BIGDATA VISUALIZATION USING GOLANG SERVICES

A Project Report Submitted
in Partial Fulfilment of the Requirements
for the Degree of

BACHELOR OF TECHNOLOGY(HONS)

in
COMPUTER SCIENCE AND ENGINEERING
by

S.SAI SIDDARDHA
Roll No. [2015BCS0032]

to

DEPARTMENT OF [COMPUTER SCIENCE AND ENGINEERING]
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
KOTTAYAM - 695017, INDIA
April 2018

DECLARATION

I, **s.sai siddardha**(**Roll No:2015BCS0032**), hereby declare that, this report entitled **"BIGDATA VISUALIZATION USING GOLANG SERVICES report"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor Technology(Hon)** in **COMPUTER SCIENCE AND ENGINEERING** is an original work carried out by me under the supervision of **DR.SHAJULIN BENEDICT** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Thiruvananthapuram-695017 **S.Sai Siddardha**
April 2018

CERTIFICATE

This is to certify that the work contained in this project report entitled **"[Bigdata Visualization using Golang Services]"** submitted by s.sai siddardha(**Roll No:2015bcs0032**) to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology(Hon)** in **[Indian institute of information technology kottayam]** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Thiruvananthapuram – 695017
April 2018

(Dr. SHAJULIN BENEDICT)
Project Supervisor

ABSTRACT

Almost all research sectors (business sector, education sector and other sectors) might lead to a bigdata problem due to huge data. Data visualization helps users to quickly identify interesting and significant events/patterns from data that are otherwise too detailed or complex to discern. Users find difficult to interpret such large data. In this report, we discussed about the challenges and opportunities for visualizing bigdata. It discloses the modern techniques and solutions applied for visualizing bigdata in various sectors. It showcases the most important charts used for bigdata visualization like composition charts, comparison charts, relationship charts, distribution charts. And, the experiments were conduct to prove bigdata visualization.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

What is BigData?

Bigdata is also a data but with huge size. Bigdata contains both structured and unstructured data exponentially increasing with time**.** Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

The main problem with bigdata is with this large amounts of data the user cannot choose data that Is useful to him.so this visualization helps to see the data clearly and he/she can predict data. Bigdata contains both structured ad unstructured data. Structured data can be easily and interpreted in database. But unstructured data cannot be easily analyzed and interpreted as like as structured data.

In recent years bigdata is growing very rapidly because of growth of social, cloud and multimedia computing. As we know traditional systems cannot store and process data companies turning into bigdata management solutions to convert unstructured data into structured data useful for the companies.

Bigdata Benefits:

The benefits of the bigdata include:

- Identifying the root causes of failures and issues in real time
- Fully understanding the potential of data-driven marketing
- Generating customer offers based on their buying habits
- increasing custmor loyalty.
- Reevaluating risk portfolios quickly
- Personalizing custmor experience.
- Adding value to online and offline customer interactions.

Big Data in Today's Business and Technology Environment

- 2.7 Zetabytes of data exist in the digital universe today.

 235 Terabytes of data has been collected by the U.S. Library of Congress in April 2011.
 The Obama administration is investing $200 million in big data research projects.
 IDC Estimates that by 2020,business transactions on the internet-business-to-business and business-to-consumer – will reach 450 billion per day.
 Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.
 Akamai analyzes 75 million events per day to better target advertisements.
 94% of Hadoop users perform analytics on large volumes of data not possible before; 88% analyze data in greater detail; while 82% can now retain more of their data.

 Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.

The Rapid Growth of Unstructured Data

- YouTube users upload 48 hours of new video every minute of the day.

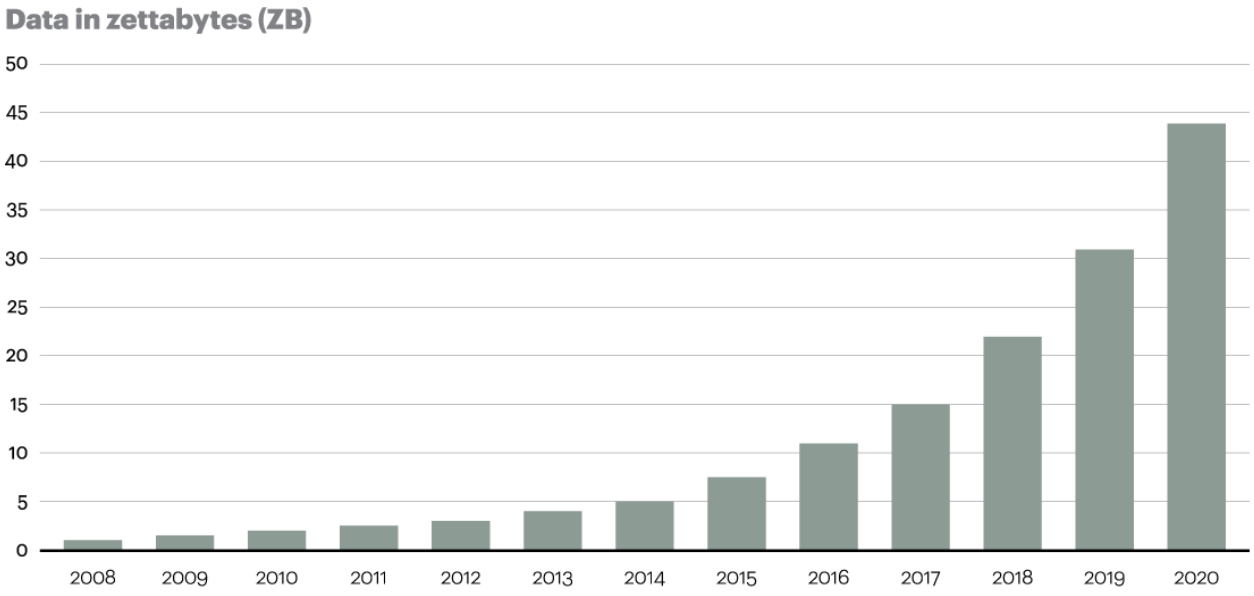
 571 new websites are created every minute of the day.

 Brands and organizations on Facebook receive 34,722 Likes every minute of the day.
 100 terabytes of data uploaded daily to Facebook.
 According to Twitter's own research in early 2012, it sees roughly 175 million tweets every day, and has more than 465 million accounts.

 30 Billion pieces of content shared on Facebook every month.

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



Source: Oracle, 2012

1.3 why bigdata visualization:

Big data visualization means implementation of visualization techniques to show contemporary relation between two variables in a dataset. If we want to know relationship between two variables in a dataset which consists of large data it is very difficult to compare those two variables by checking every row but by using some visualization techniques if we can plot the data in graphs like bar chart, scatter chart, pie chart then we can easily predict the data by seeing in a graph.

Some of the bigdata visualization techniques are like world cloud, symbol maps, connectivity charts, line charts, bar charts, map charts, 3D grpahs.

- We can visualize our data either in Tables or Graphs.
- Graphs and tables serve different purposes. Choose the appropriate data display to fit your purpose.
- Common Display Types

– Bar plots, Box plots, Heatmaps, Tree plot, Pie chart ……

– Scatterplots and in many different ways we can visualize our data.

Different kinds of visualization techniques are there for Big data and small data.

Heatmap and Scatterplot some of the basic plot examples ..

1.4 objectives of bigdata visualization

The objectives of bigdata visualization are:

1). the main objective of big data visualization is to show the graph elegantly and clearly.

2). The improved custmor visualization of data.

3). Better operational efficiency.

4). Every company uses data in its on way. the more the company uses the data the larger it will grow.

1.5 Challenges Of Bigdata Visualization:

1). Data storage and quality.

2). People who understand Bigdata analysis.

3). Good quality analysis.

4). Security and Privacy of data.

5). Various Sources of data.

