Data Exploration Project Report:

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 aimed to develop visualizations showing significant relationships and trends in the financial marketplace. In particular, the trend of companies closing the complaints in favour of or not in favour of the consumer; the relation of complaint volume to dispute rate and company's response rate and the impact of localities and neighbourhood on the complaints.

The motivation behind selecting this dataset is that it is a huge dataset with a lot of fields and an interesting area of research. Proper analysis and research of this dataset could bring out a lot of intriguing inferences which could be valuable to the concerned industry.

Data Wrangling

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.

Both the data sets are sufficiently large with the first one being ~1050k rows x 18 columns and the second one with ~42k rows x 9 columns, so the wrangling could not be done in Excel itself. So, the entire process has been performed using R Studio where libraries like "stringr" and "dplyr" have been used for strings and data manipulations whereas "ggplot2", "leaflet", "scales" have been used for creating maps and graphs.

All the column variables in the complaint dataset are read into R as factors. So, the blank values are read as a factor instead of NA, and missing values can't have a level. To look for null/NA values firstly, the blank or the missing values in the columns have to be replaced by NA and then only further wrangling and checking can be performed.

As the datasets are not completely suitable for undergoing any analysis, so they had to be cleaned and thus reformatted. The analysis for any dataset cannot be made if there are a large number of null values. So, the first step was to check the number of null values in each column of both the datasets.

Date received	0
Product	0
Sub product	166026
Issue	0
Sub issue	402362
Consumer complaint narrative	785038
Company public response	681560
Company	0
State	17187
ZIP code	79866
Tags	907102
Consumer consent provided.	21648
Submitted via	0
Date sent to company	0
Company response to consumer	0
Timely response.	0
Consumer disputed.	0
Complaint.ID	0

Zip code	0
Zip Code Type	0
City	0
State	0
Location Type	0
Lat	648
Long	648
Location	1
Decommissioned	0
Tax Returns Filed	13643
Estimated Population	13643
Total Wages	13678

A large number of null values can be seen in many columns of the datasets. The 'Consumer complaint narrative' has a large number of empty values as many consumers wouldn't have given the consent to publish their complaint narr ative thus making it an optional column. Similarly, 'Tags', 'Company public response' and 'Consumer consent provide d' are optional columns and not needed in our analysis. Similarly, in the zip codes dataset, columns like 'Location typ e', 'Zip code type', 'Location', 'Decommissioned' are not obligatory. The best way to handle this situation is to delete these particular columns which are optional and not required in the exploration of the datasets.

There are many rows with null values in the Zip code column. For analysing this particular dataset, it will be difficult to gain insights with null values in the zip code column, for this reason, the rows having null values had to be removed.

The Date received column has the date on which the complaint has been submitted by a consumer, for a wider appr oach to analyse the complaint dataset, the years have been extracted from the dates so that the analysis can be carri ed out for each year. This new variable for the years is taken and converted to a categorical variable with the years as levels.

For joining separate tables, a common key or variable is needed on which the join can be performed. The "ZIP code" column in the complaint dataset has been renamed to "Zip code" as it is the common key that will allow a mapping b etween the tables.

Data Checking

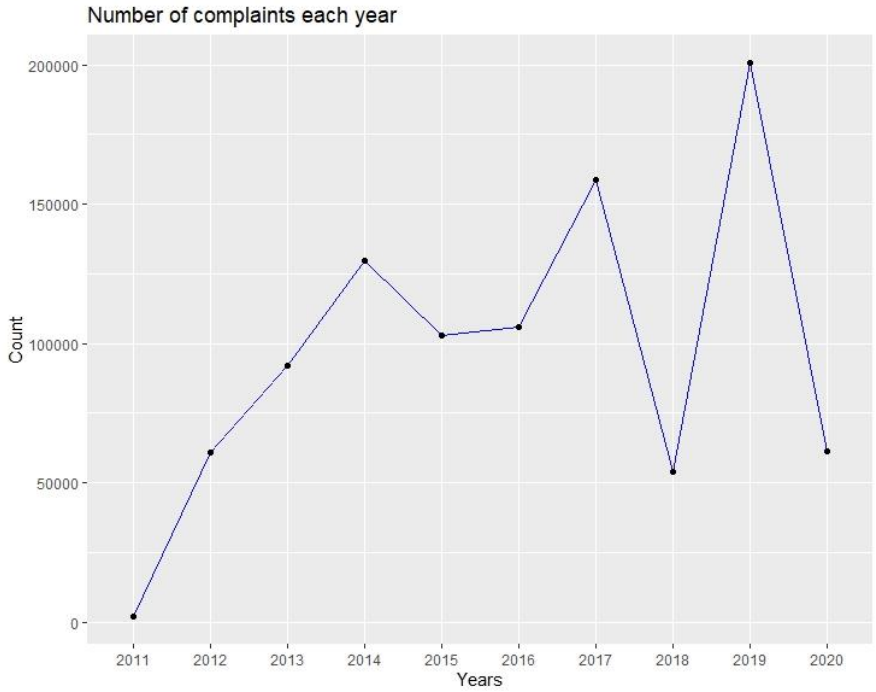
Under this process, the steps would include checking each and every column for any errors in the data sets in R. Sometimes the data in a column looks correct on the superficial level as there are no obvious datatype mismatches, duplicate records or null values but when the values and the apportioning of data in the column is examined, some values distribution don't make any relevance.

