Milestone Report

Problem

Consumer Financial Protection Bureau is an government agency,it helps consumers complaints heard by financial companies.the goal of the project is to study and identify the inappropriate practices and allowing the government to stop those before it becomes a major issue.This project focuses on the analysis of the complaints over different segments,also providing sentiment analysis of the complaints.

Data Extraction

Dataset used for analysis is US Consumer Finance Complaints data from Kaggle. Importing and Reading the csv file for further analysis,is the first step in data analysis.

There are 18 variables

1.Date received The date the CFPB received the complaint. For example, "05/25/2013."

2.Product The type of product the consumer identified in the complaint. For example, "Checking or savings account" or "Student loan."

3.Sub-product The type of sub-product the consumer identified in the complaint. For example, "Checking account" or"Private student loan."

4.Issue The issue the consumer identified in the complaint. For example, "Managing an account" or "Struggling to repay your loan."

5.Sub-issue The sub-issue the consumer identified in the complaint. For example, "Deposits and withdrawals" or "Problem lowering your monthly payments."

6.Consumer complaint narrative Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint.Consumers must opt-in to share their narrative. We will not publish the narrative unless the consumer consents, and consumers can opt-out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer.

7.Company public response The company's optional, public-facing response to a consumer's complaint. Companies can choose to select a response from a pre-set list of options that will be posted on the public database. For example, "Company believes complaint is the result of an isolated error."

8.Company The complaint is about this company. For example, "ABC Bank."

9.State The state of the mailing address provided by the consumer.

10.ZIP code The mailing ZIP code provided by the consumer. This field may: i) include the first five digits of a ZIP code; ii) include the first three digits of a ZIP code (if the consumer consented to publication of their complaint narrative); or iii) be blank (if ZIP codes have been submitted with non-numeric values, if there are less than 20,000 people in a given ZIP code, or if the complaint has an address outside of the United States).

11.Tags Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.For example, complaints where the submitter reports the age of the consumer as 62 years or older are tagged "Older American." Complaints submitted by or on behalf of a servicemember or the spouse or dependent of a servicemember are tagged "Servicemember." Servicemember includes anyone who is active duty, National Guard, or Reservist, as well as anyone who previously served and is a veteran or retiree.

12.Consumer consent provided? Identifies whether the consumer opted in to publish their complaint narrative. We do not publish the narrative unless the consumer consents, and consumers can opt-out at any time.

13.Submitted via How the complaint was submitted to the CFPB. For example, "Web" or "Phone."

14.Date sent to company The date the CFPB sent the complaint to the company.

15.Company response to consumer This is how the company responded. For example, “Closed with explanation.”

16.Timely response? Whether the company gave a timely response. For example, “Yes” or “No.”

17.Consumer disputed? Whether the consumer disputed the company’s response.

18.Complaint ID The unique identification number for a complaint.

As we examine the data most of the variables like company,product,sub_product,issue and sub_issue are categorical variables.

Cleaning up of Data

- Date field needs to be changed as per the format required
- Convert the fields to factors
- Before text mining, the comments columns needs to be removed the special characters,spaces and unwanted numbers.
- Empty values needs to be changed to NA or blank according to requirement.

Limitations

- As most of the variables are Categorical variables,the analysis is mainly limited to several categories.
- We have two date fields date_received and date_sent_company,but the exact time duration for a company to reply is not available.
- The pubic response variable in CFPB database is optional so the dataset may not contain all the comments from customer.

Preliminary Exploration

```
complaint2 <- read_csv("consumer_complaints.csv")
```

