Narrative Visualization Project:

FINANCIAL SERVICES CONSUMER COMPLAINT DATABASE

Consumer Financial Protection Bureau

By: Ananya Pandey

TABLE OF CONTENTS

INTRODUCTION

The Consumer Financial Protection Bureau works to hold financial institutions accountable in matters related to financial products & services provided by them and manages a database which is a collection of complaints about consumers' issues with the utilities provided to them by particular companies.
In recent times, the number of complaints is increasing at alarming levels as the purchasing capacity for financial products and services is increasing for the residents of the USA. It is predicted that because of the rise in awareness among people about such a platform where their voice can be heard, there will be a rise in the number of complaints.

Complaints give a better understanding of the problems faced by common people in several products and services offered by companies, helps in enforcing the laws and inform the consumers to take financial decisions only on being well informed. Also, this platform is made available to encourage impartiality and honesty in various financial services and products.

This project is an extension of the Data Exploration Project that was done previously, this Data Visualisation Project is aimed to develop interactive narrative visualizations showing significant relationships and trends in the financial marketplace. In particular, the trend of the volume of the complaints in the years 2016 to 2019, the performance of companies on the basis of their mode of submission, dispute rate, resolved rate or the products getting those complaints for the companies, and the impact of localities and neighbourhood on the complaints.

This complaints database is a rich source for the financial analysts and the companies looking for the trends about customer complaints. Thus, the target audience for this visualization project is the customers, financial analysts and the financial institutions. This can be helpful for the customers in determining which product is leading to the maximum complaints against a particular company and will help the customer to invest in companies who are performing better than others. Financial analysts look for the trends of the complaints and give advice to the financial institutions regarding the reasons for the complaints and the steps that should be taken to increase customer satisfaction which is the main motive of the companies.

DESIGN

Under this section, the description of the design process will be discussed. In the Five – Design Sheet method, the first sheet is the Brainstorming (or Ideas) sheet where any idea to the project is welcome. ("Five Design Sheet | Design Methodology for Visualisation", 2020)
For this project, a lot of brainstorming was done where many ideas for visualisation like a bar plot, stacked bar graph, scatter plot, word cloud, bubble chart, line graph, map representation of the states of the U.S. and pie chart arose, and then the ideas were filtered according to the needs of the project.
As bar graphs are really easy to understand and a good way for comparing between groups, so it is a good choice for showing the changing trend in years or to show the performance of the companies.
Line charts are a good choice when the data shows both small and large variations as in the case of the complaints for each of the months in a year.
Word cloud is a beautiful as well as an informative way of depicting textual data according to the frequency of the words and their importance.
Bubble maps, a variation of bubble chart, displays different sized circles according to the values according to their geographic location.

Pie chart represents a data having less variables in a very simple, convenient and easy to understand manner, so it is a good choice to use it for the various modes of submission.
Bubble chart is a good way of showing relationship between numbers ("1331.0 - Statistics - A Powerful Edge!, 1996", 2020).
So, the filtered ideas are then categorized where the similar ideas are kept together. Then these ideas are combined and refined in order to get closer to the bigger main design. At last, it is checked whether the visualisation is answering the original research questions or not, it is in this case.

The second, third, fourth sheets are the initial design sheets where the vision of the final visualisation is shown in three alternative ways in the three sheets accordingly. (Roberts, Headleand & Ritsos, 2015)

In the initial design 1 i.e. sheet 2, the volume of complaints for the period of 2014 – 2019, the performance of the companies over the years is shown as a bar graph as it helps in clearly seeing the trends and the patterns in that period. The dispute rate of companies and the resolved complaints are also shown as a bar graph because the changing trend can be easily observed through such graphs. To show the relation of zip code vs complaints with respect to the total wages in that area, is shown through a scatter chart as it is a good way of presenting anything related to finance. This design has its pros and cons. It is simple to understand and very clear in showing the trends but all the factors and details are not included in this design.

The initial design 2 i.e. sheet 3 shows how the complaints are distributed over the years in the form of a bar graph and the monthly distribution of complaints for each of the year can be seen through a line chart which is a good way of showing any trends over time. The performance of companies is displayed as a bar chart to compare their performance with each other. A word cloud is shown which tells about the products driving the maximum number of complaints for that particular company. Word cloud is an effective measure when analysing textual data. The states and the volume of complaints from those states is shown through a bar graph. This design has its advantages and disadvantages. It has simple graphs which are easy to interpret but this design lacks the extra details that should be portrayed through our visualisation.

The initial design 3 i.e. sheet 4 focuses on the performance of the companies which is represented through a bar plot and the performance of the states i.e. the states from which the maximum number of complaints are being lodged. A pie chart is shown which tells about the categories of the mode of submission of the complaints, it is a good way of depicting categorical data. A bubble map is displayed showing the state wise distribution of the complaints with the size of bubbles increasing with the number of complaints. A bubble chart is shown which tells about the relation of zip code of a particular state and the complaints with respect to the total wages in that area. The advantage of this design is that it is highly interactive but the companies' graph can get congested due to the large number of companies.

The last sheet, sheet 5 i.e. the Realization sheet is the sheet that is the closest to the actual visualisation where some ideas of all the designs are considered in order to reach to our final visualisation. In this a bar chart is used to show the volume of complaints over the years i.e. an yearly trend for the number of complaints received each year, a line graph to show the trends of the number of complaints for the months of the year chosen by the user, the user can see the performance of the top 20 worst performing companies for the selected year and the features i.e. the mode of submission for the complaints i.e. which medium was used for submitting the complaint whether it was submitted via email, web, fax, phone or through a referral; the products for which the complaints are registered, the dispute rate and the rate of resolved complaints can be selected for a company. A map showing the value of the complaints for each of the states is shown with the bubbles in the map representing the volume of complaints in that state. The map has a hover feature to know the exact amount of complaints from that particular state during the period of 2014 to 2019 when the user sees the map. The relationship between zip code and the

neighbourhood according to the total wages in that area for a particular state can be seen through the bubble chart where the size of the bubbles represents the total wages in that area i.e. larger the size of the bubble, more is the locality prosperous.

This final design is highly interactive and answers all the research questions as well. It also discusses about the software that is going to be used for implementation of this design. For this project, RStudio is used for making the interactive visualisation through the Shiny application. The time needed to develop the project will be around 90 to 95 working hours.

The visual variables like position as in the case of the map where the circles are plotted according to the location; size in the case of the bubble chart where the size of the circle depicts the volume/count; hue i.e. the usage of different colours that are attractive for the human eye. The colours are chosen to cater for colour blindness as well so colours like red and green are avoided in the interactive visualisation. Visual variable like value which refers to the darkness of the colour is used in the map where darker colour of the circle marker shows more is the number of complaints in that state. In this final design neither too much or too little information is provided which shows that it is highly relevant for the reader. Attention of the user is directed to the important parts by the interactions provided which keeps the user hooked on to the final design. The final design is a mix of text and visualisations that are interactive which makes it interesting and easy to understand for the user with appropriate background knowledge.

