DATA INSIGHT AND INFORMATION VISUALIZATION METHODS

[1]Prof. Jincy C Mathew , [2] Sunil L
[1] New Horizon College Of Engineering, Bangalore-560103, Karnataka, India

Abstract—There are too many visualization techniques which may confuse when and what should be used for the problem understanding. The basic purpose of visualization helps to communicate information and patterns to other, they can explain data using a small space and it can be easily shared with others. Visualization can tell more stories compelling and immediate way which can be more memorable than words. It is way of visually communicating information about quantitative and qualitative data in nature. The purpose of this document is to collect the different type of visualization techniques with the brief introduction.

Keywords: Visualization Technique, Visualization Communication, Quantitative, Qualitative.

I. INTRODUCTION

As the volume of the data continues to increase in the past few years, organizations for the first time are beginning to use visualization techniques to find insights of the data. Visualization plays a vital role in the life cycle of data analytics project. The data needs to be presented in the graphical or visual form. So that it can be reached to the audience and it can be only possible using visual, graphics, charts and plots. Visualization is a technique used to where data are represented in a pictorial way, it gives a clear insight of data in visual format by using simple charts such as graphs, bar, charts which illustrates data clearly and enable audience to understand the value of the insights. This allows people to portray data in more easily understandable way. In addition data analysts and data scientists can use data visualization technique to interact with the explored data. Although there are many types in data visualization, several fundamental types of charts to portray data and information. It is important to know when to use a particular type of chart or graphs to express a given kind of data. The objective of this is to find the best chart for expressing the data clearly.so that message can be conveyed in an intended way.

II. DATA VISUALIZATION WORKFLOW

Fig. 1 Visualization Life Cycle

2.1 Discovery / Acquisition

`The first thing you have to do before you solve the problem is to define exactly what it is. You need to transfer the data into questions into actionable. You should collect the inputs from the people who have problem. For example we need to solve a problem of an any organization. First we should start by understanding their goals and underlying why behind their data questions. Before you start thing of solution, you'll want to work with them to clearly define the problem. Once you defined the problem, you'll need data to give you

the insights needed to turn the problem around with the solution. This part of the process involves thing through what data you'll need and finding the way to get that data, whether it is from internal database or external datasets.

2.2 Cleaning / Munging

Now that you have all the raw data, you need to process it before you start the analysis. Often data can be quite messy, especially if it hasn't been well maintained. The data can contain redundant values, null values we can see the errors in data which we have extracted. We should clean the data such that it does not contain any outlawyers in it. We should go through and check the data to make sure we'll get accurate insight. You'll need to look for aggregation of your file rows and columns and ample test values to see our values make sense. If you detect something that not make any sense, you'll remove that duplicate and replace with default values. Once done with the cleaning of data, that is ready for further analysis.

2.3 Analysis / Exploratory Visualization

After getting the data cleaned we start analyzing the data set to summarize the characteristics, often with the visual methods or technique. Search for the evidence using all the tools available. Some of the applications are Tableau/Tableau public, Microsoft Office, Open Office, Gephi, Node XL(plug-in for excel),Spotfire, R, other web services like google spreadsheet, IBM ManyEyes etc. We apply statistics on the data set to analyze how accurate the data is, which focus on checking assumptions required for model building and hypotheses testing. In this phase we transform data set to an visual format of pictorial representation of graph which is easily readable by human mind. There are many visualization techniques like box plot, heat maps, pie chart, bar graph, line chart, scatter plot, histogram, word cloud(for text analysis) etc..

2.4 Publishing

After done with the exploratory visualization we create graphical user interface for the give data. This tells the story about the data. We can make better decision by looking at the data graph. We create an interactive interface or web site to publish it. Created data visualization is placed on the website for the use of end users.

III. VISUALIZATION METHODS

We know that number of visualization techniques have been developed last few years due to represent the huge amount of information and to analyze it. These techniques have many features like usability, Interactivity, interface features etc., which make them easy to use. This method has mechanism to achieve the goal of visualization as briefly. The main focus is on Data and information visualization techniques.

3.1 Data Visualization

Data visualization is one of the most important part of data analysis. It has always been important to present the data in an understandable visually appealing format. Data visualization is one of the skill that data scientist have to master in order to communicate better with the end user.

