1 Abstract

Data visualization is an essential component of data analysis. Understanding the complexity and novelty of the data is the foremost important thing before actual visualization. This paper deal with the implementation of the visualization techniques to understand the complexity and novelty of the data. The use of these techniques helps the users to prepare for the data visualization in a better way as lots of insights can be gathered.

2 Introduction

Data analytics is one of the required fields in today's era due to the growing volume of the data[1]. Due to enormous data volume, this calls for visualization techniques to be utilized for understanding the data[2]. But performing the task of visualization accompanies with the thought of imagining the visualization. This helps in the structuring of the visualization by understanding the importance of the variable in the data, the relationship of the data and complexity of the data. This paper tries to understand the complexity and novelty of the data set by providing inputs to the pre-visualization stage.

3 Task

The task of the project consists of

- Selection of dataset that has complexity, but also should be feasible to do analysis.
- Selection of visualization tool that is generic, the visualization tool should be proprietary to the field of the data set.
- Selection of visualization category, the category "type B" was selected for this project.
- Selection of technical elements for the visualization, for the project, the use of multiple coordinated views was selected.

4 Approach

The essential requirement for the task was to choose a suitable data set. For this task "diamonds.csv" from Kaggle was selected[3]. Since there is penalty of use of visualization of tools that is related to the field of the data, the use of opensource tools like "python", "plotly" and "dash" as selected. The tools that were selected and no proprietary licence and is completely generic in nature where every detail has to be built from scratch. The objective of the task was to analyze the complexity and novelty of the data. The data structure of the table is

Variable	**Type**
carat	Continuous, int
cut	Categorical, string
color	Categorical, string
clarity	Categorical, string
depth	Continuous, float
table	Continuous, float
price	Continuous, float
x	Continuous, float
y	Continuous, float
z	Continuous, float

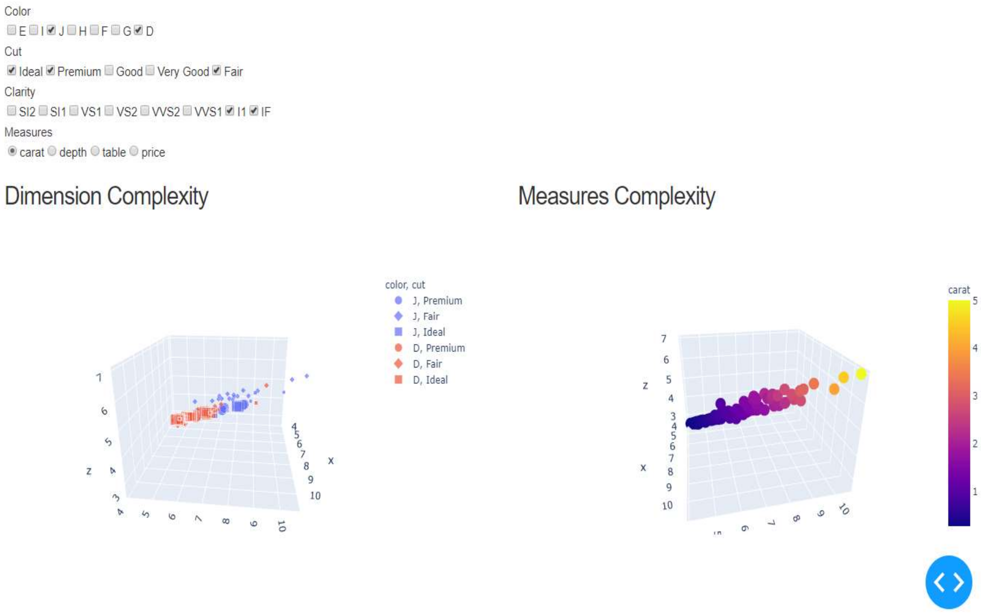
Table 1

For achieving the aim of the task, the following approach was followed to understand the variables.

- Separation of the variables of data into dimensions and measures.
- The dimensions consist of cut, colour and clarity.
- The measures consist of carat, depth, table and price.
- Though x,y,z comes under the category, but it was used to generate the three-dimensional plot.

From the above, we can see that the analysis of the variable is performed. To generate visualization the following steps were followed.

- The data consisted of x,y,z coordinates, which gives the room to perform three-dimensional visualization.
- To perform the visualization data preparation is required, which was done using python.
- The "pandas" and the "numpy" was used to perform data manipulation.
- Plotly is a fantastic visualization tool that used javascript in the background to perform the task.
- Plotly is also available in python; hence this was used to create 3-d plots.
- Since the task was to create an interactive multi coordinate visualization, so the platform to display the output was chosen in a Browser.
- Dash is another application which provides the platform to design the layout using snippets of HTML code and provides server to host the page.
- A new column was generated to quantify the clarity as the size attribute of the plot was So, after preparing the data and selecting the required to tool, the layout design was prepared.