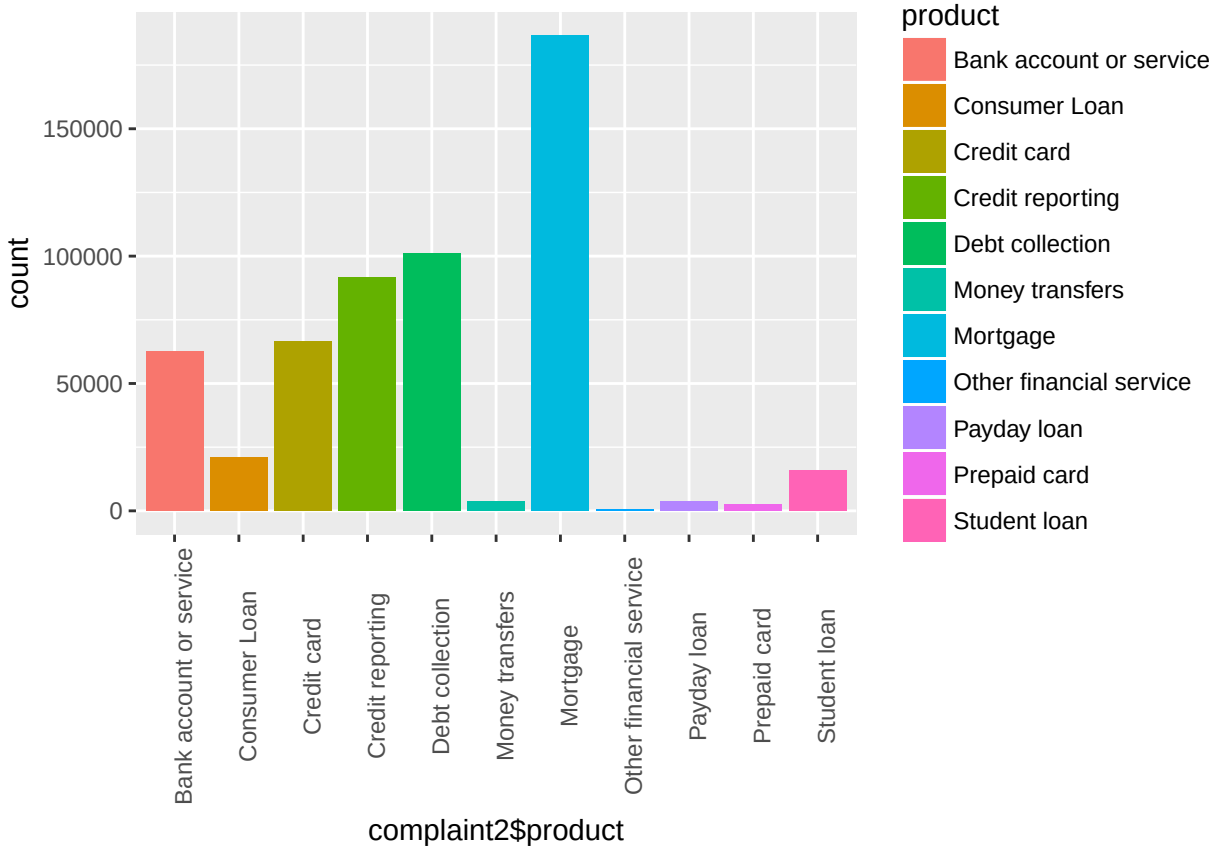
```
## Parsed with column specification:
## cols(
##   date_received = col_character(),
##   product = col_character(),
##   sub_product = col_character(),
##   issue = col_character(),
##   sub_issue = col_character(),
##   consumer_complaint_narrative = col_character(),
##   company_public_response = col_character(),
##   company = col_character(),
##   state = col_character(),
##   zipcode = col_character(),
##   tags = col_character(),
##   consumer_consent_provided = col_character(),
##   submitted_via = col_character(),
##   date_sent_to_company = col_character(),
##   company_response_to_consumer = col_character(),
##   timely_response = col_character(),
##   `consumer_disputed?` = col_character(),
##   complaint_id = col_integer()
## )
```

