Milestone 1

Team Members: Rajat Aghi, Pal Doshi, Khushal Navani, Akash Udani

Research Questions

1. **Predicting the response given by the financial institute/affiliate to a consumer complaint.**

Each observation in our data-set describes a complaint made by a consumer against any of the registered financial institutions (banks, etc.) as well as the response (if received) that the company gave regarding that complaint. For example, for a complaint made regarding bank service in Wells Fargo & Company, the response received from Wells Fargo was that this complaint was resolved with a proper explanation.

We aim to predict the response that a company would give, based on the model we train using the data-set.

2. **Predicting whether the response given by the financial institute/affiliate for a complaint would be disputed by the consumer.**

For each of the complaints in the data-set, the company has responded to the consumer. But, quite a lot of the responses have been disputed by the dissatisfied consumers. For example, one of the customers had an issue regarding his/her Citibank credit card billing statement. The bank responded with a timely response with explanation which was not up to the consumer's satisfaction level. Hence, the consumer disputed this response.

Given a complaint, we plan to calculate the likelihood that the complaint will be resolved without the customer disputing the resolution.

3. **Predicting the medium through which a new complaint is received.**

For each of the complaints recorded, we have the details of the medium (email, fax, phone, postal mail, etc.). Using this data to train our model, we plan to use it to predict the medium through which the complaint would be registered.

4. **Predicting the geographical location from where the complaint originated.**

Our data-set has information regarding the location from where the complaint originated (only the state). We wish to explore this part of the data-set and predict from which state in the United States, a new complaint originated from.

5. **Predicting the emotion of the consumer based on their complaints.**

Since 2015, the CFPB have provided consumers the option of explaining their complaints called the "consumer narratives". It is an open text field where the consumers can freely express themselves. Since these are all complaints, we expect them to have a negative sentiment. But, there are different emotions in these narratives: *anger, anticipation, disgust, fear, joy, sadness, surprise, trust.* We aim to predict the emotion of the consumer based on his/her narrative.

We are still to figure out how to go about with it. We have read some literature for the same and have considered certain lexicons like the NRC word-emotion association lexicon ("NRC emotion lexicon," n.d.) that we can use for our prediction model.

State-of-Art Review

2016 analysis of regulatory violations & consumer complaints (Harrow & Moon, 2017)

This analysis was conducted after the news of mass customer exploitation by Wells Fargo. The data set from CFPB was used to rank banks and other financial institutions based on various metrics captured in the data set such as number of complaints received and responsiveness to a consumer complaint. While this analysis cuts the data from multiple dimensions, it does not use any kind of predictive models to show temporal change or predict any metric based on the available data. Our suggested analysis aims to build analytical models that will predict certain characteristics about a complaint based on the given data set. For Example: Given the product against which the complaint was filed, what is the likelihood that the complaint will be resolved without the customer disputing the resolution. This paper instead relies on various financials experts' comments on topics such as temporal change in banking violations and consumer protections.

Calling on the CFPB for Help: Telling Stories and Consumer Protection (Foohey, 2016)

This paper focuses on the narrative of the complaints received by CFPB. It focuses on what people write about in the narrative section of the complaint. The author finds that people mainly used the narrative to either express their anger and frustration about a company's practice or to express sadness and fear about how a company's practice has impacted their lives. The author used coding frames (a method for qualitative content analysis) to analyze the narratives. The emotions were manually assigned to each narrative (with the help of research assistants). The paper then goes on to summarize the percentage of narratives categorized by emotion. While this paper categorizes narratives into emotions expressed in them, it does not use any kind of statistical or predictive model to do so. While this paper uses a qualitative method to categorize the data, we plan to use a quantitative method to perform the same task. Also, the paper only focusses on one aspect of the complaint (the narrative) while our suggested analysis aims to look at all the factors associated with a complaint.

