User Complaints Mining

gor Baranov, Michael Parravani, Ariana Biagi, Grace Tsai

February 4, 2019

Preparing data

Load and clean the customer complaints data

The data should be downloaded from Kaggle's Consumer Complaints Database.

Loading and removing rows with no complaint narrative and unnecessary colimns:

```r
if(!file.exists("./user-complaints-mining/df.Rds")) {
  df <- read_csv(file="../data/Consumer_Complaints.csv.zip",col_names = TRUE)
  df <- df[,-c(1,7,9:18)]
  df <- df[!is.na(df[,"Consumer complaint narrative"]),] #199,970
  df <- df[!is.na(df[,"Company"]),] # no NA's
  df <- df[!is.na(df[,"Product"]),] # no NA's
  df <- df[!is.na(df[,"Issue"]),] # no NA's
  df <- df[!is.na(df[,"Sub-product"]),] # 147,788 total left
  df <- df[!is.na(df[,"Sub-issue"]),]   # 81,940 total left

  # Converting all but narrative columns to factors
  df$Product <- as.factor(df$Product)
  df$`Sub-product` <- as.factor(df$`Sub-product`)
  df$Issue <- as.factor(df$Issue)
  df$`Sub-issue` <- as.factor(df$`Sub-issue`)
  df$Company <- as.factor(df$Company)
  saveRDS(df, file = "./user-complaints-mining/df.Rds")
  gc()
} else {
  df <- readRDS("./user-complaints-mining/df.Rds")
}
df
```

```
## # A tibble: 81,940 x 6
##    Product  `Sub-product` Issue `Sub-issue`  `Consumer complaint n~ Company
##    <fct>    <fct>         <fct> <fct>        <chr>                  <fct>
##  1 Debt co~ Other (i.e. ~ Disc~ Not given ~  This company refuses ~ The CB~
##  2 Debt co~ Credit card   Impr~ Talked to ~  "This complaint is in~ SQUARE~
##  3 Debt co~ Credit card   Taki~ Sued w/o p~  "I am writing to requ~ Selip ~
##  4 Debt co~ Other (i.e. ~ Cont~ Debt resul~  My identity was stole~ Southw~
##  5 Student~ Federal stud~ Can'~ Can't get ~  "I was dropped from m~ AES/PH~
##  6 Debt co~ Credit card   Disc~ Not given ~  The first communicati~ Blatt,~
##  7 Debt co~ Other (i.e. ~ Comm~ Frequent o~  "My complaint is n't ~ AR Res~
##  8 Debt co~ I do not know Fals~ Attempted ~  In a clearance interv~ SANTAN~
##  9 Student~ Non-federal ~ Can'~ Can't temp~  XXXX University, XXXX~ Navien~
## 10 Student~ Non-federal ~ Deal~ Received b~  I had attended XXXX a~ CITIZE~
## # ... with 81,930 more rows
```

