Anekant Education Society's

Tuljaram Chaturchand College of Arts, Commerce and Science,

Baramati- 413102

A Project Report on

"TO STUDY VARIOUS FACTORS WHICH AFFECT THE PRICE OF DIAMOND"

SUBMITTED TO

DEPARTMENT OF STATISTICS

Savitribai Phule Pune University,

Pune-411007

T. Y. B.Sc. (Statistics)

BY

NAME	ROLL NO.
Miss. Mahajan Sakshi Sunil	9803
Miss. Gawade Monali Balu	9800
Miss. Korade Rutuja Sukumar	9985
Miss. Komkar Vaishnavi Rahul	9984
Miss. Thengal Vrushali Rajendra	6180
Miss. Gawade Nainita Dipak	9968

Under the guidance of

Prof. N. K. Dhane

(2020-21)

Anekant Education Society"s

Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati

Department Of Statistics

CERTIFICATE

This is to certify that---------- are regular students of Department of Statistics. A project on **"Analysis of Factors Affecting on Price of Diamond"** is submitted in the partial fulfillment of the program in

T. Y. B.Sc. to the Department of Statistics, Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati.

This project has been conducted under my supervision and guidance.

Place: Baramati.

Date:

Prof. N.K. Dhane		Dr. Jagtap A. S.
Project Guide	Examiner	(Head Department of Statistics)

ACKNOWLEDGEMENT

We take this opportunity to express our sense of gratitude to Prof. N. K. Dhane for her inspiring guidance, immense motivation, constant encouragement & a critical approach at every stage, till project work complete successfully.

We are thankful to Dr. A. S. Jagtap (Head of Department of Statistics & vice principal of T. C. College, Baramati), for providing all necessary facilities & good co-operation.

We are also thankful to staff member of the Department of Statistics for their valuable discussion and guidance for completion of our project.

I express my sincere thanks to my classmates for their help in the process of this work. Last but not least, we would like to thanks non-teaching staff of the department of the statistics.

INDEX

ABSTRACT

The aim of this project is to find out which factors affect the price of diamond. Now a day"s most people also wishes to buy diamond instead of gold ornaments. The characteristics of diamond are CARAT, CUT, COLOUR, CLARITY, LENGTH, WIDTH and DEPTH.

There are many factors affecting the price of diamonds. Color, Clarity, Carat (size), Cut, Shape and Fluorescence are the major determinants of the price of diamonds. Difference in single grade or level will make the price change from 5% to 30%.

So we decide to study on which factors affecting on the price of diamond. We fit the regression model to predict the price of diamond of above characteristics. We found that the price of diamond is mainly depends on 4Cs that are carat, color, cut & clarity.

Keywords : Graphical Representation, Correlation and Regression Analysis.

INTRODUCTION

The socio-economic and political history of diamond industry is quite fascinating. Diamonds gave rise to mining industry in South Africa which is now the most advance economy the region.

One of the first things most people learn about diamonds is that not all diamonds are created equal. In fact, every diamond is unique. Diamonds come in many sizes, shapes, colors, and with various internal characteristics.

All polished diamonds are valuable. That value is based on a combination of factors. Rarity is one of those factors. Diamonds with certain qualities are more rare and more valuable than diamonds that lack them.

Jewelry professionals use a systematic way to evaluate and discuss these factors. Otherwise, there would be no way to compare one diamond to another. And there would be no way to evaluate and discuss the qualities of an individual diamond. Diamond professionals use the grading system developed by GIA in the 1950s, which established the use of four important factors to describe and classify diamonds: Clarity, Color, Cut, and Carat Weight.

These are known as the 4Cs. When used together, they describe the quality of a finished diamond. The value of a finished diamond is based on this combination. A diamond"s value is often affected by the rarity of one or more of the 4Cs.

Along with rarity, quality is the principal reason that sets the price of a diamond. However, determining the quality of a diamond is quite complex, and involves the combination of 4 different key factors, known as the 4Cs – carat weight, color, cut, and clarity.

