San José State University

Department of Applied Data Science

DATA 230: Data Visualization

Project Report

Group 2: Nupur Pathak, Revathi Boopathi, Sree Divya Cheerla, Vani Bhat

Abstract — **With the global pandemic led business shutdowns and relaxations in the financial regulations, there has been a surge in consumer complaints with the financial institutions. Consumer Financial Protection Bureau (CFPB), a federal government agency, assesses the compliance management of these financial institutions. It collects these consumer complaints and facilitates their resolution.**

The CFPB dataset collected consists of over 300 thousand complaints registered from January 2018 to December 2022. In this dataset, we have analysed the complaints from consumers across multiple financial products and service offerings.

The outcome of the insights derived from the product analysis can help companies improve their products and services and identify new offerings, thereby enhancing consumer experience. It will also help the CFPB to revise the regulations and strengthen the compliance controls based on the systemic problems in the financial offerings.

Keywords - Visualization, Tableau, CFPB, Complaints

I. Introduction

Consumer Financial Protection Bureau (CFPB) is an independent bureau within the Federal Reserve System that educates and empowers consumers with the necessary information they would need to make financial decisions. In reference to [6], the CFPB will set and enforce clear, consistent rules that allow banks and other consumer financial services providers to compete on a level playing field and those let consumers clearly see the costs and features of products and services. The 'Consumer complaint database' is CFPB's repository of complaints filed by the consumers about various consumer financial products and services that were sent to companies for response. The typical process for the complaint data accumulation is the complaint submitted is directly routed to the company (Banking and Financial Institutions) which provides response to the issues in the complaint. Then the information is published in the public Consumer Complaint Database on or after 15 days which the consumer can review upon notification by CFPB. After a consumer receives the company's response to their complaint, the consumer can provide feedback on the company's response by completing an optional survey. This information is made available to companies via the Company Portal. The CFPB makes a subset of complaint data publicly available in the Consumer Complaint Database.

We are aware that consumers are facing numerous challenges on a day-to-day basis with banking products and services. So, we think it is important to address the issues faced by them to know and educate people make informed financial decisions. This motivated us to select this dataset for the project. Consumer complaints can act as a powerful tool to provide insights for the companies to identify whether the customers are satisfied with their product and service offerings.

The number of complaints by the consumers in the consumer complaint database is an indication of lack of information of the various financial products to the consumers and possible weaknesses or other deficiencies, such as violations of laws or regulations in its financial products/services. Hence, there is a need to understand and analyse this data both from consumer and company perspective as to why, what and where the issues are being faced and address them appropriately.

In this project, we have explored the consumer complaint database which acts as a powerful tool to provide insights for the companies to identify whether the consumers are satisfied with their product and service offerings. The story from analysis to business insights has been brought out by different visuals using Tableau as the visualization tool.

The main objective of this project is to identify the major areas of the banking/ financial product services where the consumers are facing maximum issues and address them for the respective product categories offered by the companies. Our analysis can help CFPB to regulate consumer financial products and services and educate and empower consumers to make informed financial decisions.

Note: Complaint volume is considered in the context of company size and/or market share and population information. For example, companies with more customers may have more complaints than companies with fewer customers may have less complaints.

II. Related Works

[1] Organizations need to accelerate the pace with which they can realize business value from data. The focus is on improving "time to value," which is the length of time it takes from the beginning of a project to the delivery of anticipated business value. Today, as organizations grow more reliant on data to support strategic and operational decision making, delays in developing and implementing business intelligence and analytics are becoming less tenable.

[2] This paper explores pioneers in data research and shows how their work helped revolutionize visualization techniques. We will examine the different styles and barriers of data visuals, and the methods used to overcome them. The key features exemplified through the white paper are that trust is a key design issue, visuals must be expressive and convey data accurately, make Effective use of your visuals by exploiting human perception, the importance of context to help your audience to better understand your data view.