CFPB's consumer complaint database: Analysis reveals valuable insights (Deloitte, 2013)

In this paper, Deloitte highlights some patterns/trends that they have observed after analyzing the CFPB's consumer complaint database. The paper gives us a brief idea about how the consumer complaint database works. Their analysis has produced a number of valuable insights about the nature and sources of complaints, including:

- Troubled mortgages are behind majority of the complaints
- Customer misunderstandings may create more complaints than financial institution error
- Affluent, established neighbourhoods were more likely source of complaints
- Complaint resolution times have improved

The paper does not use any predictive models; however it uses statistical methods to determine the above patterns. The paper mentions how financial institutions can use this data to improve their internal processes and meet the expectations of its customers. The paper further mentions, how by mining their own internal complaint database in addition to the Consumer Complaint database, firms might better understand the factors contributing to the complaints. By leveraging complaint data along with advanced analytical procedures, institutions could develop advanced analytics aimed at identifying and rectifying potential compliance issues before they escalate to a higher level of regulatory and reputational risk. It also mentions that since the database consists of only complaints

that have been submitted to the CFPB, it may not fully reflect a financial institution's own complaint records. A closer look at the observations can help financial institutions understand their own internal complaint issues and databases as they relate to data from Consumer Complaint Database, and use the resulting insights to potentially improve their regulatory compliance efforts, customer experience, and their own operational effectiveness.

Analysis and Study of CFPB Consumer Complaint Data Related to Mortgage Servicing Activities (Daganhardt, n.d.)

In this report, Black Knight Financial Services (BKFS) and the Five Star Institute (FSI) seek to provide context and insight to complaints received by the CFPB by comparing the Bureau's two predominant mortgage complaint categories, servicing and default, with loan trends. The report includes publicly available data from both the CFPB and mortgage servicers, as reported from the CFPB database. They have briefly mentioned the challenges and the approach taken to evaluate the trends of complaints versus trends of loan inventories. The data in the paper shows that on average, mortgage servicers received only 13 complaints for every 10,000 non-performing loans and only one complaint for every 10,000 performing loans as of Q4 2014. According to the report, complaints related to general servicing of performing loans appeared to be running at less than one $1/100^{th}$ of a percent. The report uses statistical models to help servicers understand how they are performing vs the average and whether they have the same pain points as their peers. The report provides context relative to both the overall level of complaints and the positive impact of the efforts of the servicing industry has had on reducing the level of complaints filed with CFPB, particularly associated with non-current loans.

Data Collection

Our data set contains the list of consumer complaints received by the Consumer Financial Protection Bureau. We downloaded this dataset from the official website of Consumer Financial Protection Bureau. They are complaints about financial products and services received by the bureau from December 2011 till January 2017.

Our consumer complaint dataset includes the following details:

- Dates on which the complaint was received and was further forwarded to the respective institutes/affiliates
- Type of product and the issue related
- Consumer's narrative (optional)
- Company's name with state and zipcode
- Company's response to the complaint
- Consumer feedback
- Consumer's consent for publishing the complaint details

Data Cleaning

The dataset was quite clean. There were a few missing values in the state and the zip code attributes (5594 zipcode, 5576 state). For about 2 of them, only the state value was missing, which we managed to fill based on the zip code.

There are 40638 missing values for the consumer dispute attribute. It constitutes about 6% of the entire dataset and we plan to omit these complaints for our project.

There are no missing values for the attributes we would be predicting: the medium of complaint submission and response sent to consumer.