Diamond prices can vary hugely depending on a diamond's shape, cut quality, clarity and color (4Cs). In short, the better a diamond's quality, the more you'll need to pay to purchase it.

METHODOLOGY

Now a day's most people also wishes to buy diamond instead of gold ornaments. The characteristics of diamond are carat, cut, color, clarity, length, width and depth.

There are many factors affecting the price of diamonds. Color, Clarity, Carat (size), Cut, Shape and Fluorescence are the major determinants of the price of diamonds. So we decide to study on which factors affecting on the price of diamond.

For this project, we take diamonds data set from **kaggle website**. This classic dataset contains the prices and other attributes of almost 54,000 diamonds. The data contains 53,940 observations on 10 variables like Index counter, price, carat, cut, color, clarity, table, x, y and z.

For this study, we take **sample of size 400** from the above data by **simple random sampling without replacement method**.

Here we coded the characteristics of Diamond

Cut Type	code
Fair	1
Good	2
Very Good	3
Premium	4
Ideal	5

Colour Type	code
J	1
I	2
H	3
G	4
F	5
E	6
D	7

Clarity Type	Code
I3	1
I2	2
I1	3
SI2	4
SI1	5
VS2	6
VS1	7
VVS2	8
VVS1	9
IF	10
FL	11

OBJECTIVE

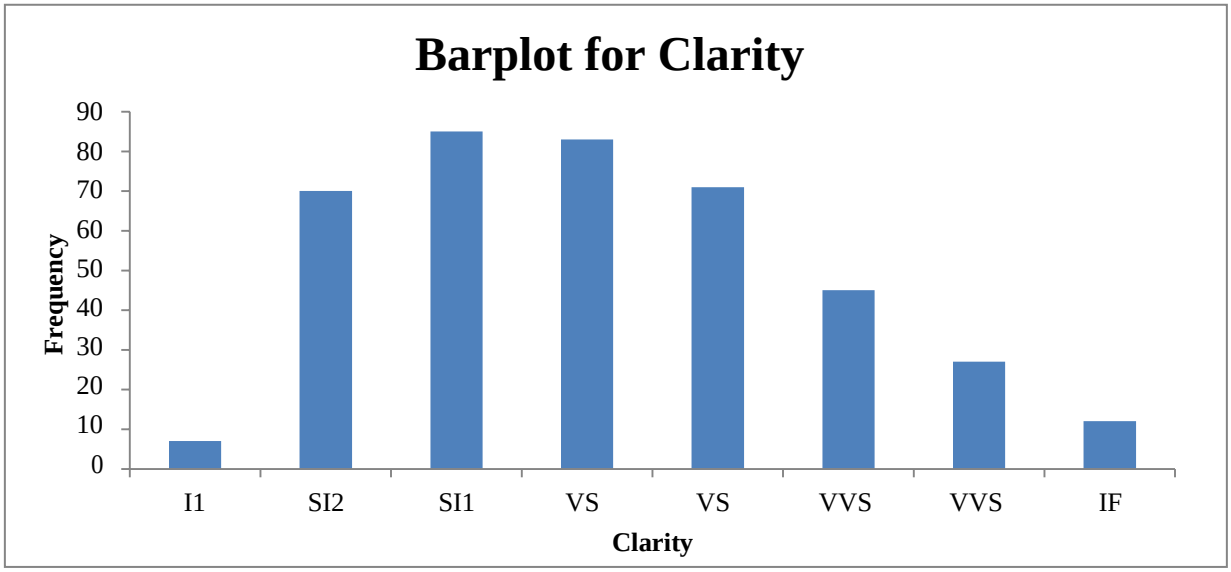
- To check whether there is relation between price and factors of diamond.
- To study various factor which affect the price of diamond.
- To predict the price of diamond.