Cleaning up the data

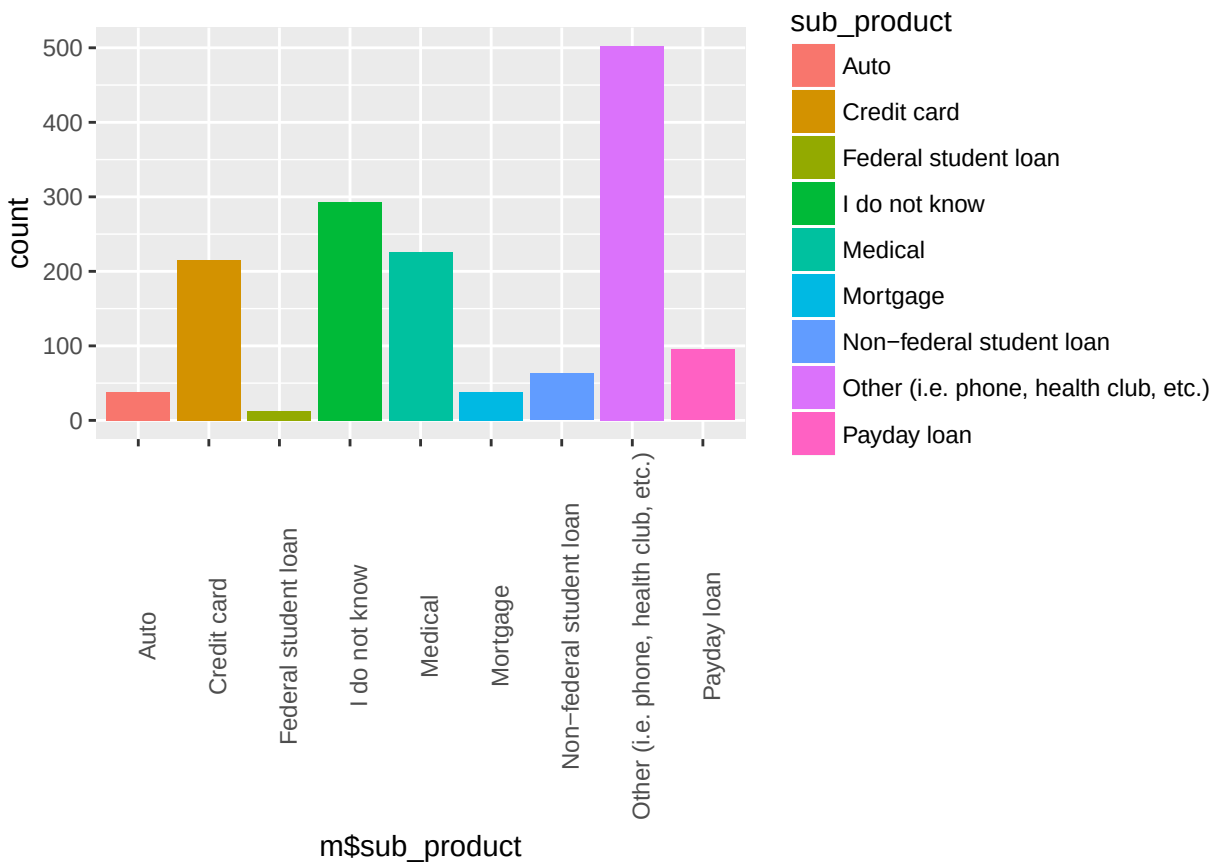
```
complaint2$date_received <- mdy(complaint2$date_received)
complaint2$date_sent_to_company <- mdy(complaint2$date_sent_to_company)

complaint2$product<-as.factor(complaint2$product)
complaint2$company<-as.factor(complaint2$company)
complaint2$sub_product<-as.factor(complaint2$sub_product)
complaint2$issue<-as.factor(complaint2$issue)
complaint2$sub_issue<-as.factor(complaint2$sub_issue)
complaint2$submitted_via<-as.factor(complaint2$submitted_via)

m<-na.omit(complaint2)
```

