ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Data Visualization Tools: Insights Collection and Decision Making

Akash Rajendrakumar Patil[1]

M.S. Student, Department of Information Technology and Management, University of Texas at Dallas, Texas, United States of America[1]

ABSTRACT: Data visualization is a proficient way to convey information using different visual encoding like dimensions of the plane, size, value, texture, color, orientation, shape and visual elements like charts, graphs, maps, animation, box plot, scatter plot, small multiples and network diagrams. It helps us to gather insights on our data which leads to new interesting observations and patterns or sometimes triggers some actions. It plays a significant role in decision making. In this paper we will systematically understand how various data visualization tools can be used to generate diverse visualizations and how we can use them in the decision-making process.

KEYWORDS: Tableau; Power BI; D3; R Studio; insights

I. INTRODUCTION

Data visualization has a procedural approach. The first step is to define the problem domain or the objective behind the visualization. This makes it easy to understand the scope of the visualization. This step is imaginative and cannot be automated. The second step is to gather and structure the data. It is very important to format the raw data efficiently as this data is the source of the visualization. Quality data acts as a medium to form great data visualizations.
The third step is to adopt a proper processing technique which will ensure maximum visual efficacy and mobility of image. The fourth step is to provide simplification using people context. This step can be automated. It ensures that the visual elements used in data visualization are easily interpreted by the people. The final step is communication. This is a very crucial step as it helps people to gather insights. We must make sure that our audience are able to gather information quickly.
For example, let us consider stock market, we can denote the decrease in prices of stock as red color and increase in prices of stock as green color. This acts as a metaphor to the traffic signal where green is to go and red is to stop. In this example we have used color encoding. Let us take another example. Consider an interesting topic shared on social media worldwide and many people all over the world are providing their opinion on that topic. People from United States of America have provided many more suggestions as compared to people from any other country with India in the second spot and Australia in the third spot. So, we will make the size of United States of America the biggest. Here we have used size encoding. Hence, by using different visual encoding along with people context, data can be displayed efficiently and effectively to the audience.

II. DATA VISUALIZATION USING TABLEAU

Tableau uses an exploratory data visualization approach. Drag and drop operations are used to create data visualizations. Tableau is available in multiple forms Tableau Desktop, Tableau Server, Tableau Mobile, Tableau Public and Tableau Online. In Tableau Desktop we create the worksheets, dashboards and stories. Connect to multiple data sources and flat files. Tableau Server is used to publish the visualizations created in Tableau Desktop. It is also used to merge and allocate answers for business intelligence. Tableau Mobile can only be used to manipulate the visualizations created on Tableau Desktop. Tableau Public is used to publish visualizations for the public. You can provide your views on any blog or website. Tableau Online is basically a cloud solution provided by Tableau. In this

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

paper we will be discussing about Tableau Desktop. In Tableau there are large number of file types and server available for connection.

A. *Analysis of a Super Market*

The data source is an Excel file. It has information about geographic data (city and country), time data (date of observation), product information (name, quantity) and price (standardized quantities, prices converted to comparable currency).

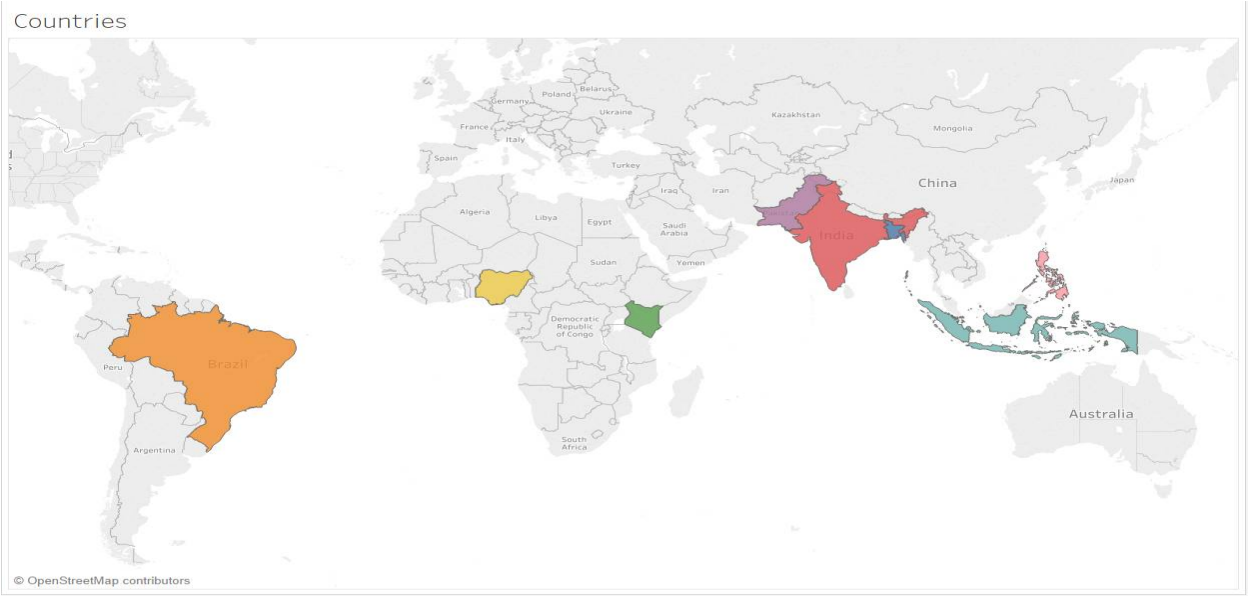

Fig.1: Countries where supermarket is present

Fig.1 talks about the different countries where the supermarket is located. Here there is color encoding, every country is denoted with a different color.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

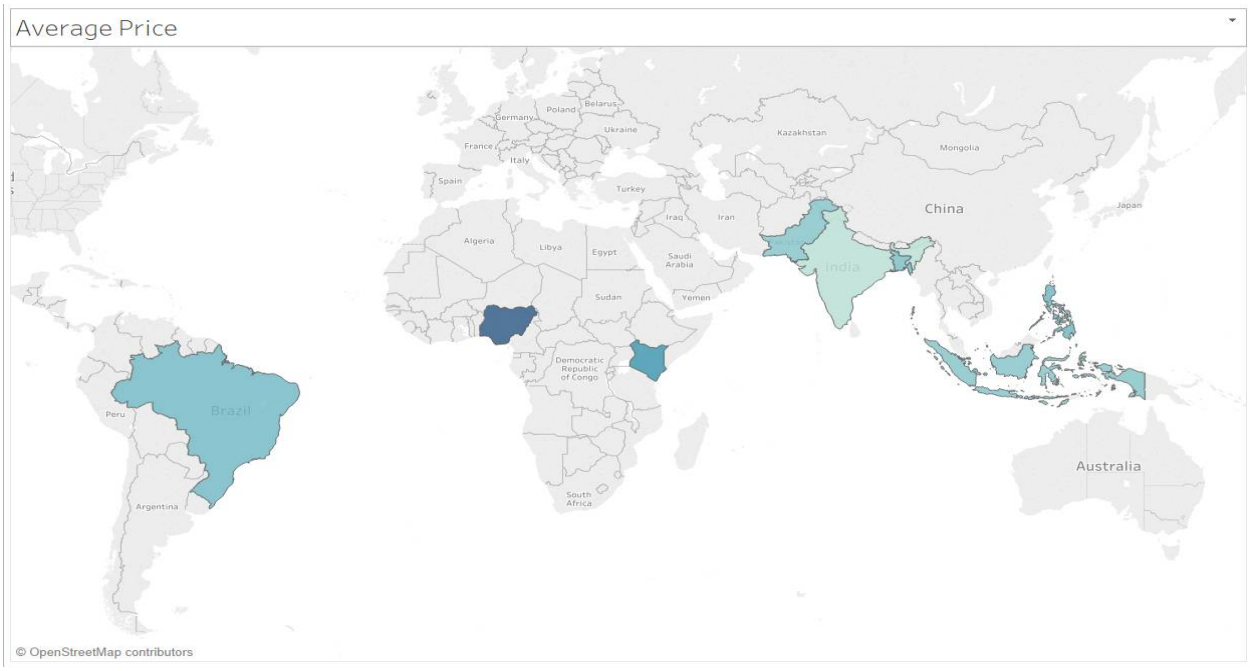

Fig.2: Average Price Variation of Products

Fig.2 shows the average price variation of products. Here there is value encoding. Darker the color, higher the average price. Nigeria has the highest average price of products as it is the darkest.

