Exploratory Data Analysis on Diamond Dataset using R

K .DurgaCharan, K.Sandhya Rani, SomulaRamasubbareddy, K.Govinda, E.Swetha

Abstract: ***The socioeconomic and political history of the diamond industry is fascinating. Understanding diamonds are important because each diamond is unique in its way. Even an expert cannot incorporate as much information about price as a picture of the entire market without analysing the characteristics of the diamonds. Hadley's ggplot2 ships with a data set that records the carat size and the price of more than 50 thousand diamonds, from Diamond Search Engine collected in 2008. In this paper we perform an exploratory data analysis using R on the diamond dataset to understand the diamond market trends, quality and price by analysing factors for Market Research. The difficulties that may arise with the analysis include, improper dataset, Presence of Outliers, Faulty Data etc. Also since the dataset dates back to 2008, the estimates may not be used with today's market trends.***

Keywords: Diamond, Exploratory Data analysis, R, Market Research

I. INTRODUCTION

The starting step in exploring any big amount of data is Exploratory Data Analysis(EDA). It starts with a basic understanding of data and the framing of which factors to focus on. This is done with a broad view of the patterns and quantitative techniques which give a basic understanding of what the dataset depicts. Our quest for clues that help in framing the future steps and the question that arise from the data set is what truly drives the Exploratory Data Analysis, as there are no standard set of rules which tell the user how to approach data. That aside EDA gives life to a set of statistical methods which help to define the purpose of the data. Almost all EDA practices are graphical in nature with a few quantitative techniques. The motive for the high dependence on graphics is that by its very nature the main role of EDA is to liberally investigate, and graphs gives the examiners unparalleled chances to do as such, tempting the information to uncover its basic mysteries, and being constantly prepared to increase some new, regularly unsuspected, understanding into the information. In blend with the characteristic example acknowledgment abilities that we as a whole have, design gives, obviously, unparalleled energy to do this.

Revised Manuscript Received on March 26, 2019.

K .DurgaCharan, Dhanekula Institute of Engineering, Technology, Vijayawada, India.
K.Sandhya Rani, Dhanekula Institute of Engineering &Technology ,Vijayawada, India.
SomulaRamasubbareddy, VNRVJIET, Hyderabad,India.
K.Govinda, VIT University, Vellore,India.
E.Swetha, SV College of Engineering, Tirupati, India.

II. LITERATURE SURVEY

The scope of statistics has widened in the past decade which has led to contain the areas of exploratory data analysis and visualising the data - going beyond the standard paradigms of estimation and testing, to look for patterns in data beyond the expected, according to Tukey 1972, 1977; Chambers, Cleveland, Kleiner, and Tukey 1983; Cleveland 1985, 1993; Tufte 1983, 1990; Buja, Cook, and Swayne 1996; and Wainer 1997 among others[2-3]. Moreover there has been significant development in that; several models have been developed that can be used to explore complex and large data. The complex modelling approaches include all types of parametric, semi parametric, regression, tree-based models according to Hastie, Tibshirani, and Friedman review in 2000[1]. Advances in calculation have encouraged progresses equally in EDAwith an emphasis onmodelling. There has been made available high clarity graphics, more reined GUIs and much faster accessible software which helps in much more evaluation and exploration of data to promote exploratory data analysis and data pattern generations.

For demonstrating, new calculations running from neural systems to hereditary calculations to Markov affix re-enactment enable clients to fit models that have no shut shape articulations for evaluations, vulnerabilities, and back dispersions. What's more, obviously, the two illustrations and demonstrating have profited from the sheer increment in the speed and capacity limit of desktop PCs. The associations between measurements, illustrations, and calculation seem even in the title of this diary. Sadly, there has not been much association made between look into in the two ranges of exploratory information examination and complex demonstrating. On one hand, exploratory investigation is frequently considered without models. From the other heading, in Bayesian induction, exploratory information examination is normally utilized just in the beginning periods of model plan, however appears to have no place once a model has really been fit. This article contends that (an) exploratory and graphical techniques can be particularly viable, when utilized as a part of conjunction with models, and (b) display based surmising can be particularly compelling, when checked graphically. Our key advance is to detail (basically) all graphical shows as model checks, with the goal that new models and new graphical techniques go as one.

III. DATA ANALYSIS

A. Dataset

There are a total of 53,940 diamonds in the dataset with

10 features (carat, cut, colour, clarity, depth, table, price, x, y, and z)[4]. The variables cut, colour, and clarity, are ordered factor variables with the following levels.

