

Data Visualization Project Report

Gadde Venkata Sai Kumar.

Url for the webpage: <http://128.119.243.147:5000/>

Motivation for this project:

The major motivation for this project is to help companies identify problems and issues so that they can provide good customer service. Companies are crossing boundaries to become multinational companies and therefore are required to provide goods for millions of people. With scaling of company comes the huge number of customers. Companies have to prioritize problems being faced by the customers so that they can solve the issues which are being faced by large number of customers first. So, in order to solve this problem, we can use data visualization to find the weak links in the functioning of the company and the companies can concentrate upon those weak areas. This analysis of data helps in bottoms up approach rather than the top down approach.

Goals of the Visualization:

- Drill down complaints to specific issues and prioritizing the issues.
- Visualize sentiment of the complaints to find the satisfaction level of the consumers. The sentiment analysis is done by state in USA, so that companies can focus upon their weak areas.
- Find the most persistent complaints based on the state.
- For Further research (If time permits): Twitter streaming API can be used to monitor the sentiment analysis of the consumers in real time social media.

Datasets: Datasets are collected from the two sources.

1. **Dataset1:** FDIC: Federal Deposit Insurance Corporation. FDIC preserves and promotes public confidence in the U.S financial system by insuring deposits in banks and thrift institutions for at least \$2,50,000, by identifying, monitoring and addressing risks to the deposit insurance funds and by limiting the effect on the economy and the financial system when a bank or thrift institution fails.
 - a. Dataset Size: [149385 , 18]
2. **Dataset2:** The dataset is about the distribution of different financial institutions across the USA. This is used to observe the density of financial institutions across the states. This can be used to compare with the number of complaints from the people.
 - a. Dataset Size: [93432 , 30]

Data Visualizations included in this project:

1. Histogram
2. Choropleth of USA
3. Word -cloud
4. Sparkline (Bar Chart and Pie Chart)
5. Row Charts
6. Chernoff faces

Methodology:


Data Auditing:

Step1 is about data cleaning. Data is not clean in the raw datasets. There are many nan values and those values are removed. As the number of rows became very less after removing the nan values from the dataset, the whole dataset is imported into the mongodb. Mongodb is a NoSQL database and the data is stored in the json format.

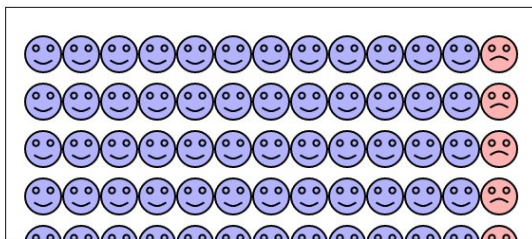
Spark Line:

In data visualization, context plays a crucial role. Sparklines can be used for “Data intense, design simple and word sized graphic is used to visualize the general shape of a visualization” and I found this claim of sparkline to be true, as it compared crucial statistics and can keep the user involved in the text.

can also be used to categorize the compl
I Credit Card. So, Companies need to cor
panies.  The plot shows

is which assists companies i
companies can't solve the maj
to a person  .41 billion\$ i
blems can be classified broad
customers can receive better

Chernoff faces: I have two important objectives for this visualization. This visualization is a primer or is very significant which conveys the importance of proper customer care for the companies. This is a very interesting statistic made by *Ruby Newell-Legner*, as it takes 12 positive experiences to make up for one unresolved negative experience. The second objective is the waiting time for the visualizations to load (In case of mongodb). This small animation can help in iterating the statistic which is significant.



Source: 'Understanding Customers' by Ruby Newell-Legner

Choropleths:

As evident, choropleths helps in obtaining the overall view of distribution of the parameter in consideration across a country. As given in the draft layout, initial layout is done. But after the initial data auditing, I have observed many of the companies received few number of complaints and few of the companies received many of the complaints. So, I have done some exploration of data for these companies. So, I have created a second web page which contains detailed specific analysis of 5 companies (Bank of America, Morgan Stanley, Capital One, Wells Fargo, JPMorgan and Chase.) which are considered as the largest banks in USA. I have observed that the visualizations can solve a real problem. The observations and conclusions are listed in the webpage. Good insights can be observed from the visualization which cannot be observed by viewing tables.

Sentiments of the complaints: The sentiments of the complaints are calculated using the TextBlob library in python and the data values are appended to the database.

[illegible]

1. Dc.js
2. D3.js
3. Pandas, Crossfilter.js
4. TextBlob, NLTK
5. PyMongo

Two approaches are done for building website for this project. Mongodb (NoSQL) database is used in one approach and csv files are imported in the other approach. As the size of the dataset is large with high dimensions([149385 , 18]), this approach is not used for the final edlab hosting. This is tried and tested in the local host and the code is uploaded for grading. I would like to show this full stack as a demo. Please follow the instructions in the readme.txt for hosting this full stack in the local host.

The code for the second approach is in the edlab account. And a zip file is uploaded.

Future Work:

I would like to work on visualizing datasets to find the correlations in the data. I would like to use visualization tools to get an intuitive idea of the data to use it in the machine learning, which can help reduce the number of trial and errors generally done for tuning machine learning algorithms.

Course feedback:

I am a student from mechanical department and I converted to computer science for my masters. This course has been a wonderful introduction for me to the web development and on top of that, to the data wrangling and its visualization. George Grinstein has been a wonderful teacher and his intuitive teaching has helped me to obtain an intuitive feel of the subject and its importance in the world of big data rather than implementing a few algorithms.

Thank you.