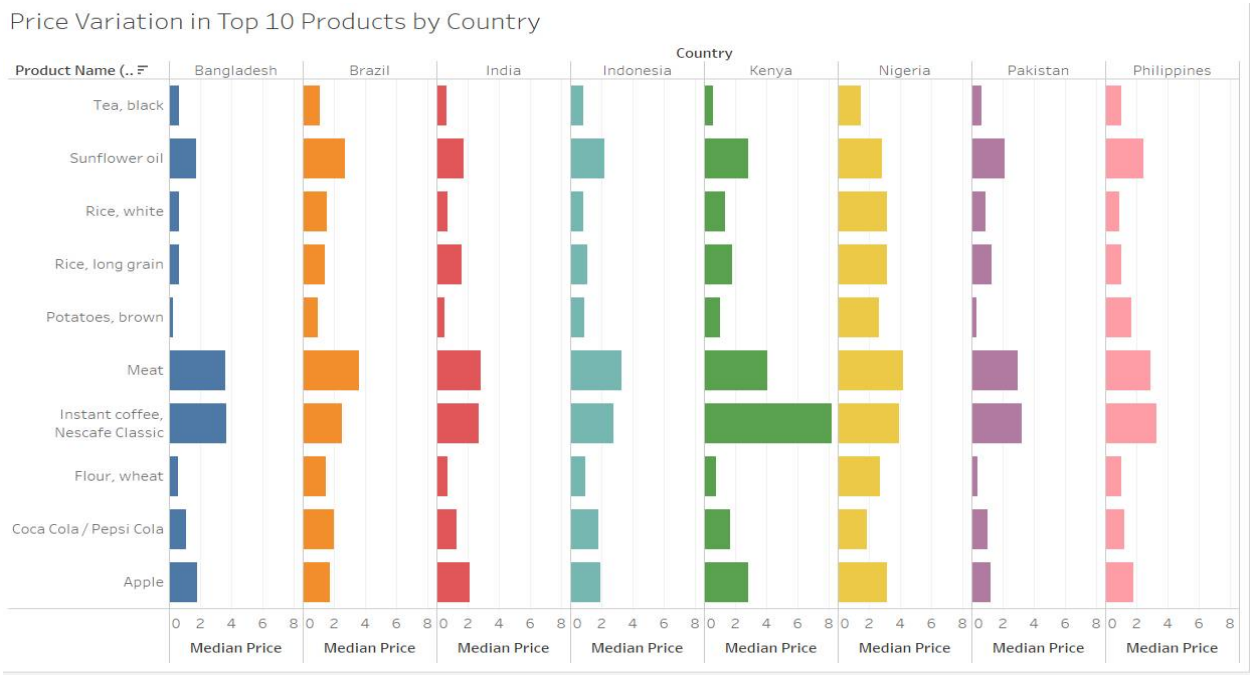

Fig.3: Median Price Variation

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Fig.3 gives us information about median price variation. Here there is color encoding. Every country is represented with a different color. Here, the size of the bar denotes the average products sold in that country which is size encoding. Here we gather insights about which products sells the most in which country. For example, the product Instant coffee, Nescafe Classic is in high demand in Kenya.

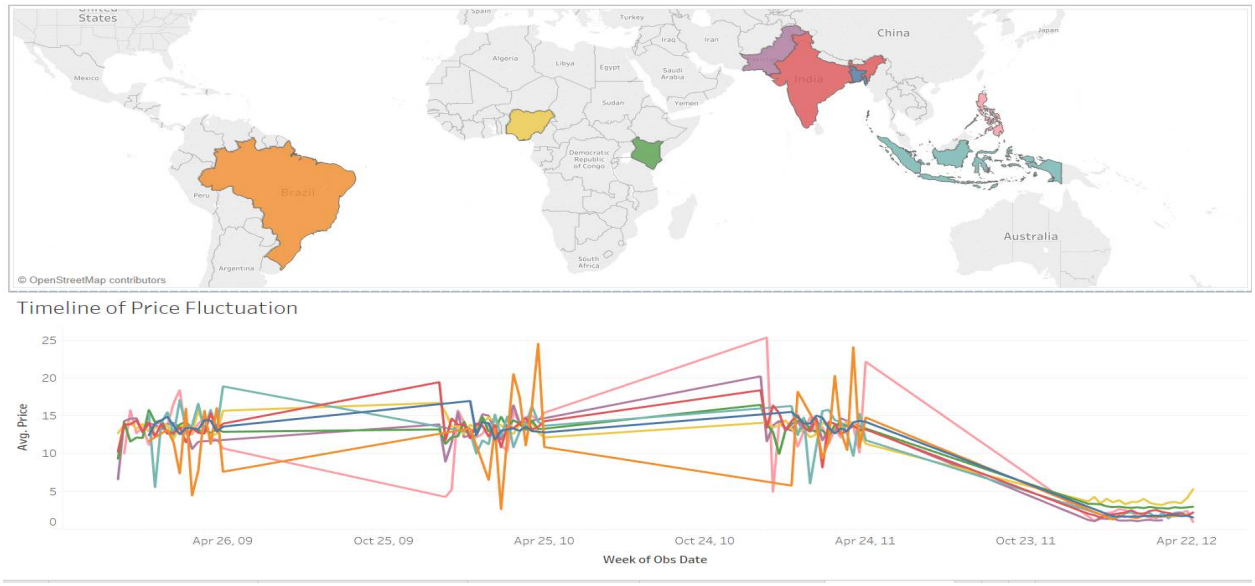

Fig.4: Price Fluctuation

Fig.1, Fig.2 and Fig.3 were worksheets. Fig.4 is a dashboard. Dashboards are created using the worksheets. It depicts the countries where the products are sold and the how the price fluctuation takes place over time from 2009 to 2012. Here also there is color encoding as every country is represented with a different color. This can be used in trend analysis to understand how prices of products have changed over a course of time.

Fig.5: Average Products Sold

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Fig.5 is a story. A story is created using the worksheets and dashboards. There is color encoding as every country is represented with a different color. There is size encoding as well. The size of the bar denotes the average products sold in that country. Here we can easily interpret which product is sold on a larger scale in which country as we have highlighted them.

B. *Impact of Natural Disasters*

The data source is an Excel file. It contains information about the refinery name, state, disaster type, title, county and date in the United States of America. Here we will be majorly focusing on shell refinery.