IMPLEMENTATION

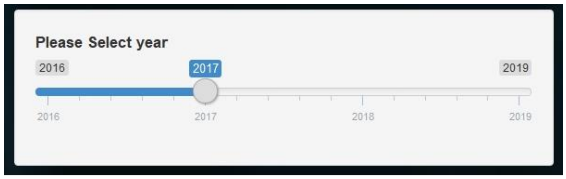
This section includes the description of the implementation of the design. The first data set is a “Consumer Complaint Database” which is a collection of complaints about consumers’ financial products & services that are sent to companies for response by the Consumer Financial Protection Bureau. This dataset includes the date of the lodgement of the complaint, the product, sub-product, issue, sub-issue, the complaint narrative of the consumer, the company’s public response, the company against whom the complaint is issued, the state provided by the customer, the area’s zip code, whether the consumer has given consent for publishing the narrative or not, tags of whether the complaint was submitted by the consumer or on behalf of the consumer, the medium used for submitting the complaints, the date the complaint was sent to the company, the company’s response, whether the response was timely or not, whether the consumer disputed or not and the identification number of the complaint.

The second dataset is of “U.S. Federal Government zip codes database” which includes all the zip codes of America, the type of zip code whether standard, PO box or unique, city, state, latitude, longitude, number of tax returns filed in that particular area, the estimated population of that area and the total wages for each of the zip codes.

Initially, implementing the design with such a huge dataset was a difficult task, but the data is wrangled in order to make it ready for the further visualisations. But even after significant wrangling and filtering the complaints only for the years 2016 to 2019, the data is still very huge, which is a big challenge faced during the implementation process.

Implementation of the final design as highlighted in the design process was not a simple task. It required a lot of time, efforts and hard-work to come up with the final interactive visualisation.

In order to implement the final design, the first thing is that a slider input widget has been created for the user to select the year that he/she wants to see the features and details for.



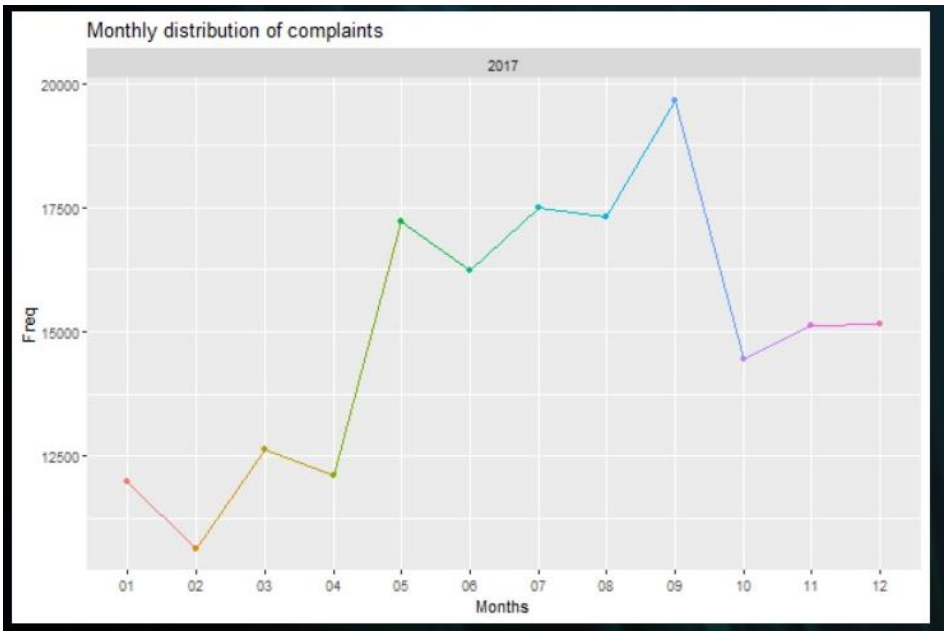
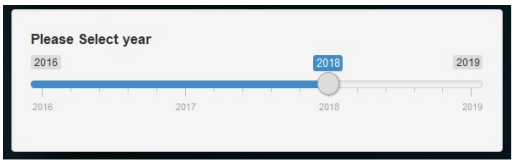


Figure 1: Month-wise number of complaints for the selected year of 2017.

A month-wise breakdown is shown for each of the years to show how complaints are increasing and decreasing over time, this also shows that 2018 had a huge decline in the number of complaints because of the stock-market crash in the United States which led to a rapid decline in people buying products and thus lesser and lesser complaints were registered during that time.



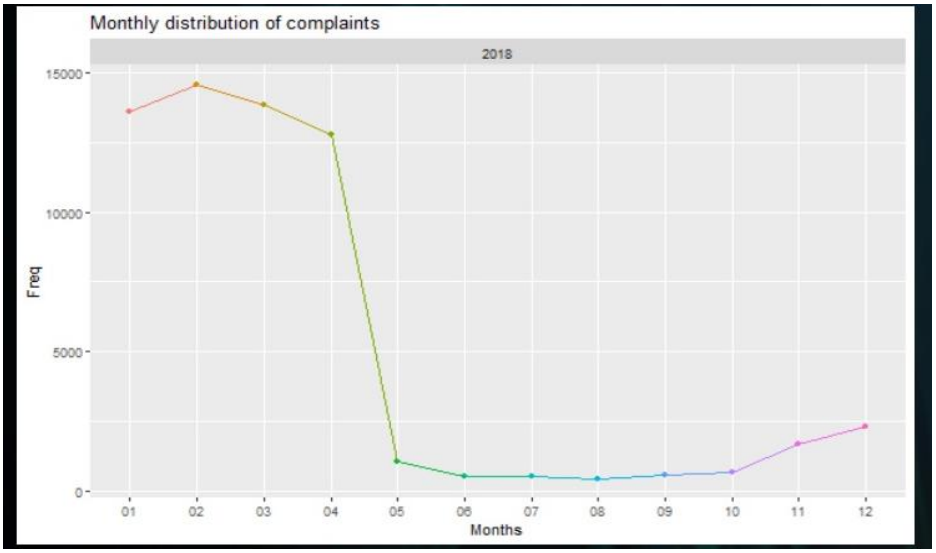
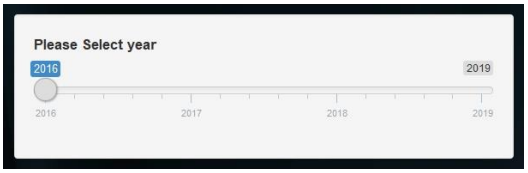


Figure 2: Month-wise number of complaints for the selected year of 2018 showing the decline in complaints' volume.

This is presented in the form of a line graph through the “ggplot2” library as line graphs represent changes with time in a very sophisticated manner.

For the same selected year, the top 20 worst performing companies are shown with the help of a bar graph using the “ggplot2” library. This bar chart shows the performance of the companies in general with respect to the number of complaints, and then an input widget is created to select the company for which the user wants to see the performance with respect to the list of features which can be selected from another input widget, in the form of a drop-down menu, created for the features which include the mode of submission of the complaints, the percentage of resolved complaints, the products for which the complaints are registered or the dispute rate of the company which are made using the “ggplot2” package in R.