Preliminary data analysis

We conducted the following exploratory data analysis to get a better understanding of the dataset:
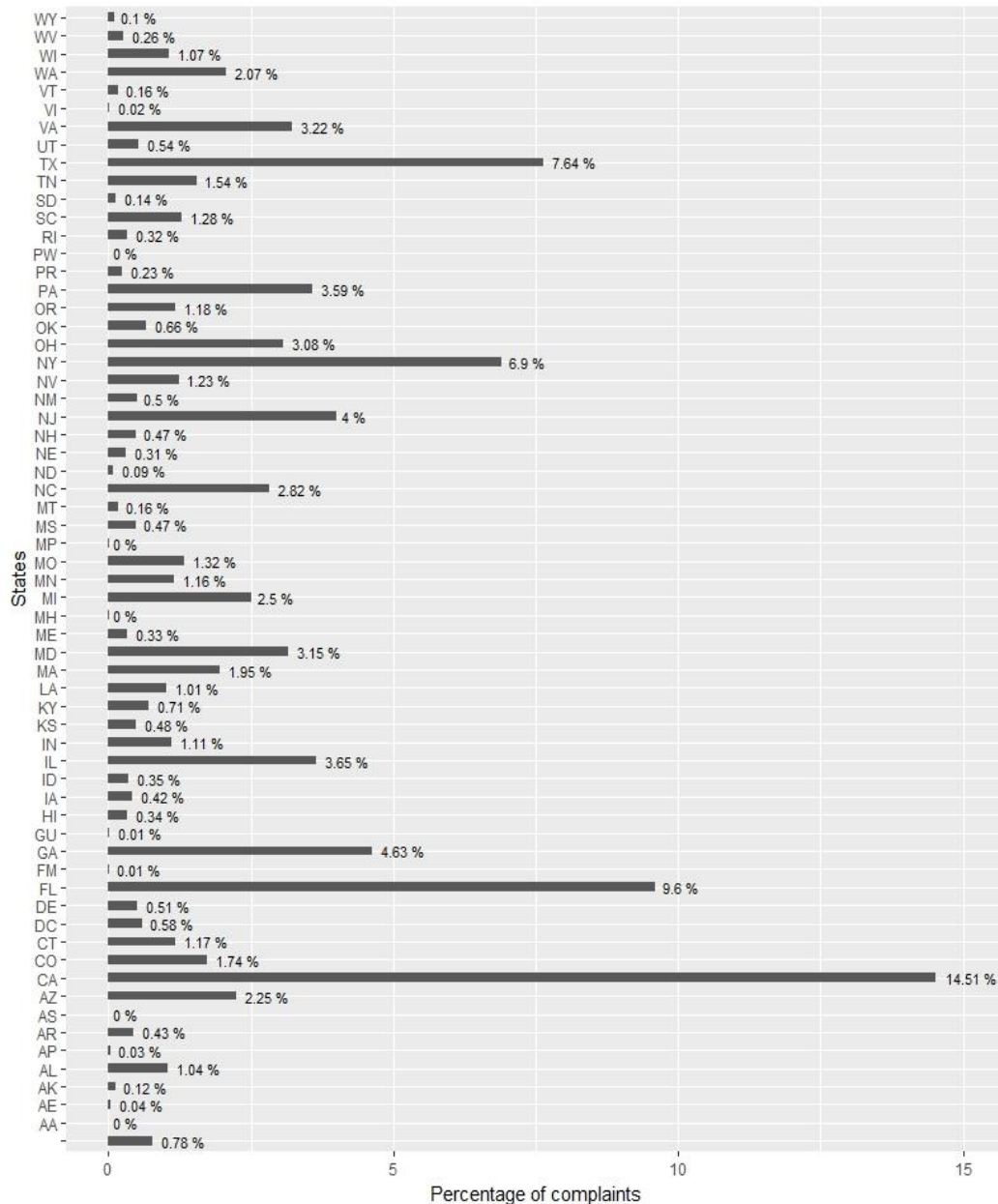

Figure 1 State wise distribution of consumer complaints

This bar graph represents the percentage of complaints originating from different states and territories in the United States. Maximum number of complaints originate from California, followed by Florida, Texas, New York and so on.

We have used percentages instead of the actual count of complaints because percentages are a better representation when it comes to relative comparison between the states.
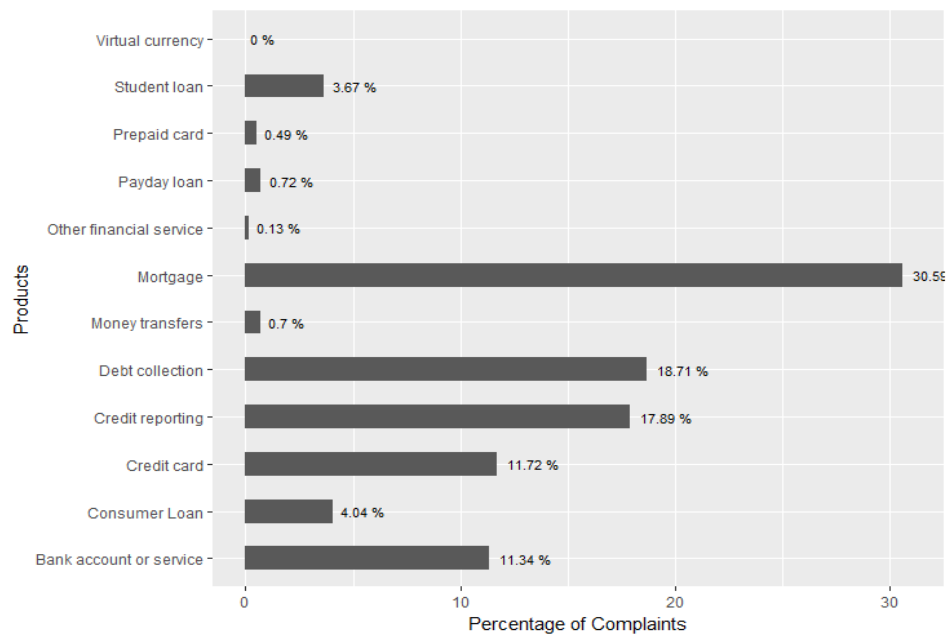
Figure 2 Product wise distribution of consumer complaints