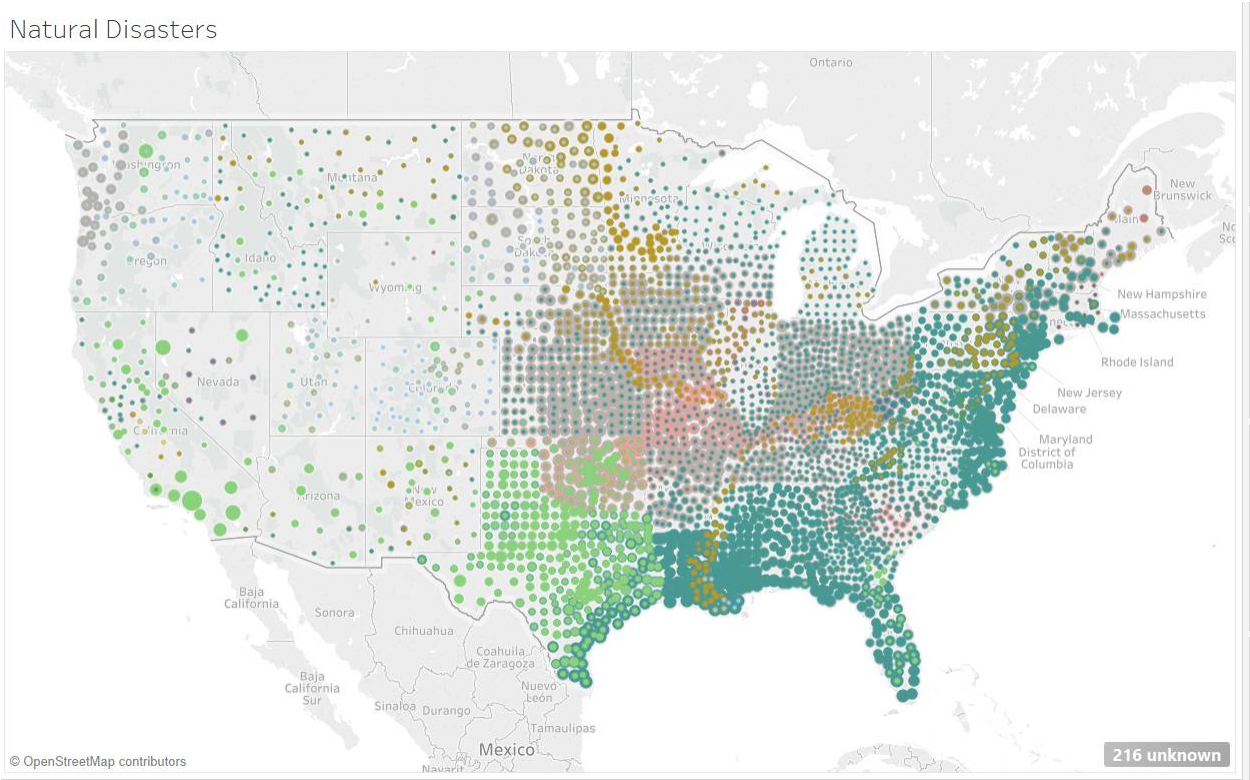

Fig.6: Different Natural Disasters in the United States of America

Fig.6 shows all the natural disasters which have been taking place in the United States of America over the course of ten years. Here there is color and size encoding. Every natural disaster is represented with a different color. The size of the circle denotes the impact of natural disaster in that area.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

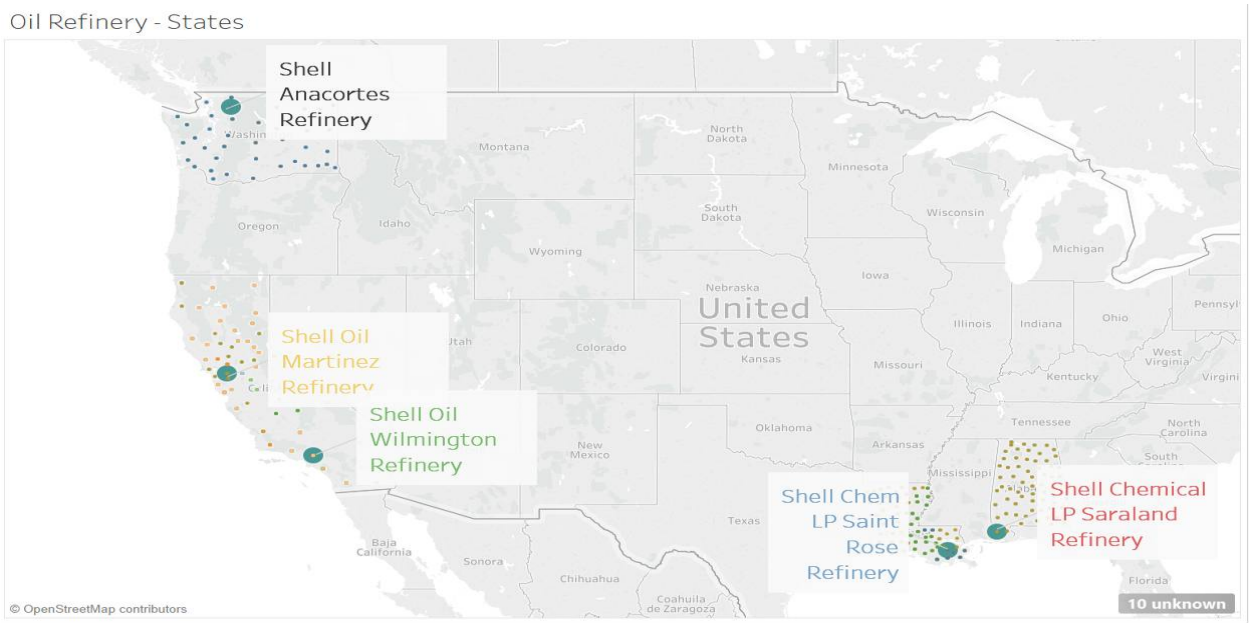

Fig.7: Location of Refineries

In fig.7 we have listed 5 of the shell refineries. This figure gives us insights about where the refineries are located and the disasters by which they are affected. It helps to understand where it is profitable to continue business and which area is affected the most by which natural disaster.

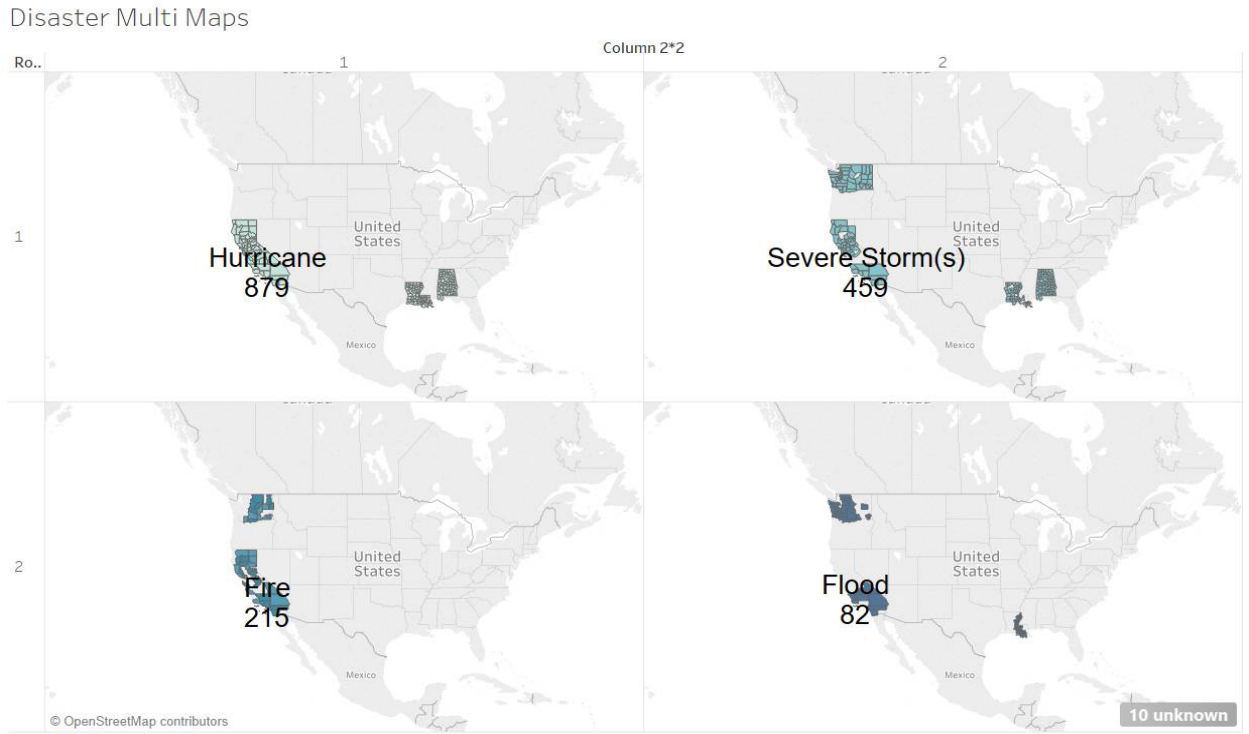

Fig.8: Top Disasters

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Fig.8 is a small multiple. Small multiples have same visual encoding and are different partitions of same dataset. They are commonly used for sharp side by side comparison. Here we come to know about the four disasters and which areas they predominantly affect.