Univariate Analysis

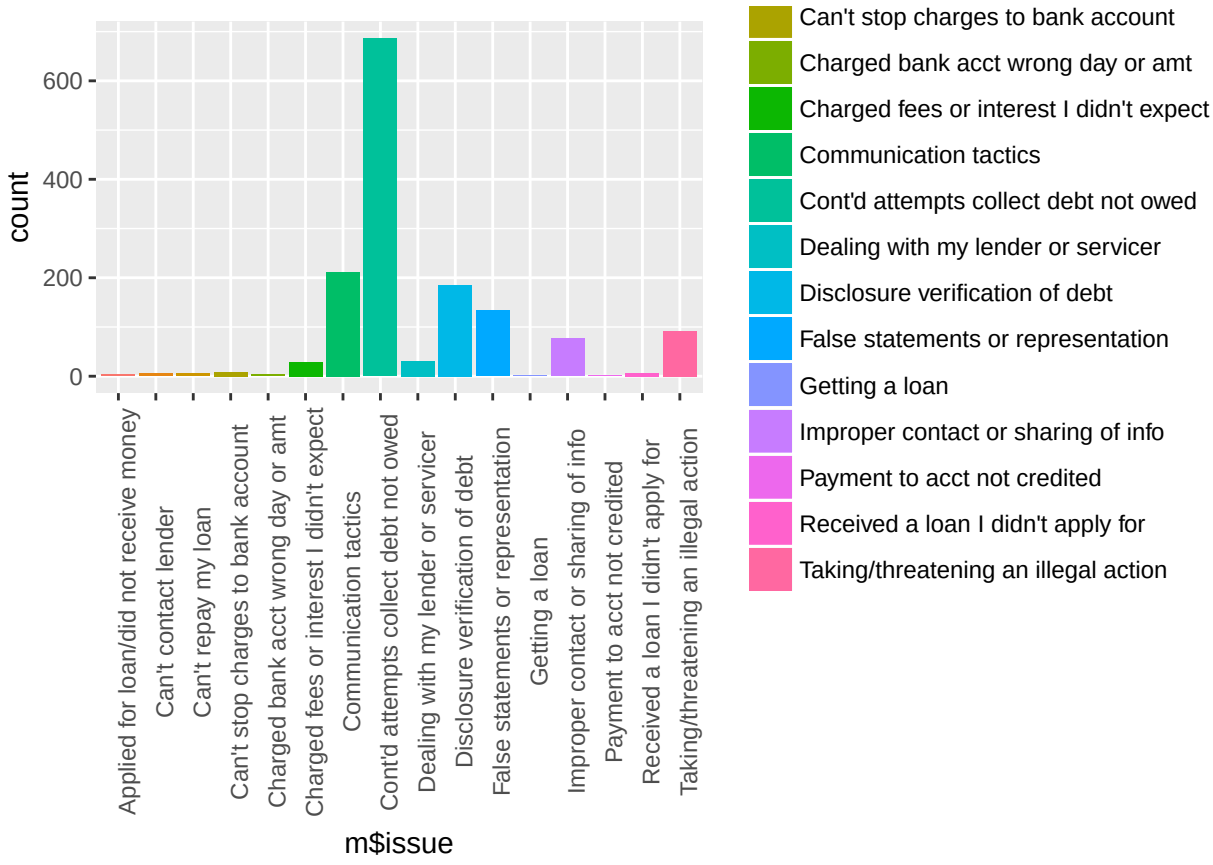
```
ggplot(complaint2,aes(x=complaint2$product,fill=product))+
  geom_bar()+
theme(axis.text.x=element_text(angle=90))
```



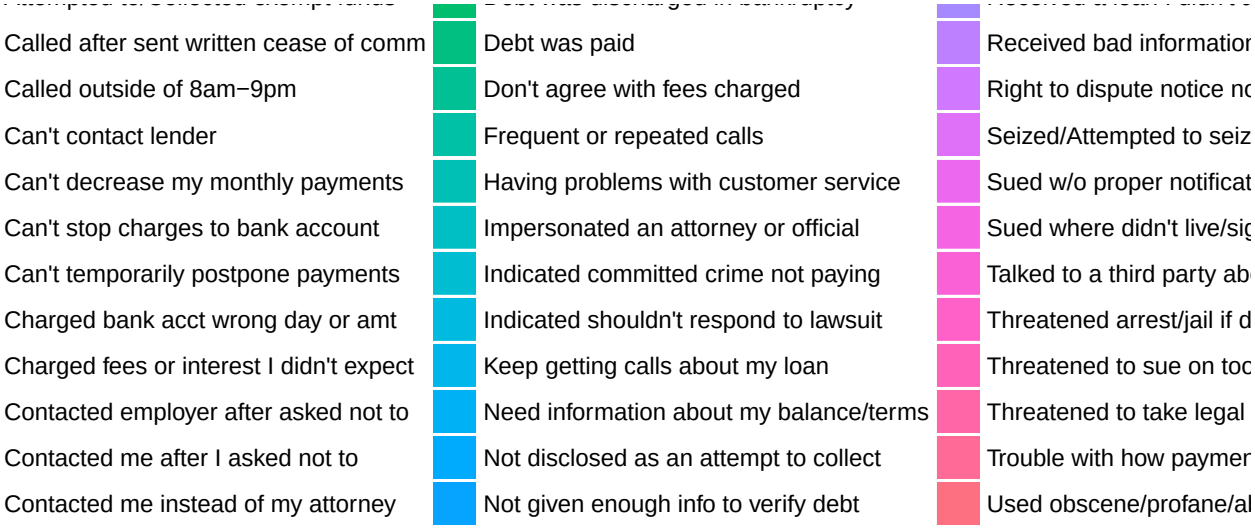

```
ggplot(m,aes(x=m$sub_product,fill=sub_product))+
  geom_bar()+
  theme(axis.text.x=element_text(angle=90))
```