So firstly, there were a lot of extra white spaces in the values of certain columns, the trim function has been used to get rid of this error.
The second step is to check if the dates fall in their expected periods i.e. the date of receiving the complaint should not be larger than the date the complaint is sent to the company by the Consumer Financial Complaints Bureau.
The dates are also checked whether they all fall in the same format for maintaining consistency of data and thus are arranged in a consistent *dd/mm/yyyy* format to eliminate any error.
The third step involves reviewing the states whether the states mentioned in the data are valid U.S. states.
The fourth step is to check that the length of the zip codes mentioned is equal to five. The zip codes of America are 5-digit numbers.
Furthermore, checking for any duplicates at Complaint ID level in the complaints database is performed and are removed if found.
Similarly, column-wise checking was performed for the zip codes database.
In the first place, the zip codes had to be checked whether they are valid or not. Some of the zip codes mentioned were 3-digit long so they had to be converted to 5-digit by adding two zeroes before the digits.
Lastly, the columns for 'City' and 'State' had to be checked for genuine entries by plotting the latitude and longitude on the map of USA. Some outliers found were most probably due to a typing error, where the latitude is given to be negative, this is corrected by changing the negative latitude to positive. Similarly, for longitudes, some values are written as positive and are thus corrected by making them negative.

Data Exploration

Under this section, the exploration of the variables of both the datasets will be carried out to answer the questions presented in the project proposal using R Studio.

Figure 1: Complaints registered per year



The line graph drawn using "ggplot2" library in R shows that initially, a rise in the number of complaints can be seen per year with the main reason of people getting aware of such a platform available. But a great dip in the number of

complaints in 2018 can be due to the market crash held in that year. 2019 sees a boom in the market with people purchasing more financial products and thus more complaints being registered.

- Complaints are registered for all types of financial products and services.