C. *Titanic Viz*

Thedata source is an Excel file. It gives us information about the passengers, survived, name, sex, age, ticket, fare, cabin and embarked. Here we will be forming clusters, which will help us determine how many people survived and by how much in each cluster. So, for clustering purpose we first write the code in R Studio for kmeans clustering.We will connect R Studio to Tableau using the Rserve package. Calculated fields are created in Tableau so that we can invoke the functions and models in R Studio.

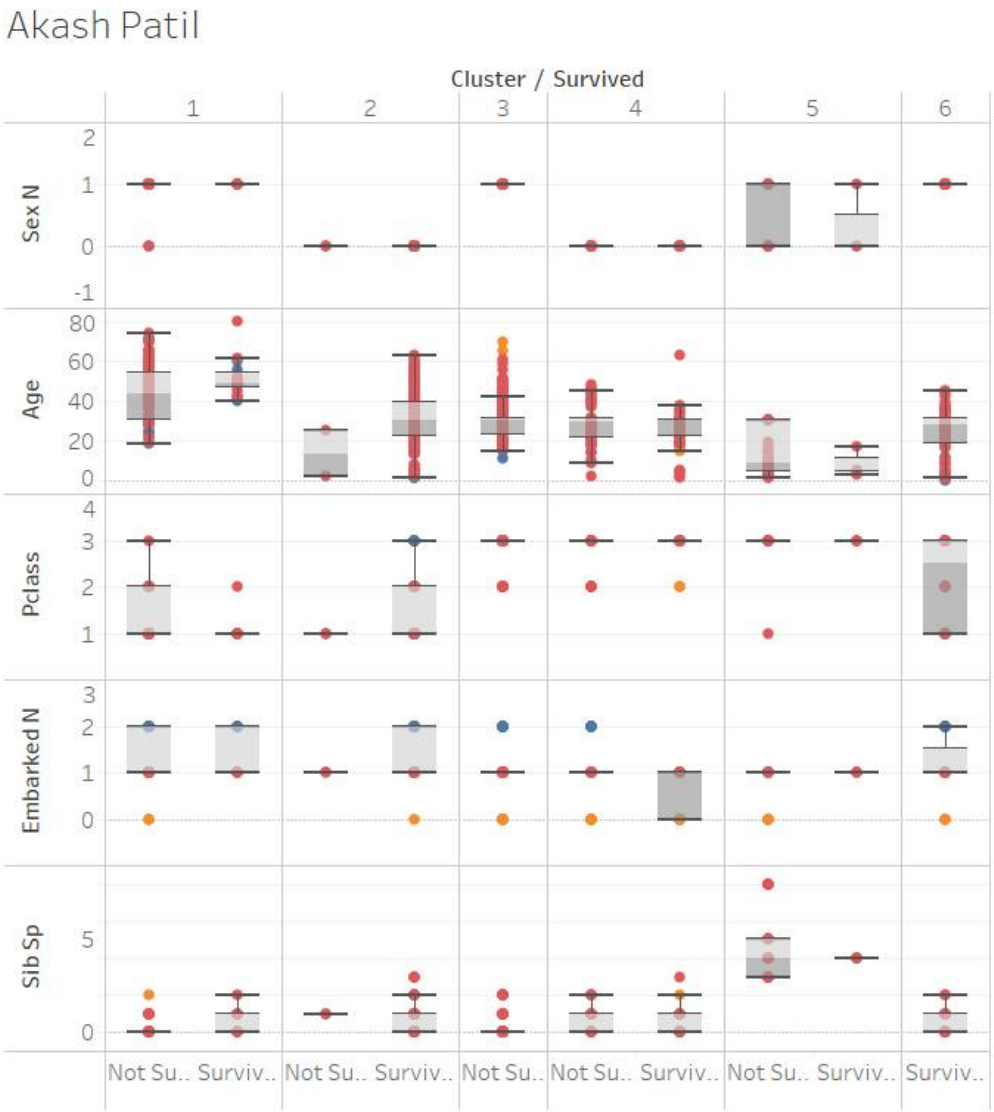

Fig.9: Cluster and Parameter Relationship

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

In fig.9 we come to know that there are 6 clusters. It shows us how many people survived and not survived by using different parameters like Embarked N, Siblings, Passenger Class, Age and Sex. This is a box plot. It is predominantly used to calculate the interquartile range. Here we can easily determine the minimum, maximum and average(median) number of survivors in each category.