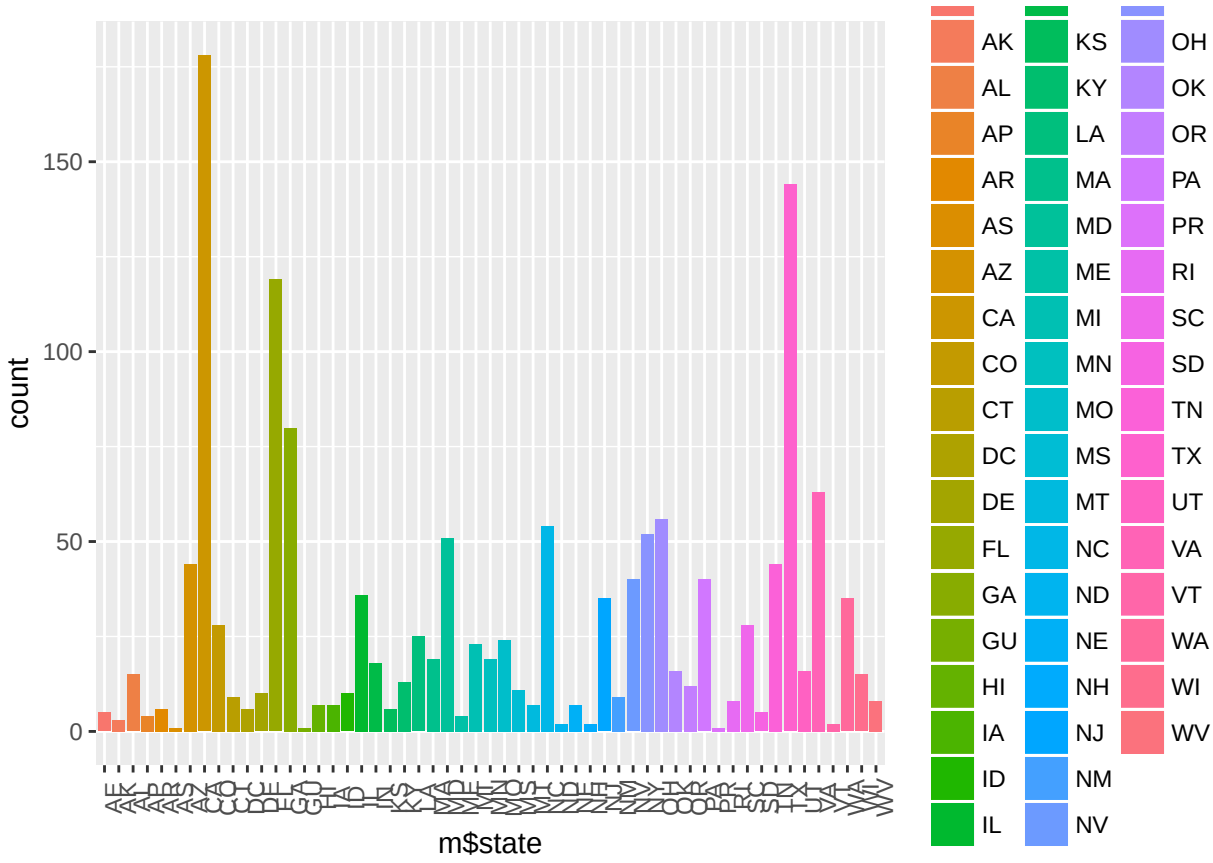

```
ggplot(m,aes(x=m$issue,fill=issue))+
  geom_bar()+
  theme(axis.text.x=element_text(angle=90))
```