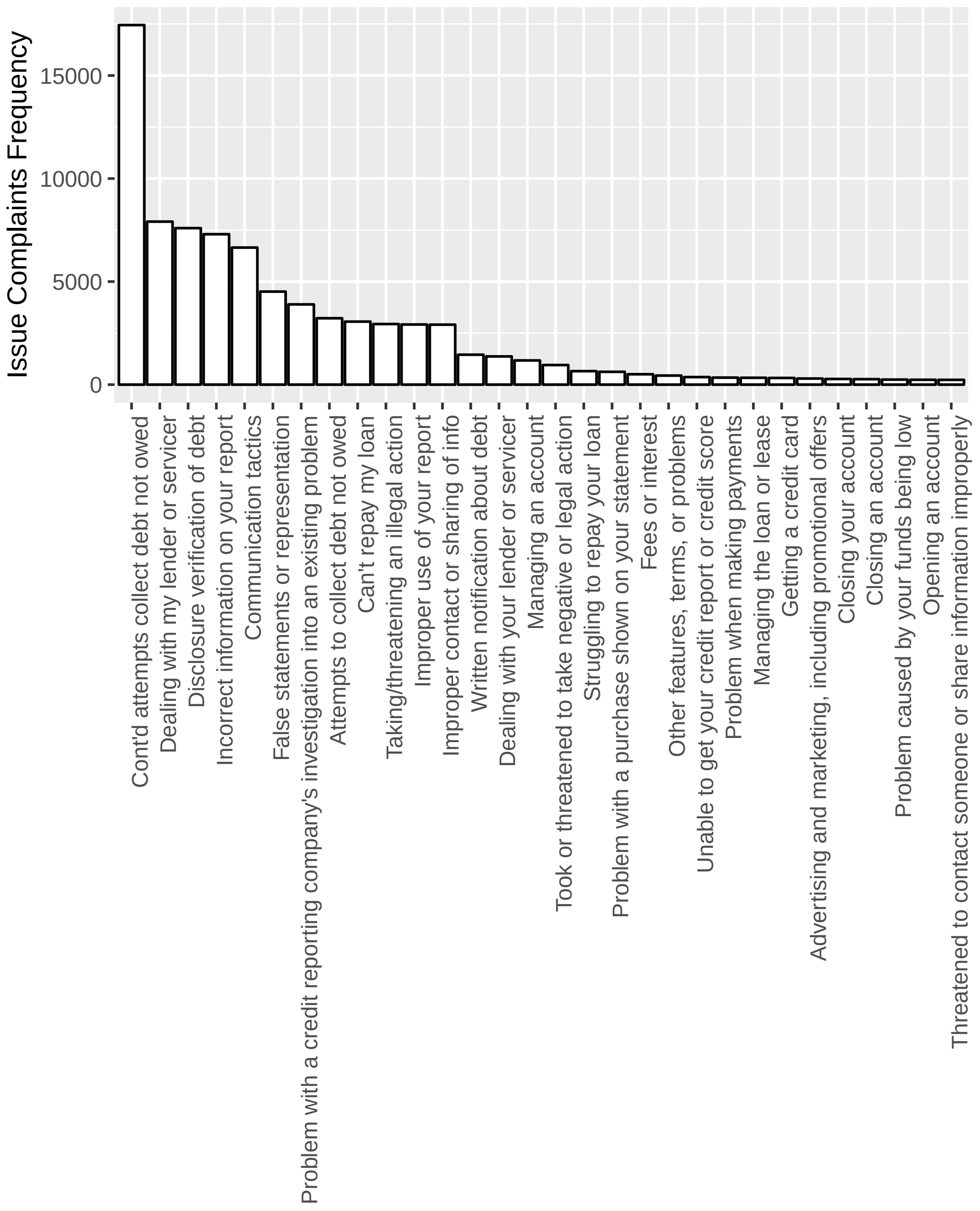
Converting all but narrative columns to factors:

```r
df$Product <- as.factor(df$Product)
df$`Sub-product` <- as.factor(df$`Sub-product`)
df$Issue <- as.factor(df$Issue)
df$`Sub-issue` <- as.factor(df$`Sub-issue`)
df$Company <- as.factor(df$Company)
```