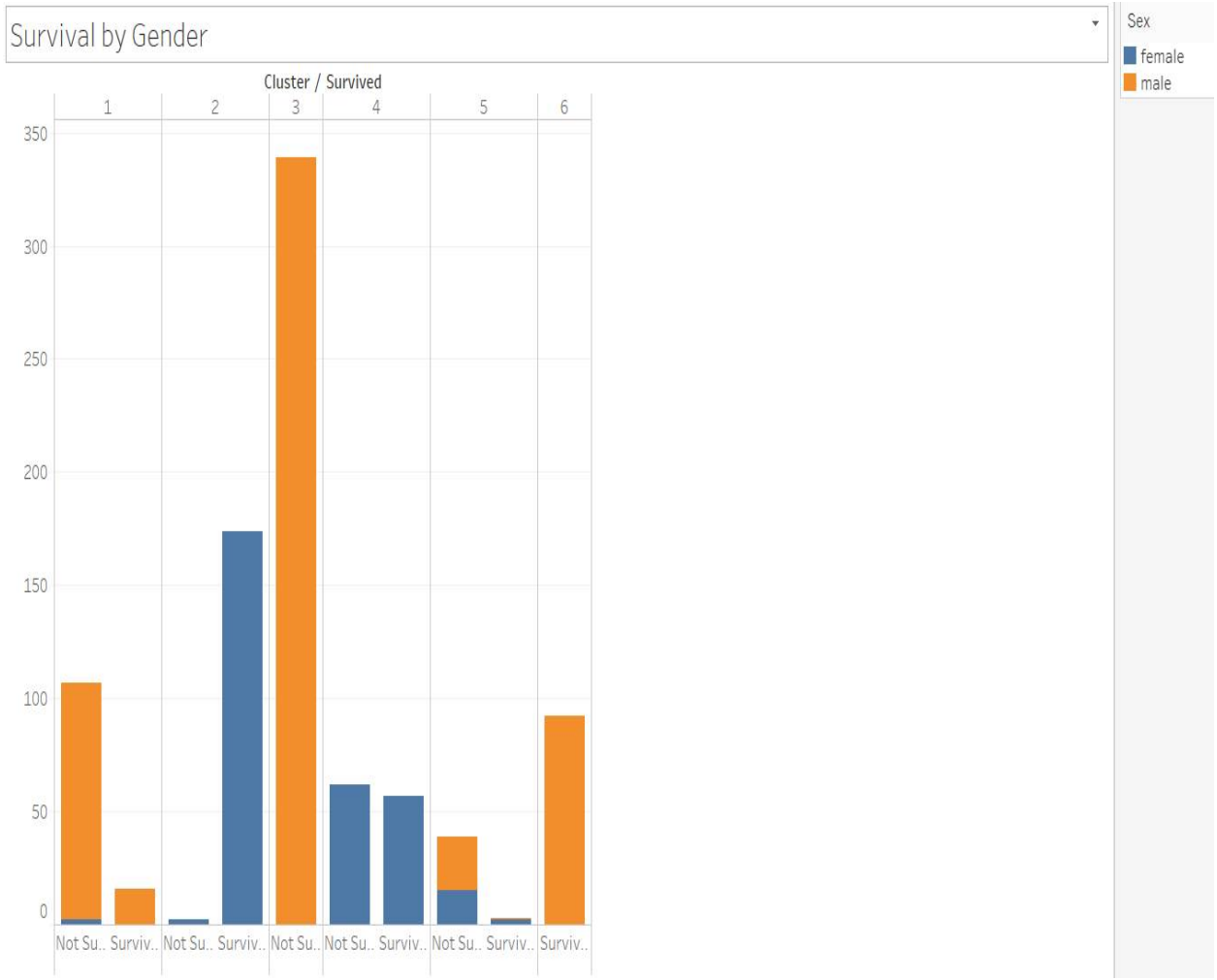

Fig.10: Observation using Sex Parameter

In fig.10 we are displaying the number of people survived and not survived using only the sex parameter. The blue color denotes female and the yellow color denotes male. Here we have used color encoding. From this figure we can gather many insights. Cluster 2 and cluster 4 has only female survivors. Cluster 1 and cluster 6 has only male survivors.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

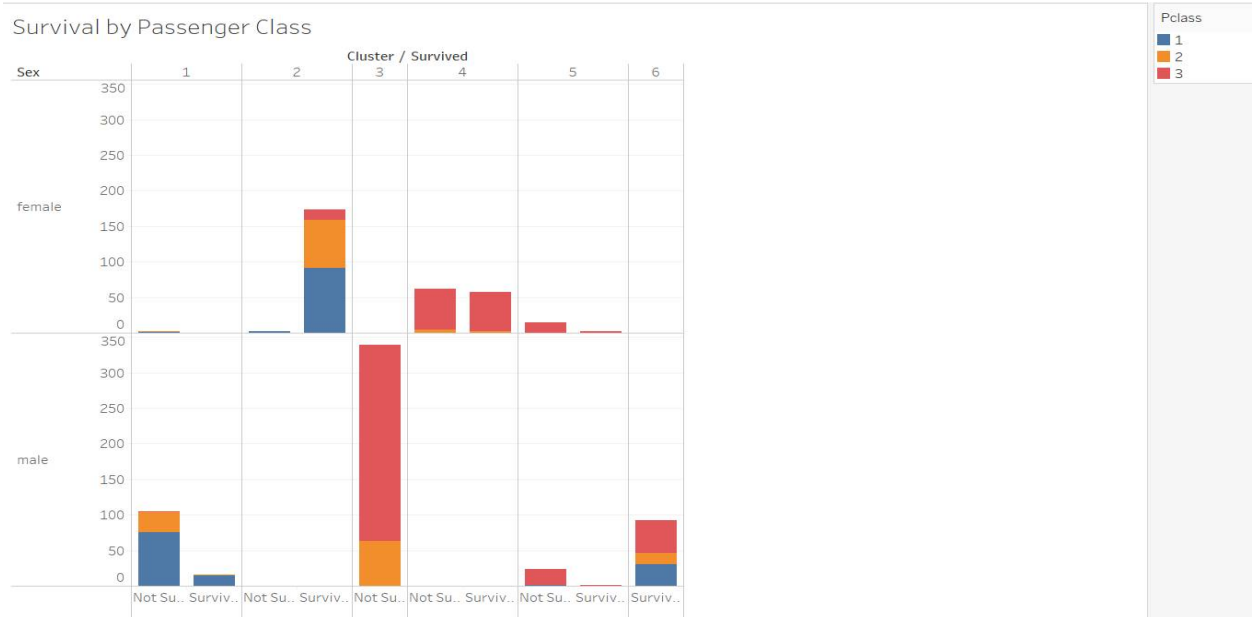

Fig.11: Observation using Passenger Parameter

In fig.11 we are displaying the number of people survived and not survived using only the passenger class parameter. The blue color denotes first class, yellow color denotes second class and red color denotes third class. Here we have used color encoding. In cluster 2 there are no male passengers. There are no survivors in cluster 3. Maximum number of female survivors are present in cluster 2. Maximum number of male survivors are present in cluster 6. Cluster 6 has no female survivors.

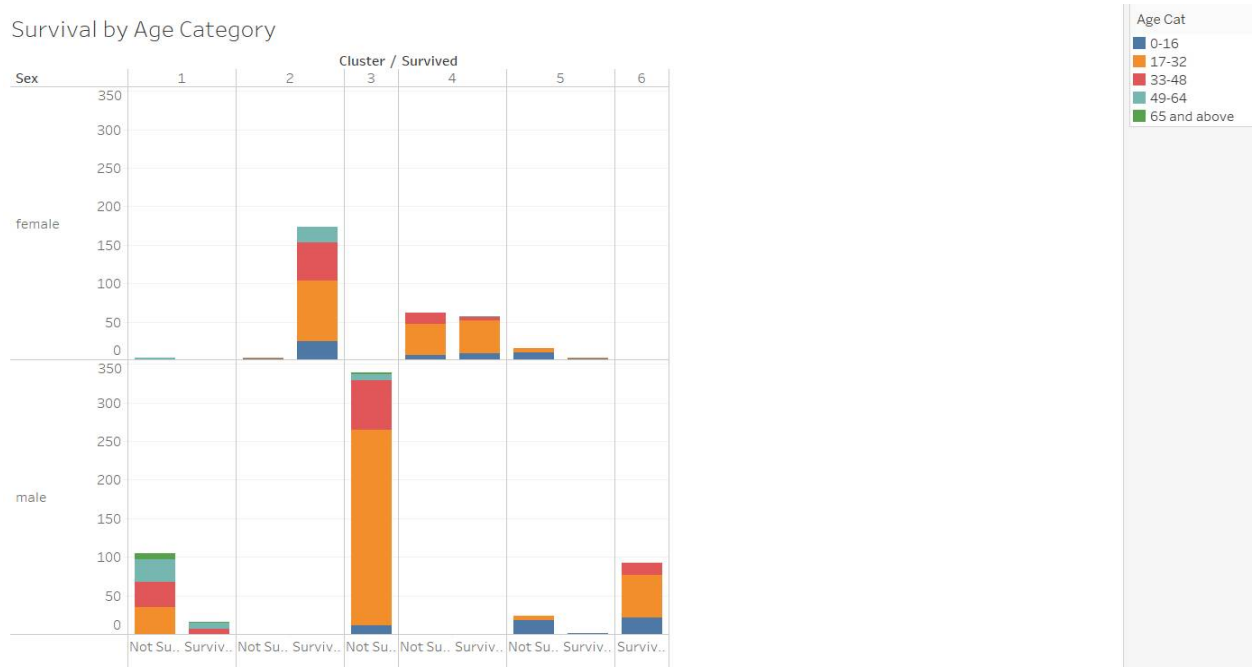

Fig.12: Observation using Category Parameter

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

In fig.12 we are displaying the number of people survived and not survived using only the age category parameter. The blue color denotes 0-16 years, yellow color denotes 17-32 years, red color denotes 33-48 years, light blue color denotes 49-64 years and the green color denotes 65 and above. Here we have used color encoding. Maximum number of survivors for the age category 17-32 years are female in cluster 2. The maximum number of survivors for age category 33-48 years are also female in cluster 2.

III. DATA VISUALIZATION USING MICROSOFT POWER BI

MicrosoftPower BI is a business analytics tool. It is used to create reports, dashboards as well as animation. It provides proficient interactive data visualization. Power BI is available in multiple forms Power BI Desktop, Power BI Service, Power BI Mobile Apps, Power BI Gateway, Power BI Embedded, Power BI Report Server and Power BI Visuals Marketplace. In our paper we will be talking about Power BI Desktop.