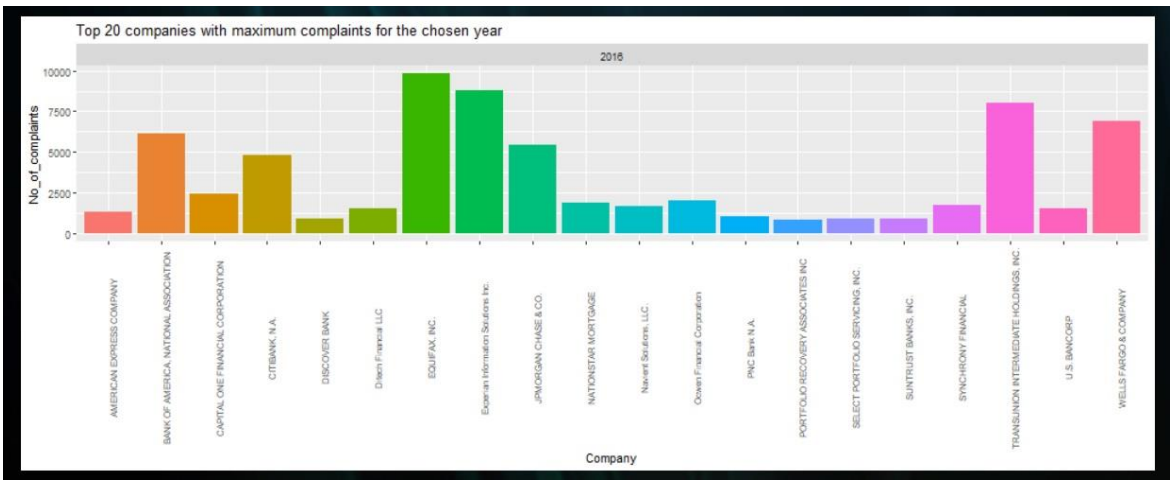
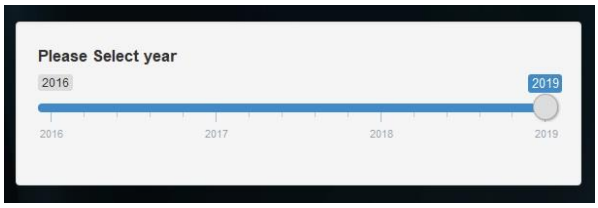


Figure 3: Top 20 companies with maximum complaints for 2016.



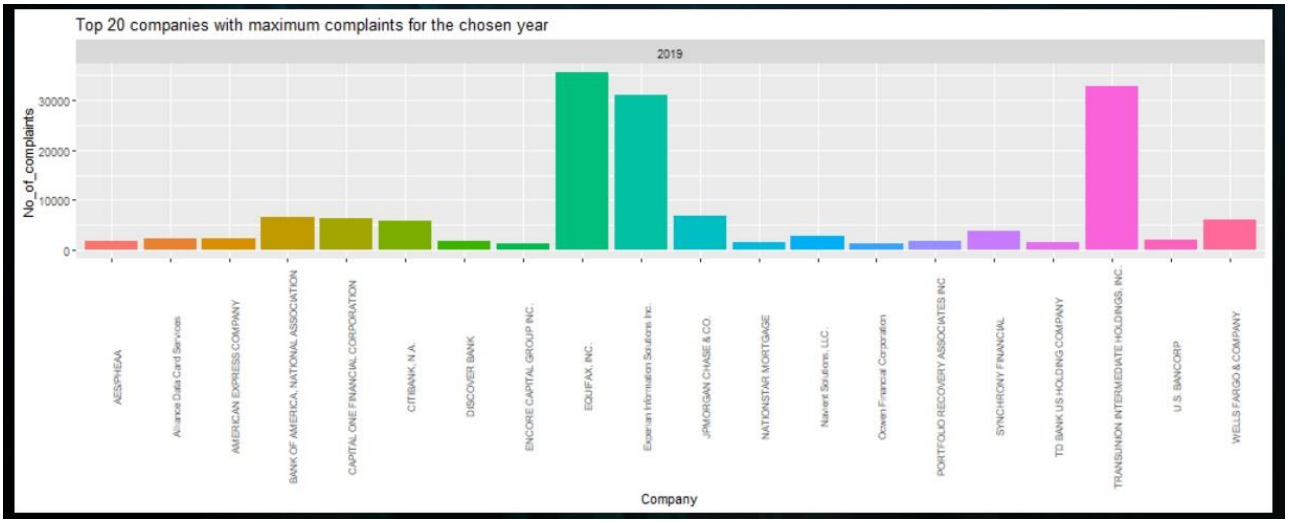


Figure 4: Top 20 companies with maximum complaints for 2019.

Here, the mode of submissions is shown in the form of a pie chart which shows the distribution of the mediums used for submitting a complaint. Pie charts are extremely simple to understand and is a good choice when there are less categories in a particular variable.

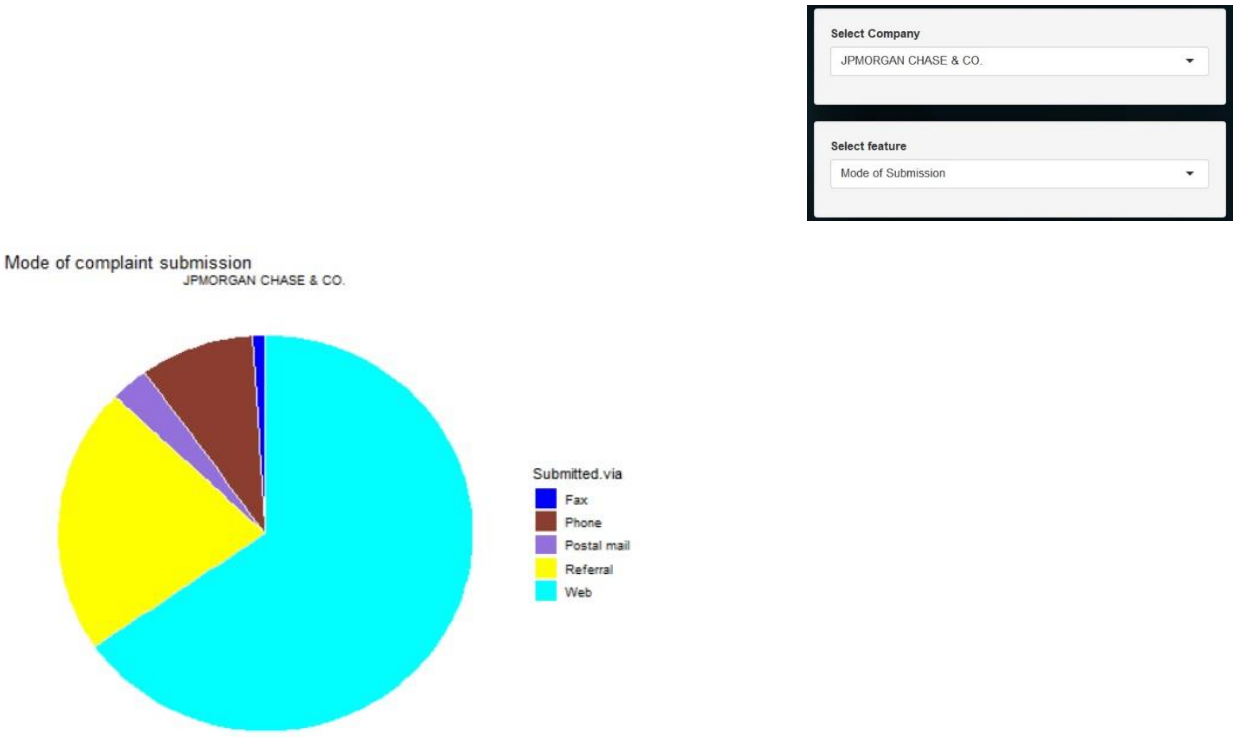


Figure 5: A pie-chart showing the mode of submission distribution for JPMorgan Chase & Co.