Let me first give you an idea of the kind and the amount of data that we are dealing with. According to silicon angle, there was 2.5 zetabytes of stored data world over in the year 2012 and it is set to hit more than 50 zetabytes mark by the end of 2017. To put things in perspective, this data has been largely been produced by websites and cross platform transactions. Fact that there would be a total of 20 billion "smart" devices connected to the internet by the end of 2020 and the number can be baffling!

Data visualization allows data scientist to converse with their end users. The outcome of the data analysis is not immediate comprehensible to people who do not directly deal with data. Data visualization bridge the gap and make people appreciate the possibility of data analysis.

3.1.1 Pie Chart

Pie chart is also called as circular graph, in which the circle is divided into sectors. Compare proportion of each category, where each category is show as the slices of the pie. A pie chart is also most commonly misused kind of chart. I situation calls for using a pie chart, employ it only when showing only 2-3 items in a chart, and only for sponsor audiences. Pie chart is used to determine the size of data compared to other data wedges. There are two common variation i.e. Standard or 3d pie chart and Exploding pie charts.

In 3d pie char we can see in three-dimensional view, which is more attractive than the normal pic chart. In Exploding pie chart the divided segments are fare from each other. Color usability make it easier to understand, different color is used for slices and segments.

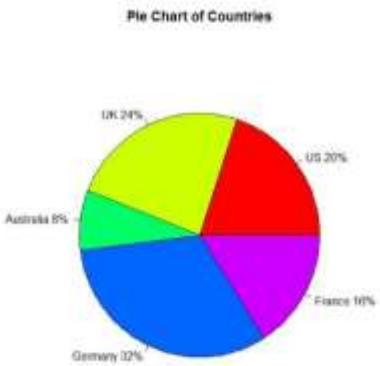

Fig.2 Normal Pie Chart

Fig.3 Standard(3D) Pie Chart

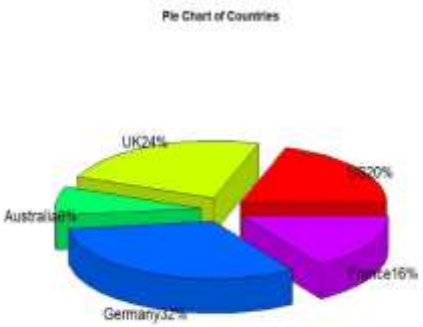

Fig.4 Exploding Pie Chart

3.1.2 Bar Graphs

One of the most frequently use data visualization method is bar graph, it is also called as bar chart. Bar charts are used much more often and are useful for showing comparison and trends over time. Even though people use vertical bar chart more often, horizontal chart allows an another more room to fit the text labels. Vertical bar charts tend to work well when the table are small. Such as when showing comparison over number of gears using frequency. There are many variations i.e. Vertical Bar graph, Horizontal Bar graph, Stacked Bar plot, Grouped bar plot and many more. They can use according how they are going to represent their data in more precise way.