EXPLANATORY DATA ANALYSIS

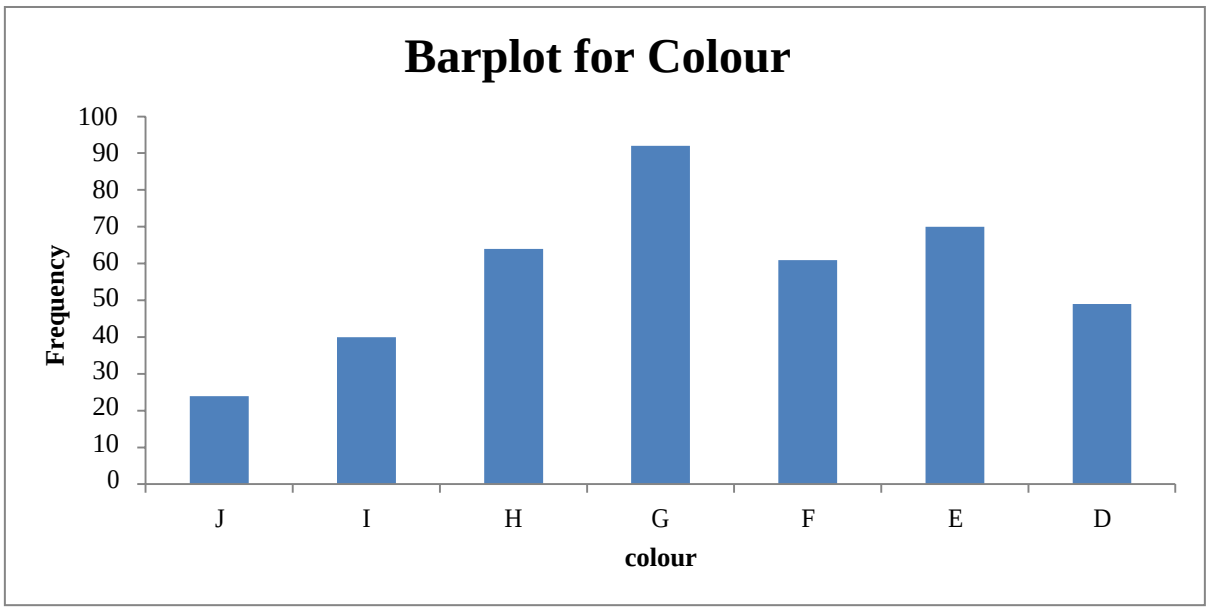
➢ Descriptive Statistics:

	carat	Depth	table	Price	x	y	z
Mean	0.826225	61.64075	57.50075	4210.08	5.790475	5.7975	3.558075
Standard Error	0.024671	0.070343	0.108193	211.971	0.058463	0.058109	0.03714
Median	0.71	61.8	57	2560	5.695	5.73	3.535
Mode	0.3	61.9	56	781	6.43	5.11	2.67
Standard Deviation	0.493428	1.406863	2.163858	4239.42	1.169259	1.162174	0.742802
Sample Variance	0.243472	1.979263	4.68228	17972680	1.367167	1.350648	0.551754
Kurtosis	-0.04047	1.974796	0.39916	1.31847	-0.97857	-0.99146	0.117445
Skewness	0.862468	-0.3758	0.595677	1.437201	0.325726	0.315783	0.039757
Range	2.09	11	13	17776	4.61	4.61	5.27
Minimum	0.23	55.8	53	373	3.9	3.94	0
Maximum	2.32	66.8	66	18149	8.51	8.55	5.27
Sum	330.49	24656.3	23000.3	1684032	2316.19	2319	1423.23
Count	400	400	400	400	400	400	400