The resolved complaints showing the percentage of complaints resolved in fav our and not in favour of the consumer for the company for the chosen year and the chosen company. A bar chart is used for showing this because it is a great option for showing discrete values and easily compares the percentage of complaints resolved in favour and not in favour of the customer. ("What Type of Chart Is Used to Show Discrete Data | WebDataRocks", 2020)

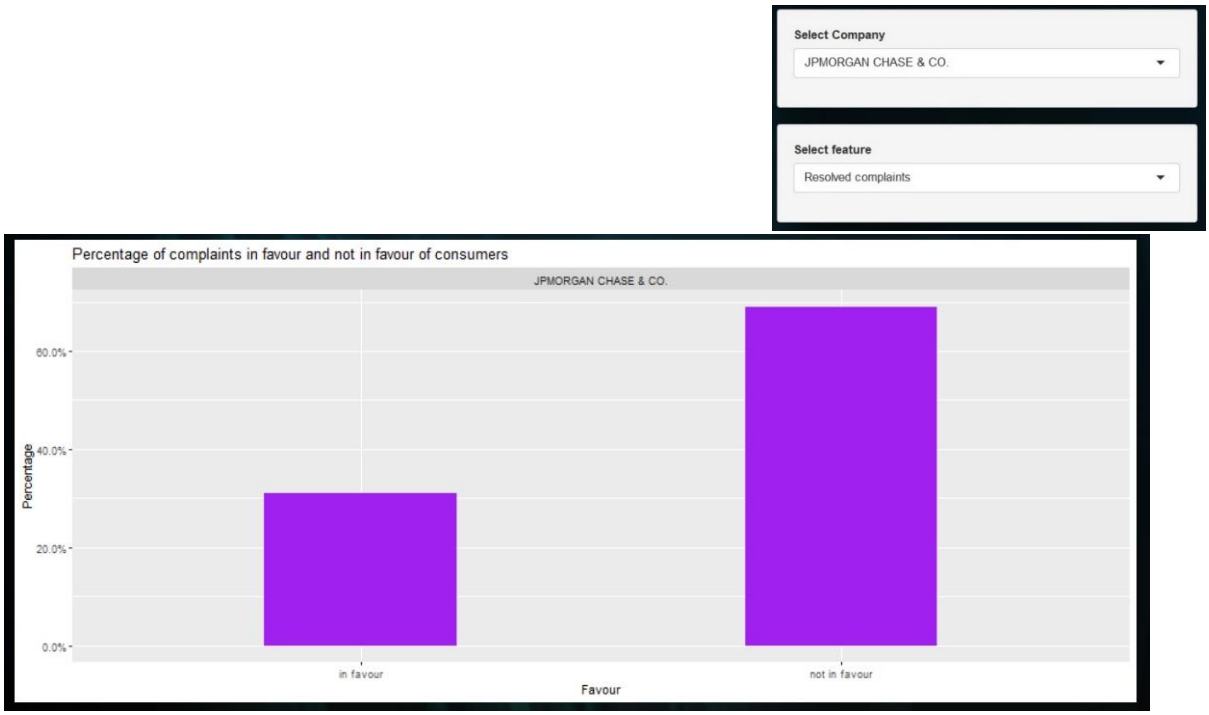


Figure 6: A bar chart showing the percentage of complaints resolved in favour and not in favour of the consumer for JPMorgan Chase & Co.

The products are displayed as a word-cloud which tells about the products that drive the maximum complaints for a company for the chosen year. Word clouds are a great pick when textual data is analysed by highlighting the word with the maximum frequency. Libraries like “tm”, “wordcloud”, “RColorbrewer” and “SnowballC” are used for generating this word cloud. ("Text mining and word cloud fundamentals in R : 5 simple steps you should know - Easy Guides - Wiki - STHDA", 2020)

Figure 7: A word-cloud showing the products attracting the maximum number of complaints from U.S. BANCORP.

The company’s response to the consumer is displayed in the form of a bar chart. A bar chart drawn with the help of the “ggplot2” package is used as it is an ideal choice when comparison between different groups is to be shown.

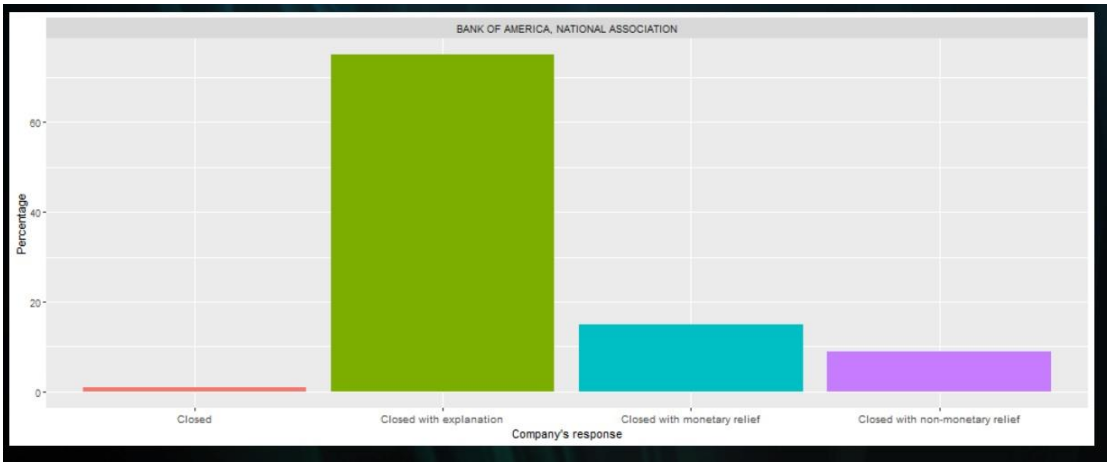


Figure 8: A bar chart showing the percentage of the various responses of the Bank of America to its consumers.

Then, a bubble map is implemented using the “leaflet” package which shows the total number of complaints from each of the states of the U.S. A tooltip showing the state and the number of complaints from it is shown when the user hovers over the states.

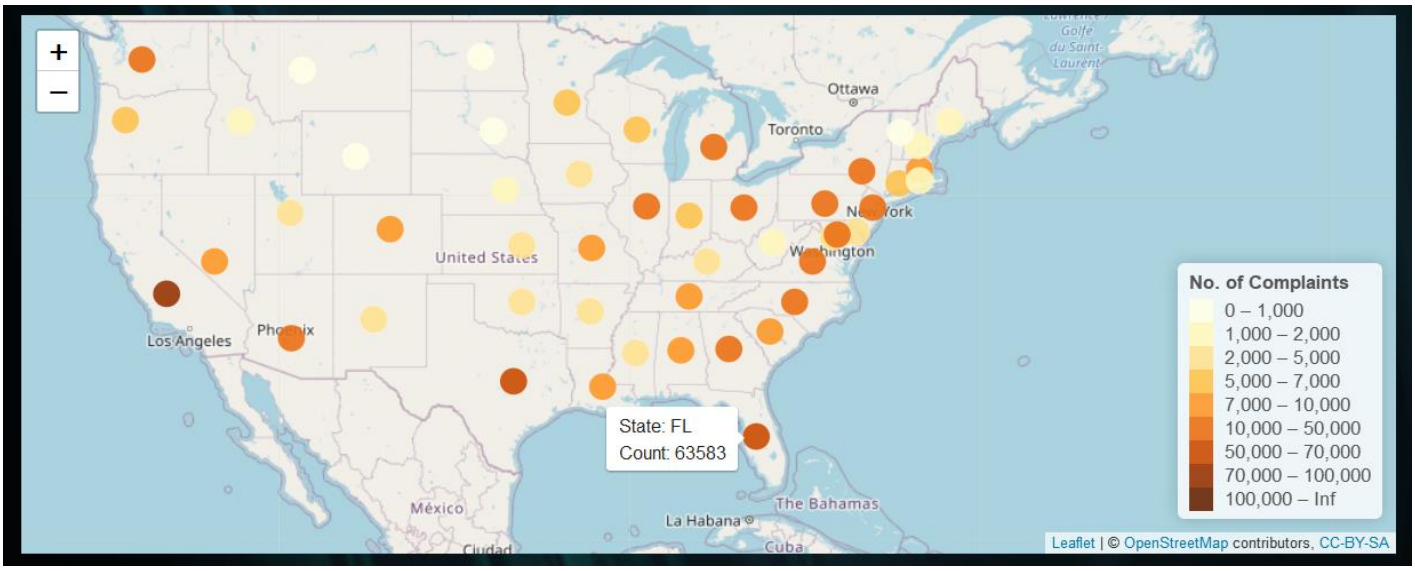


Figure 9: A leaflet showing the number of complaints received from each of the states.