CHAPTER 2

LITERATURE SURVEY

Data visualization is the main focusing concept in big data analysis for processing and analyzing multi variate data, because of rapid growth of data size and complexity of data. Basically data visualization may achieve three main problems, i.e.

1. Structured and Unstructured pattern evaluation in big data analysis.

2. Shrink the attributes in data indexed big data analysis.

3. Rearrange of attributes in parallel index based data storage.

Big Data analytics plays a key role through reducing the data size and complexity in Big Data applications. Visualization is an important approach to helping Big Data get a complete view of data and discover data values. Big Data analytics and visualization should be integrated seamlessly so that they work best in Big Data applications.

Data visualization transforms data into images to aid the understanding of data; therefore, it is an invaluable tool for explaining the significance of data to visually inclined people. Given a (big) dataset, the essential task of visualization is to visualize the data to tell compelling stories by selecting, filtering, and transforming the data, and picking the right visualization type such as bar charts or line charts. Our ultimate goal is to automate this task that currently requires heavy user intervention in the existing visualization systems.

An evolutionized system in the field faces the following three main challenges: (1) Visualization verification: to determine whether a visualization for a given dataset is interesting, from the viewpoint of human understanding; (2) Visualization search space: a "boring" dataset may become interesting after an arbitrary combination of operations such as selections, joins, and aggregations, among others; (3) On-time responses: do not deplete the user's patience.

Our services are different from other services like given a bigdata set ant the essential form is visualizing by selecting right chart(composition chart, comparison chart, relationship charts, distribution charts) and plot it by using some R packages and show it in by golang services.

CHAPTER 3

DATA VISUALIZATION

There are four important ways you can present the bigdata the ways are:

1) Composition charts

2)Comparison charts

3)Relationship charts

4)Distribution charts

TO SELECT A RIGHT CHART TYPE:

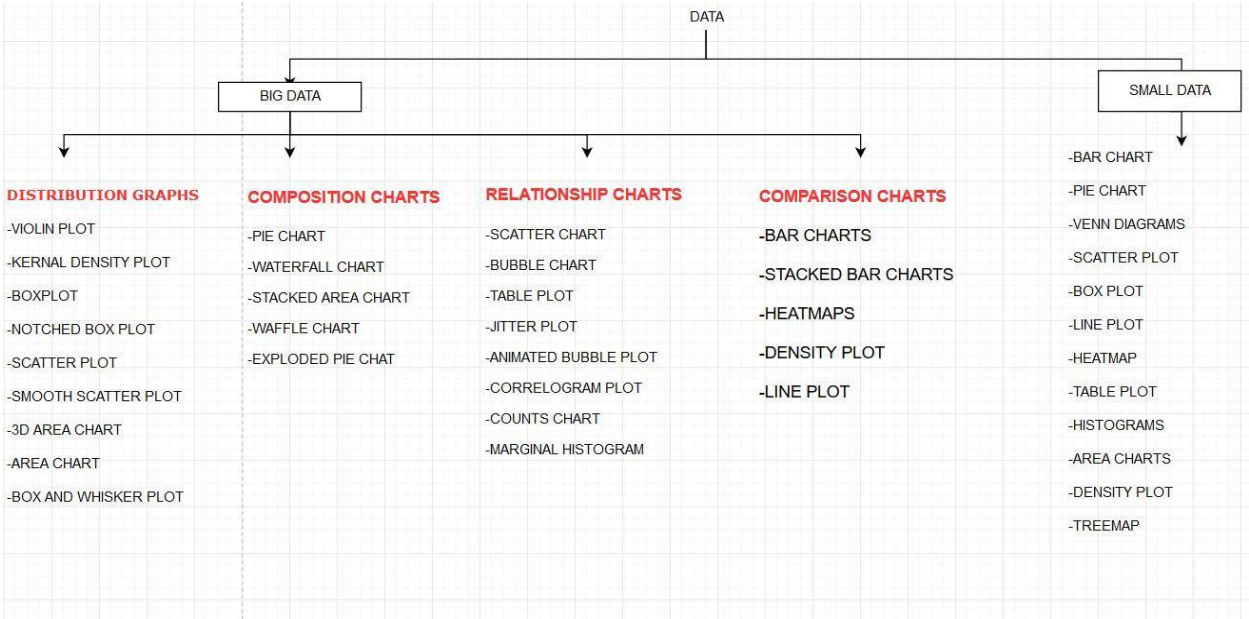
To determine which chart is best suited for each of those presentation types, first you must answer a few questions:

- How many variables do you want to show in a single chart? One, two, three, many?
- How many items (data points) will you display for each variable? Only a few or many?

- Will you display values over a period of time, or among items or groups?

As every chart plots the bigdata some graphs are good for some type. Like Bar charts are good for comparisons, while line charts work better for trends. Scatter plot charts are good for relationships and distributions, but pie charts should be used only for simple compositions — never for comparisons or distributions. These are some restriction we need to follow.

Taxonomy for above charts be like:



3.1 COMPOSITION CHARTS:

Composition of data is probably the most misused method in data representation endeavors. The idea is to show how individual parts make up the whole by combining them together and displaying them as a sum. Composition can also be used to show how a total value can be divided into parts or to highlight the significance of each part relative to the total value.

Use data composition charts to show

- Company market share and a few key players in the market
- Total country population by TOP religions, languages, or ethnical groups
- Total revenue, by TOP product lines, divisions, or regions

The more important composition charts include:

- **Area chart**

  Area chart is used to show continuity across a variable or data set. It is very much same as line chart and is commonly used for time series plots.

- **Bar chart**

A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable.

- **Pie Chart**

  A pie-chart is a representation of values as slices of a circle with different colors.

- **Stacked area chart**

  An area chart displays the development of quantitative values over an interval. It resembles a line chart as it uses lines connecting data points with each other.

- **Waffle Chart**

Waffle charts is a nice way of showing the categorical composition of the total population.

- **Scatter chart**

  Scatter Plot is used to see the relationship between two continuous variables.

3.2 COMPARISON CHARTS:

Comparison of data points is probably the most common and easy-to-understand method for data analysis. As the name suggests, we use comparison to evaluate and compare values between two or more data points. With comparison you can also easily find the lowest and highest values in the chart.

Usually comparisons are made to accomplish one of the following goals:

- To list key values to quickly find and read them (i.e., revenue per month)
- To rank several data categories from best to worst or the other way around
- To show pattern recognition by visually highlighting gaps, spikes, outliers, or trends

Some of the important comparison charts are:

- **Bar Plot :**

  A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

- **Density Plot :**

Density plots are usually a much more effective way to view the distribution of a variable.

- **Heatmaps :**

  A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a bit like looking a data table from above.

- **Line Charts :**

  A line chart is a graph that connects a series of points by drawing line segments between them. These points are ordered in one of their coordinate (usually the x-coordinate) value. Line charts are usually used in identifying the trends in data.

3.3 RELATIONSHIP CHARTS:

These types of charts show the relationship, correlation, or connection of two or more variables and their properties. A good use of relationship graphs would be to demonstrate how something does or does not affect another variable positively or negatively.

Use a relationship chart to:

- Spot flaws in effectiveness by evaluating expenses vs. income by store or region.
- pot flaws in effectiveness by evaluating expenses vs. income by store or region.

Some of the important relationship charts are:

- **Scatter Plot with Encircling :**

  Scatter Plot is used to see the relationship between two continuous variables. It is type of scatter plot that encircles set of points.

- **Jitter Chart :**
  It is a type in scatter plot with overlapping values.jo jitter chart uses the same function as scatter plot.

- **Counts Chart :**
  To overcome the problem of data points overlap is to use what is called a counts chart.

- **Bubble Plot :**

While scatterplot lets you compare the relationship between 2 continuous variables, bubble chart serves well if you want to understand relationship within the underlying groups based on:

1. A Categorical variable (by changing the color) and

2. Another continuous variable (by changing the size of points).

- **Animated Bubble chart :**
  It is same as the bubble chart, but, you have to show how the values change over a fifth dimension (typically time).