[3] Data visualization is a key component to business and data analytics, allowing analysts in businesses to create tools such as dashboards for business executives. Various software packages allow businesses to create these tools to manipulate data for making informed business decisions.

[4] Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends, and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Data visualization is the presentation of quantitative information in a graphical form. In other words, data visualizations turn large and small datasets into visuals that are easier for the human brain to understand and process. Data visualizations are surprisingly common in our everyday life, but they often appear in the form of well-known charts and graphs. It can be used to discover unknown facts and trends. Good data visualizations are created when communication, data science, and design collide. Data visualizations done right offer key insights into complicated datasets in ways that are meaningful and intuitive. In this article, we would like to discuss data visualization, importance of data visualization, data visualization tools etc.

[5] To meet the requirements of the high-dimensional data processing in the information field, this paper attempts to explore the methods and techniques of clustering data visualization for general data resources. Through dimensionality reduction and visual mapping of high-dimensional data, this paper established a visualization system model which will visualize the influencing factors. The visual system model approach was tested by using the IRIS dataset from the University of California Irvine (UCL) database. The results demonstrate that this model can effectively analyse the data set, visualize the features of IRIS data in real time and achieve the expected results.

III. The Consumer Complaints Dataset

Choosing the right dataset is an important task in data analysis and building visualizations to tell the story. To derive actionable insights, the dataset must be large enough to play with it and observe trends /seasonality/ common phenomenon.

The first dataset is about the Consumer Complaint Database of complaints received by the Consumer Financial Protection Bureau (CFPB) about financial products and services from 2018 - 22. This dataset comprises of 18 columns and over 300k records which gives us the details on when, where and against which company the complaints were raised, and the status on the resolution of the complaints. The complaint narratives explain the consumer's experience with the financial product.

The data is compiled from the CFPB website which is https://tinyurl.com/mrxdzwes

Below is the metadata of the data source:

Field name	Type	Description
Date received	Numeric	The date the CFPB received the complaint
Product	Categorical	The type of product the consumer identified in the complaint
Sub-product	Categorical	The type of sub-product the consumer identified in the complaint
Issue	Categorical	The issue the consumer identified in the complaint
Sub-issue	Categorical	The sub-issue the consumer identified in the complaint
Consumer complaint narrative	Categorical	Consumer-submitted description of "what happened" from the complaint
Company public response	Categorical	The company's optional, public-facing response to a consumer's complaint
Company	Categorical	The complaint is about this company
State	Categorical	The state of the mailing address provided by the consumer
ZIP code	Categorical	The mailing ZIP code provided by the consumer
Tags	Categorical	Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers
Consumer consent provided	Categorical	Identifies whether the consumer opted in to publish their complaint narrative
Submitted via	Categorical	How the complaint was submitted to the CFPB
Date sent to company	Numeric	The date the CFPB sent the complaint to the company
Company response to consumer	Categorical	This is how the company responded
Timely response	Categorical	Whether the company gave a timely response
Consumer disputed	Categorical	Whether the consumer disputed the company's response
Complaint ID	Numeric	The unique identification number for a complaint

Figure 1: Metadata – Complaints database

The second dataset is from the Census. We have used the US state wise population data from 2018-22. The dataset shows the yearly geographical distribution of population across different region/state

Field name	Type	Description
State	Categorical	The state for which the population is indicated
Year	Numeric	The year for which the population is recorded
Population	Numeric	The population for a state for a given year
State_Code	Categorical	The state code for which the population is indicated
Region	Categorical	Geographical areas categorised into regions
Division	Categorical	Geographical areas categorised into divisons

Figure 2: Metadata - Population Database

The source link to the data is https://tinyurl.com/23t4atwt

IV. Methods

A. Data Cleaning and Wrangling

The 'Consumer Complaint Database' dataset was pre-processed in Python to handle the required records.