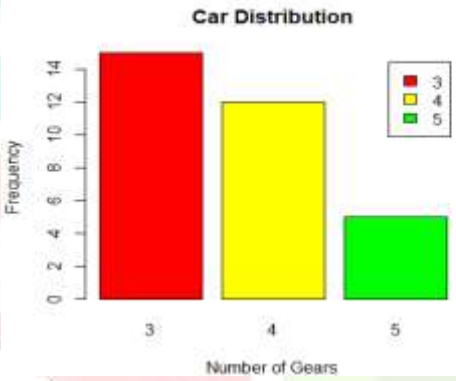

Fig.5 Vertical Bar Graph

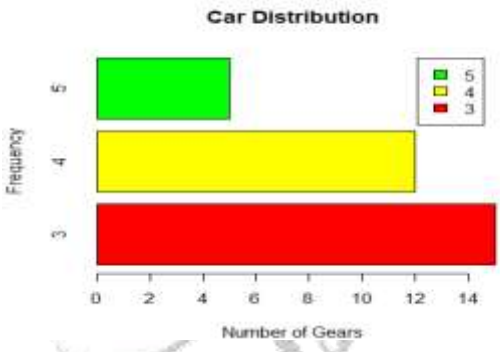

Fig.6 Horizontal Bar Graph

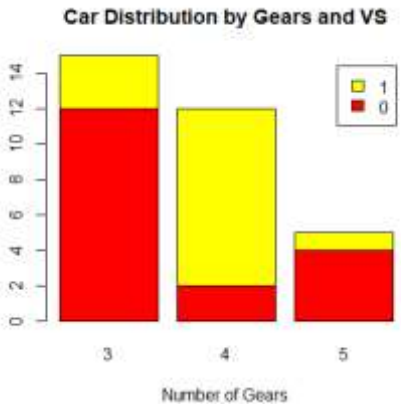

Fig.7 Stacked Bar Plot

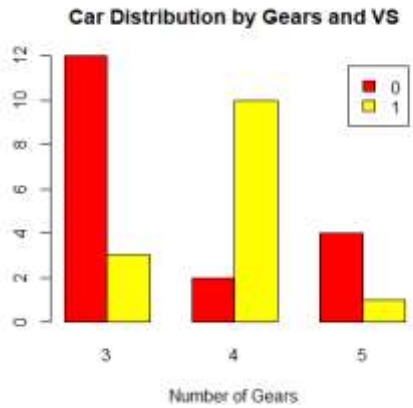

Fig.8 Grouped Bar Plot

3.1.3 Histogram and Kernel Density Plot

Histogram is common and vital technique used in statistics and data analysis. Histogram are useful for demonstrating the distribution of data to an analyst audience or to data scientist. It is use for the distribution of the continues variable or data. Data distribution are typically one of the first steps when visualizing the data to prepare for model planning.

Kernel Density plot is another way of representation histogram in the form of curve. It is plotted on the continues variable. In Fig.10 we can see the highest peak where most of the cars lie between 15 to 20 in the plot.

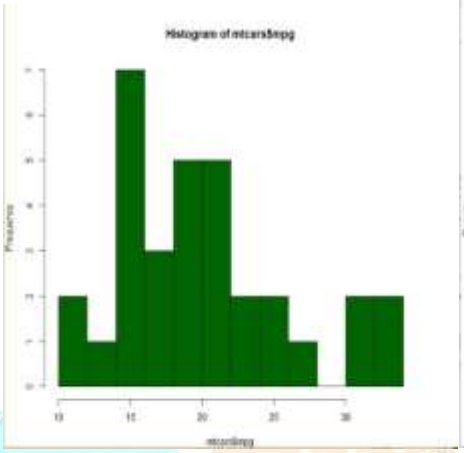

Fig.9 Histogram

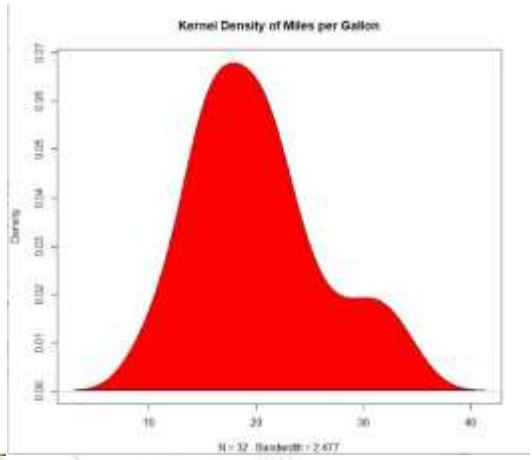

Fig.10 Kernel Density Plot

3.1.4 Box Plot

Tukey proposed the idea of boxplot to visualize the distribution of values. Another useful univariate graphical technique is the boxplot. The box plot described here in its vertical format, which is the most common, but a horizontal format also is possible. Box plots are very good at presenting information about central tendency, symmetry and skew, as well as outlier, although they can be misleading about aspects such as multimodality. One of the best use of box plot is in the form of side-by-side boxplot.