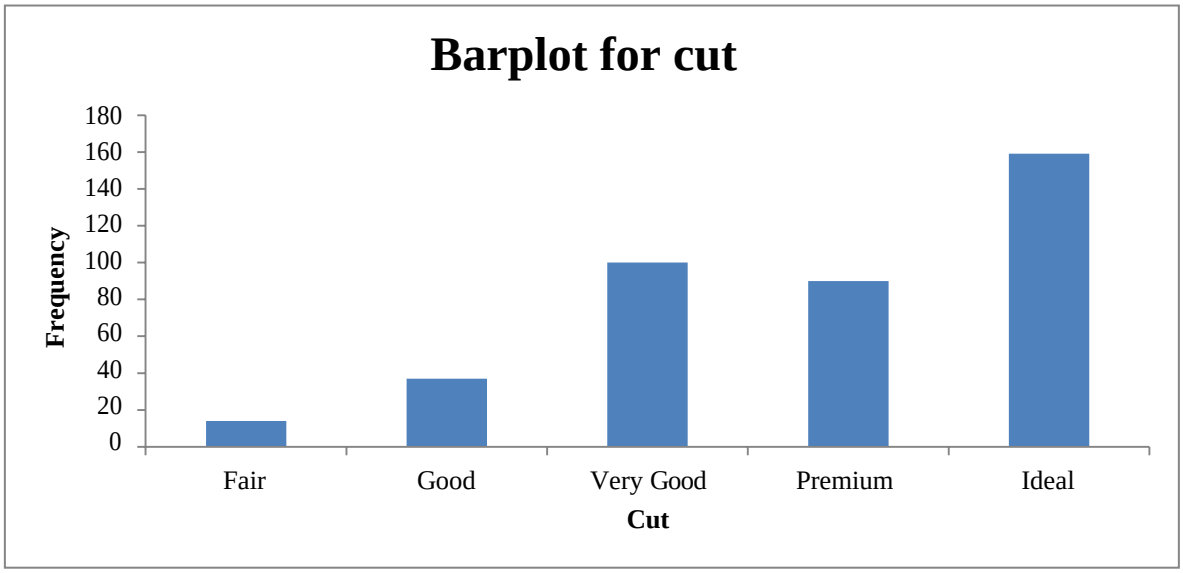
Graphical Representation



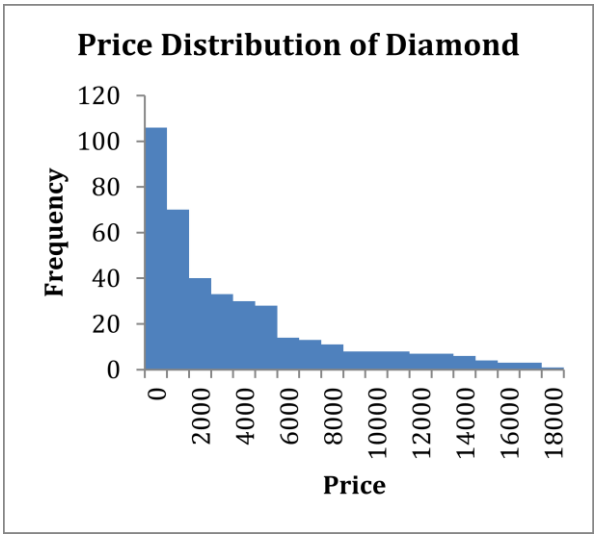
Conclusion: Here, we observed that, „SI1" type of clarity is maximum, where as „I1"type of clarity is minimum.



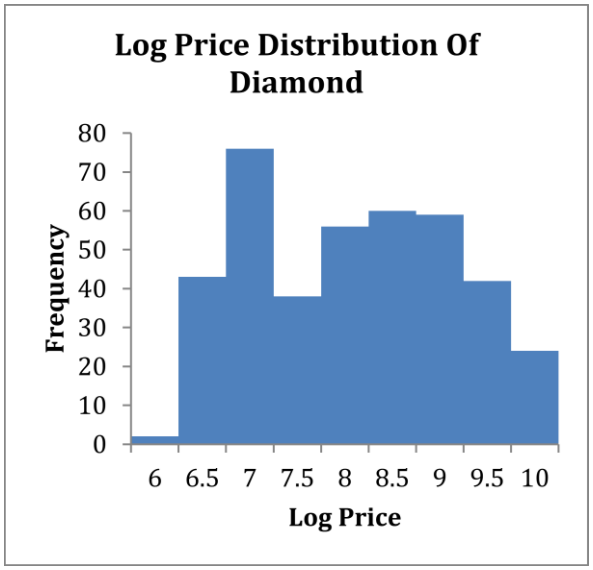
Conclusion: Here, we observed that "G" type of color is maximum, whereas "J" type of Color is minimum.