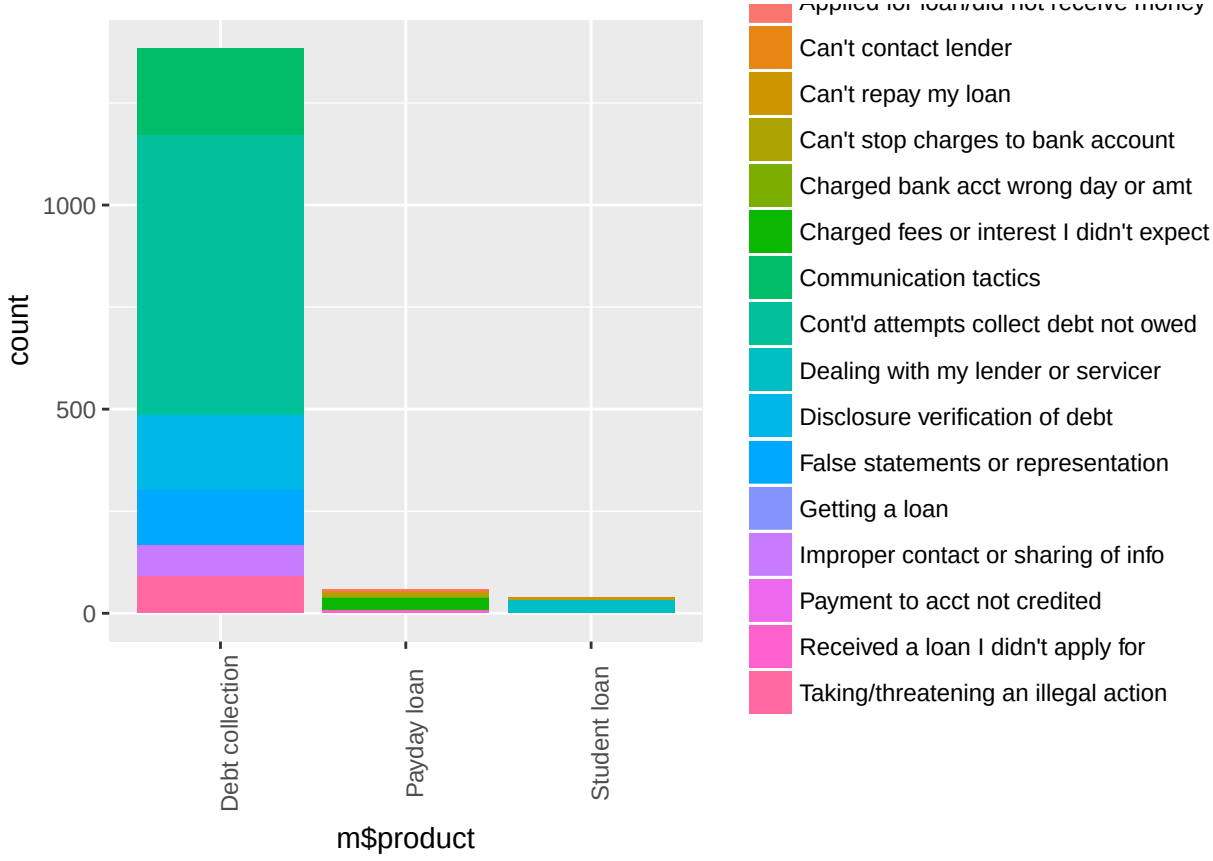

```
ggplot(m,aes(x=m$sub_issue,fill=sub_issue))+
  geom_bar()+
  theme(axis.text.x=element_text(angle=90))
```




```
ggplot(m,aes(x=m$state,fill=state))+
  geom_bar()+
  theme(axis.text.x=element_text(angle=90))
```




```
ggplot(m,aes(x=m$product,fill=m$issue))+
  geom_bar()+
  theme(axis.text.x=element_text(angle=90))
```