A. *Student Location Analysis:*

The data source is an Excel file. It gives us information about the State, Year, Population and Rank. Here we will be making an animation which will depict how student choice of studying in a state varies over a course of time.

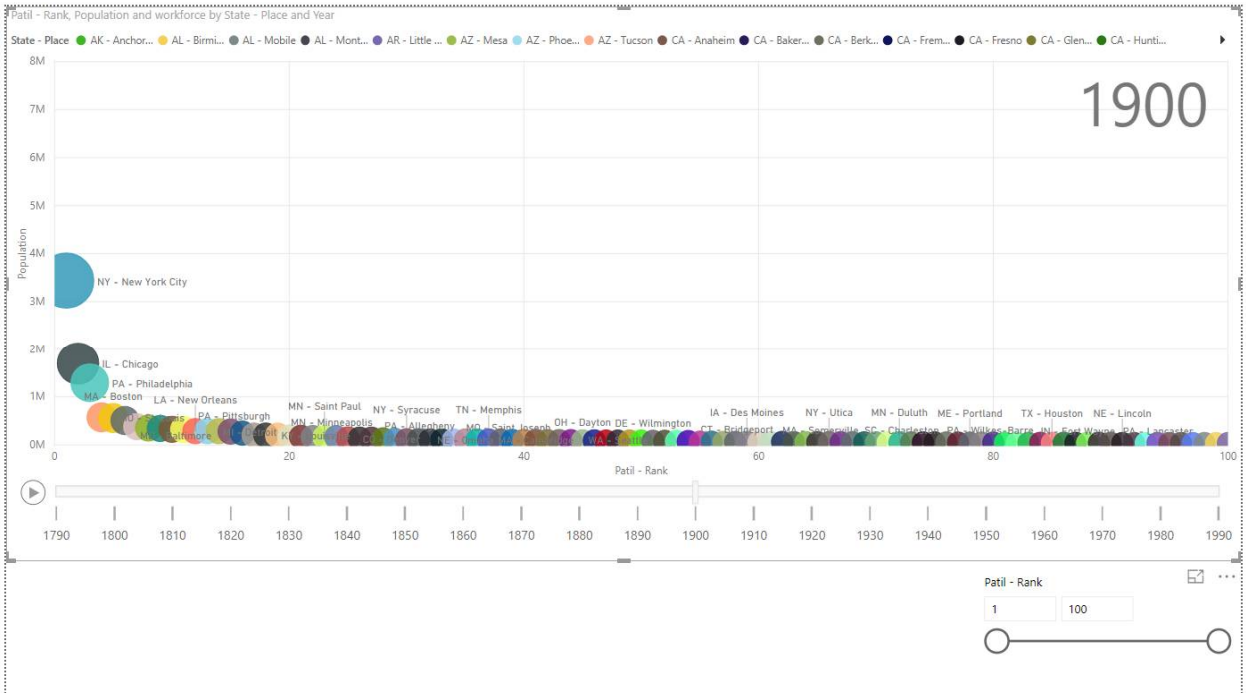

Fig.13: Start of Animation (Without Filtering)

Fig.13 talks about rank which varies from 1 to 100. Here there is a lot of overplotting and it is very difficult to gather insights. Patil-Rank acts as a filter with the help of which we can select only the states on which we want to focus on.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

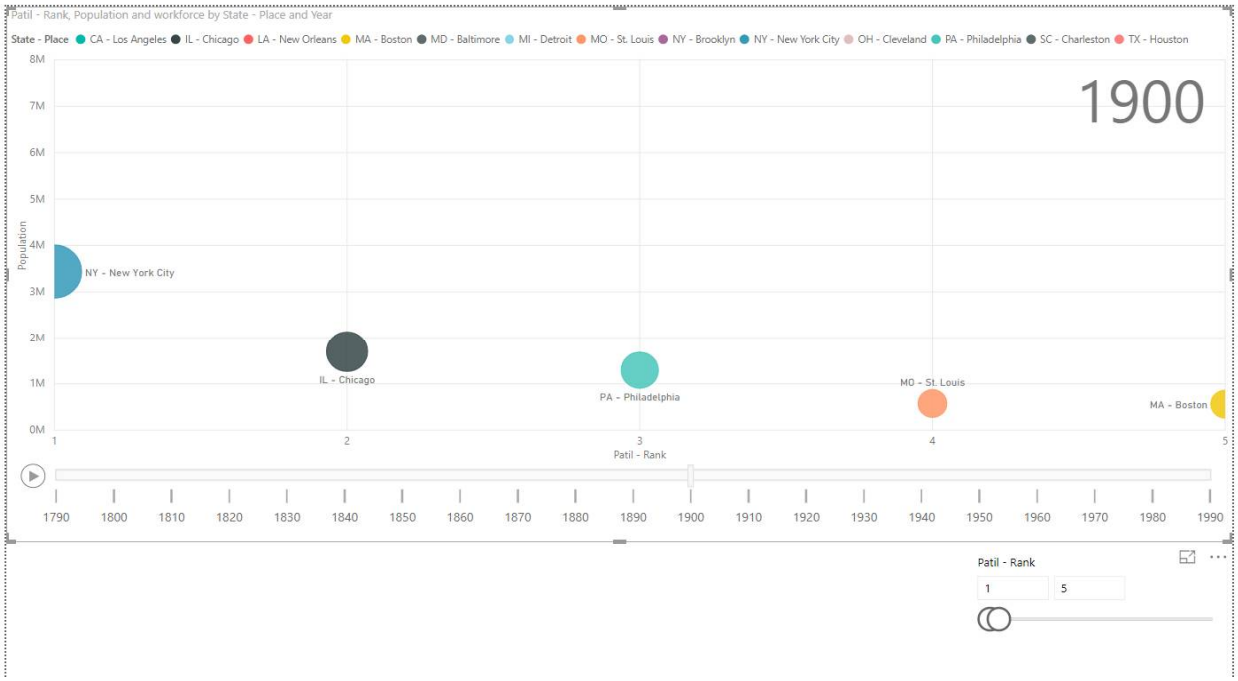

Fig.14: Start of Animation (With Filtering)

In fig.14 we are focusing only on top 5 states. 1900 is the beginning year or we can say that the initial state of the animation. Here we can see that the top state is NY- New York City followed by IL-Chicago, PA- Philadelphia, MO-St. Louis and MA-Boston.