Color
□ E □ I ☑ J □ H □ F □ G ☑ D
Cut
☑ Ideal ☑ Premium □ Good □ Very Good ☑ Fair
Clarity
□ SI2 □ SI1 □ VS1 □ VS2 □ VVS2 □ VVS1 ☑ I1 ☑ IF
Measures
● carat ○ depth ○ table ○ price

Dimension Complexity

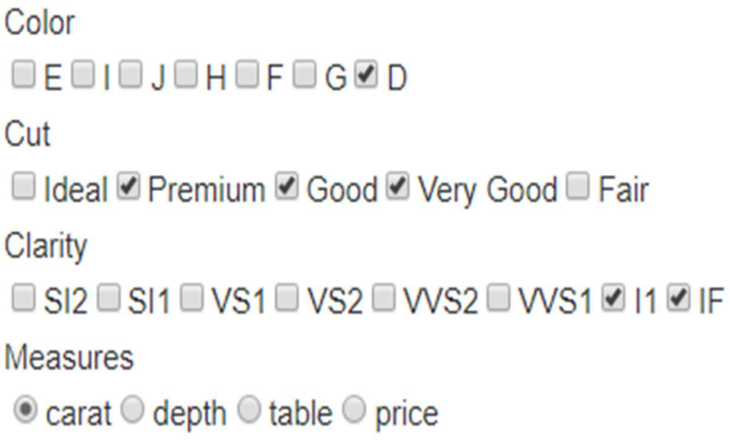
Measures Complexity

Figure 1

The above figure shows the complete layout of the dashboard of the visualization. The following approach was followed to implement the above layout.

- The uni-page single layout was chosen to analyze the complexity and novelty of the data.
- Three HTML "div" tags were created one to hold checkbox and the two plots.
- The two plots were placed row-wise adjacent to each other so that the end-user doesn't have to scroll the page.

- There are three checkbox rows for dimension variables and one radio button rows for measures.



Color

□ E □ I □ J □ H □ F □ G ☑ D

Cut

□ Ideal ☑ Premium ☑ Good ☑ Very Good □ Fair

Clarity

□ SI2 □ SI1 □ VS1 □ VS2 □ VVS2 □ VVS1 ☑ I1 ☑ IF

Measures

● carat ○ depth ○ table ○ price

Figure 2

- Each plot was rendered by "plotly" using "scatte3d" plot function.
- The left plot was named as "Dimensional Complexity" because it visualized about the complexity of the dimension. The color gradient represents the "color" attribute; the symbols represent "cut" attribute", and the size of the symbols represents "clarity". Below is the sample of the plot.

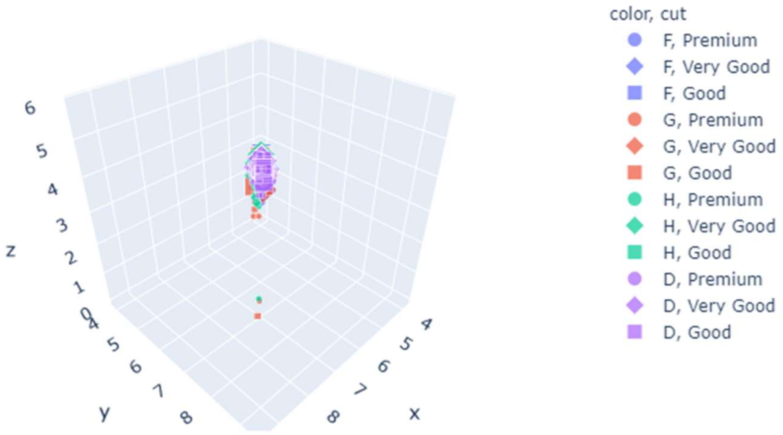


Figure 3

- The right plot was named as "Measures Complexity", as it visualizes the the measures.

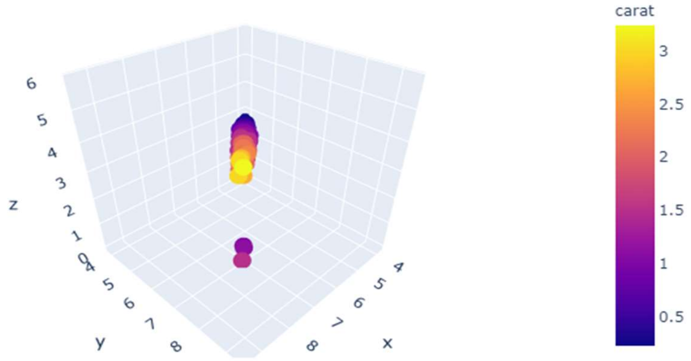


Figure 4

Upon analyzing the visualization, it was found that the dataset has a complex relationship among the variables. Maximum of the diamonds follow in a particular range.

5 Conclusion

The visualization was created using open source tools. The novelty about the visualization is usually the visualization is performed in 2-D canvas, which has certain limitations. Certain aspects of the value get missing. The visualization is interactive in which we can select the parameters, and we can study the change in the complexity of among the variables. Both plots updates automatically when the different settings are set in the checkboxes and radio button. The main striking features of this visualization is that it is three dimensions. So three-dimensional plots open up new findings of the data. The features of pan and zoom in the plots help to analyze the local points, which helps to capture the hidden information. As the visualization is interactive, so it opens up a whole new world of possibilities and provides greater customization to the end-user.

Please visit https://github.com/Ashishvikram/A3_Noval_Visualization for the full source code and visualization.

6 References

[1] J. F. Tripp, “Data Visualization,” in *International Series in Operations Research and Management Science*, 2019.

[2] R. Earnshaw, “Visualization,” in *Advanced Information and Knowledge Processing*, 2019.

[3] S. Agrawal, “Diamonds,” *2017*. [Online]. Available: https://www.kaggle.com/shivam2503/diamonds. [Accessed: 01-Apr-2020].