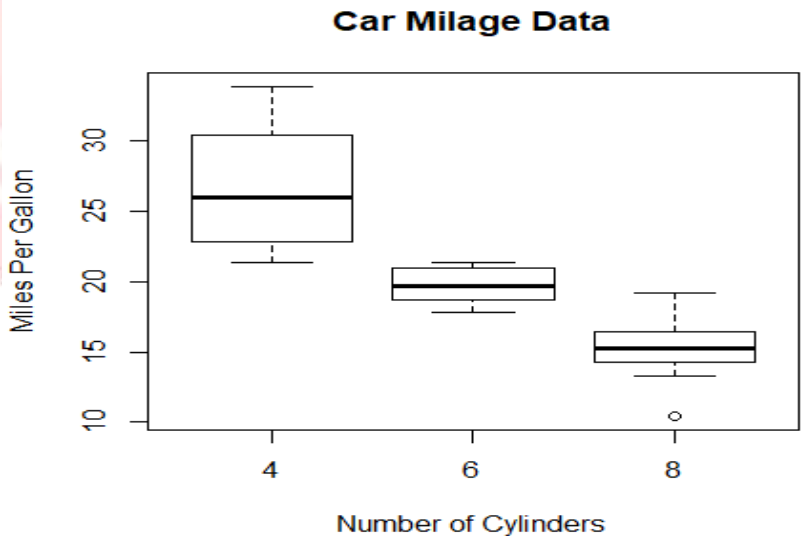

Fig.11 Box Plot

3.1.5 Heat Map

Heat map is a graphical representation of data where the values of individual in a matrix are represented by colors. This is used to check how frequently the data are present in a row. In the Fig. 12 we can see that Paul Millsap is good in all the sports, so we see dark colors in the 12th row of heat map. Then he might be hired in the selection because his good in all the respective games.

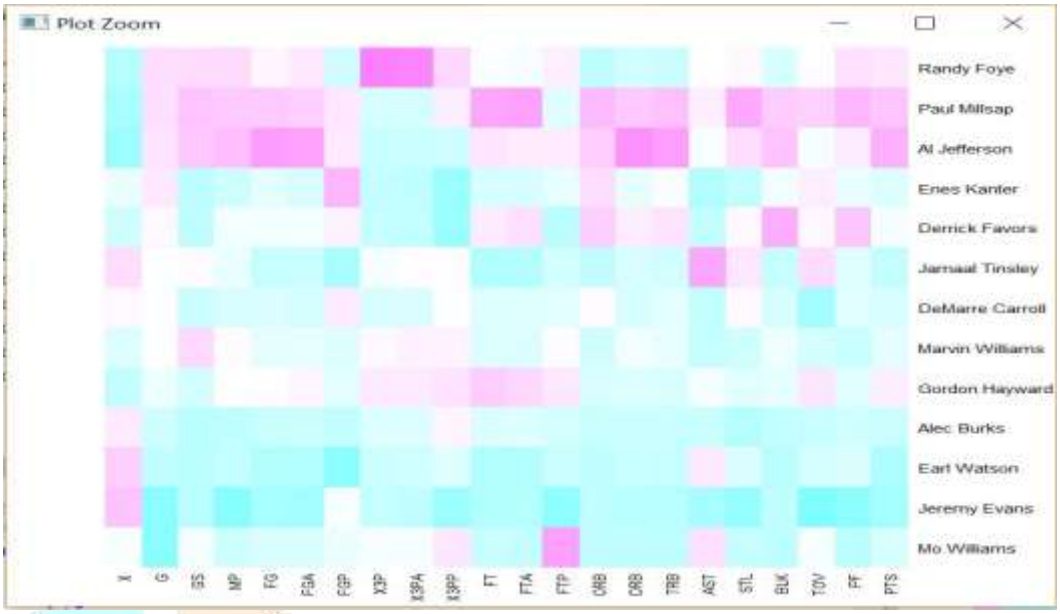

Fig.12 Heat Map

3.1.6 Word cloud

Word cloud is also called as tag cloud or weighted list, is a visual depiction of frequency of word in any selected written material, such text book notes or internet website. In word cloud we see how frequently the words are repeated, the more repeated word is the bigger in size and placed in the center of the word cloud. In Fig.12 we can see that will is repeated more number of times. This word cloud is developed on Narendra Modi's speech.

