Develop Quantifying Understanding of Diamond Table: A data analysis Approach

Ishaan Arora, Rushil Agarwal
Cluster innovation Centre, University Of Delhi
3rd Floor, Rugby Sevens Building, University Stadium, GC Narang Road, University Enclave, New Delhi – 110007
Ishaanarora.cic@gmail.com
rushil0195@gmail.com

***Abstract*— This research paper aims to quantify the commerce of Diamond trade by statistical inference and data analysis. The entire approach of deriving the model has been explained in a detailed fashion.**

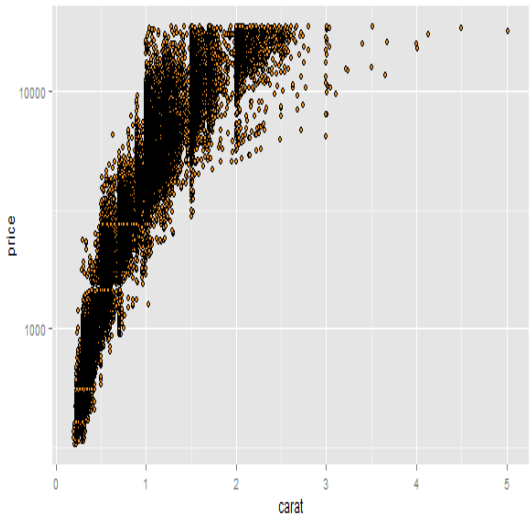
***Keywords*— Statistical Inference, Data Analysis**

I. INTRODUCTION

The data set under consideration has been extracted from a certified diamond information portal [1] .The data frame has over 54000 observations and 10 variables (carat, cut, color, quality, depth,
Table, Length of Diamond (x),Width of Diamond(y),Height of Diamond(z)).The aim of this paper is to consider all the important parameters like color, cut and carat to determine the price of a particular diamond.
R, which is an open-source data analytics compiler has been used throughout the research.

II. INITIAL SCATTER PLOT

In order to gain an initial idea about the behavior of the data set .A scatterplot was plotted between carat(X- axis) and price (Y-axis)

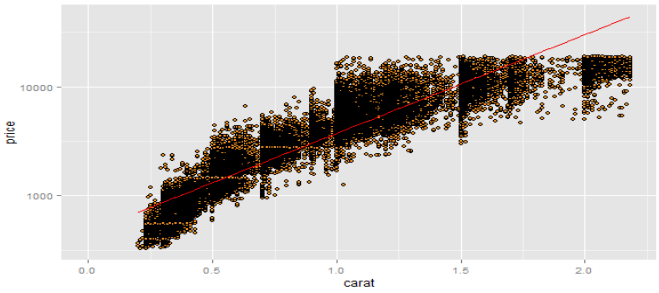


A. Problems with the initial Plot

Majority of points lie within 3 carats but the axis extends to 6 carats due to very few outliers.

B. Solution

Quantile of 0.99 of the points were taken to remove the outliers it was observed majority (99%) of data points lie within 3 carats .Also suspect of a linear relationship between price and carat was gone on plotting a linear line.



C. Observations from the Initial Plot

- Variance tends to increase with an increase in carat size
- The relation between carat and price is certainly not linear.

III. TRANSFORMATION OF CARAT AND PRICE

A. Carat Transformation

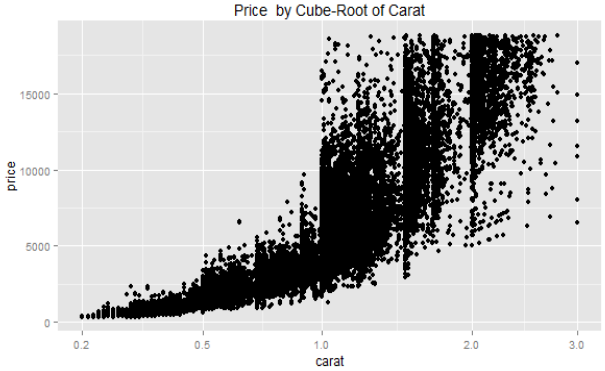
Factually carat of a diamond depends on Volume of the diamond .Mathematically speaking

$$\text{Carat} \sim f(\text{Volume})$$

Volume is a product of height(x), width(y) and breadth (z).So intuitively a possibility of transforming carat to cube root of carat is high.

