)
sum((res1/(1-hii2))^2) #1728.387


sum
fitt1=predict(model_lasso)

#Plot of Residuals
ggplot(NULL,aes(x=fitt1,y=res1))+
  geom_point(size=1,col=5)+
  labs(title="Scatterplot of Residuals",
       x="Fitted values",
       y="Residuals",
       col="Index")

#Detection of three outliers
outliers=which(res1>2|res1< -2)
outliers
out1=c(6750,7230,16739,17664,30325,31012)
```

```r
train_data1=train_data[-out1,]


data_new1=train_data1[,-c(2,3,4,8,9,10,11,15,22,23)]
model_lasso=lm(y_p~.,data_new1)
summary(model_lasso)
res1=residuals(model_lasso)
hii2=hatvalues(model_lasso)
sum((res1/(1-hii2))^2) #1728.387

res2=resid(model_lasso)
fitt2=predict(model_lasso)

#Plot of Residuals
ggplot(NULL,aes(x=fitt2,y=res2))+
  geom_point(size=1,col=2)+
  labs(title="Scatterplot of Residuals",
       subtitle="Lasso Regression",
       x="Fitted values",
       y="Residuals",
       col="Index")

durbinWatsonTest(model_lasso)
ncvTest(model_lasso)




View(data_new1)
model_lasso1=lm(y_p~.,data_new1,weights = (1/res2)^2)
summary(model_lasso1)
res3=resid(model_lasso1)
fitt3=predict(model_lasso1)

ggplot(NULL,aes(x=fitt3,y=res3))+
  geom_point(size=1,col=3)+
  labs(title="Scatterplot of Residuals",
       x="Fitted values",
       y="Residuals",
       col="Index")

```

```
#Comparing Density Plot in Test Set
data_new_test=test_data[,-c(2,3,4,8,9,10,11,15,22,23)]
View(data_new_test)
val1=cbind(fitted=predict(model_lasso,data_new_test),actual=
   data_new_test[,1])

meltdata1=melt(val1)
p1 = ggplot(data=meltdata1,aes(value,fill=X2))+
  geom_density(alpha=.6)+
  labs(title="Density Plot of Actual value and Fitted value of
      Price",
       subtitle = "Test-Set",
       col="Index")

p1

rSqr=sum((val1[,1]-val1[,2])^2);#2124.267
rSqr
```