Figure 2: Number of complaints for each type of product

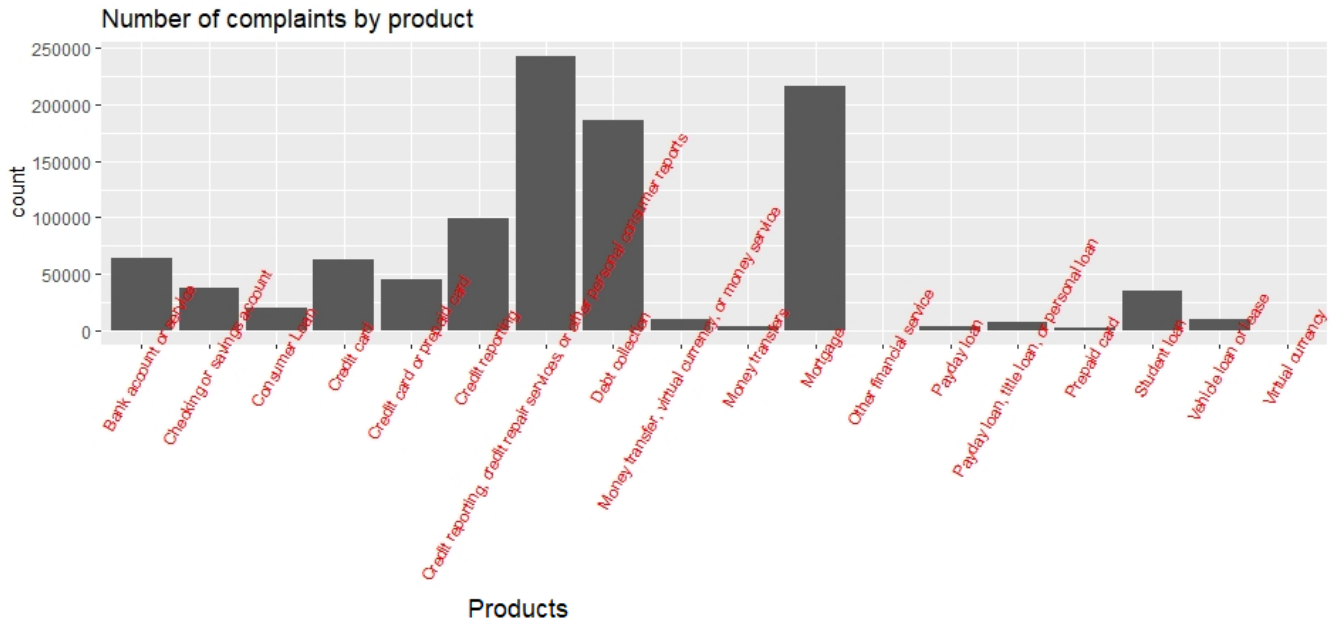
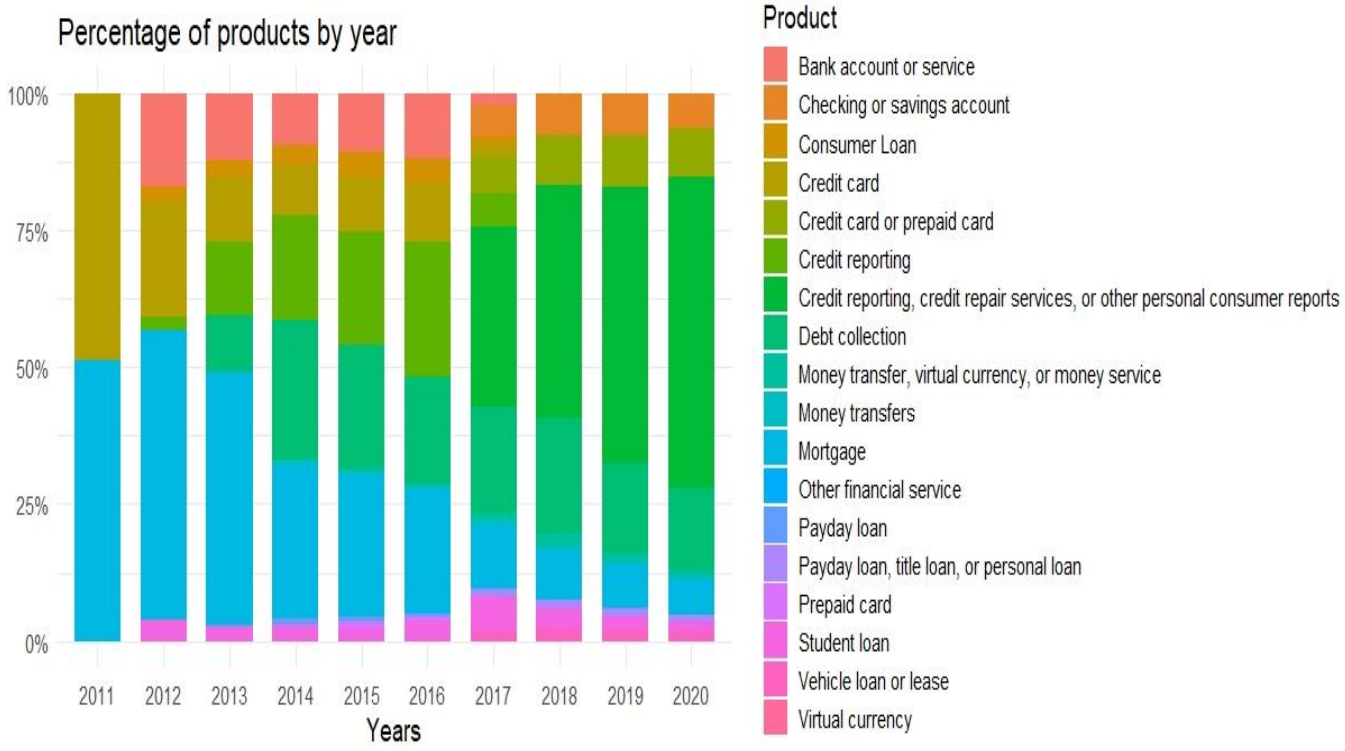