Now, there is an input widget in the form of a drop-down menu to select the state for which the user wants to see the effect of neighbourhood on the complaints. In order to implement it, the top 1000 complaints are considered to design a bubble chart using the “ggplot2” library, to observe the effect of localities on the number of complaints.

Figure 10: A bubble chart showing the effect of the neighbourhoods on the number of complaints.

Various packages and libraries have been used in order to implement the final design.

`library(shiny):` The “shiny” library is used to run the shiny server in order to make interactive web applications using R where the analysis is put forward in the form of interactive visualisations.

`library(ggplot2):` The “ggplot2” is a plotting package used for creating graphs showing the details of the data frame It is used for data visualisation using the grammar of graphics.

`library(dplyr):` The “dplyr” library is used for manipulating the data i.e. in our case, filtering the data using ‘select()’ which choses the variables according to the needs of the user, summarizing according to some conditions and mutating the data according to the user requirements.

`library(shinyWidgets):` The “shinyWidgets” package is used for customising the input widgets of the shiny application. This package provides a variety of input options where the classic input method can be changed to a slider input or to a drop-down menu according to the choice of the designer.

`library(data.table):` The “data.table” library is used for conversion of the data to a data table, data table is very convenient when any type of sub-setting or merging of the data has to be carried out as it does not convert the columns with character variables to factors.

`library(stringr):` The “stringr” package is used for string manipulation, white space manipulation or the pattern matching functions.

`library(tm):` The “tm” is the text mining package which helps in the conversion to lower case, remove numbers and punctuations and extra stop-words as well.

`library(SnowballC):` The “SnowballC” package performs the word stemming algorithm which gives common root to words for further comparison of words. (Bouchet-Valat, 2020)

`library(wordcloud):` The “wordcloud” package is used for word cloud generation which helps in textual analysis and the frequent words are visualised as a word cloud. It is appealing as compared to scatter plots which can get extremely congested. ("Text mining and word cloud fundamentals in R : 5 simple steps you should know - Easy Guides - Wiki - STHDA", 2020)

`library(RColorBrewer):` The “RColorBrewer” is used for loading various colour palettes for creating visualisations with beautiful colours. ("RColorBrewer function | R Documentation", 2020)

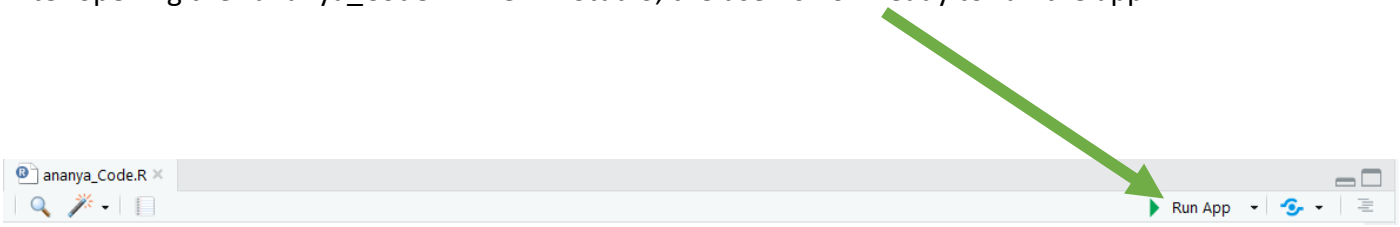
`library(leaflet):` The “leaflet” library is used for creating maps with interactive features of adding markers, lines etc. according to our data frame values.

USER GUIDE

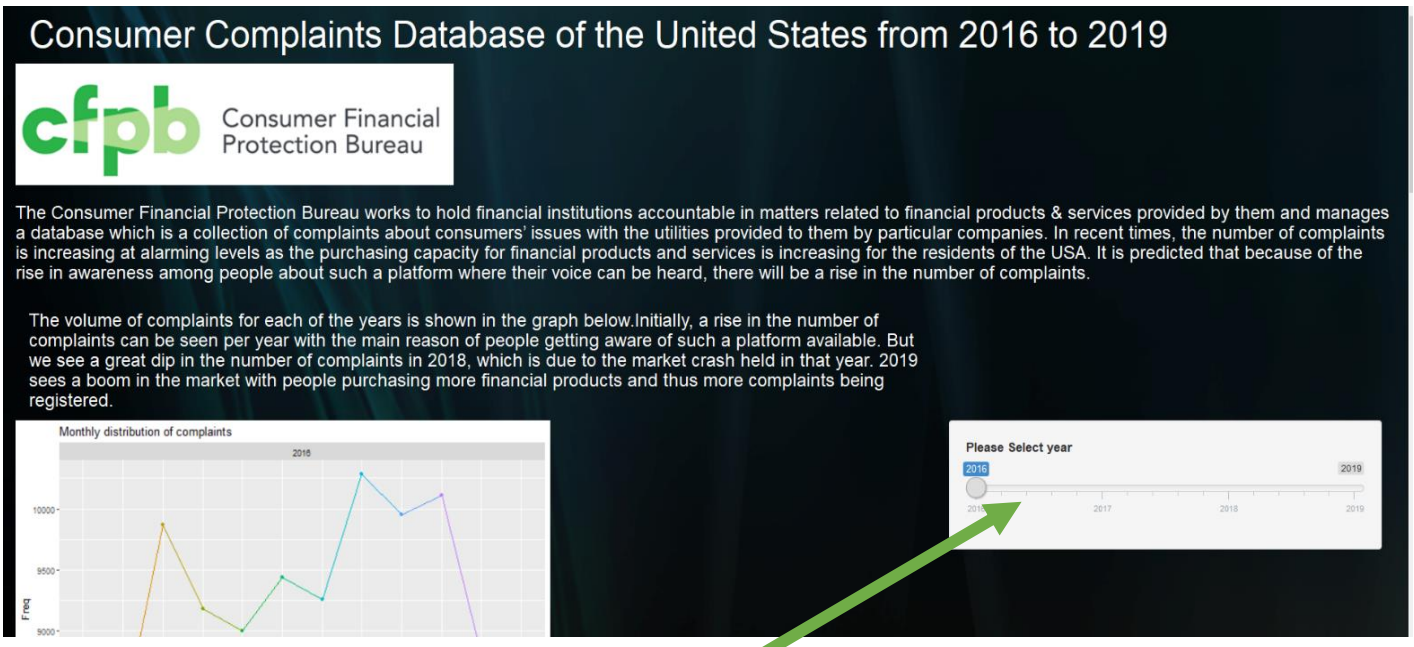
For running the app, the user needs to have RStudio with the required libraries installed in his/her device. Then all the files related to the project (as provided) must be placed under their RStudio's working directory.

Note: A file called "www" is in the folder and should definitely be there in the working directory in order to get the desired layout of the application.

After opening the "ananya_Code.R" file in RStudio, the user is now ready to run the app.