This bar graph represents the percentage of complaints received for different types of financial products. It is evident that Mortgage has received the highest number of complaints followed by debt collection, credit reporting and so on.
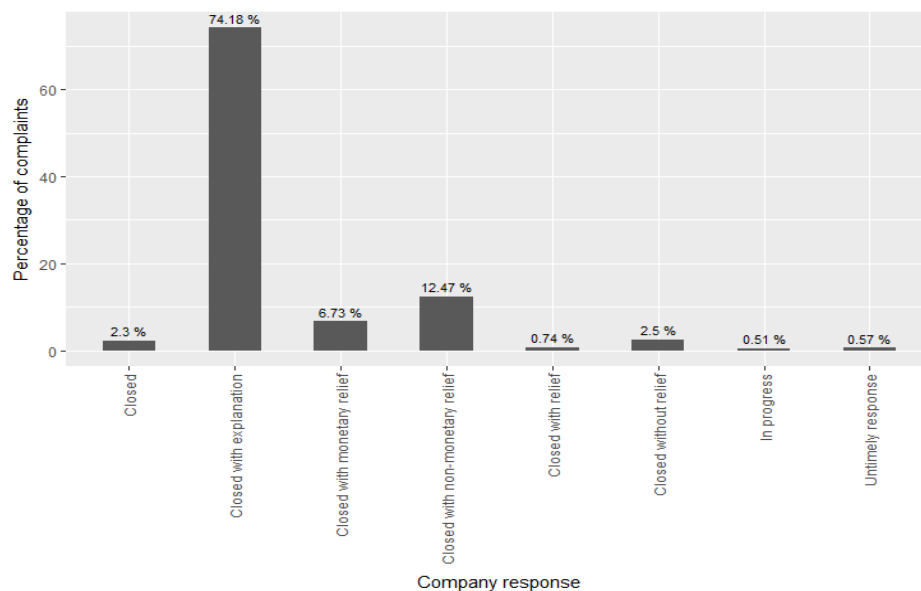

Figure 3 Complaint vs Response Distribution

This bar graph represents the percentage of complaints for different types of companies' responses to the complaints . The response 'Closed with explanation' has the highest number of complaints.
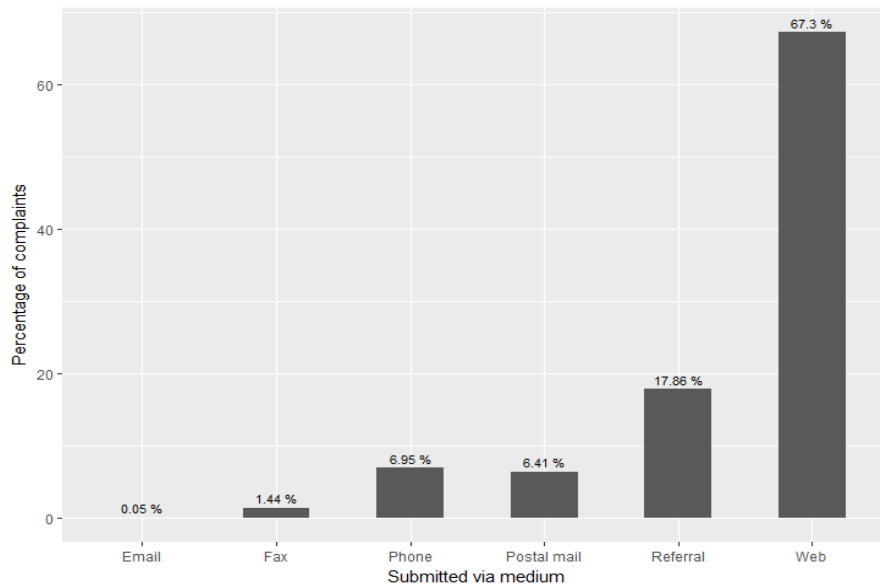
Figure 4 Communication medium vs complaint distribution