1. The dataset imported in pandas data frame consists of 18 columns as seen in Fig. 3.

```
dataset = pd.read_csv('master_complaints-2022-04-07_19_07.csv')
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 18 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   Date received                   1048575 non-null  object
 1   Product                         1048575 non-null  object
 2   Sub-product                     1048575 non-null  object
 3   Issue                           1048575 non-null  object
 4   Sub-issue                       1048575 non-null  object
 5   Consumer complaint narrative    447115 non-null   object
 6   Company public response         1048575 non-null  object
 7   Company                         1048575 non-null  object
 8   State                           1048575 non-null  object
 9   ZIP code                        1048575 non-null  object
 10  Tags                            1048575 non-null  object
 11  Consumer consent provided?      902522 non-null   object
 12  Submitted via                   1048575 non-null  object
 13  Date sent to company            1048575 non-null  object
 14  Company response to consumer    1048575 non-null  object
 15  Timely response?                1048575 non-null  object
 16  Consumer disputed?              156911 non-null   object
 17  Complaint ID                    1048575 non-null  int64
dtypes: int64(1), object(17)
memory usage: 144.0+ MB
```

Figure 3: Data Description

2. We observed that the datatype for the date columns such as 'Date received' and 'Date sent to company' had different date formats such as 05-11-2020 and 06/29/19. We have updated the datetime format for these date fields using python strptime() function as seen in Fig 4.a.

```
alldates_received = []
for dt in dts1:
    try:
        d = datetime.datetime.strptime(dt, '%m-%d-%Y').date()
    except Exception as err:
        print(dt)
        pass
    try:
        d = datetime.datetime.strptime(dt, '%m/%d/%y').date()
    except Exception as err:
        print(dt)
        pass

    alldates_received.append(d)
```

Figure 4: Date modification

```
dts2 = dataset['Date sent to company']
alldates_sent = []
for dt in dts2:
    try:
        d = datetime.datetime.strptime(dt, '%m-%d-%Y').date()
    except Exception as err:
        pass
    try:
        d = datetime.datetime.strptime(dt, '%m/%d/%y').date()
    except Exception as err:
        pass
    alldates_sent.append(d)
```

Figure 5: Date field creation

3. These dates are then added to the data frame. The final list of columns is as below:

```
dataset.columns
```

```
Index(['Date received', 'Product', 'Sub-product', 'Issue', 'Sub-issue',
       'Consumer complaint narrative', 'Company public response', 'Company',
       'State', 'ZIP code', 'Tags', 'Consumer consent provided?',
       'Submitted via', 'Date sent to company', 'Company response to consumer',
       'Timely response?', 'Consumer disputed?', 'Complaint ID',
       'Date_received', 'Date_sent'],
      dtype='object')
```

Figure 6: Columns available

4. We have also added a year column and filtered the records from 2018 onwards. These records are written to the output excel file as seen in Fig.

```
a = dataset[dataset['Year']>=2018]
```

```
a['Year'].unique()
```

```
array([2020, 2019, 2022, 2021, 2018], dtype=int64)
```

```
a.to_csv('complaints_2018-22.csv')
```

Figure 7: Data filtering

5. This file is uploaded in Tableau for data visualization. The string fields of date sent, and date received are dropped and we have retailed the fields with datetime format. Following are the dimensions and measures of the dataset:

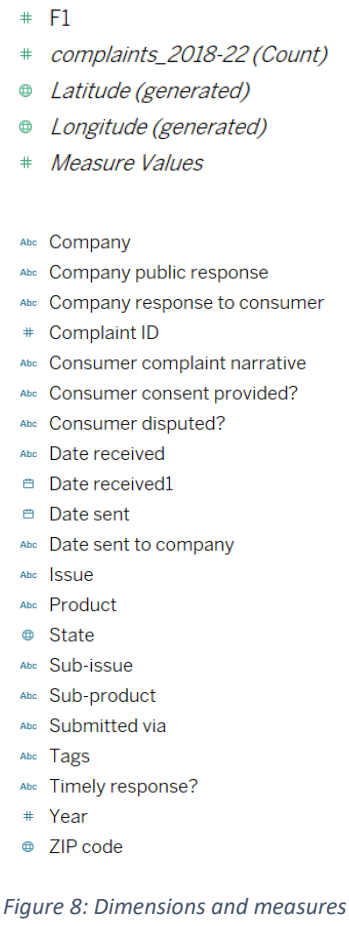

Figure 8: Dimensions and measures

B. Data Joins

We have performed an inner join operation of the Consumer Complaint dataset with the US population dataset and Hex Spatial file (to help generate hex maps for US states) with the intention of bringing out the state wise correlation between the number of complaints registered and the population density. The schema for the same in tableau is as shown below.

Figure 9: Joins

We identified various aspects to be considered and thereby formulated various business questions that will substantiate our requirements and address our problem statement. The key aspects to be considered are various products, sub-products and issues on which complaints were raised and response from each of the companies.

We chose Tableau as our visualization medium because of a wide variety of visuals and analytical features it offers with an ease of use. We created the necessary calculated fields and measures required to plot the visuals.

V. Visualization Design & Results

Our method involves creating different visualisations in tableau which helps solve and address our problem statement.

Dashboard 1: Executive Summary
Dashboard 2: Product Analysis
Dashboard 3: Company Product Analysis
Dashboard 4: Company Response Analysis

Dashboard 1: Executive Summary

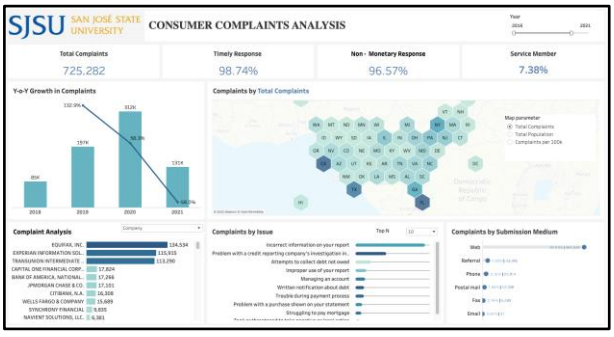

Figure 10: Executive Summary Dashboard

The **Executive summary dashboard** gives us an overview of the consumer complaints against the financial institutions in the United States from 2018- 2021.

- **Total Complaints KPI -** Indicates the total number of complaints filed from 2018-2021.
- **Timely Response KPI -** Indicates the percentage of complaints received response in time from the financial institutions.
- **Non - Monetary Response KPI -** Indicates the percentage of complaints closed with non-monetary relief.
- **Y-o-Y Growth -** In this visual, from the bars we can see the growth / drop in the number of complaints from 2018- 2021 and from the line chart we see the percentage growth/ drop in complaints.
- **Complaints by State** - The Hex map visual gives us an overview of geographical distribution of complaints. From the Map parameter filter at the top right corner, we can also see the population of each state and complaints per 100k information in the same visual by selecting the respective parameters. The gradient of the colour tells us which state has the highest complaints/population/density of complaints per 100k.
- **Complaint Analysis** - From the Bar chart, by choosing product/company from the filter on top-right corner, we can see which product/company is receiving the highest number of complaints.
- **Complaints by Issue** - From this visual, we can know the Top N issues faced by the consumers, here the N can be changed using the Top N filter at the top right corner and we can also see the sub - issues faced in the tooltip for more information.
- **Complaints by Submission Medium** - From this visual, we can know through which medium the consumers are filing complaints the most.
- **Year Filter** - Using this filter, we can see the consumer complaints analysis specific to the year /range of years selected.

From this dashboard we can answer some of the questions like:

A. Which state has the highest complaints? What is the population of that state and the density of complaints per 100k population in that state?

B. In which year are the complaints filed the most?

C. By selecting any year, we can know which product/company had the highest complaints, what are the Top N issues faced for that product/company.