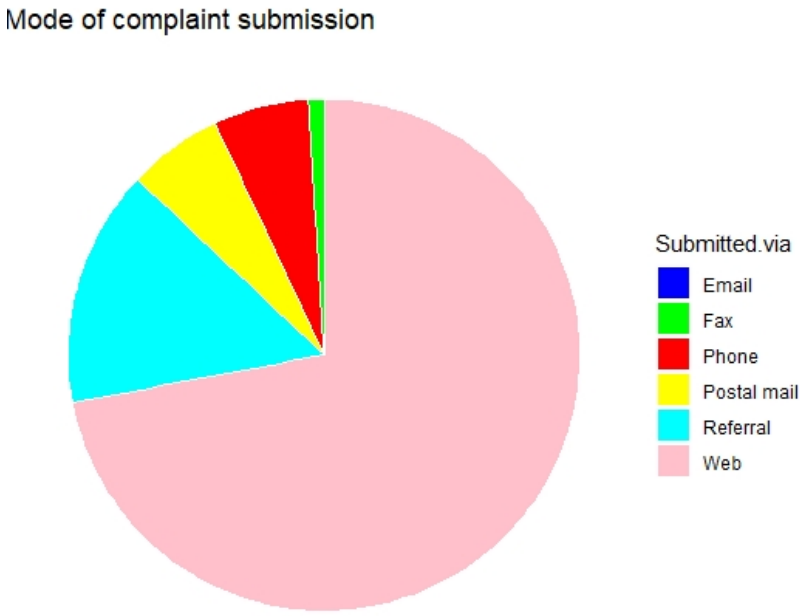
Figure 3: Mortgage, debt collection and credit reporting drive the majority of complaints



The reason major proportion of complaints are for mortgages, debt collection and credit reporting can be because of the banks' increased monthly payments and people not being able to pay back on time as they were completely unprepared for such a move.

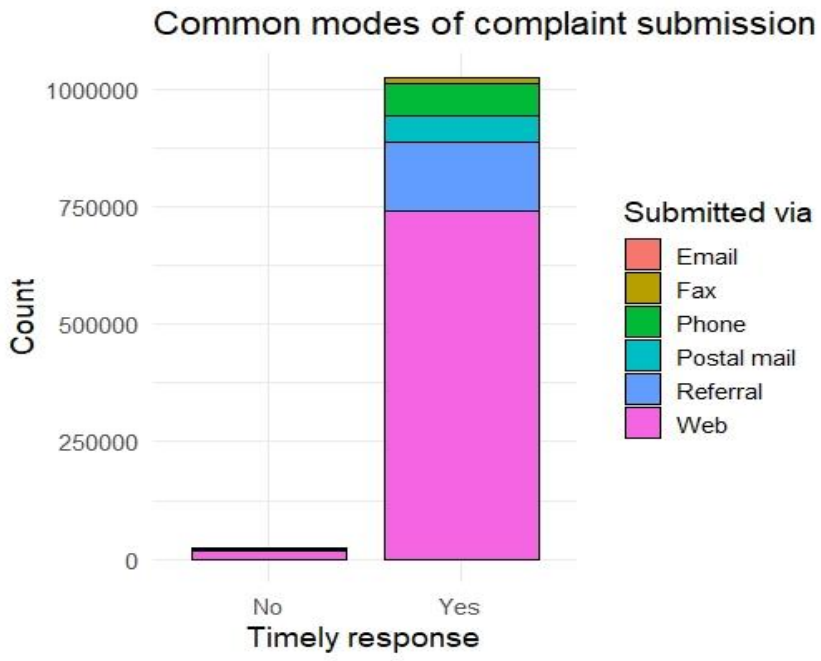
- Complaints can be submitted via Email, Fax, Phone, Web, Referral or Postal mail. It is found that the most common medium by which complaints are submitted to the CFPB is Web as it is extremely convenient and the complaint is submitted without any hassle, in no time.

Figure 4: Mediums used for submitting the complaints to CFPB



- Consumers can be provided with a timely response or an untimely response. The consumers who lodged their complaints through web got a faster and a timely response from the company against whom the complaint was issued.

Figure 5: Complaints submitted through web got a faster response



- Complaints can be closed *in favour* of the consumer by offering any type of relief, or they can close the complaint *not* in favour of the consumer, perhaps providing an untimely response, closed without explanation or closed without relief.