This bar graph represents the percentage of complaints submitted via different mediums. Web is the preferred mode for registering complaints. All the other mediums are relatively less used.
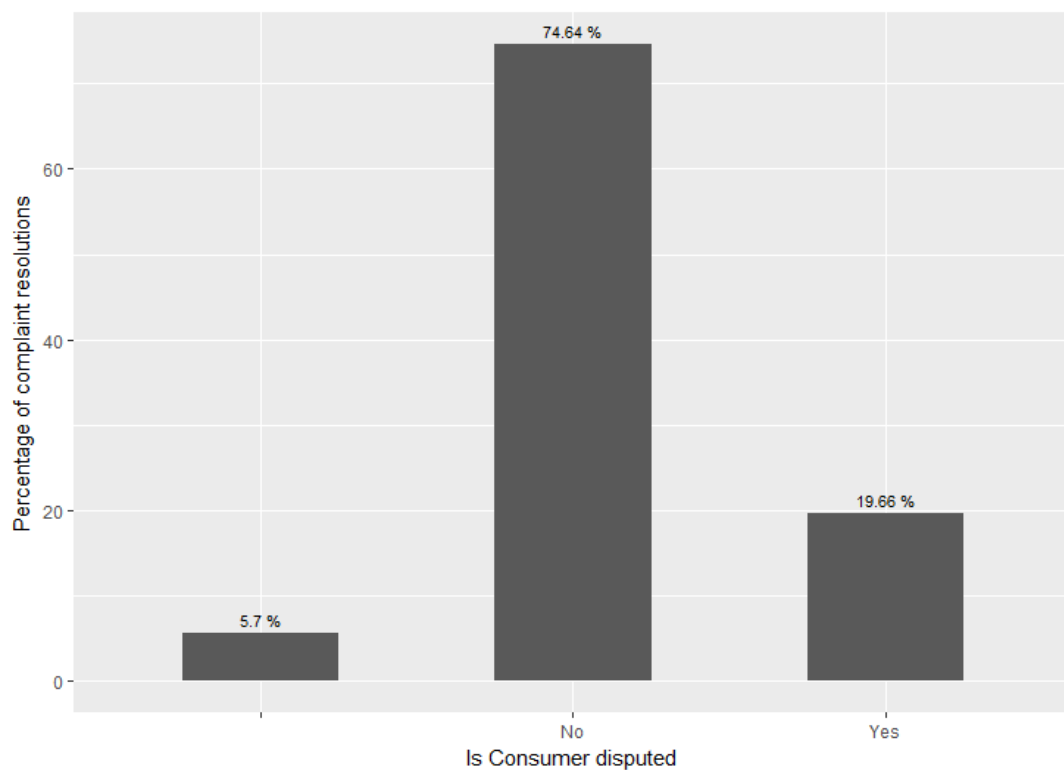