Conclusion: Here, we observed that „"Ideal"" type of cut is maximum, whereas "Fair"" type of cut is minimum.



Conclusion:
Price of diamond is heavily positively skewed. So we use log-transformation to see distribution type of price of diamond.



Conclusion:
The graph do not seems like to bell shape but slightly look likes normal curve.

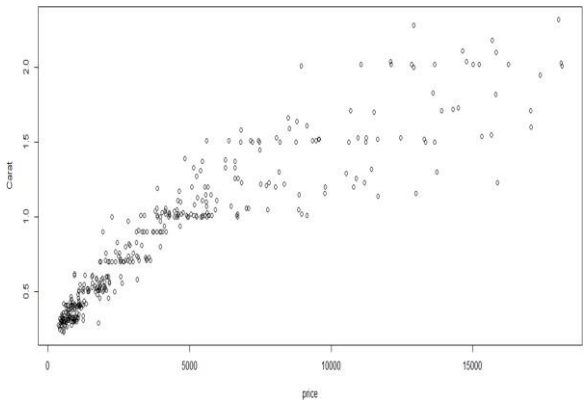
.

STATISTICAL ANALYSIS

➢ Regression Analysis:

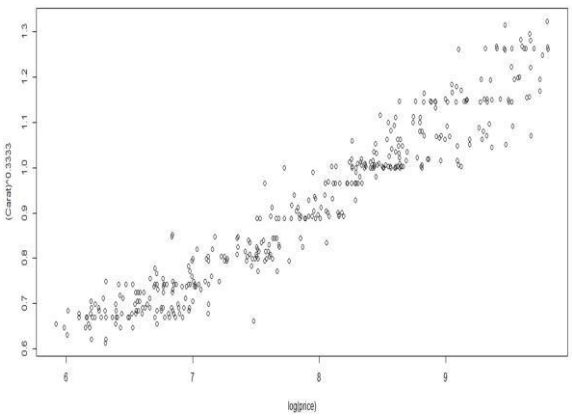
➢ Scatter Plot:

To check whether there is linear relationship between the response variable (i.e. price) and independent variables (carat).



Conclusion:

There is positive (non-linear) relationship between the both variables Price and Carat of diamond.



Conclusion:

With this transformation (carat ^0.333), we can see linear relationship.

➢ Correlation Coefficients:

Factors	price
Carat	**0.92486**
Cut	-0.06701
Color	-0.14198
Clarity	-0.16888
Depth	-0.00243
Total	0.096484
X	0.889375
Y	0.89021
Z	0.856656

Factors	log price
carat	**0.930904**
cut	-0.08452
color	-0.10778
clarity	-0.24716
depth	-0.00368
Total	0.128537
X	0.961387
Y	0.961512
Z	0.920184

Factors	log price
carat^0.3333	**0.962818**
Cut	-0.08452
Color	-0.10778
Clarity	-0.24716
Depth	-0.00368
Total	0.128537
X	0.961387
Y	0.961512
Z	0.920184

➢ Simple Linear Regression model of log(price) on cube root of carat:

R-Software-

lm(formula = log(price)~ carat^1/3)

Residuals:

Min	1Q	Median	3Q	Max
-0.72478	-0.18553	-0.00438	0.18651	0.97834

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.84147	0.07163	39.67	<2e-16 ***
Carat^1/3	5.53291	0.07782	71.1	<2e-16 ***

Signif.codes: 0 „***" 0.001 „**" 0.01 „*" 0.05 „." 0.1 „ " 1

Residual standard error: 0.2837 on 398 degrees of freedom

Multiple R-squared: 0.927, Adjusted R-squared: 0.9268

F-statistic: 5055 on 1 and 398 DF, **p-value: < 2.2e-16**

Conclusion: We conclude that, **92.7%** of total variation in the fitted model explained by the response variable **Price** and independent variable **Carat**.