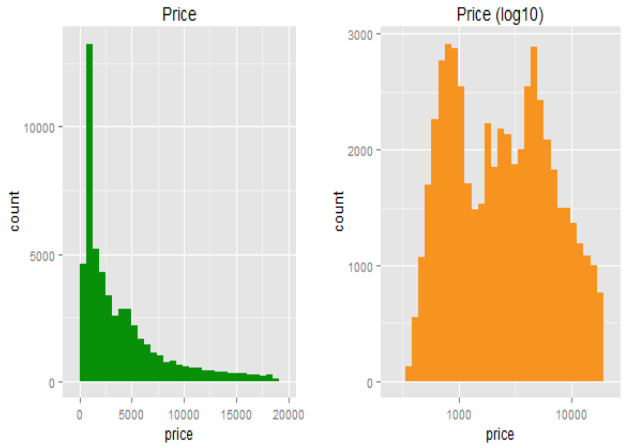
$$\sqrt[3]{\text{carat}} <- \text{carat}$$

We transformed the carat into cube root of carat by explicitly writing a function in R. On making the transformation we get the below curve

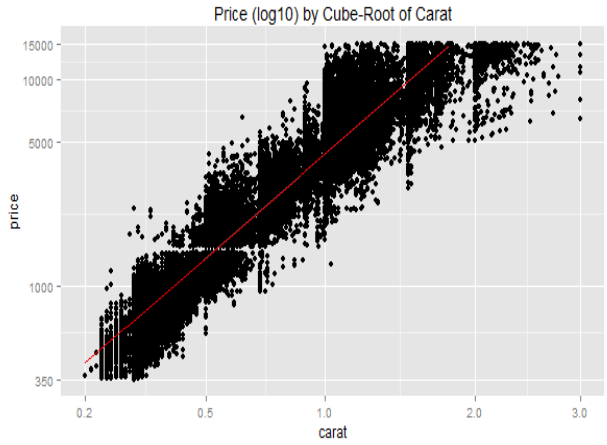


B. Price Transformation

Price is a monetary variable and is always positive. For a diamond price definitely will be above a certain threshold. So it is logical to transform price into $\log_{10}$ (Price.) .When we make this transformation and plot price's histogram and compare it with untransformed Price. Results consolidate our intuition. Below is the graph



We see the log transformed graph has 2 maxima points which implies there are two instances of increase in price first at lower price and then at higher price and this phenomenon is also known as "Poor Buyer and Rich Buyer" in social sciences.
When we make this transformation in price addition to the previous carat transformation .We get the scatterplot shown below



C. Observations of Scatter Plot after transformations

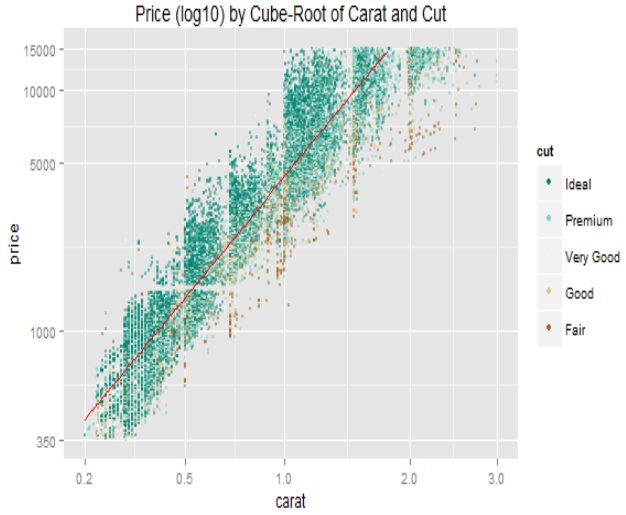
Behavior of cube root of carat and log of Price is certainly linear .As established by the above plot.

IV. CONSIDERING EFFECT OF COLOR AND CUT WITH PRICE KEEPING CARAT CONSTANT

There are other parameters which play an important role in determining the final cost (Price) of diamond.

A. Role of Cut in determining the Cost

To determine the effect of color on cost we plot the similar scatterplot but this time color data points with different color depending on cut the diamond possess. Below is the scatter graph divided cut wise.