Figure 6: Lesser complaints are resolved in favour of consumers

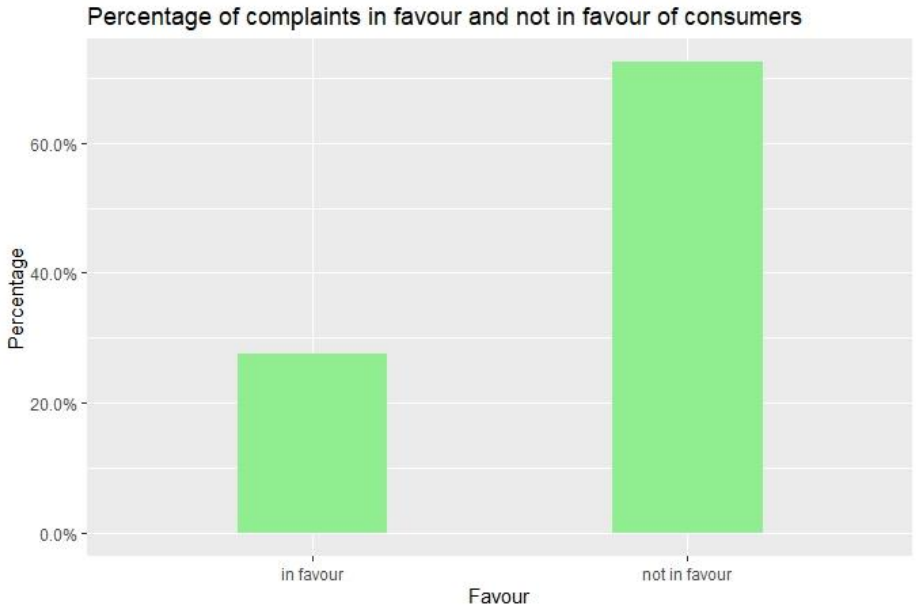
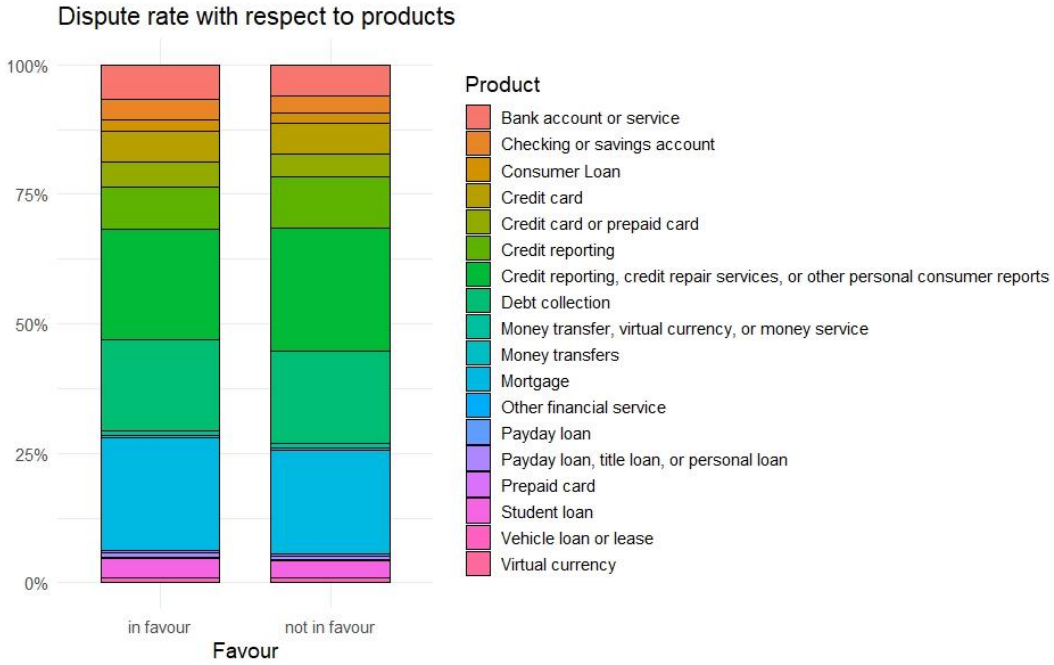


Figure 7: Percentage of complaints in favour of and not in favour of consumer with respect to the type of product



- Taking worst 50 companies ranked by the number of complaints. EQUIFAX, Experian Information Solutions, Bank of America and TRANSUNION Intermediate Holdings are the companies with the largest amount of complaints. Mostly, companies closed the complaints with a mere explanation