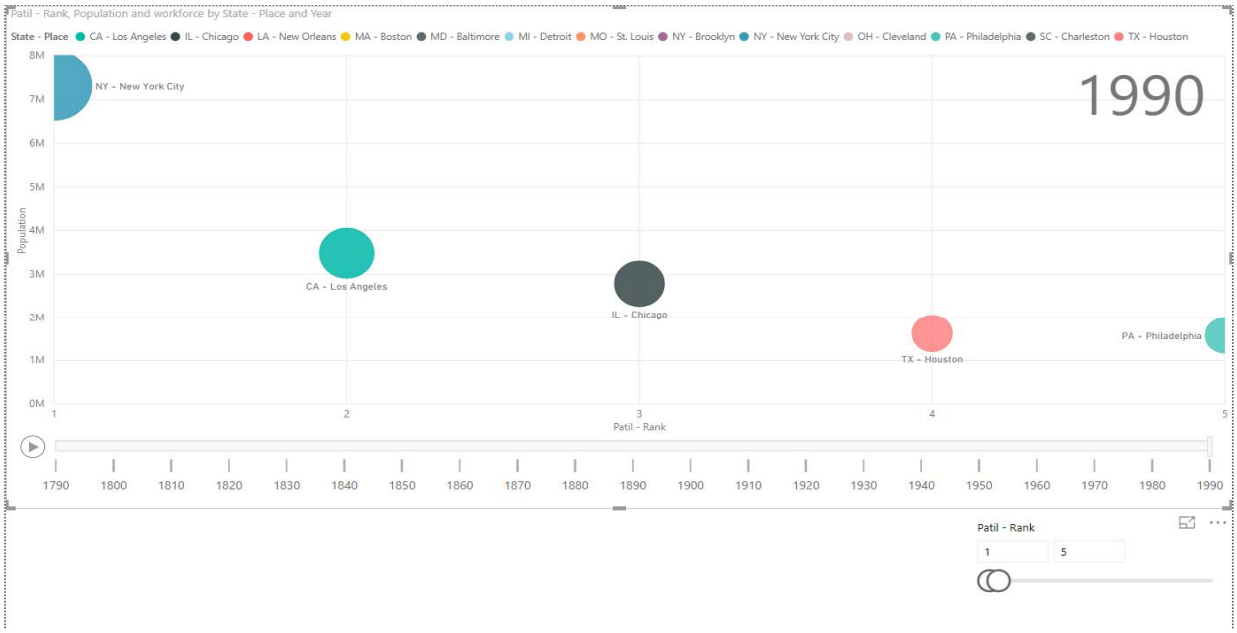

Fig.15: End of Animation (With Filtering)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

In fig.15 also we are focusing on top 5 states. 1990 is the end year of our dataset or we can say that the final state of animation. Here we can see that the top state is NY- New York City followed by CA- Los Angeles, IL- Chicago, TX- Houston and PA- Philadelphia

IV. DATA VISUALIZATION USING D3

D3 stands for Data Driven Document. Here the code is written in HTML, CSS and JavaScript. Data is driven from a source to the HTML file. DOM (Document Object Model) refers to hierarchical structure of the HTML and plays a vital role in data binding. With the help of D3 we bind or attach the input values to the elements of DOM. D3 is most useful when used to generate and manipulate visuals as Scalable Vector Graphics (SVG). SVG focuses on vectors rather than pixels. It also ensures that the quality of the image is not deteriorated when re-sized. SVG has 6 basic shapes rectangle, circle, ellipse, polyline, polygon and line.

A. *Food Data Analysis*

The data source is an Excel file. It gives us information about the food items, store names, sales, gross sale, quantity, discount on food items and net sales of the food available in my university- The University of Texas at Dallas (UTD).

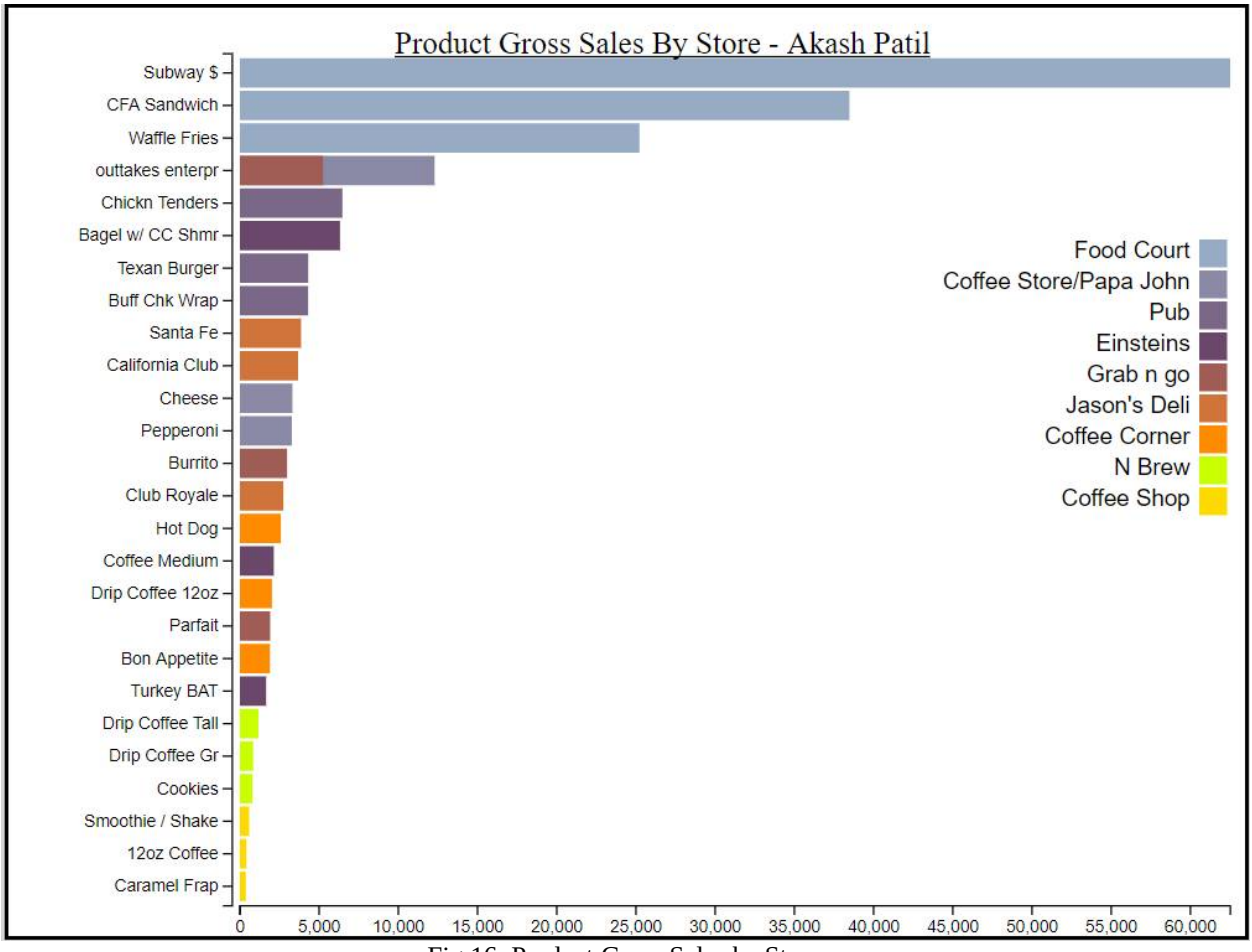

Fig.16: Product Gross Sales by Store

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Fig.16 is horizontal bar chart. We have applied bar padding to ensure that there is proper spacing between the bars. We have sorted the bars according to product gross sales by store. We have also provided legends on the right-hand side which gives us information about which food store or fast food restaurant is a part of which food court area. We can derive many insights from this bar chart. We come to know that Subway which is a part of Food Court has the highest product gross sales. Caramel Frap which is a part of Coffee Shop has the lowest product gross sales. We can also see that the color becomes brighter as the product gross sales by store decreases. This can also be considered as value encoding which shows that darker color has higher impact. There is also size encoding. The stores with larger product gross sales have bars having higher length.

V. DATA VISUALIZATION USING R STUDIO

R Studiois one of the best tools when we want to deal with statistical models. However, it can also be used for data visualization purpose. In R Studio for plotting data we make use of the ggplot2 package. It is also known as grammar of graphics. It does not return a value, it creates a new plot object. Print method is used to display this object on the screen. One of the easiest way to use ggplot2 is with qplot(quick plot). In this paper we will be pulling data from the wooldridge2 database for making plots.