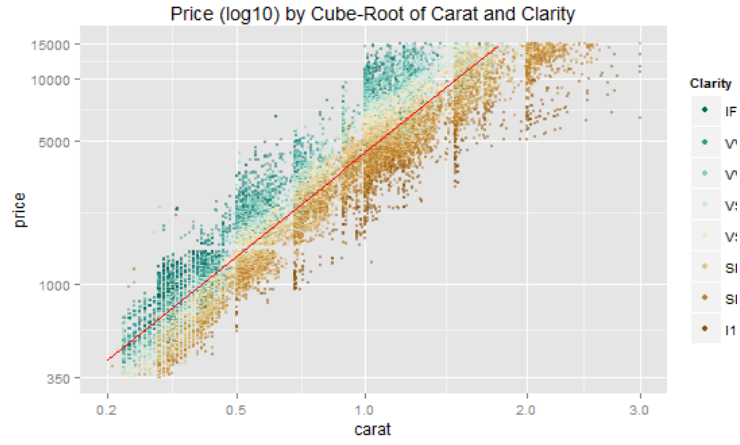


Observations from above graph

a. *Change in various cuts is not visible to the human eye on graph but price varies depending on color*
b. *Moreover the given data set has mostly ideal cut diamonds*
c. On smoothening the above colored scatter plot the behavior between log price and cut is linear
d. At constant carat ,quality of cut increases with increase in price

B. Role of Clarity in determining the Cost

To determine the effect of cut on cost we plot the similar scatterplot but this time color data points with different colors depending on different clarity the diamond possess. Below is the scatter graph divided clarity wise



Observations from above graph

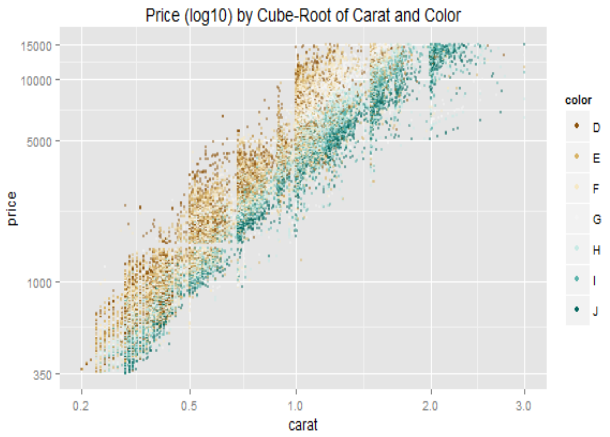
a. *Change in various clarity is visible as color changes from brown to magenta*
b. On smoothening the above colored scatter

plot the behavior between log price and clarity is linear

c. At constant carat ,quality of clarity with increase in price

C. *Role of Color in determining the Cost*

To determine the effect of color on cost we plot the similar scatterplot but this time color data points with different colors depending on different colors the diamond possess. Below is the scatter graph divided color wise



Observations from above graph

a. *Change in various color is visible as color changes from brown to magenta*

b. On smoothening the above colored scatter plot the behavior between log price and color is linear

c. At constant carat ,quality of color increases with increase in price

d. ,quality of color increases with increase in price

V. MODELLING THE ABOVE STATISTICAL INFERENCES

Having figured out the relationship between various parameters which turned out to be linear after certain transformations. Hence the model developed is

$$A*\text{Price (log10)} + B*\sqrt[3]{\text{carat}} + C*\text{Cut} + D*\text{Clarity} + E*\text{Color F=0}$$

IV. CONCLUSIONS

The study has successfully developed a linear model to predict the price of a diamond depending upon the various parameters like clarity, cut, color and carat.

V.ACKNOWLEDGMENT

We acknowledge the support provided by the administration of Cluster Innovation Centre, University of Delhi and allowed us to use their technological resources and infrastructure which helped us to pursue our research work

.

VI. REFERENCES

[1] http://diamonds.org/
[2] http://hci.stanford.edu/publications/2013/invisibleaudience/invisibleaudience.pdf
[3] http://cran.r-project.org/
[4] http://appliedresearch.cancer.gov/archive/cognitive/immt.pdf
[5] https://rdotnet.codeplex.com/
[6] http://gastonsanchez.com/Handling_and_Processing_Strings_in_R.pdf
[7] http://www.statmethods.net/graphs/line.html
[8] http://www.statmethods.net/stats/power.html