Figure 8: Companies with maximum number of complaints

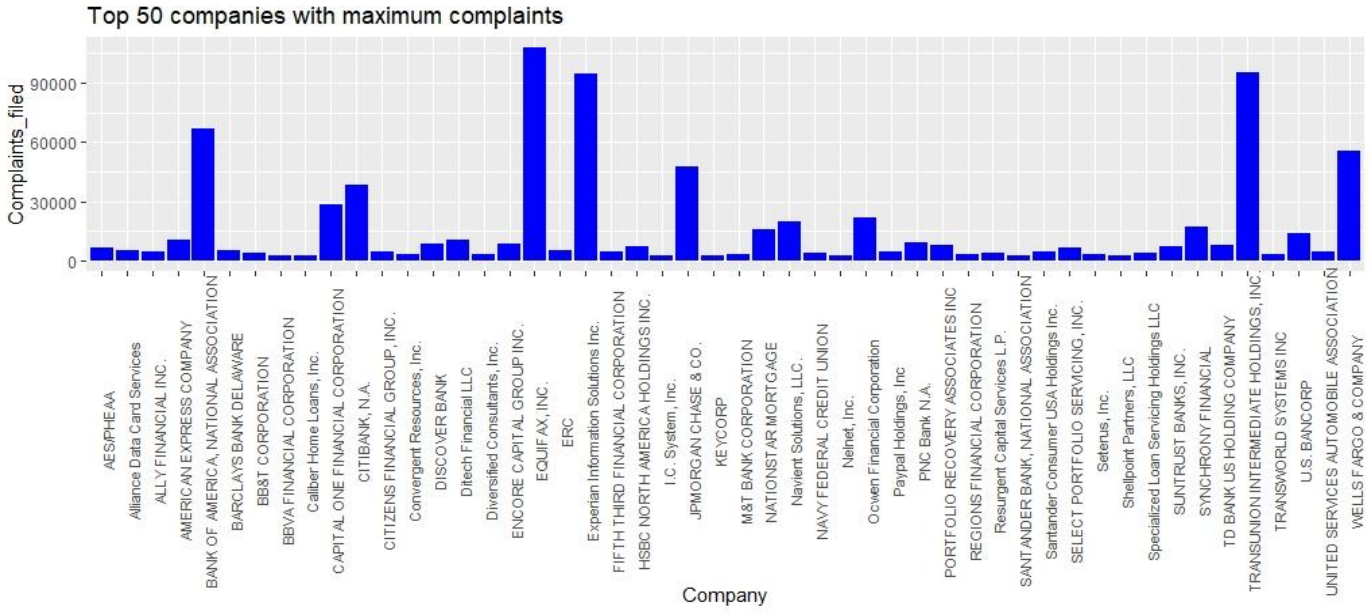
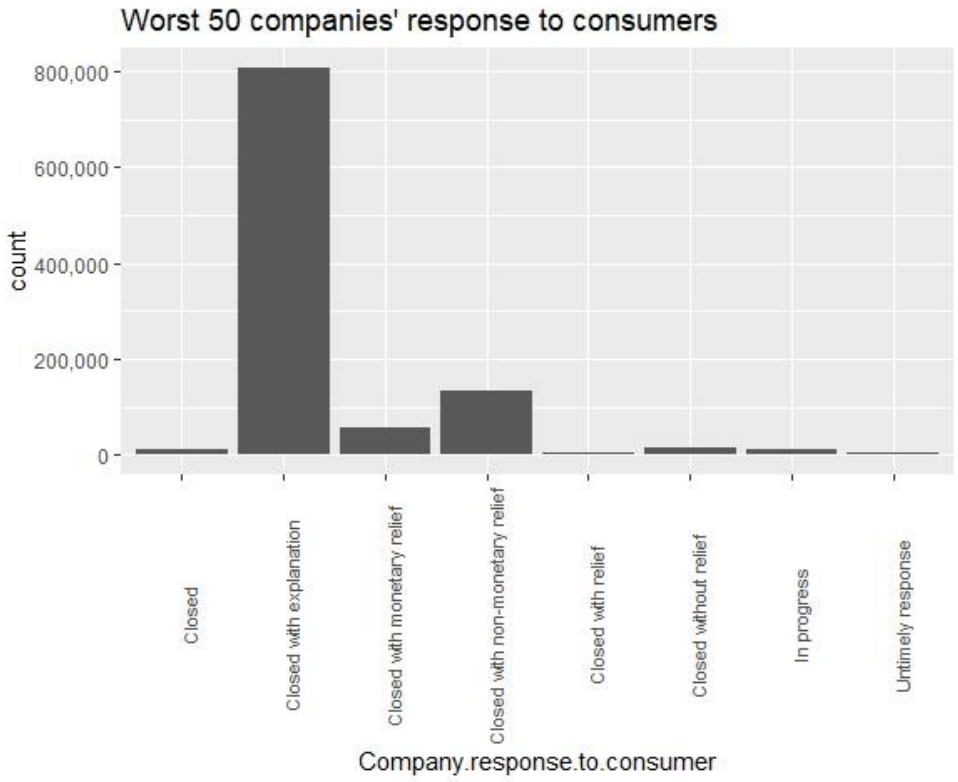