(Worst) ——————————> (best)
cut: Fair, Good, Very Good, Premium, Ideal
colour: J, I, H, G, F, E, D
clarity: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF
Other observations include:
Most diamonds are of ideal cut.
The median carat size is 0.7.
Most diamonds have a colour of G or better.
About 75% of diamonds have carat weights less than 1.
The median price for a diamonds $2401 and the max price is $18,823.

IV. RESULTS

The primary characteristics in the diamond set are carat and price. The task is anticipating which characteristics drive the price and finally determining which of them can be used to determine price.
Carat, colour, cut, clarity, depth, and table are seemingly responsible for the determination of price. Further it can be said, carat and clarity are the primary factor responsible for variation of price.

The following graphs show the variation of diamond count vs. the individual variable.

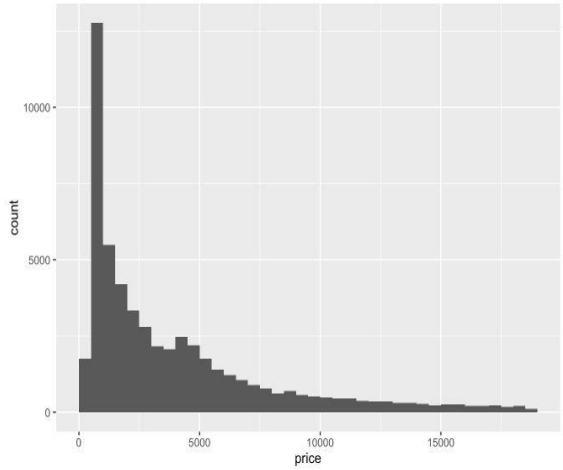

Figure 1.Variation of Price of diamonds with respect to the count attribute

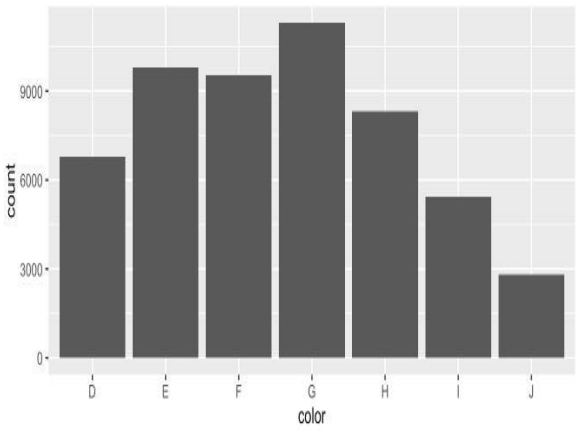

Figure 2. Visualization of number of dimonds with respect to their unique colours

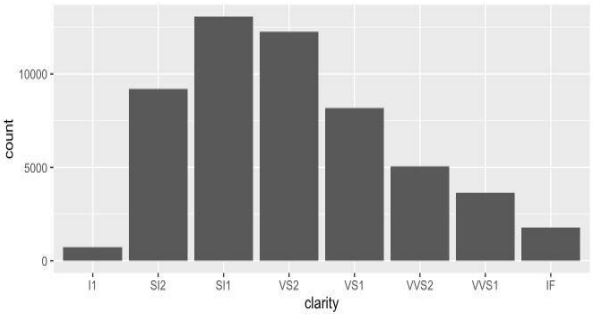

Figure 3. Visualization of count of diamonds present in the dataset with respect to Clarity