D. By which submission medium, the complaints are being filed the most.

Dashboard 2: Product Analysis

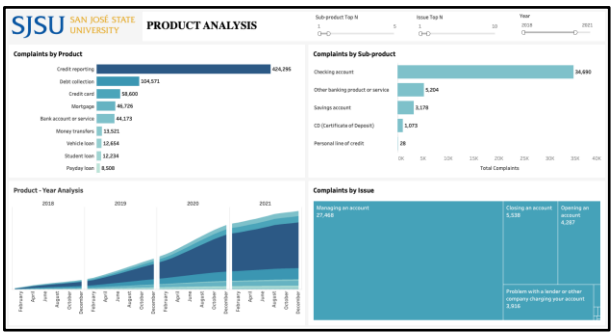

Figure 11: Product Analysis Dashboard

The **Product Analysis dashboard** gives us a detailed analysis of the count of consumer complaints for each product, sub-product and also the top issues the complaints are filed against for each product.

- **Complaints by Product -** From the Bar chart, we can see which products the consumers have highest complaint.
- **Complaints by Sub-product -** In this visual, we can see the Top N sub products the consumer has highest complaints with respective to each product. We can change the N Value using the *Sub product Top N filter* located at the top bar.
- **Product - Year Analysis - From** the Area chart, we see product wise growth of complaints for each month from 2018 - 2021.
- **Complaints by Issue** - From the Tree Map visual, we can see the Top N issues faced by the consumer for each product. We can change the N Value using the *Issue Top N filter* located at the top bar.
- **Year Filter** - Using this filter, we can see the consumer complaints analysis specific to the year /range of years selected.

From this dashboard we can answer some of the questions like:

A. What are the products people are facing highest problems? Which are the top N sub-products based on the complaints for each product?

B. In which month and year, the product has the highest complaints.

C. What are the Top N issues faced by the consumer for each product?

Dashboard 3: Company Product Analysis

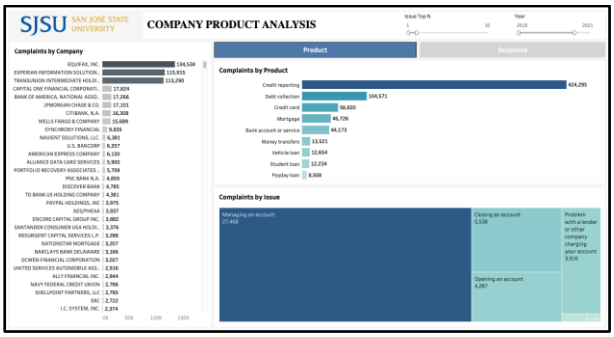

Figure 12: Company Product Analysis

The **Company Product Analysis dashboard** gives us a detailed analysis of the count of consumer complaints for each company, product and also the top issues the complaints are filed against for each company.

- **Complaints by Company -** From the Bar chart, we can see the total number of complaints filed for each company.
- **Complaints by Product -** In this visual, we can see which products the consumers have highest complaints with.
- **Complaints by Issue -** From the Tree Map visual, we can see the Top N issues faced by the consumer for each product. We can change the N Value using the *Issue Top N filter* located at the top bar.
- **Year Filter** - Using this filter, we can see the consumer complaints analysis specific to the year /range of years selected.

From this dashboard we can answer some of the questions like:

A. Which companies are receiving the highest complaints?

B. What are the products with which the consumers are filing complaints for each company?

C. For each company, what are the Top N issues with which the consumers are facing problems that should be resolved.

Dashboard 4: Company Response Analysis

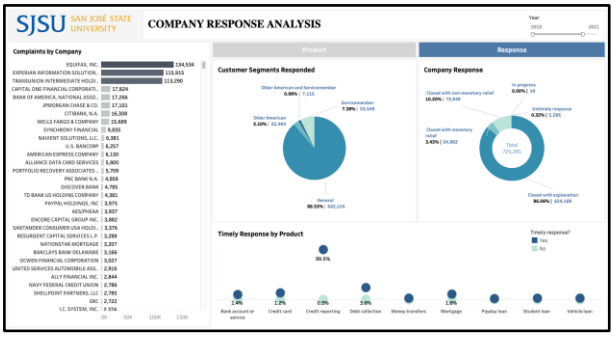

Figure 13: Company Response Analysis

The **Company Response Analysis dashboard** gives us a detailed analysis of the count of consumer complaints filed for each company, the company response, consumer segments responded and also product wise timely response percentage given by each company.

- **Complaints by Company -** From the Bar chart, we can see the total number of complaints filed for each company.
- **Company Response -** From the Donut chart, we can see how each company has responded to the complaints filed.
- **Consumer Segments Responded** - From Pie chart, we can see the various consumer segments who have filed the complaints against each company.
- **Timely Response -** From the Scatter plot, we can see product wise, the percentage of complaints responded in time by each company.
- **Year Filter** - Using this filter, we can see the consumer response analysis specific to the year /range of years selected.

From this dashboard we can answer some of the questions like:

a. What is the percentage of complaints closed with monetary / non- monetary relief by each company?
b. How many complaints are still in 'In progress'?
c. Which consumer segments are filing more complaints?
d. How much percentage of complaints are being resolved in time for each product by each company?

VI. Use Case & Insights

The Consumer Financial Complaints Analysis dashboard's end-to-end functionality and interactivity could further be explained by taking the example of an analysis that a CFPB officer can perform using the visuals present in the dashboard.

The total complaints in the year 2021 seems to be 131k. On understanding this further geographically, it is evident that the total complaints in most of the states is a direct proportion of the total population of the number of people of that corresponding state. Just considering the complaints statistics show that the highly populous states i.e., California, Texas and Florida have the highest number of complaints registered.

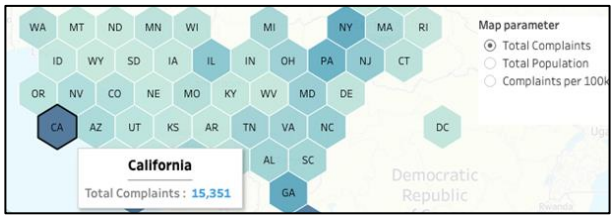

Figure 14: Geographical complaints analysis

However, on normalizing the population factor and considering complaints per 100k population in each state, it is evident that California and Texas seem to have fewer complaints/100k whereas Florida has the highest complaints per capita. So, the official can further investigate the most common issues faced by people of Florida when it comes to financial complaints.

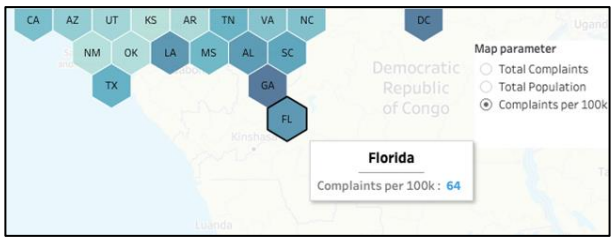

Figure 15: Geographical complaints per 100k

It is evident from the complaints analysis graph that 'Credit Reporting' seems to show the greatest number of pain points for the consumers and specially those issues related to incorrect information related to information on the credit report (e.g., a report of some other person).

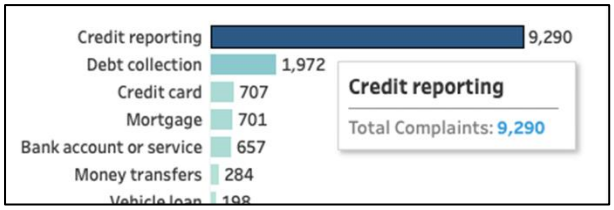

Figure 16: Florida Product Analysis

Now that the pain point is understood, the official needs to understand which companies are present in the region that show the highest complaints to intimate them to take actions to improve the customer satisfaction.