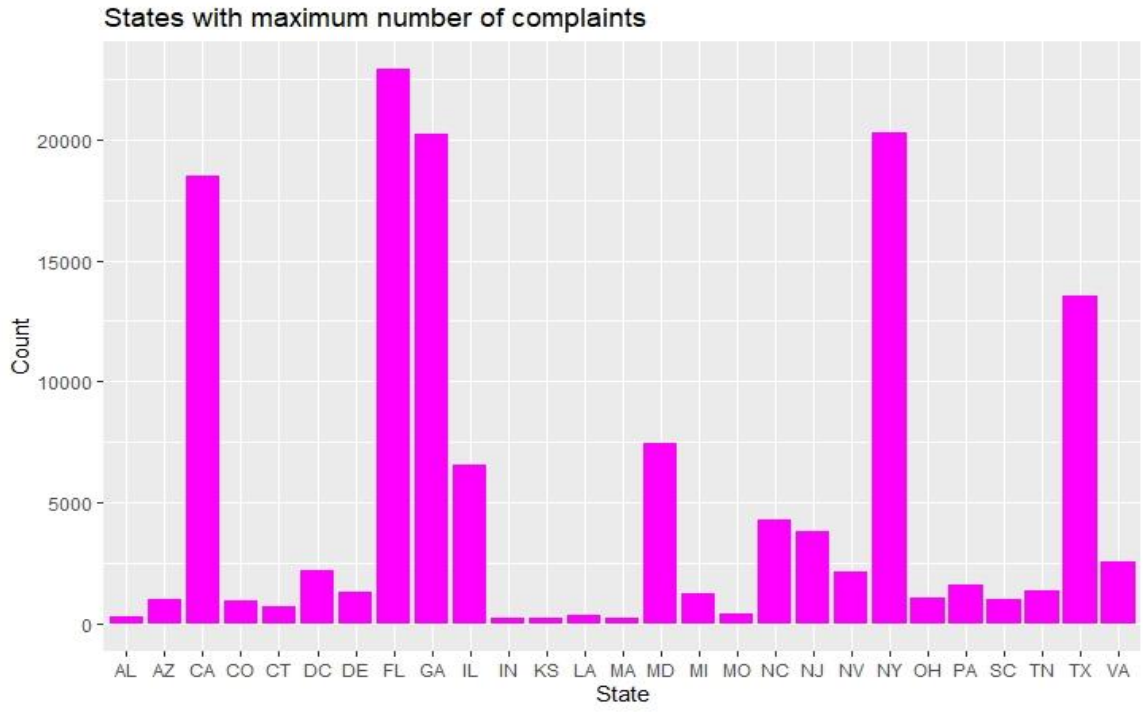
Figure 9: Response of the top 50 companies with maximum complaints to their consumers. S



For further analysis of the volume of complaints state-wise and the influence of the neighbourhood on the number of complaints, the rows having null zip codes and the rows having incomplete zip codes i.e. “325xx” have been removed for a clearer understanding of the data and to gain a better insight of the relation between the variables of the data sets.