As soon as the app is run, the basic structure of the app is visible in no time, but as the dataset is pretty huge, it takes a few seconds in loading the data and presenting the visualisations.

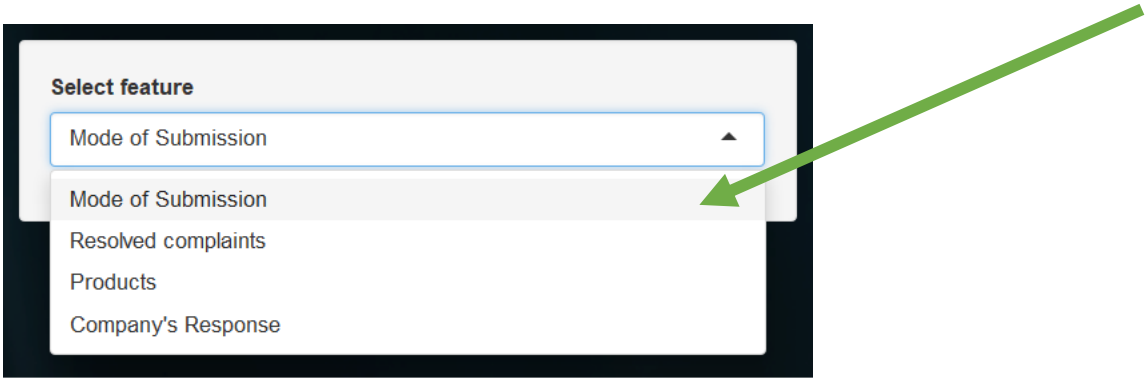


The first thing that the user sees is a slider bar asking the user for selecting the year. The user can play around with it to see the month-wise breakdown of the number of complaints registered to the CFPB for the selected year.

In the next graph presented, the user can see the top 20 companies with the maximum number of complaints for the year selected by the user.

For exploring the visualisation, the user now selects the company for which he/she wants to know the features for.

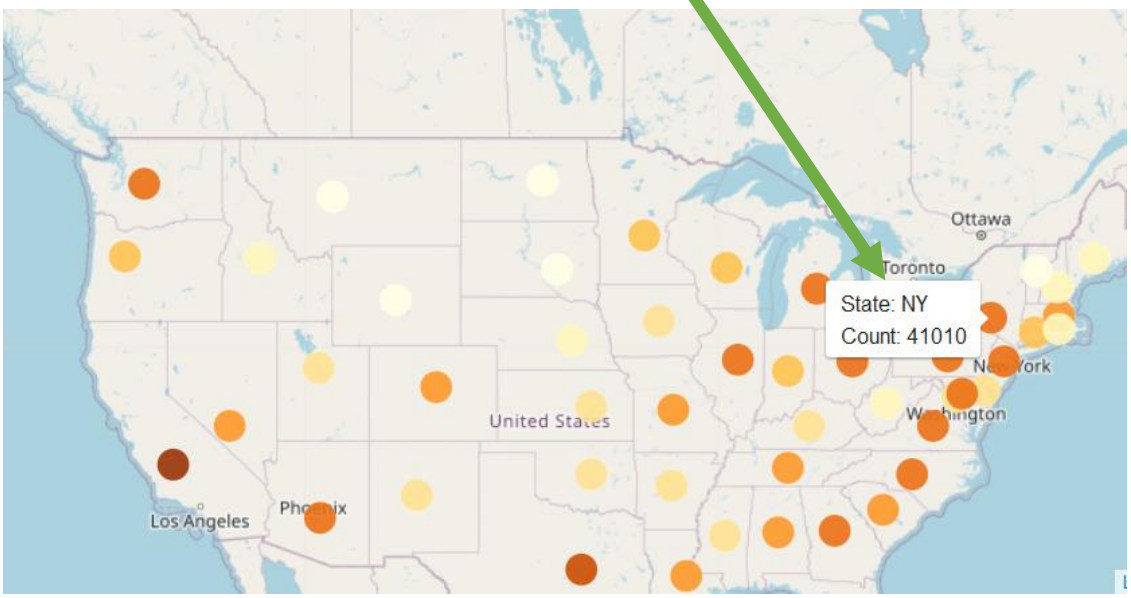
Now, the user can choose the feature he/she wants to see for the chosen company.



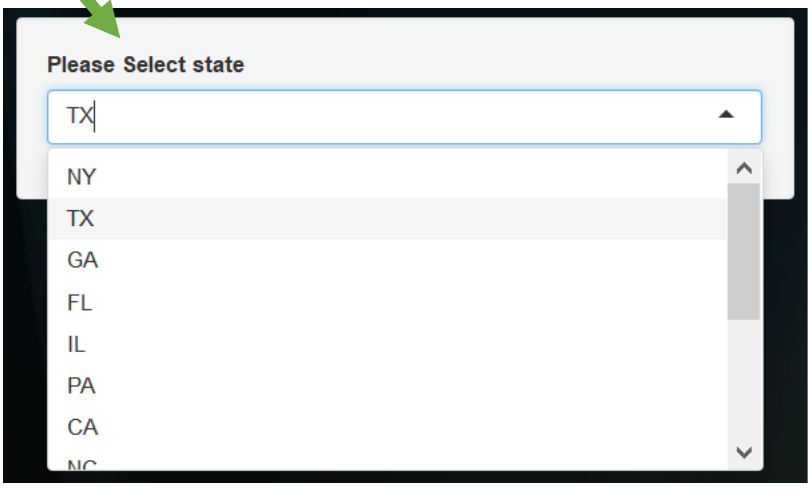
The features are: the mode of submission i.e. how the complaint against a company has been submitted to the CFPB, whether it was by web, fax, referral, phone or postal mail; resolved complaints i.e. how many complaints were resolved in the favour of the consumer or not in favour of the consumer; products i.e. the products that are driving the maximum complaints against the chosen company for the year selected by the user; and company's response i.e. how the company responded to the consumer complaint by closing with explanation or providing some monetary relief or no explanation was provided.

Next, the user sees a map which shows the states of the United States and the number of complaints received from those states. The volume of complaints is shown by colour i.e. darker the colour, more is the number of complaints.

The user can hover the mouse over the circles to see a tooltip that shows the states and the total number of complaints registered by that state in the period of 2016 to 2019.



Now, the user has a drop-down menu from where the state can be selected.



The bubble plot now shows how the neighbourhood affects the number of complaints. The size of the bubbles is with respect to the total wages in that zip-code i.e. larger the size, higher is the total wages of people in that area. ("How to Create a User Guide", 2020)

CONCLUSION

This narrative visualization project fulfils the research questions specified in the initial project proposal. Initially, the project begins with identifying the target audience for whom the project can be beneficial. The focus group for this project are the customers, financial analysts and financial institutions. Then the five-design sheet method is used in order to design and consider alternative options, and then a realization sheet which is then coded in order to implement the design produced.

The implementation part of the project shows how the complaints are distributed for each month of the year that is selected. For the selected year, the worst 20 companies based on the number of complaints against them are shown. For the entire period, EQUIFAX, Experian Information Solutions, Bank of America and TRANSUNION Intermediate Holdings are the companies attracting the largest amount of complaints. Now, the user can select the company for which he/she wants to see the features for.

From the visualisations it can be seen that web is the most preferred medium of complaint submission by the consumers; around 60-65% complaints are resolved not in the favour of the consumers; credit, debt and mortgage are the products driving the maximum number of complaints and maximum companies try to close their complaints just with an explanation. From the leaflet, it is clearly visible that states like California, Texas and Florida generate more number of complaints as they are economically better than other states. Lastly, the bubble chart shows how the established and prosperous localities of the chosen state tend to register a higher number of complaints i.e. localities with people having a higher wage tend to buy and invest more in financial products.