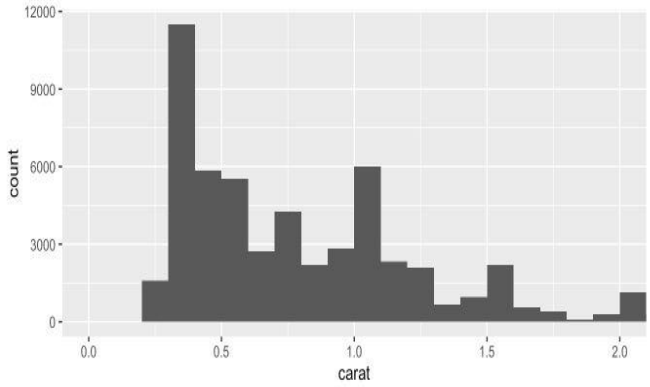

Figure 4. Visualization of count of dimonds based on the carat attribute

The dimensions of a diamond are seemingly in correlation with each other. As the dimension increases, the size of the diamond also increases. Since the diamonds correlate with dimensions, its correlation with carat is also inferable. Thus price has a good correlation with x, y, z and carat size.
Let us focus on price against carat and also the prices of the diamonds for the best feature in a category.There exists a strong correlation factor amidst carat and the dimensional factors x,y,z. Similarly with the increase of carat weight there is an increase in price. Thus it is safe to estimate the relation between carat vs. price to exponential. Diamonds with good clarity, cut and colour often occur at lower prices and vice versa otherwise.
The least median price is associated with Ideal diamonds and this is quite unexpected as the estimate that ideal diamonds are high priced is negated. Moreover the best colour D also has the least price which is marked as an outlier. Further prices go up with cut and falls at Ideal kind of cut.The dimensions of a diamond (x, y, and z) tend to correlate with each other. The longer one dimension, then the larger the diamond. The dimensions also correlate with carat weight which makes sense.The one thing that is same as our guesstimate is that price increases proportionally with carat and also has a good correlated factor with the dimensions x, y and z.

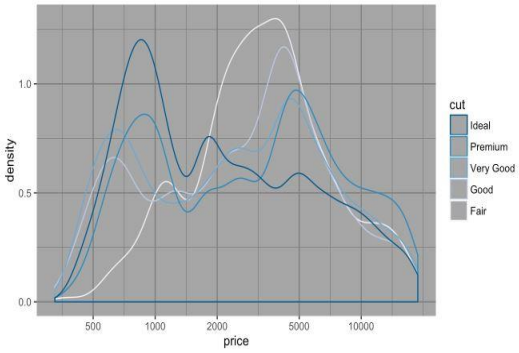

Figure 5. Estimation of prices with guesstimate analysis on density attribute based on cut

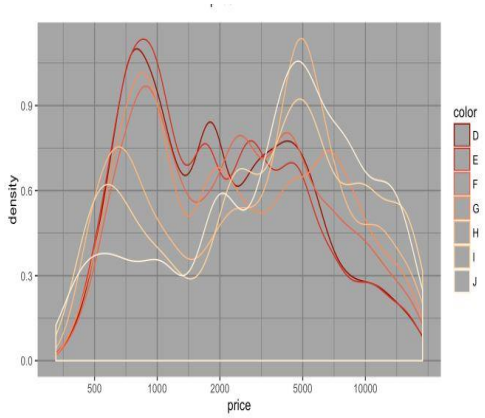

Figure 6. Estimation of prices with guesstimate analysis on density attribute based on Color

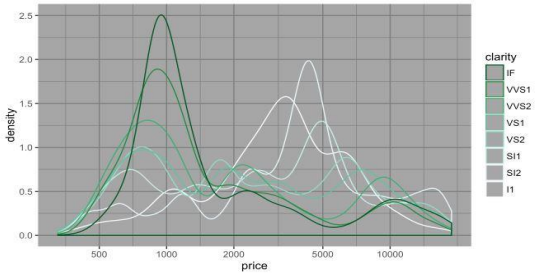

Figure 7. Estimation of prices with guesstimate analysis on density attribute based on density

The above graphs give us a detailed view of the peculiar patterns earlier visualised in the box and scatter plots. Often the diamonds that are on the better scales of colour cut and clarity are the ones that are priced lower compared to the diamonds whose colour, cut and clarity are on the less desired side. The price/carat vs. clarity plot seems to be on the more considerate side as the least clarity diamonds have least price and this price increased as the clarity increased and then once again decreased as the clarity went near the best clarity. This partly can be due to the fact that people prefer impurities in the diamond as a factor for price increase.

Our attempt to find patterns that are unique did not really yield any results with respect to the table variable and thus we can say table is completely independent of the price determination. Thus we shall try to evaluate the plots after transforming the variables. i.e. log transformation of price against the cubic root of the variable under evaluation against another dependent variable.

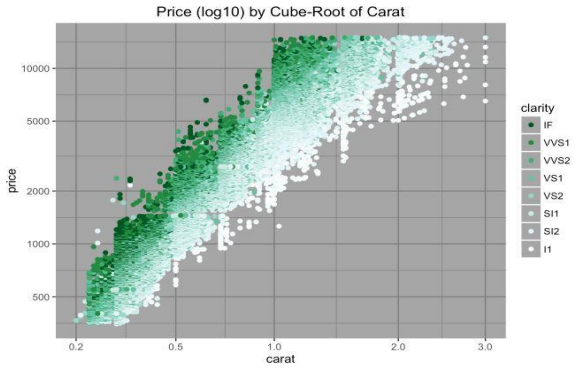

Figure 8. Graphical Scatter plot Visualization of carat vs. Price for cube root analysis on distribution of clarity