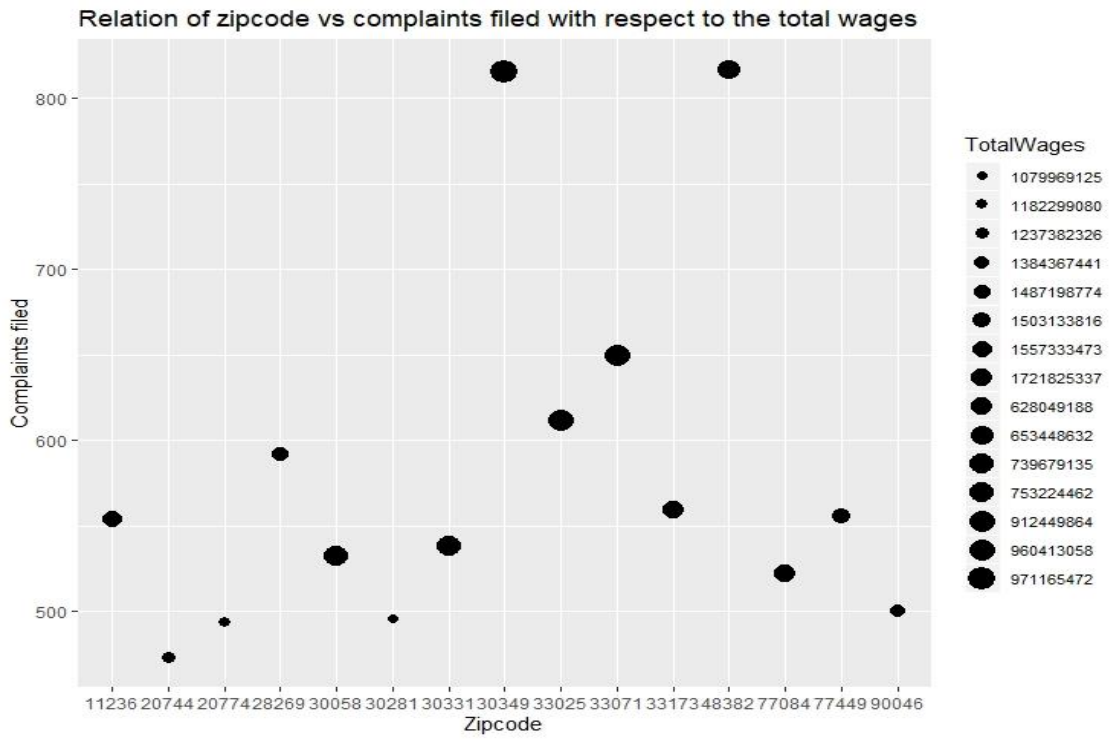
- California, Gainesville, New York, Florida and Texas are the states to receive the highest number of complaints from consumers.

Figure 10: Distribution of complaints across states



- To observe the effect of localities on the number of complaints, both the complaints and the zip-code data set have been joined on the zip code column. Established and prosperous localities 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.

Figure 11: The effect of localities and neighbourhoods over the volume of complaints



Conclusion

Each of the visualizations developed in the report fulfils the criteria specified in the initial project proposal. Initially, the report projects the exploration of individual columns and trends the variables have produced from 2011 to 2020. The exploration process started with the trend of complaints as a whole and then kept exploring further all the variables i.e. the issues, the products, the disputes and then according to the location what is the trend of the complaints being filed. The trend of the complaints being registered is shown by how the complaints are increasing with each year. The product column has been explored by visualizing a graph that shows mortgages, credit, and debt drive the majority of complaints from consumers, the dispute rate of consumers with their products, the medium through which the maximum number of complaints are submitted is the web, and the states with the maximum number of complaints are Florida and New York showing that a large part of complaints come from few states which are economically better than other states.

Figure 6 depicts how many complaints are closed in favour and not in favour of the consumers i.e. closing the complaint with any type of relief or closing by just explaining.
Figure 7 shows the complaint volume against the consumer's dispute rate i.e. whether the complaint was closed in favour or not in favour of the consumer with respect to the variety of products and services.
Figure 9 shows how complaints are related to the company's response rate i.e. whether the complaint was closed with an explanation or closed with monetary relief or closed with non-monetary relief, closed with relief or closed without any relief or untimely response or the complaint is in progress.
Figure 11 shows the impact of neighbourhood on the number of complaints i.e. wealthier the locality, higher is the complaint volume.

Reflection

This project helps in understanding the importance of exploratory data analysis with the help of univariate visualisations where only one variable is involved as well as bivariate visualisations which are performed to show relationships and trends between variables. This assists in building familiarity with the existing data and makes finding solutions easier and simpler. This project also supports in developing a better insight about data wrangling, which is a process of transforming data from a raw form to another form that is appropriate for further usage, and also about data checking which undergoes to check the correctness of the data. The data analysis performed by exploring the data and then presenting it through visualisations shows how data can be manipulated and reformatted to give beneficial insights that might prove effective in decision making.

Although this might not be a perfect way to wrangle the data and checking the correctness of the data, with more resources and a better understanding, the data could have been checked in a more efficient way. 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 results, 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. Tomar, S., 2020. *A Comprehensive Introduction To Data Wrangling - Springboard Blog*. [online] Springboard Blog. Available at: <https://www.springboard.com/blog/data-wrangling/> [Accessed 28 April 2020].
5. Flores, W., 2020. *Six Validation Techniques To Improve Your Data Quality | Transforming Data With Intelligence*. [online] Transforming Data with Intelligence. Available at: <https://tdwi.org/articles/2016/02/19/six-validation-techniques-data-quality.aspx> [Accessed 28 April 2020].