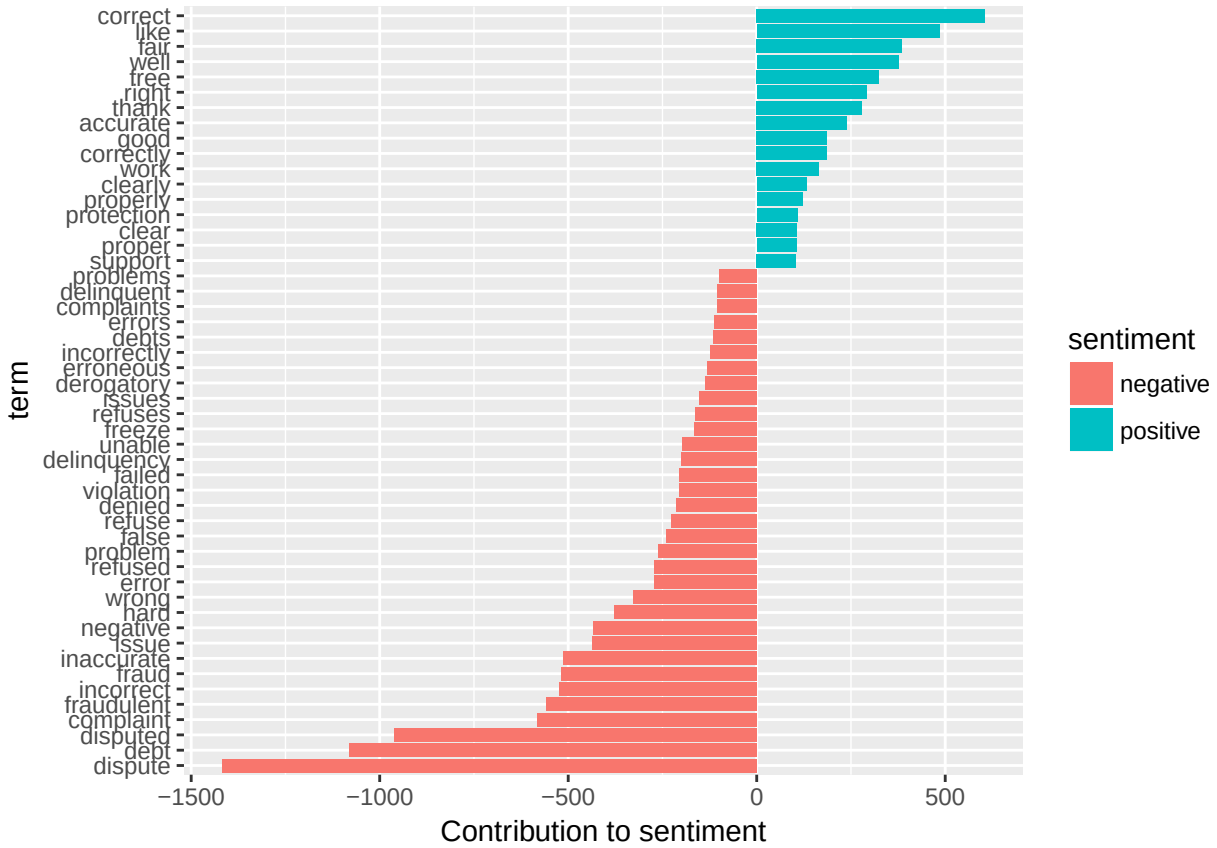


Sentiment Analysis

```
data <- complaint2 %>%
  filter(company == "Equifax") %>%
  select(consumer_complaint_narrative)%>%
  na.omit()

qdap_clean <- function(x){
  x <- replace_abbreviation(x)
  x <- replace_contraction(x)
  x <- replace_number(x)
  x <- replace_ordinal(x)
  x <- replace_symbol(x)
  x <- tolower(x)
}



tm_clean <- function(corpus){
  tm_clean <- tm_map(corpus,removePunctuation)
  corpus <- tm_map(corpus,stripWhitespace)
  corpus <- tm_map(corpus,removeWords,c(stopwords("en"),"xxxx","xx"))
  return(corpus)
}
```

```
data %>%
    unnest_tokens(word,consumer_complaint_narrative) %>%
    inner_join(get_sentiments("bing"))%>%
    count(word,sentiment,sort = TRUE) %>%
    ungroup()%>%
    filter(n >=100 ) %>%
    mutate(n = ifelse(sentiment == "negative", -n, n)) %>%
    mutate(term = reorder(word, n)) %>%
    ggplot(aes(term, n, fill = sentiment)) +
    geom_bar(stat = "identity") +
    ylab("Contribution to sentiment") +
    coord_flip()
```

```
## Joining, by = "word"
```



##narr_sentiments %>% count(sentiment, term, wt = count) %>% ungroup() %>% filter(n >= 3000) %>% mutate(n = ifelse(sentiment == “negative”, -n, n)) %>% mutate(term = reorder(term, n)) %>% ggplot(aes(term, n, fill = sentiment)) + geom_bar(stat = “identity”) + ylab(“Contribution to sentiment”) + coord_flip()

“‘

Findings and Further Approach

- From initial EDA,distribution of complaints based on company,product and issues are performed. This may be extended to Multivariate analysis like number of complaints based on sub-products among each product,
- As we have state details of the complaint originated, this leads to the approach of considering states with maximum number of complaints.
- Sentiment Analysis was performed for overall consumer complaints,this can be further analysed based on each product.