Figure 5 Dispute-complaint distribution

This bar plot shows the percentage distribution of complaint resolutions that have been disputed and accepted by the customer. Majority of the complaint resolutions have not been disputed by the consumer.
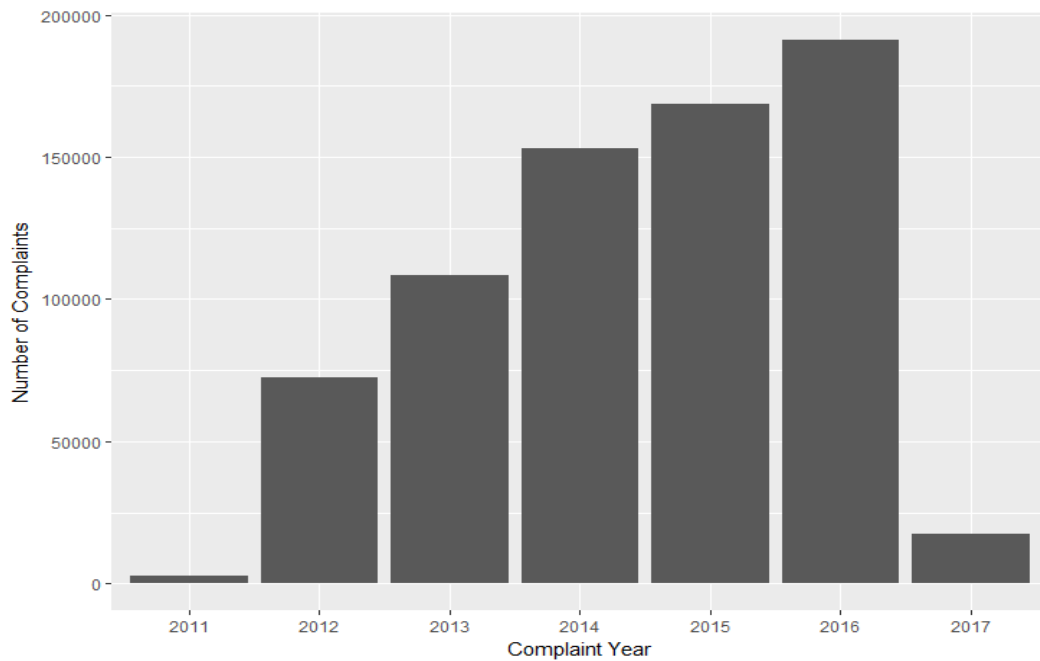
Figure 6 Year wise complaint distribution

This graph shows the distribution of the complaints over the years ranging from 2011 to 2017. The number of complaints are less in 2011 as we only have data from the month of 1st December 2011, and the data for 2017 is just up to 8th February 2017. We can see that the complaints are increasing steadily over the years.
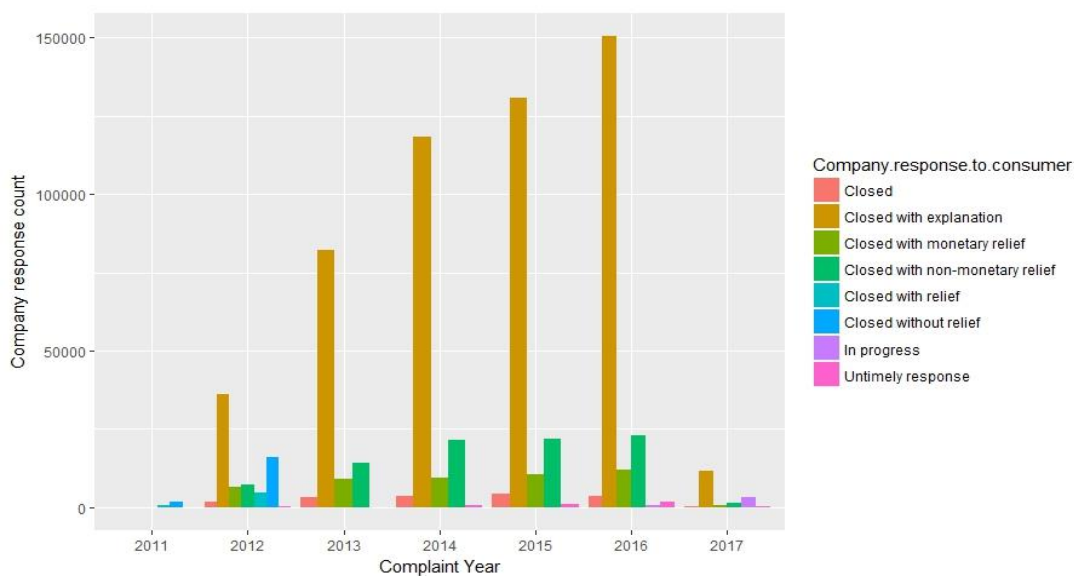