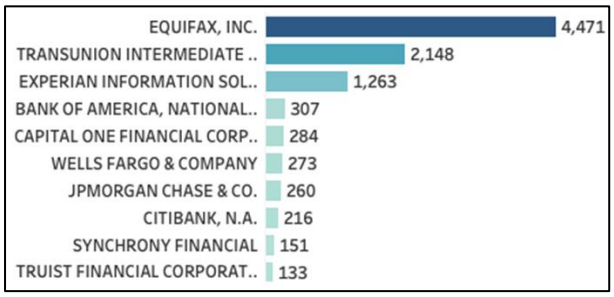

Figure 17: Company Analysis

It is evident that Equifax, Inc & TransUnion Intermediate holdings are the 2 major companies that contribute to ~47% of the overall complaints from all the companies. In this scenario, the official can intimate the Equifax and the TransUnion Intermediate company regarding the points they need to further consider and act on.

VII. Discussions

Based on the insights derived from the product analysis the key takeaway for the companies is that they can improve their products and services and identify new offerings, thereby enhancing consumer experience.
Based on the geographical distribution analysis of complaints and population density, companies can analyse each state whether to increase consumer support as well as further investigate the issues in their products.

VIII. Conclusion

1. In 2020, we see a rise in volume of complaints of ~312K due to global pandemic and credit relaxations. However, in 2021, the number of complaints dipped to ~131k (-58%)
2. Florida is to be the highly populous state with the highest number of complaints per 100k (64)
3. ~98% of the time companies have responded to the customers on-time
4. Complaints seem to be closed with no cost ~97% of the time
5. 90% of the complaints are raised through web
6. Credit Reporting" product accounted for 58% of the complaints followed by "debt collection" and "credit card"
7. Equifax, Inc & TransUnion Intermediate holdings are the 2 major companies that contributed to ~29% of the overall complaints raised

VIII. Future Work

Consumer pain points can be inferred through sentimental analysis from the narrative of those consumers who have consented to give their feedback. These would help the companies to take necessary actions to bring down consumer financial complaints in future. Future analysis on consumer segments like Service members, Olden American groups, etc. will help in better profiling of consumers and thereby providing targeted financial offers for impacted consumers

IX. Discussions

Through this project, we learnt how to solve a business problem end-to-end and how to furnish it to a stakeholder. Here are a few key questions we learnt answer for. What research needs to be performed while selecting the right dataset and the dataset to which we can correlate to? How to understand and evaluate the data? How to sensibly perform joins and integrate various dataset? How to formulate solution to any business problem for example, we understood what kind of issues exist in the finance industry and tried to formulate various business KPIs to be measured. We learnt functional implementation and details for the visualization platform.

X. References

[1] David Stodder, "Accelerating the Path to Value with Business Intelligence and Analytics", June 2017 (White Paper). Available: https://tinyurl.com/hn26ex9s
[2] Tableau, "Tracing the history and best practices of visual representation: From early cave drawings to the data visualization revolution", (White paper) Available: https://tinyurl.com/2s3rkej
[3]Michael Diamond, Angela Mattia, "Data visualization: an exploratory study into the software tools used by businesses", Journal of Instructional Pedagogies, Available: https://files.eric.ed.gov/fulltext/EJ1151731.pdf
[4] Mohaiminul Islum, Shangzhu Jin, "An Overview of Data Visualization", 2019 IEEE International Conference, Available: https://ieeexplore-ieee-org.libaccess.sjlibrary.org/document/9012031
[5] Yipeng Li, Siyuan Hou, "Methods and Techniques in Data Visualization Model", 2017 IEEE International Conference,Available: https://ieeexplore.ieee.org/document/8789106
[6] Consumer Complaint Database https://www.consumerfinance.gov/data-research/consumer-complaints/