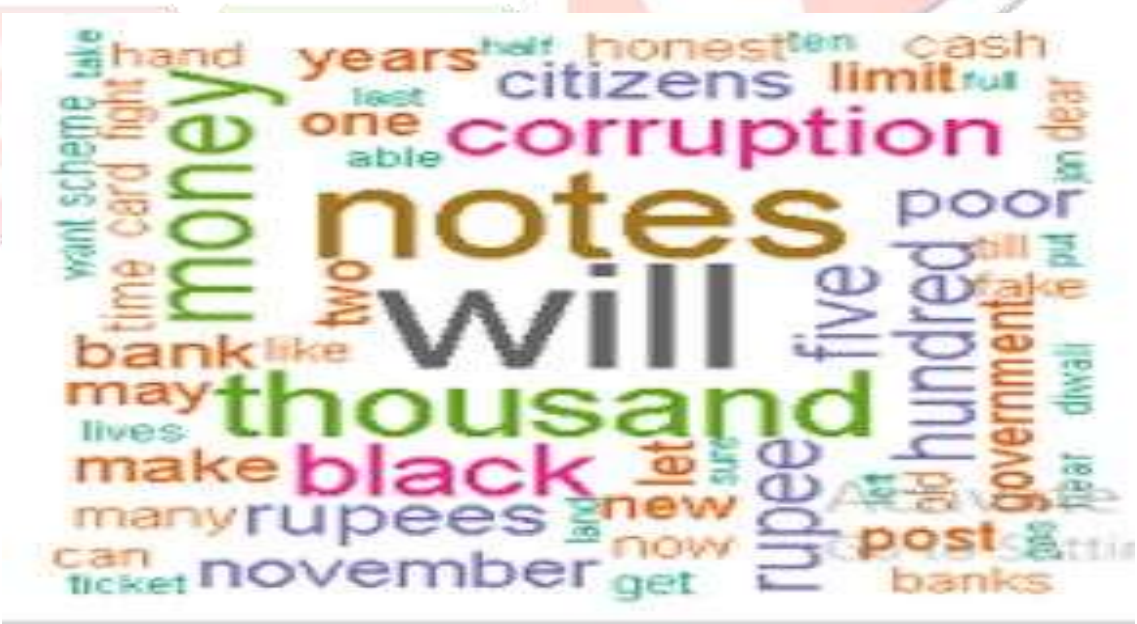

Fig.12 Word Cloud

IV. CONCLUSION

The main aim of this documentation is to initiate the basic idea of visualization for further generation. In business organization they mostly ignore, misunderstand, used in insufficient data and information visualization, which leads to less production. So, need to concentrate more on visualization technique in order to get maximum benefit of visualization. Another important thing is to specify what data type need to represent to which visualization method which can be more understandable. As data mining is one of the leading field of computer science, largely used in business organization to produce results in the form of visual data mining. Interactive and

animative is more precise to convey the conclusion of huge and very huge amount of data set that is not easily understood by traditional approach.

REFERENCES

[1] P.C. Wong et al. (2003). "Dynamic Visualization of Transient Data Streams", Proc. IEEE Symp. Information Visualization, IEEE CS Press, 2003, pp. 97-104.

[2] Patrick Mader, Jane Cleland-Huang, "From Raw Project Data to Business Intelligence", *IEEE Software*, vol. 32, no. , pp. 22-25, July-Aug. 2015, doi:10.1109/MS.2015.92

[3] C. Chen, (2004). "Information Visualization: Beyond the Horizon", 2nd ed., Springer, 2004

[4] C. Johnson, (2004). "Top Scientific Visualization Research Problems", IEEE Computer Graphics and Applications, vol. 24, no. 4, pp. 13-17, July/Aug. 2004.

[5] E.R. Tufte, (1997). "Visual Explanations: Images and Quantities, Evidence and Narrative", Graphics Press, 1997.

[6] Stephen Few, (2007). "Save the Pies for Dessert", Perceptual Edge Visual Business Intelligence Newsletter August 2007.

[7] Robert Kosara, Helwig Hauser, and Donna L. Gresh, (2003). "An Interaction View on Information Visualization". Published in 2003.

[8] P.E. Hoffman, (1999). "Table Visualizations: A Formal Model and Its Applications", doctoral dissertation,

[9] D.M. Butler, J.C. Almond, R.D. Bergeron, K.W. Brodlie, and A.B. Haber, (1993). "Visualization Reference Models", Proc. Fourth IEEE Conf. Visualization, G.M. Nielson and D. Bergeron, eds., pp. 337-342.

[10] North, C. (2005). "Toward Measuring Visualization Insight. IEEE Computer Graphics and Applications", vol. 11, no. 4, pp. 443-456

[11] S. Card, J. MacKinlay, and B. Shneiderman, (1998). "Readings in Information Visualization: Using Vision to Think". Morgan Kaufmann.

[12] Sandro Fiore, Cosimo Palazzo, Alessandro D"Anca, Ian Foster, Dean N. Williams, Giovanni Aloisio "A big data analytics framework for scientific data management" 2013 IEEE International Conference on Big Data, 978-1-4799-1293-3/13/$31.00 ©2013 IEEE.

[13] Umeshwar Dayal, Malu Castellanos, Alkis Simitsis, Kevin Wilkinson, "Data Integraion Flows for Business Intelligence", ACM 978-1-60558-422-5/09/0003, March 24-26, 2009.

[14] Plaisant C, (2004). "The Challenge of Information Visualization Evaluation", Proceedings of AVI 2004: 6th International Conference on Advanced Visual Interfaces, ACM Press, New York, 2004, pp. 109-116