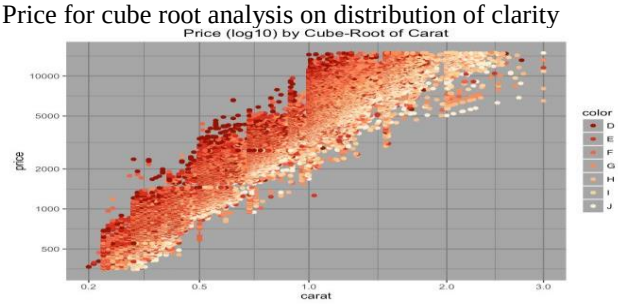

Figure.9. Graphical Scatter plot Visualization of carat vs. Price for cube root analysis on distribution of colour

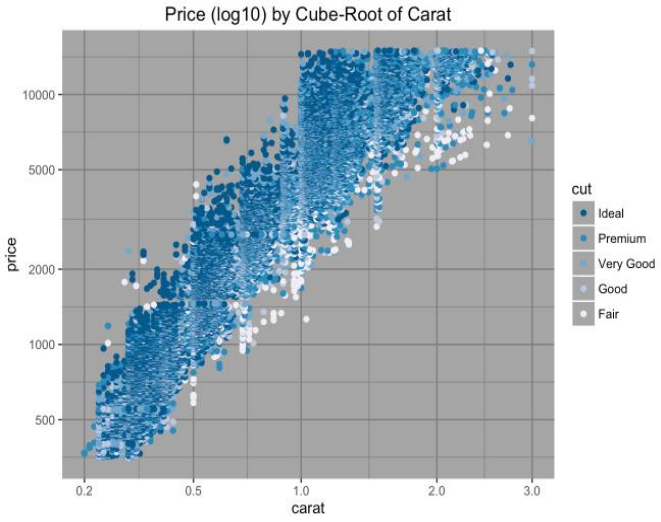

Figure 10. Graphical Scatter plot Visualization of carat vs. Price for cube root analysis on distribution of cut

Diamonds with better colour are priced higher, however due to less number of diamonds sold they are at a disadvantage. Similar pattern is also seen in case of clarity and cut. However not much variation in price is seen due to these factors.

V. CONCLUSION

Diamonds with D colour have less median price and J has highest median price. This is partly due to the fact that the carat plays a major role in the diamond price and we don't have a diamond with D colour and having good carat size. The fact that D has low median price doesn't necessarily mean it is cheap.This is the sole factor which independently determines the price. The bigger the carat sizes the more the price of the diamond.The diamond with the idea cut have low price but this

doesn't mean all the diamonds with good cuts having low price. The price increases as the cut rating increases and doesn't decrease until it comes close to Ideal cut diamond.Diamonds with the ideal cut are the ones with the least median price for a unit carat. Moreover we can see the price increase as the cut becomes better and finally decreases as the cut becomes best.Under constant carats it is seen diamonds with less clarity are in general cheaper than the diamonds with better clarity. In case of colour we have a slight variation in terms of colour D whose price although high has a less median price which states D colour are not the most bought diamonds.

REFERENCES

1. Hastie T, Tibshirani R, Friedman JH. 2001 The Elements of Statistical Learning. Springer: New York.
2. Hoaglin DC, Mosteller F, Tukey JW. 1983 Understanding Robust and Exploratory Data Analysis. Wiley: New York.
3. Morgenthaler S. 2009. Exploratory data analysis. Wiley Online Library. WIRES Computational Statistics DOI: 10.1002/wics.
4. http://mathalope.co.uk/2015/03/11/what-is-the-diamonds-dataset/
5. rstudio-pubs-tatic.s3.amazonaws.com/247043_f40699caa0e24463b11f72a3737b1df8.html
6. https://pankajnath.shinyapps.io/Diamond-ShinyApp/
7. http://www.inside-r.org/packages/cran/ggplot2/docs/diamonds
8. http://diamondse.info- Diamond Search Engine
9. https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Timothy.Thesis.pdf

AUTHORS PROFILE

K .DurgaCharan,
Dhanekula Institute of Engineering & Technology,
Vijayawada, India.

K.Sandhya Rani,
Dhanekula Institute of Engineering & Technology
Vijayawada, India.

SomulaRamasubbareddy,
VNRVJIET, Hyderabad,India.
svramasubbareddy1219@gmail.com

K.Govinda,
VIT University, Vellore,India.

E.Swetha,
SV College of Engineering, Tirupati, India.