This project helps in understanding the importance of communicating the insights found by the exploratory analysis of the data through narrative interactive visualisations built from the scratch. This project also helps in understanding how the visual representation of data in the form of interactive graphs and charts is important to identify trends and patterns easily and further assists in building familiarity with the analysis of the data by showing interesting insights through interactive visualisations. ("What is Data Visualization and Why Is It Important? | Import.io", 2020)

This project assists in developing a deeper knowledge on how interactive applications are made in RStudio using the shiny package. This project helps us in distinguishing which type of graph should be used with what type of data, this is an extremely important task in order to show visualisations for presenting the data in a simple interactive manner.

Although this might not be the perfect way of showcasing an interactive narrative visualisation using this data, with more resources and better understanding better features, some animations and a better layout could have been added to the visualisations. A lot more inferences could be extracted out of this data by expanding and going beyond the scope of the project proposal and it would be interesting to see how different variations and factors could improve the quality of visualisations, thereby fetching more information from the dataset.

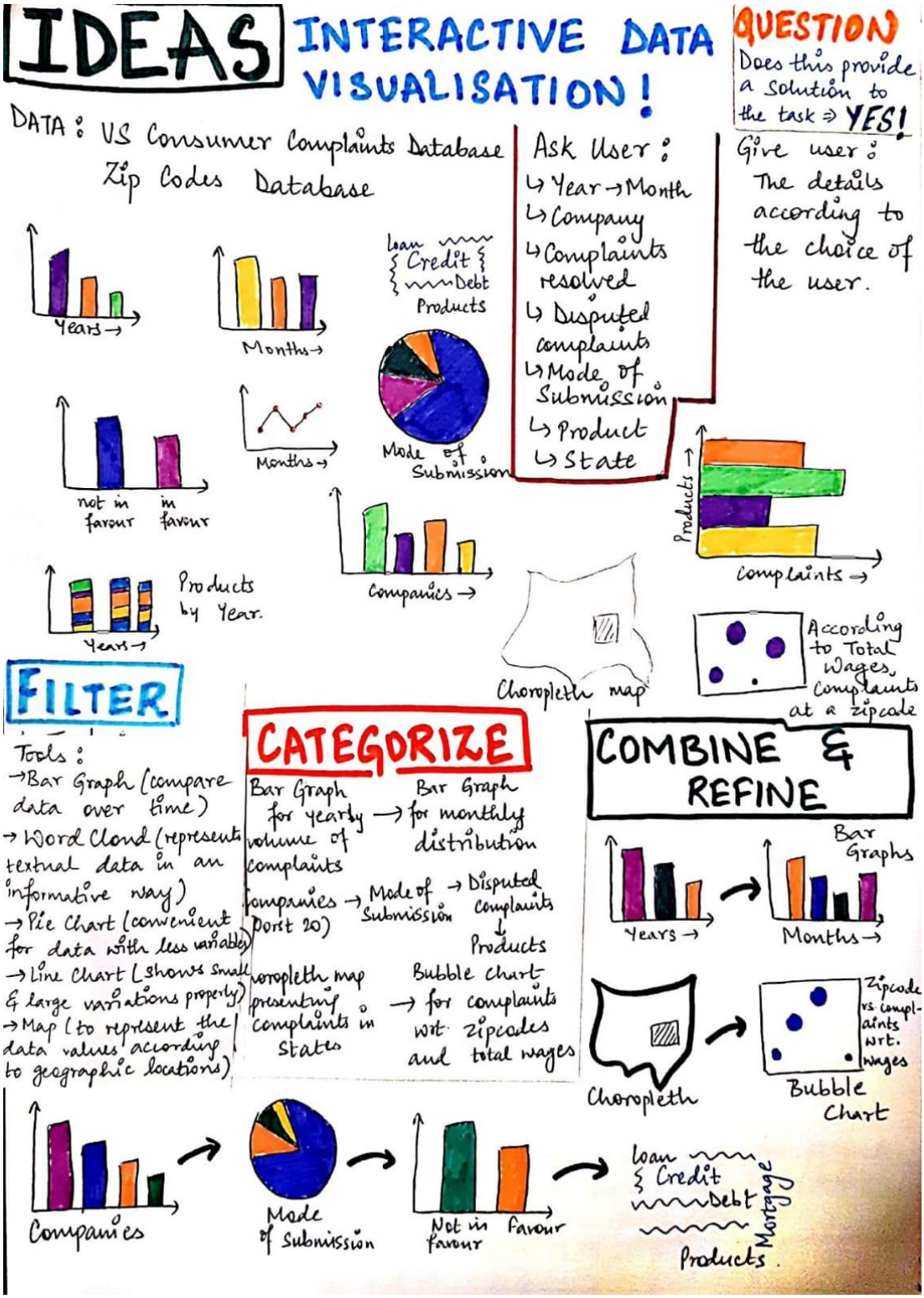
BIBLIOGRAPHY

1. Consumer Complaint Database. (n.d.). Retrieved from https://www.consumerfinance.gov/data-research/consumer-complaints/Consumer Financial Protection Bureau
2. CFPB Open Tech. (n.d.). Retrieved from https://cfpb.github.io/api/ccdb/fields.html
3. U.S. Government's Zip codes dataset. Retrieved from https://github.com/MacHu-GWU/uszipcode-project/blob/master/dataset/federalgovernmentzipcodes/federalgovernmentzipcodes.zip
4. 1331.0 - Statistics - A Powerful Edge!, 1996. (2020). Retrieved 15 June 2020, from https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/DEFA265B565EBF93CA2571FE007D69DA?opendocument
5. Bouchet-Valat, M. (2020). Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library [R package SnowballC version 0.7.0]. Retrieved 17 June 2020, from https://cran.r-project.org/web/packages/SnowballC/index.html#:~:text=SnowballC%3A%20Snowball%20Stemmers%20Based%20on,to%20aid%20comparison%20of%20vocabulary.
6. Text mining and word cloud fundamentals in R : 5 simple steps you should know - Easy Guides - Wiki - STHDA. (2020). Retrieved 17 June 2020, from http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know#:~:text=The%20text%20mining%20package%20(tm,keywords%20as%20a%20word%20cloud.
7. RColorBrewer function | R Documentation. (2020). Retrieved 17 June 2020, from https://www.rdocumentation.org/packages/RColorBrewer/versions/1.1-2/topics/RColorBrewer
8. How to Create a User Guide. (2020). Retrieved 17 June 2020, from https://stepshot.net/how-to-create-a-user-instruction-manual/
9. Five Design Sheet | Design Methodology for Visualisation. (2020). Retrieved 17 June 2020, from http://fds.design/
10. What is Data Visualization and Why Is It Important? | Import.io. (2020). Retrieved 17 June 2020, from https://www.import.io/post/what-is-data-visualization/
11. What Type of Chart Is Used to Show Discrete Data | WebDataRocks. (2020). Retrieved 17 June 2020, from https://www.webdatarocks.com/blog/best-charts-discrete-data/
12. Roberts, Headleand, & Ritsos. (2015). Retrieved 18 June 2020, from https://www.cs.odu.edu/~mweigle/courses/cs725-s17/FdS-overview.pdf

APPENDIX

Five design Sheets-

Sheet 1:



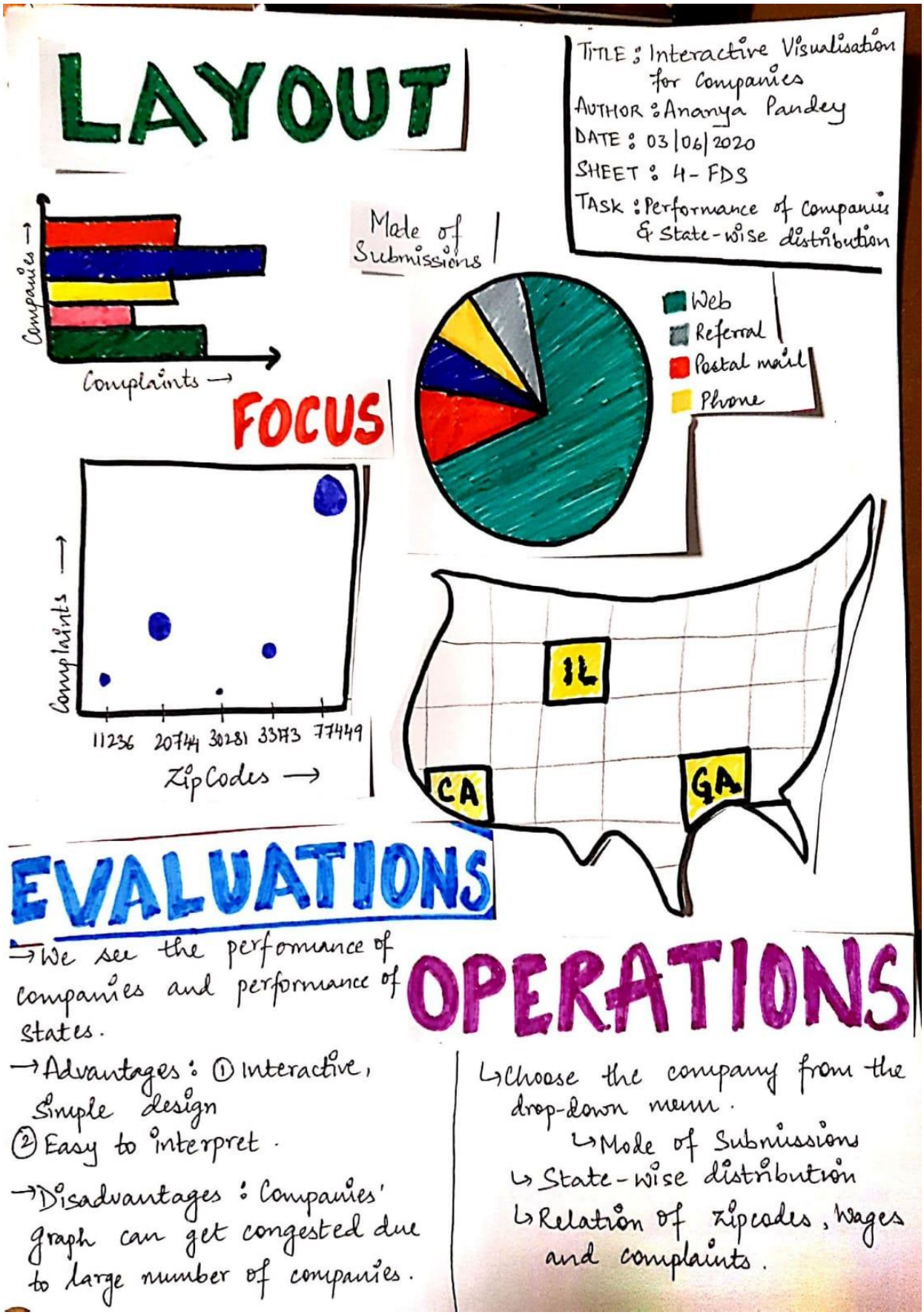
Sheet 2:


LAYOUT
TITLE: Interactive Visualization of complaints
AUTHOR: Ananya Pandey
DATE: 03/06/2020
SHEET: 2 - FDS
TASK: Timeline of Complaints.
FOCUS:
No. of complaint
2020 2019 2018 2017
No. of Complaints
Equifax
Experian Information Sln.
Bank of America
Transunion
Companies
Resolved complaints
Dispute rate
Relation b/w wages, zipcode & complaints
Zipcodes
OPERATIONS
→User can select from the drop-down menu:
↳ Year
↳ Company
↳ Features: Resolved complaints
Dispute rate
Zipcodes, wages vs. complaints.
EVALUATIONS
↳ We get yearly number of complaints, we see performance of companies and their features for that year.
↳ Advantages: ① Simple design
② See year-wise trends
↳ Disadvantages:
① All details are not included in this design.
FOCUS
→ To get the details for the companies year-wise.
→ To join the datasets to get relation of wages with respect to neighbourhoods and complaints.

Sheet 3:


LAYOUT
TITLE: Interactive Visualisation for Companies
AUTHOR: Ananya Pandey
DATE: 03/06/2020
SHEET: 3- FDS
TASK: Performance Chart of Companies.
FOCUS:
No. of complaints
2020 2019 2018 2017
Years →
No. of Complaints
Equifax
Experian Information Sln.
Bank of America
Transunion
Companies →
No. of complaints
Months →
Loan
Repair
Credit
Debit
Debt
Mortgage
Word cloud for products
No. of complaints
TX CA NY
States →
OPERATIONS
EVALUATIONS
↳ We get yearly number of complaints, We see performance of companies.
Month-wise distribution of complaints.
State-wise distribution of complaints
↳ Advantages: ① Simple design
② Easy to interpret
↳ Disadvantages: Lacks extra details and features.
↳ User can select from the drop-down menu:
↳ Year
↳ Companies
↳ Products
↳ States
FOCUS
→ Get details for the companies year-wise.
→ Get products for each of the companies.

Sheet 4:


LAYOUT
TITLE : Interactive Visualisation for Companies
AUTHOR : Ananya Pandey
DATE : 03/06/2020
SHEET : 4 - FDS
TASK : Performance of companies & State-wise distribution
Companies →
Complaints →
Mode of Submissions
Web
Referral
Postal mail
Phone
FOCUS
Complaints →
11236 20744 30281 33173 77449
Zip Codes →
IL
CA
GA
EVALUATIONS
→ We see the performance of companies and performance of states.
→ Advantages : ① Interactive, Simple design
② Easy to interpret.
→ Disadvantages : Companies' graph can get congested due to large number of companies.
OPERATIONS
↳ Choose the company from the drop-down menu.
↳ Mode of Submissions
↳ State-wise distribution
↳ Relation of zipcodes, wages and complaints.

Sheet 5:


LAYOUT
Years →
Complaints
Months →
TITLE: Interactive Data Visualisation
AUTHOR: Ananya Pandey
DATE: 03/06/2020
SHEET: 5 - FDS
TASK: Couple Visualisation - Realisation
Companies
Mode of Submission
Disputed complaints
loan
Debt
Credit
Student
Mortgage
Products
OPERATIONS
Complaints registered in that state
→ User chooses by:
↳ Year
↳ Company
↳ State
→ Count Measures / Options available:
↳ Mode of Submission
↳ By product
↳ Disputed complaints
↳ Resolved complaints
Size wrt Total Wages for that zipcode.
Zipcodes
FOCUS
DETAIL
→ complaints in a given year
2020
→ complaints in a given month
Jan
→ no. of complaints registered against a company
Equifax
↳ Technology: Displayed in RShiny app.
↳ Estimated time: 90-95 hrs for designing visualisation
↳ Hardware / Software needed: RStudio