➢ Multiple Linear Regression model of log(price) on cube root of carat and cut:

R-Software-

lm(formula = y ~ x11 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-0.7252	-0.1845	0.0039	0.1814	1.1058

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6491	0.09081	29.172	< 2e-16 ***
Carat^1/3	5.56804	0.07752	71.825	< 2e-16 ***
Cut	0.04165	0.01234	3.376	0.000807 ***

Signif.codes: 0 „***" 0.001 „**" 0.01 „*" 0.05 „." 0.1 „ " 1

Residual standard error: 0.2801 on 397 degrees of freedom

Multiple R-squared: 0.9291, Adjusted R-squared: 0.9287

F-statistic: 2599 on 2 and 397 DF, **p-value: < 2.2e-16**

Conclusion: We conclude that, **92.9%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat and color** .

➢ Multiple Linear Regression model of log(price) on cube root of carat, cut and color:

R-Software-

lm(formula = y ~ x11 + x2 + x3)

Residuals:

Min	1Q	Median	3Q	Max
-0.73445	-0.14992	0.00218	0.15419	1.12479

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.098196	0.094937	22.101	< 2e-16 ***
x11	5.753512	0.070408	81.717	< 2e-16 ***
x2	0.051426	0.010902	4.717	3.32e-06 ***
x3	0.079876	0.007419	10.767	< 2e-16 ***

Signif.codes: 0 „***" 0.001 „**" 0.01 „*" 0.05 „." 0.1 „ " 1

Residual standard error: 0.2466 on 396 degrees of freedom

Multiple R-squared: 0.9451, Adjusted R-squared: 0.9447

F-statistic: 2273 on 3 and 396 DF**, p-value: < 2.2e-16**

Conclusion: We conclude that, **94.5%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat, Cut and color**.

- **Multiple Linear Regression model of log(price) on cube root of carat, cut , color and clarity:**

R-Software-

lm(formula = y ~ x11 + x2 + x3 + x4)

Residuals:

Min	1Q	Median	3Q	Max
-0.67602	-0.1096	0.01544	0.12293	1.27634

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.041203	0.085972	12.111	< 2e-16 ***
x11	6.188088	0.054573	113.392	< 2e-16 ***
x2	0.025677	0.007844	3.273	0.00116 **
x3	0.092336	0.005302	17.417	< 2e-16 ***
x4	0.116249	0.005875	19.787	< 2e-16 ***

Signif.codes: 0 „***" 0.001 „**" 0.01 „*" 0.05 „." 0.1 „ " 1

Residual standard error: 0.175 on 395 degrees of freedom

Multiple R-squared: 0.9724, Adjusted R-squared: 0.9722

F-statistic: 3484 on 4 and 395 DF**, p-value: < 2.2e-16**

Conclusion: We conclude that, **97.2%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat and color** .

Overall conclusion:

From the above all R-square's best fitted model for our data is

Y =1.041203+6.188088*X11+0.025677*X2+0.092336*X3+0.116249*X4

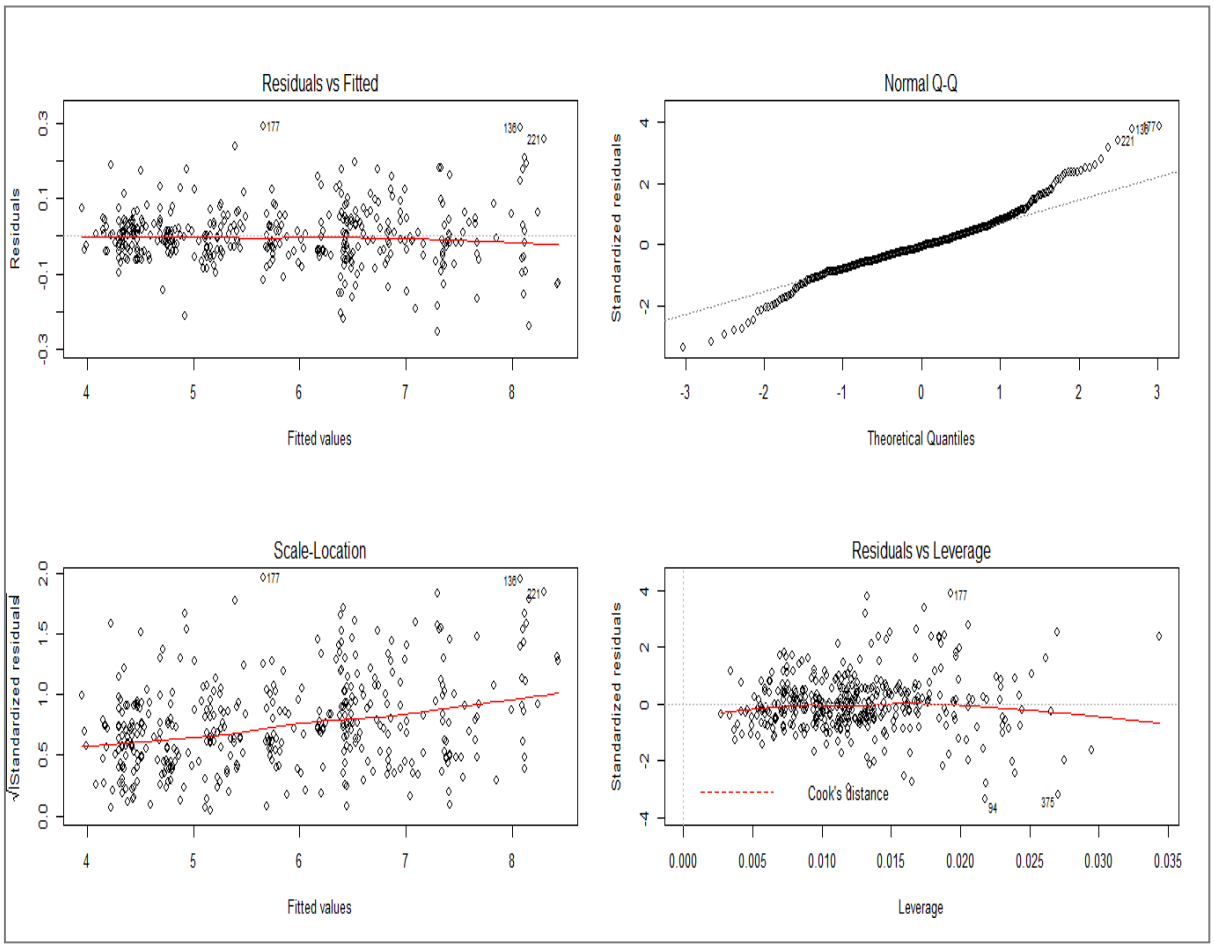
➢ Normality Assumptions for fitted model:

R-Software-

m=lm(y~x11+x2+x3+x4)

>Par(mfrow c = (2,2))

> plot (m)



Conclusion:

Here, residual plot shouldn‟t show any pattern, so Error variance is constant. From QQ plot, we can see a straight line, so errors are normally distributed. Thus both the assumptions of Constant variance and Normality of Errors are satisfied.

Conclusions

- We conclude that, **92.7%** of total variation in the fitted model explained by the response variable **Price** and independent variable **Carat** .

- We conclude that, **92.9%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat and colour**.

- We conclude that, **94.5%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat,Cut and color**.

- We conclude that, **97.2%** of total variation in the fitted model explained by the response variable **Price** and independent variables **cube root of carat and color** .

- The price of diamond is mostly affected by **carat** (**92.7%**) but other factors like cut, colour, clarity also affect the price of diamond by some percent.

Scope and Limitation

➢ Scope:

- In this project we have use only multiple linear regression. We can use other regression techniques like Multicolinearity, Detecting Outliers etc. and predictive modeling to include several data mining techniques.

➢ Limitations:

- We have study this on sample of 400, so we can increase sample size and analyzed again it.

References

- kaggle website:

  https://www.kaggle.com/shivam2503/diamonds

- Introduction to Linear Regression Analysis
  - Douglas C.Montgomery, Elizabeth A. Peck, G .Geoffrey Vining

-