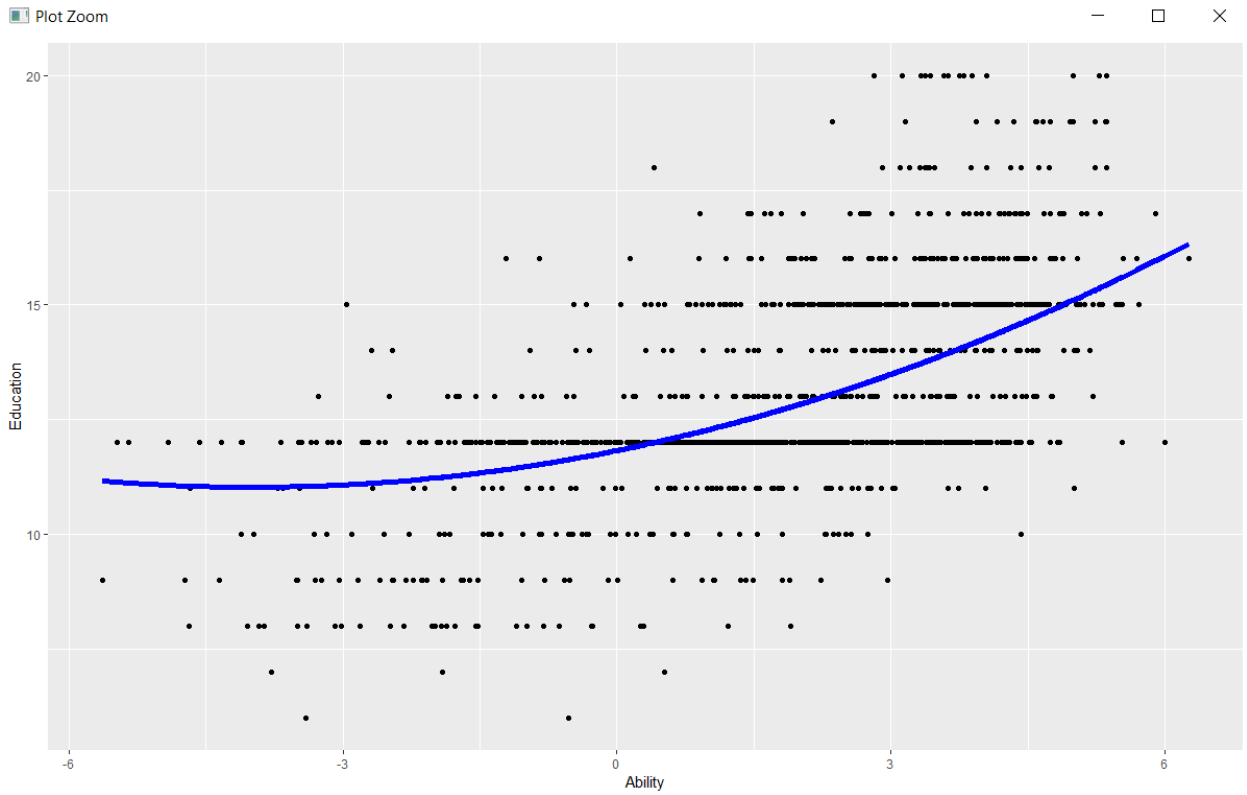

Fig.17: Education vs Ability

In Fig.17 we have Ability on x axis and Education on y axis. From the plot we come to know that Ability is directly proportional with Education. So, we can say that as a person's Ability to study increases his Education level also increases.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

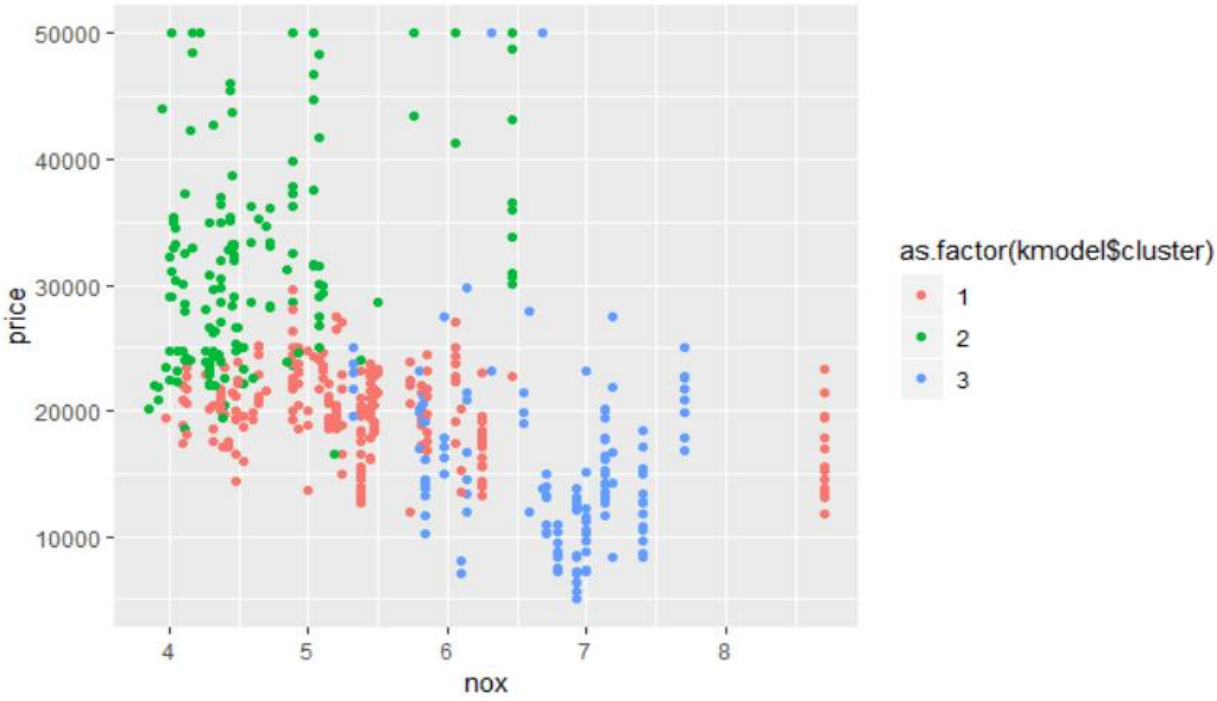

Fig.18: Kmeans Clustering

In fig.18 we have nox which is fuel on x axis and price on y axis. Here we have performed kmeans clustering using the kmeans function in R Studio. We can see that there are 3 clusters formed. Cluster 1 is denoted by red color, cluster 2 is denoted by green color and cluster 3 is denoted by blue color. So, we can say that members of cluster 2 pay higher prices for nox.

VI. CONCLUSION

Data Visualization provides a pictorial or animated representation of our data and makes it easier to study the variation, make comparison and understand several patterns which may arise from our data. It makes our work easy. We do not have to look at the entire data set and all those tedious numbers. Just by having a look at the visualization we can efficiently understand what is happening with our data. It makes the decision-making process quick, easy and effective.

.

REFERENCES

1. Danyel Fisher – Julie Steele & Noah Illinsky, “Beautiful Visualization”, O’Reilly Media, Inc., 2010.
2. Edward Segel& Jeff Heer, “Narrative Visualization: Telling Stories with Data”, IEEE Transactions on Visualization and Computer Graphics, Vol.16(6), pp.1139-1148, 2010.
3. Gordon Shaw, Robert Brown and Philip Bromiley, “Strategic Stories”.
4. Jessica Hullman, “Understanding of Sequence in Narrative Visualization”, IEEE Transactions on Visualization and Computer Graphics, Vol.19(12), pp.2406-2415, 2013.
5. Isabel Meirelles, “Design for Information”, Rockport, 2013.
6. Joseph Adler, “R in a Nutshell”, Sebastopol, CA: O’Reilly Media, 2012.
7. Daniel Putler, “Customer and business analytics applied data mining for decision making using R”, Boca Raton, FL: CRC Press, 2012.
8. Tamara Munzner, “Visualization Analysis & Design”, Boca Raton, FL: CRC Press, 2015.
9. Scott Murray, “Interactive Data Visualization for the Web”, Sebastopol, CA: O’Reilly Media, 2013.
10. Ben Fry, “Visualizing Data”, Sebastopol, CA: O’Reilly Media, 2008.
11. Jeffrey M. Wooldridge, “INTRODUCTORY ECONOMETRICS – A MODERN APPROACH”, 2012.
12. Nathan Yau, “Visualize This”, Wiley Pub, 2011.