Figure 7 Year wise complaint-response distribution

This bar plot represents the number of complaints corresponding to the different types of complaint responses that the companies give, from 2011 to 2017. It is interesting to note that most of the complaints are closed with explanation over all these years.
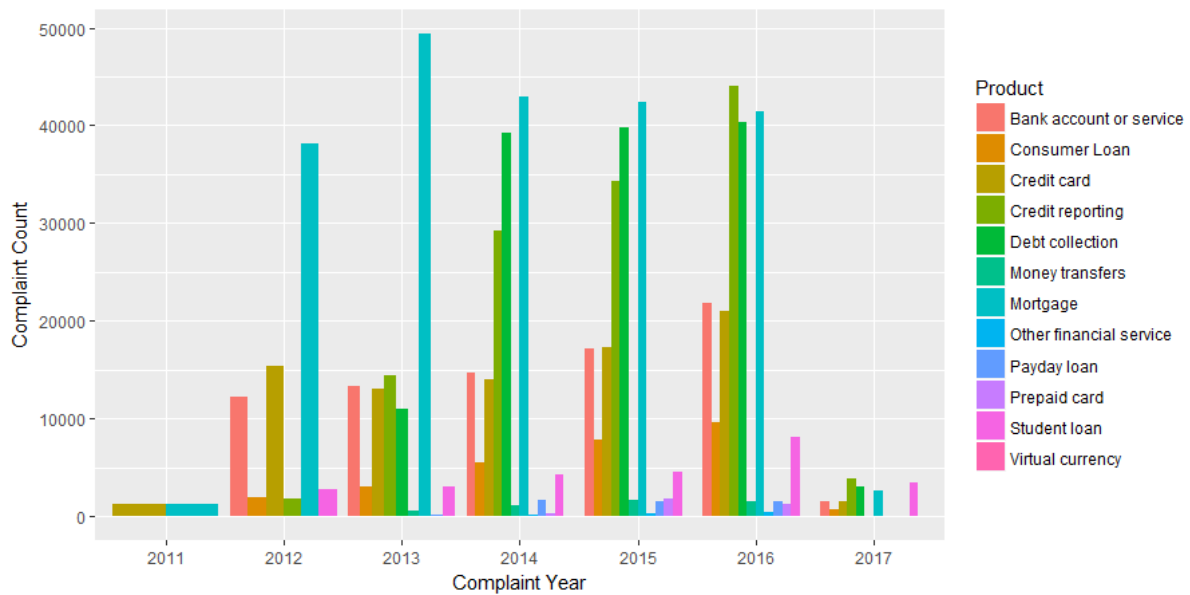
Figure 8 Year wise complaint-product distribution

This bar plot represents the number of complaints received for different types of products from 2011 to 2017. It is interesting to note that the complaints for products 'Bank account or service', 'Credit Card', 'Credit reporting' have been increasing over the years.

Notes

Given the quarterly information about each **financial institute's total assets** (CFPB, 2014), we want to find a correlation between the number of complaints (either total or of a type like mortgages and student loans) and the total count of assets in possession of any financial institution. The number of complaints along with associated asset information over each quarter might yield interesting insights based on which we can predict a company's performance in the future.

Work Distribution

Through the course of Milestone 1, we have maintained a uniform work distribution amongst all four members of the group. We majorly worked on four research questions, with each team member analysing the dataset in context of the research question taken up by them. All four of us did some research and came up with a related resource (mentioned in the second section of this report). Together we came up with two new prospective research questions but since we are not certain at this moment if we should proceed with them, we have not done any exploratory analysis with respect to those questions.

References

[1] NRC emotion lexicon. Retrieved February 25, 2017, from
http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[2] Harrow, R., & Moon, C. (2017). 2016 analysis of regulatory violations & consumer complaints. Retrieved February 25, 2017, from https://www.valuepenguin.com/analysis-regulatory-penalties-consumer-complaints

[3] Foohey, Pamela, Calling on the CFPB for Help: Telling Stories and Consumer Protection (December 9, 2016). 80 Law & Contemporary Problems (Forthcoming); Indiana Legal Studies Research Paper No. 356. Available at SSRN: https://ssrn.com/abstract=2883242

[4] Retrieved February 25, 2017, from https://www2.deloitte.com/content/dam/Deloitte/se/Documents/financial-services/CFPBConsumerComplaintDatabase091913US_FSI_.pdf

[5] Consumer Financial Protection Bureau. (2015). *Consumer Complaint Database*. Retrieved from https://catalog.data.gov/dataset/consumer-complaint-database

[6] Daganhardt, D. *Analysis and Study of CFPB Consumer Complaint Data Related to Mortgage Servicing Activities*. . Retrieved from http://dsnews.com/wp-content/uploads/sites/25/2015/04/WP_DNA_CFPBComplaintTrackingWhitePaper_FullReport_FINAL.pdf

[7] Institutions subject to CFPB supervisory authority | consumer financial protection bureau. (2014, March 31). Retrieved from https://www.consumerfinance.gov/policy-compliance/guidance/supervision-examinations/institutions/