- **Correlogram :**

  Correlogram is used to test the level of co-relation among the variable available in the data set. The cells of the matrix can be shaded or colored to show the co-relation value.

- **Marginal Histogram :**

If you want to show the relationship as well as the distribution in the same chart, use the marginal histogram. It has a histogram of the X and Y variables at the margins of the scatterplot.

- **Table Plot :**
  The table plot is a visualization method that is used to explore and analyse large datasets. Table plots are used to explore the relationships between the variables, to discover strange data patterns, and to check the occurrence and selectivity of missing values.

3.4 DISTRIBUTION CHARTS :

- **Area Chart :**
  Area chart is used to show continuity across a variable or data set. It is very much same as line chart and is commonly used

for time series plots. Alternatively, it is also used to plot continuous variables and analyze the underlying trends.

- **Box Plot :**

  Box plot is an excellent tool to study the distribution. It can also show the distributions within multiple groups, along with the median, range and outliers if any.

- **Dot+Box Plot :**

  On top of the information provided by a box plot, the dot plot can provide more clear information in the form of summary statistics by each group. The dots are staggered such that each dot represents one observation. So, in below chart, the number of dots for a given manufacturer will match the number of rows of that manufacturer in source data.

- **Candle Stick Plot :**

  It is like a combination of line-chart and a bar-chart: each bar represents all four important pieces of information for that day: the open, the close, the high and the low. Being densely packed with

information, they tend to represent trading patterns over short periods of time, often a few days or a few trading sessions.

- **Kernal Density Plot :**

  Kernal density plots are usually a much more effective way to view the distribution of a variable.

- **Notched Box Plot :**

  The boxplot compactly displays the distribution of a continuous variable. It visualizes five summary statistics (the median, two hinges and two whiskers), and all "outlying" points individually .for a notched box plot.

- **Stripe Plot :**

  A strip plot is a graphical data analysis technique for summarizing a univariate data set. The strip plot consists of:

  Horizontal axis = the value of the response variable;

  Vertical axis = all values are set to 1.

- **Tufte Box Plot :**

Tufte's Box plot is just a box plot made minimal and visually appealing.

- **Violin Plot :**

  A **violin plot** is a method of plotting numeric data. It is similar to box **plot** with a rotated kernel density **plot** on each side. The **violin plot** is similar to box **plots**.

Now after understanding everything about graphs we can plot every graph using some R packages. Now using golang as service every plot mentioned above are plotted in localhost. As most graphs use ggplot2 package the packages that this project include:

- **Ggplot2**
- **Hexbin**
- **Ggthemes**
- **Tabplot**
- **Ggcorrplot**
- **ggExtra**
- **ggalt**
- **hexbin**
- **RColorBrewer**
- **Magrittr**
- **Leaflet**

CHAPTER 4

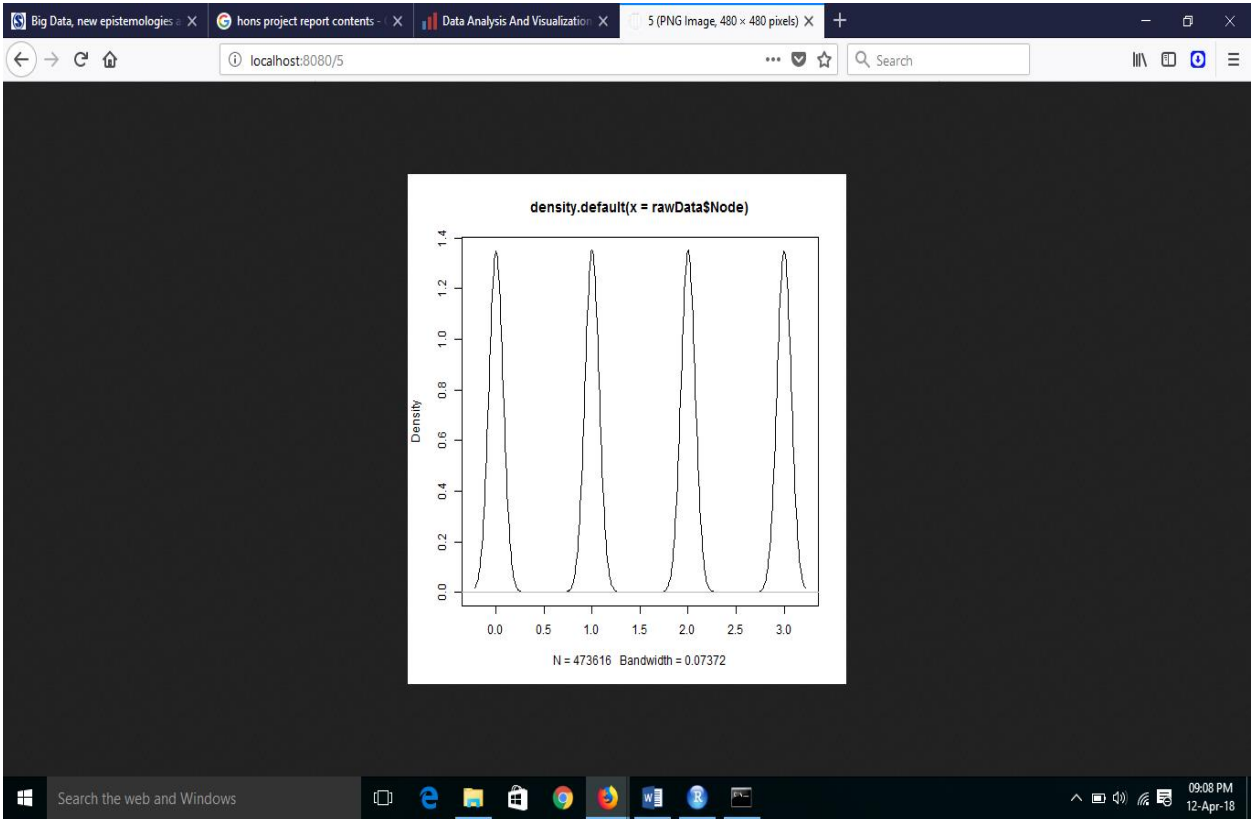
EXPERIMENTAL RESULTS

Box plot:

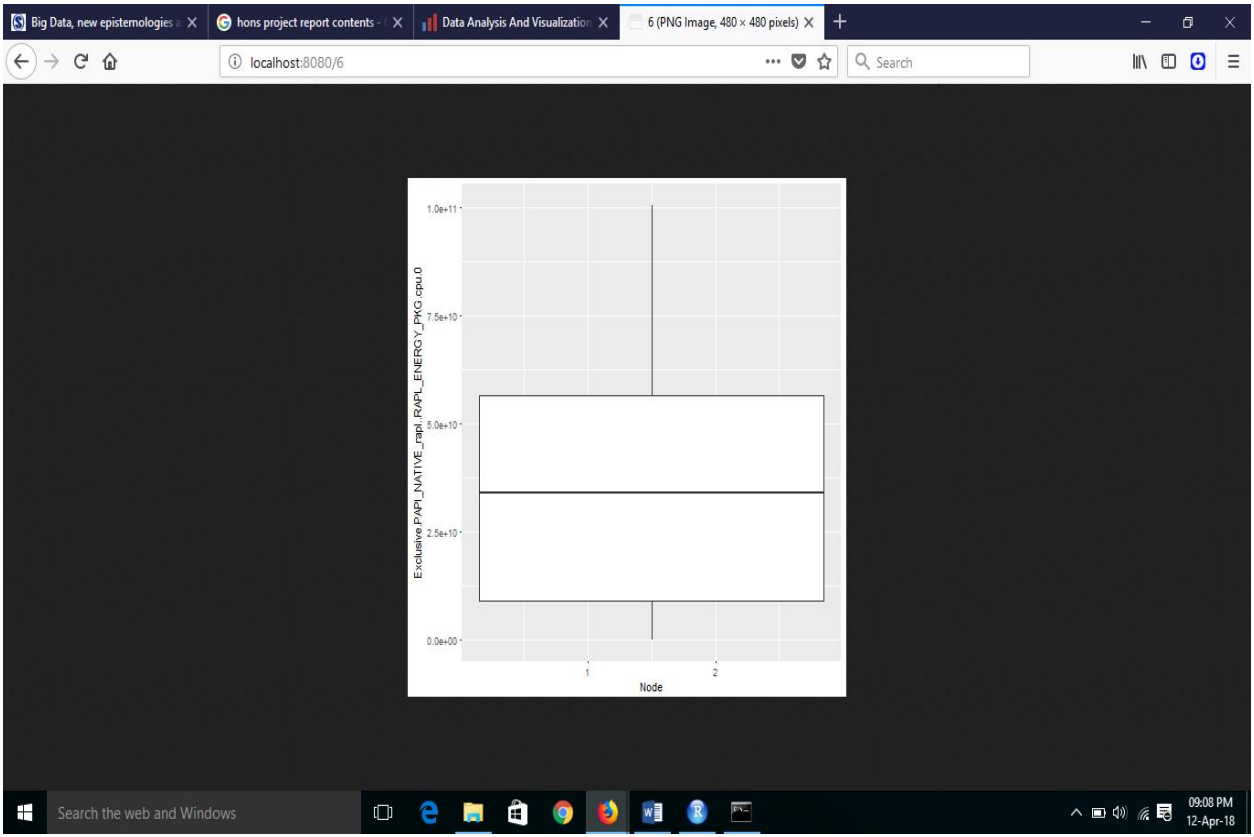
Dot+Box plot:


Big Data, new epistemologies
hons project report contents -
Data Analysis And Visualization
4 (PNG Image, 480 × 480 pixels)
localhost:8080/4
Search
Box plot + Dot plot
1.0e+11
7.5e+10
5.0e+10
2.5e+10
0.0e+00
Exclusive PAPI_NATIVE_rapl::RAPL_ENERGY_PKG:cpu:0
Node
Search the web and Windows
09:08 PM
12-Apr-18

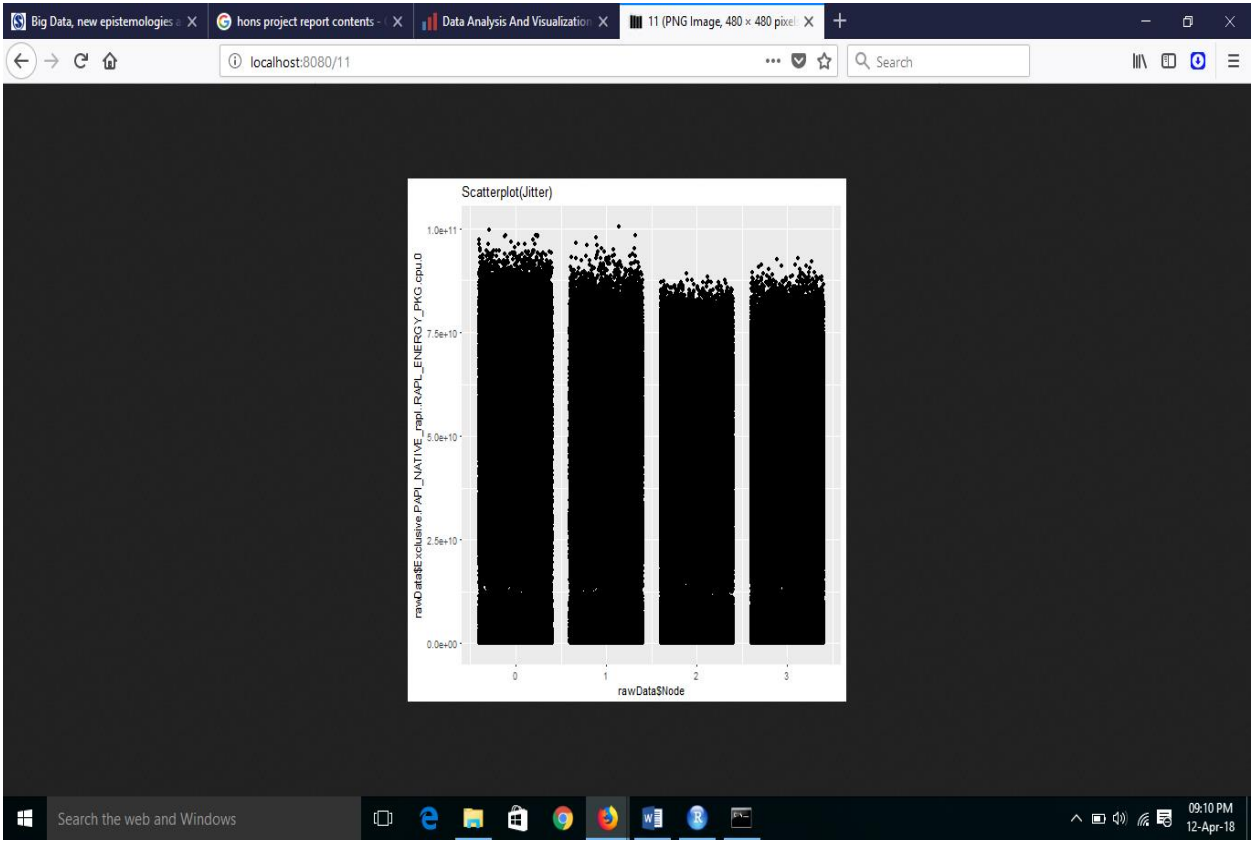
Density plot:



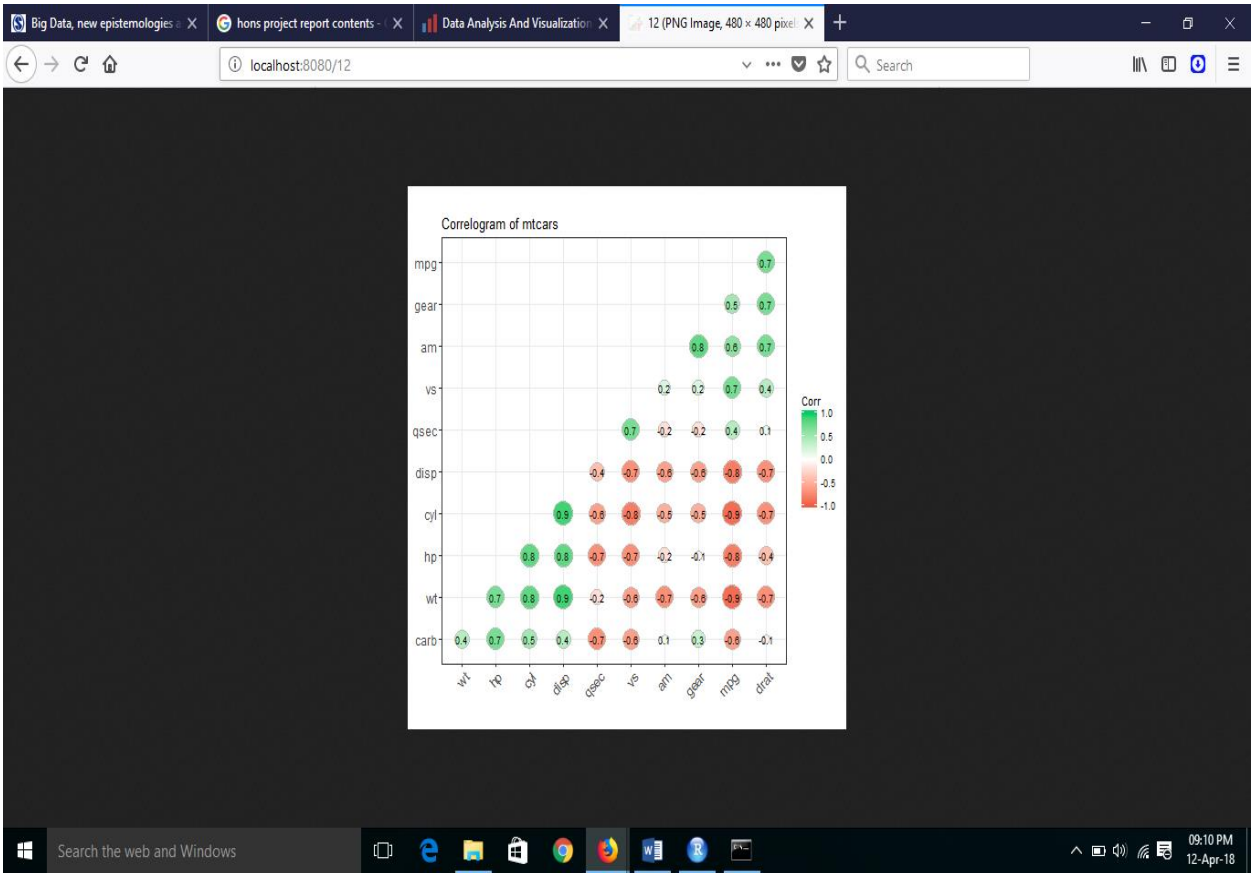
NotchedBox plot:



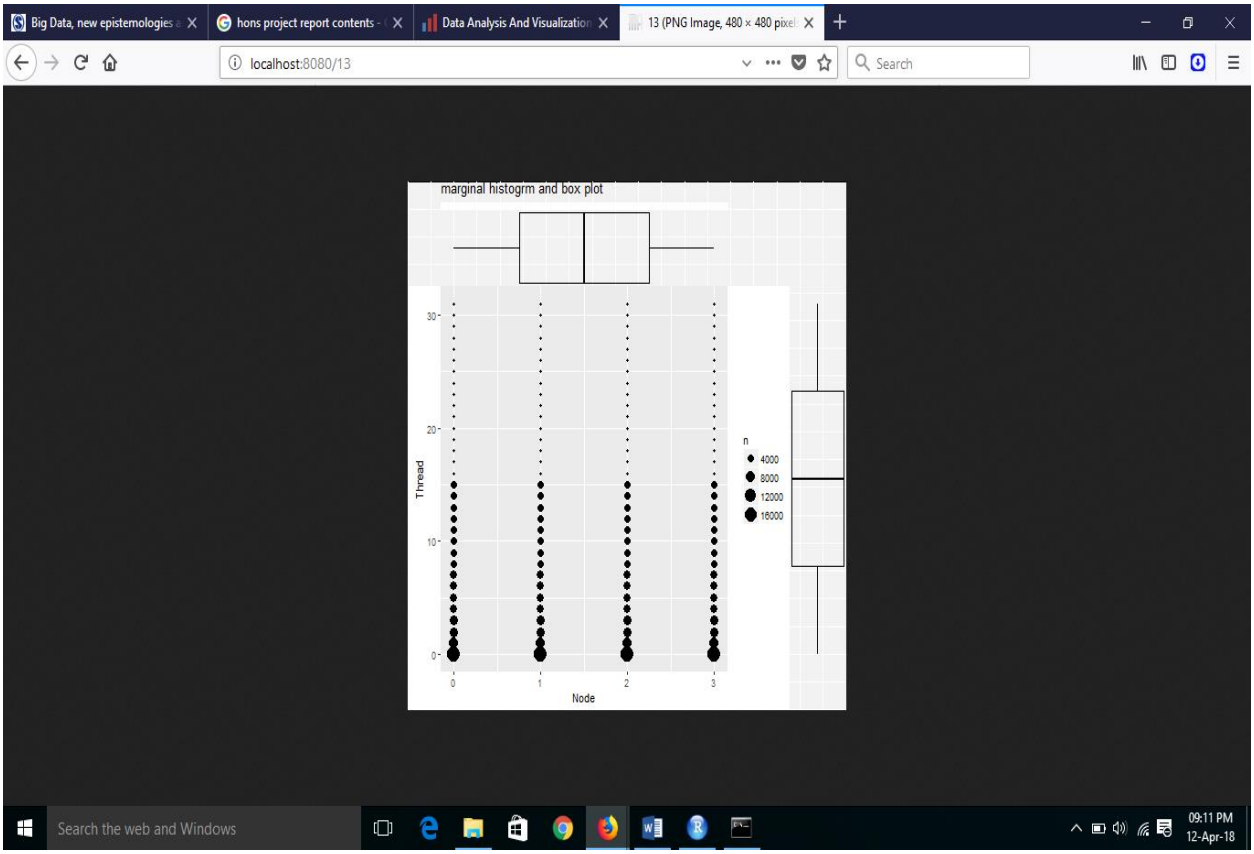
Jitter plot:


Big Data, new epistemologies
hons project report contents -
Data Analysis And Visualization
11 (PNG Image, 480 × 480 pixel
localhost:8080/11
Search
Scatterplot(Jitter)
rawData$Exclusive.PAPI_NATIVE_rapl::RAPL_ENERGY_PKG:cpu:0
1.0e+11
7.5e+10
5.0e+10
2.5e+10
0.0e+00
0
1
2
3
rawData$Node
Search the web and Windows
09:10 PM
12-Apr-18

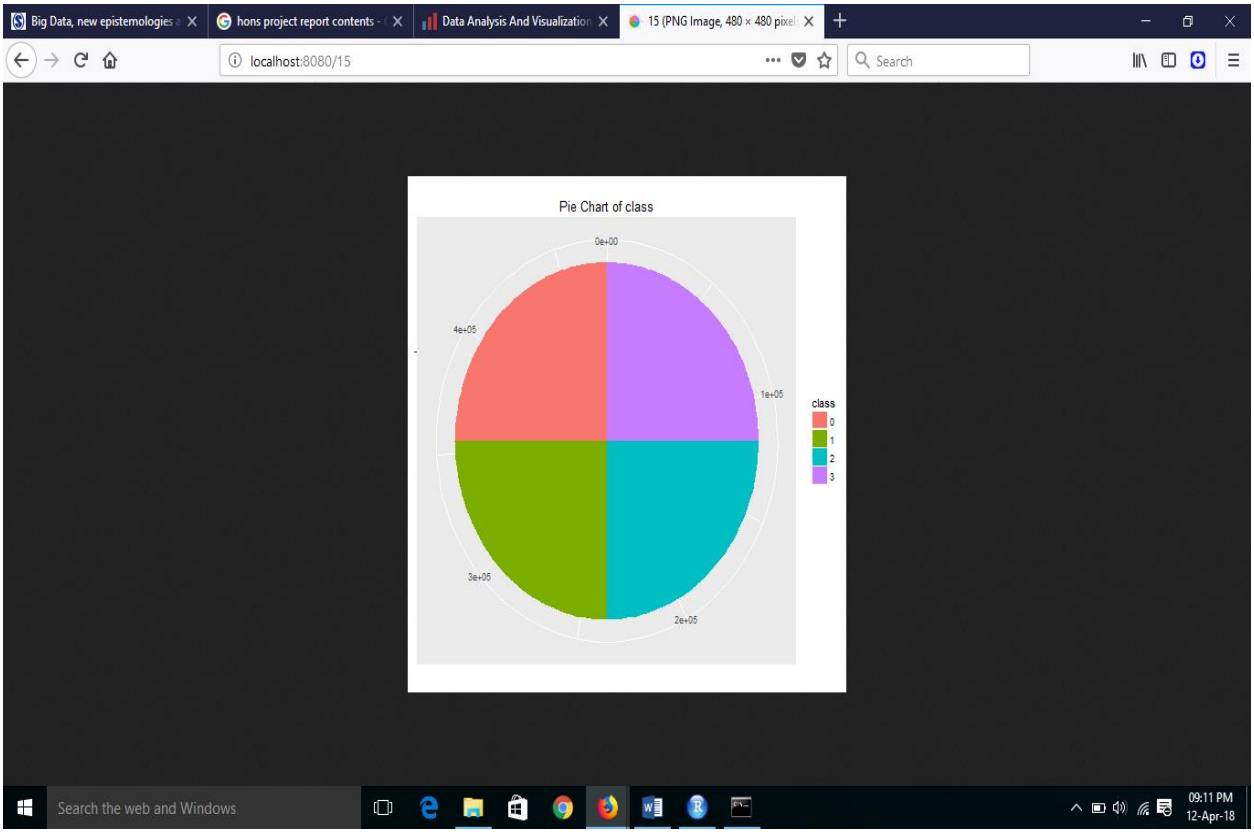
Correlogram:



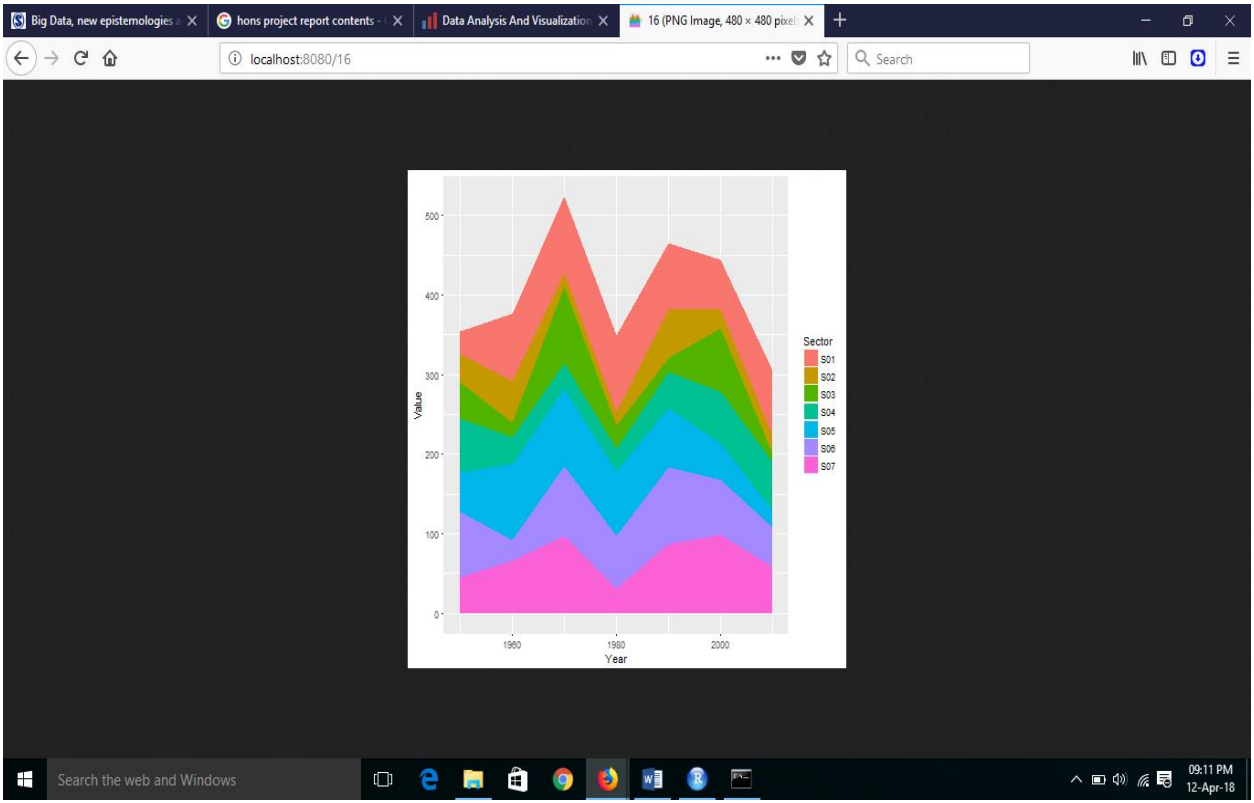
Marginahistogram+boxplot:



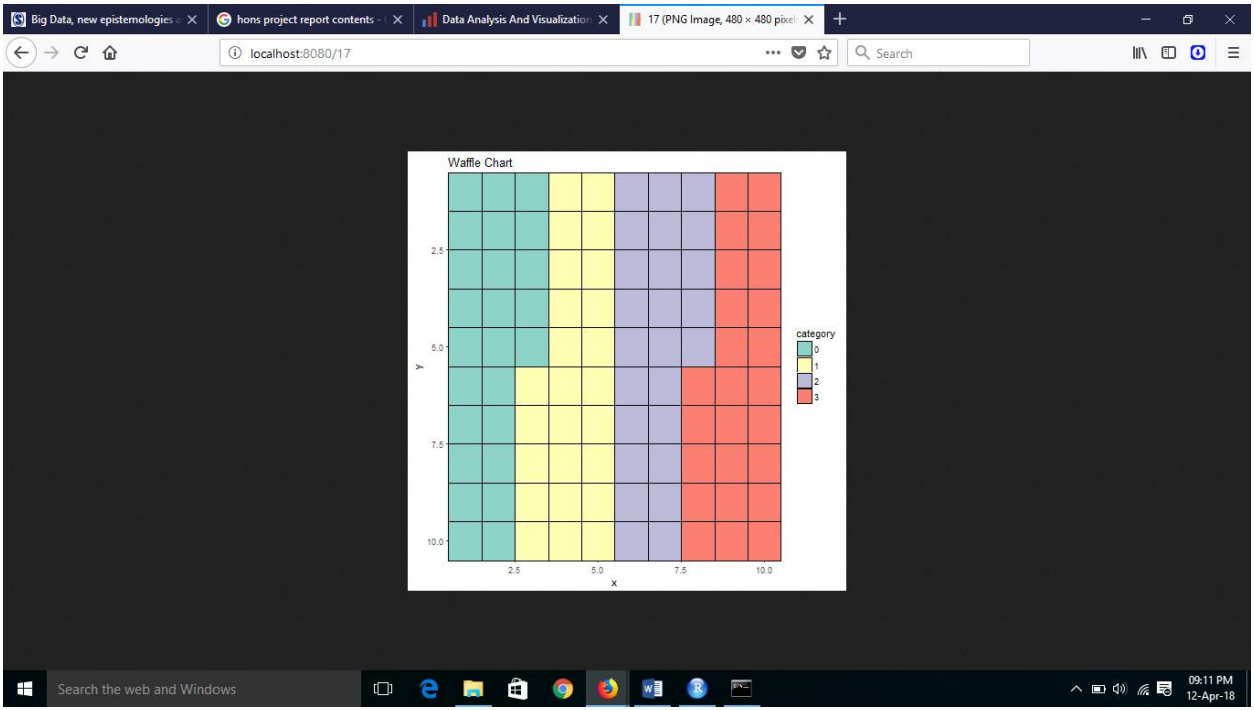
Pie chart:


Big Data, new epistemologies
hons project report contents
Data Analysis And Visualization
15 (PNG Image, 480 × 480 pixel
localhost:8080/15
Search
Pie Chart of class
0e+00
1e+05
2e+05
3e+05
4e+05
class
0
1
2
3
Search the web and Windows
09:11 PM
12-Apr-18

Stacked area chart:

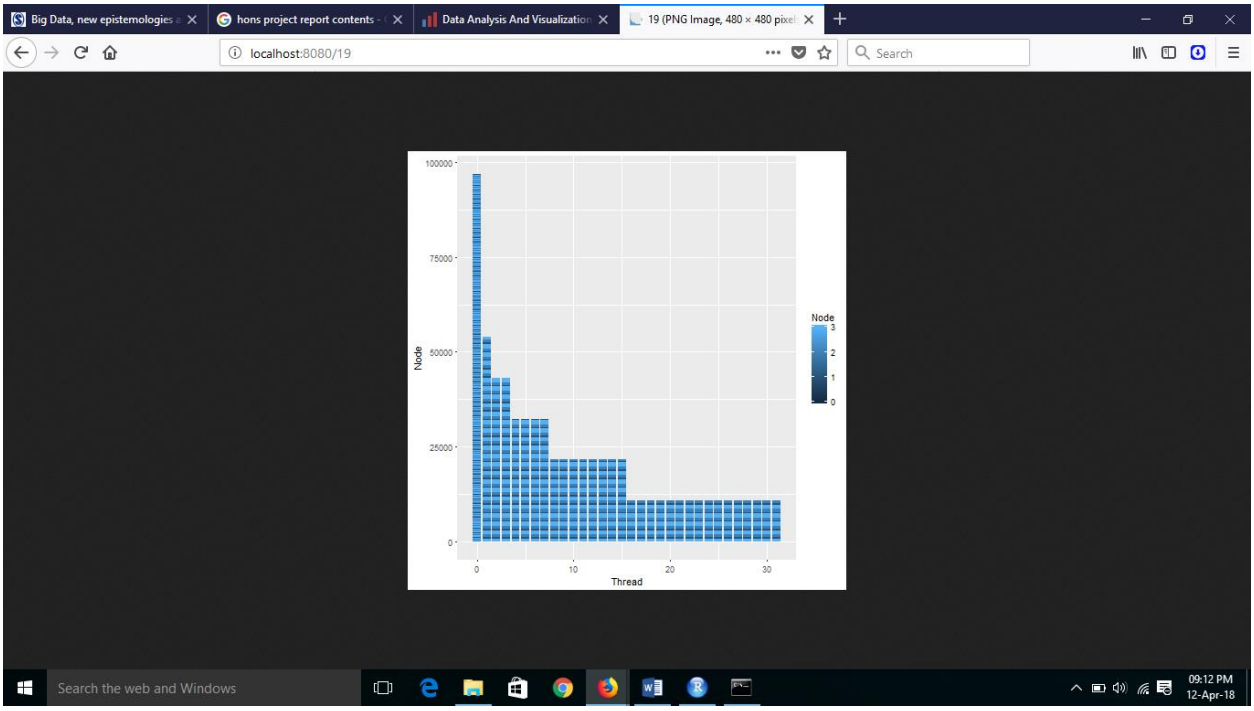


Waffle chart:

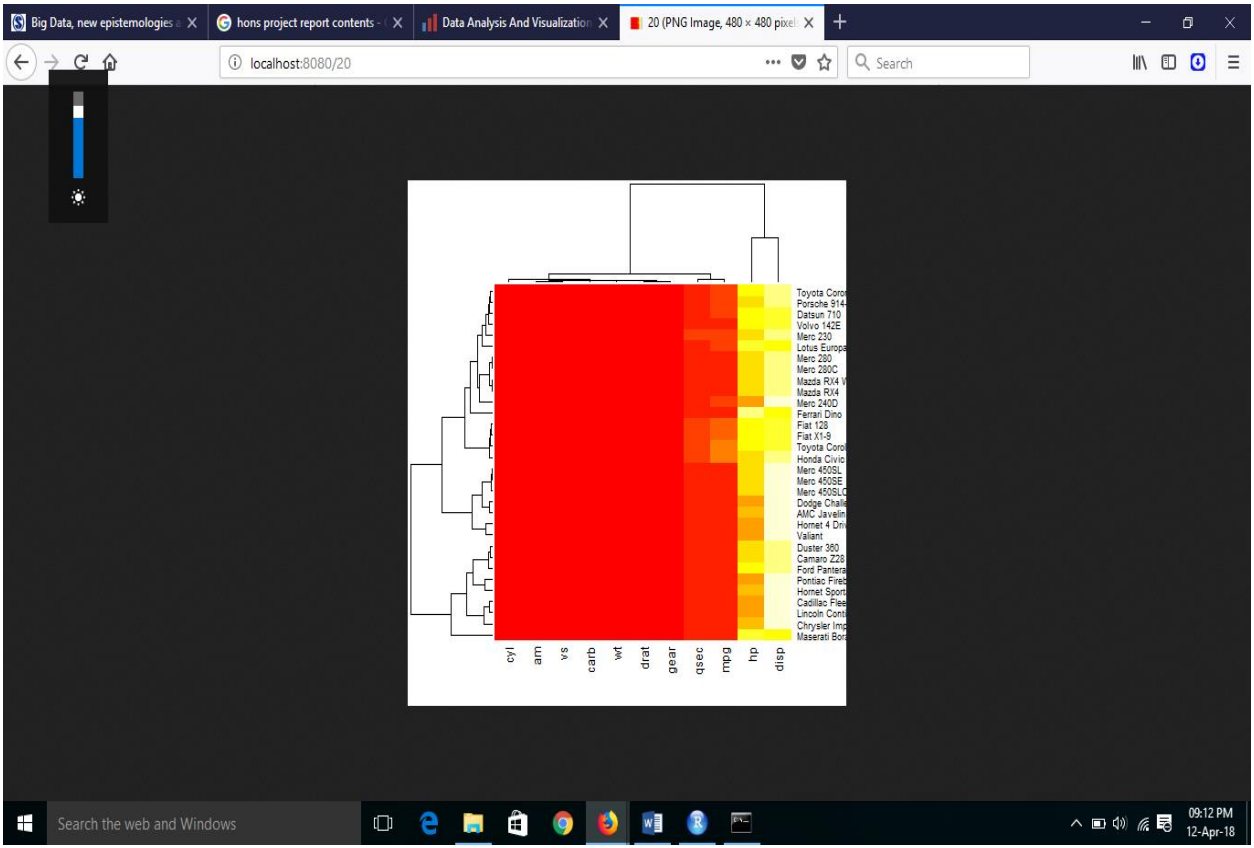


Histogram chart:

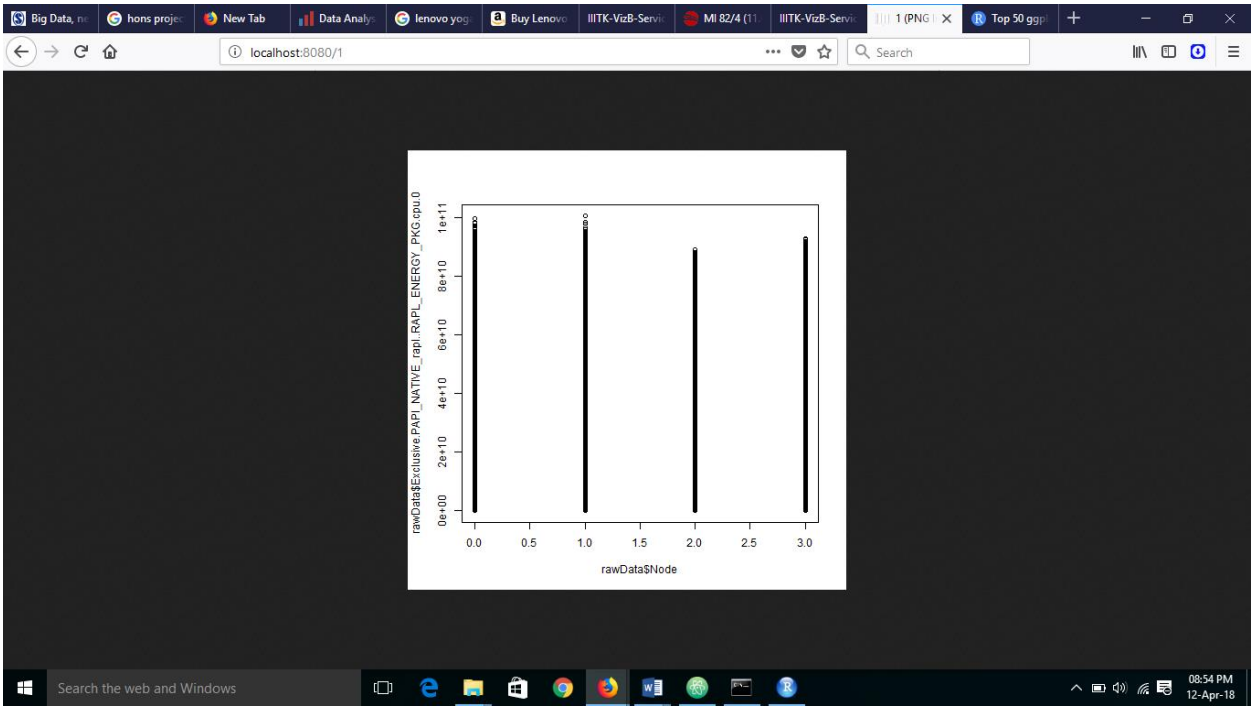
Bar chart:


Big Data, new epistemologies
hons project report contents
Data Analysis And Visualization
19 (PNG Image, 480 × 480 pixel
localhost:8080/19
Search
100000
75000
50000
25000
0
Node
0
10
20
30
Thread
Node
3
2
1
0
Search the web and Windows
09:12 PM
12-Apr-18

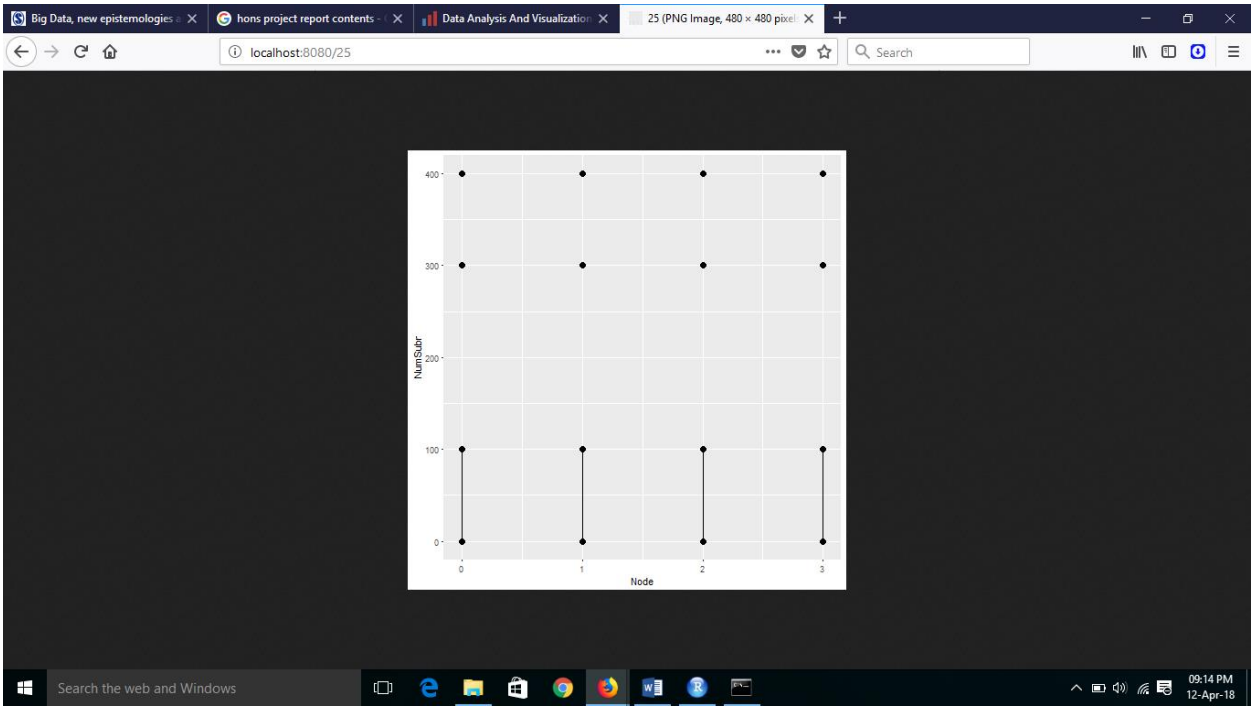
Heat map:


Big Data, new epistemologies
hons project report contents -
Data Analysis And Visualization
20 (PNG Image, 480 × 480 pixel
localhost:8080/20
Search
Toyota Coro
Porsche 914
Datsun 710
Volvo 142E
Merc 230
Lotus Europa
Merc 280
Merc 280C
Mazda RX4 W
Mazda RX4
Merc 240D
Ferrari Dino
Fiat 128
Fiat X1-9
Toyota Coro
Honda Civic
Merc 450SL
Merc 450SE
Merc 450SLC
Dodge Challe
AMC Javelin
Hornet 4 Dri
Valiant
Duster 360
Camaro Z28
Ford Pantera
Pontiac Fired
Hornet Sport
Cadillac Flee
Lincoln Cont
Chrysler Imp
Maserati Bor
cyl
am
vs
carb
wt
drat
gear
qsec
mpg
hp
disp
Search the web and Windows
09:12 PM
12-Apr-18

Scatter chart:



Lollipop chart:



CHAPTER 4

CONCLUSION AND FUTURE WORK

From this study we have found that Visualizations can be static or dynamic. Interactive visualizations often lead to discovery and do a better job than static data tools. Do interactive visualization for more analyzing of data. Interactive brushing and linking between visualization approaches and networks or Web-based tools can facilitate the scientific process. Web-based visualization helps get dynamic data timely and keep visualizations up to date. More new methods and tools of Big Data visualization should be developed for different Big Data applications. Big Data analytics and visualization can be integrated tightly to work best for Big Data applications. More new graphs and More new visualizations could try in the future for better enhancement of visualization. A lot of technologies are developing for better visualization of bigdata and in future this can be developed by asking the user to enter data and ask like to which points data should be plotted.

CHAPTER 5

BIBLIOGRAPHY

These are the some references :

1. https://www.statmethods.net/graphs/index.html
2. https://www.rdocumentation.org/packages/ggalt/versions/0.1.1/topics/ggalt
3. https://www.rdocumentation.org/packages/gganimate/versions/0.1.0.9000/topics/gganimate
4. https://www.rdocumentation.org/packages/gapminder/versions/0.3.0/topics/gapminder
5. https://www.rdocumentation.org/packages/ggExtra/versions/0.7/topics/ggExtra
6. https://www.rdocumentation.org/packages/ggcorrplot/versions/0.1.1
7. https://www.rdocumentation.org/packages/quantmod/versions/0.4-11
8. https://www.rdocumentation.org/packages/scales/versions/0.4.1
9. https://www.rdocumentation.org/packages/ggthemes/versions/3.4.0

10. https://www.rdocumentation.org/packages/lubridate/versions/1.7.0
11. https://www.rdocumentation.org/packages/dplyr/versions/0.5.0
12. https://cran.r-project.org/web/packages/tabplot/vignettes/tabplot-vignette.html
13. https://www.rdocumentation.org/packages/hexbin/versions/1.29.0/topics/hexbin

https://www.rdocumentation.org/packages/rbokeh/versions/0.5.0

14. https://ieeexplore.ieee.org/abstract/document/8268737/
15. http://journals.sagepub.com/doi/abs/10.1177/2053951714528481
16. https://link.springer.com/chapter/10.1007/978-981-10-3223-3_44
17. http://journals.sagepub.com/doi/abs/10.1177/1558689816651015
18. https://dl.acm.org/citation.cfm?id=3078883
19. https://pdfs.semanticscholar.org/2975/4e4295a9ce4d51937c0712d6482634474628.pdf