Feature engineering

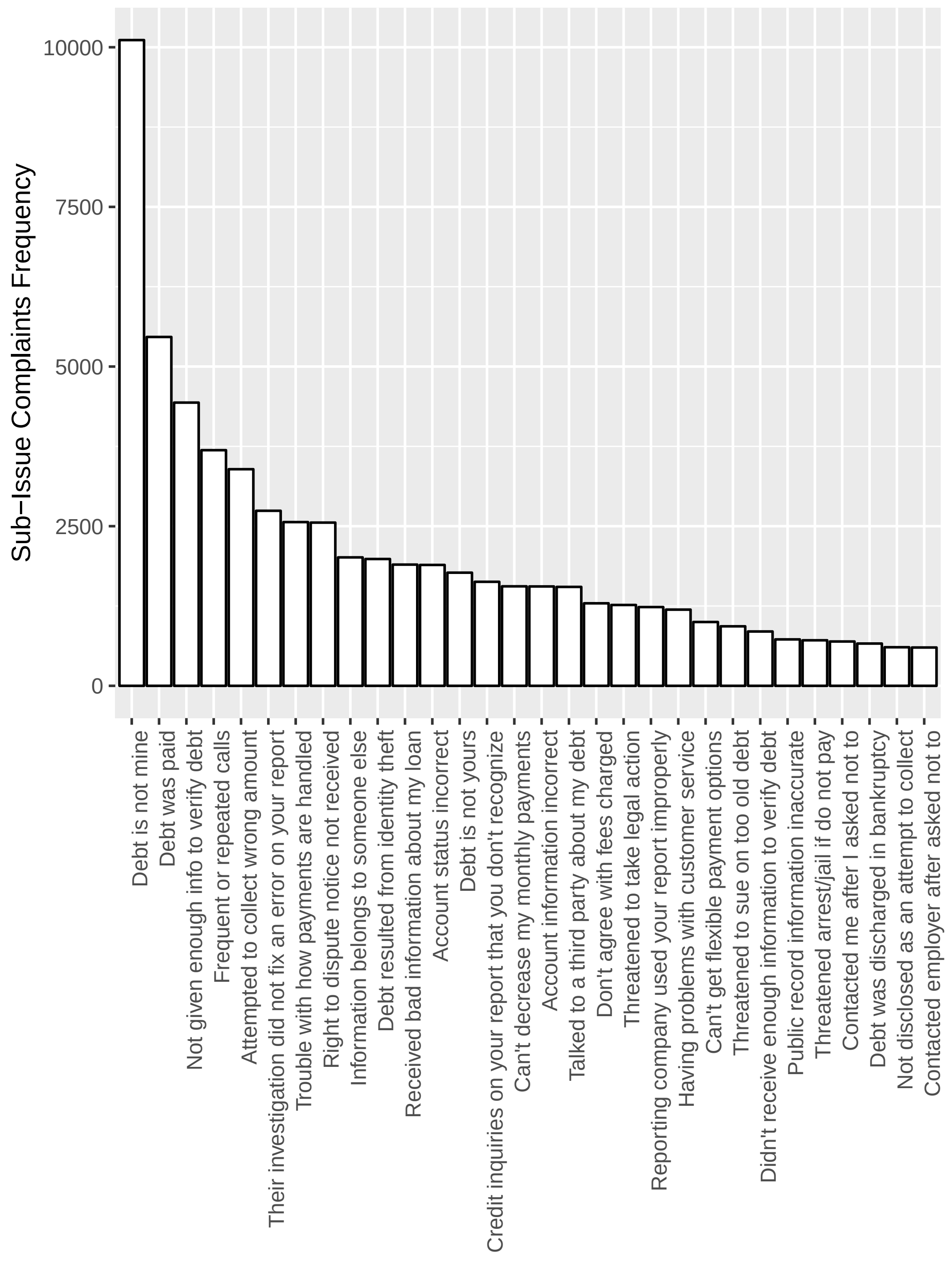
Distribution of the most frequent "Issue" complaints

```r
most_freq_issues_list <- levels(fct_infreq(df$Issue))[1:30]
ggplot() + aes(fct_infreq(df[df$Issue %in% most_freq_issues_list,]$Issue))+
  geom_histogram(colour="black", fill="white", stat = "count")+
  ylab("Issue Complaints Frequency") + xlab("")+
  theme(axis.text.x = element_text(angle =90, hjust = 1))
```


Issue Complaints Frequency
15000
10000
5000
0
Cont'd attempts collect debt not owed
Dealing with my lender or servicer
Disclosure verification of debt
Incorrect information on your report
Communication tactics
False statements or representation
Problem with a credit reporting company's investigation into an existing problem
Attempts to collect debt not owed
Can't repay my loan
Taking/threatening an illegal action
Improper use of your report
Improper contact or sharing of info
Written notification about debt
Dealing with your lender or servicer
Managing an account
Took or threatened to take negative or legal action
Struggling to repay your loan
Problem with a purchase shown on your statement
Fees or interest
Other features, terms, or problems
Unable to get your credit report or credit score
Problem when making payments
Managing the loan or lease
Getting a credit card
Advertising and marketing, including promotional offers
Closing your account
Closing an account
Problem caused by your funds being low
Opening an account
Threatened to contact someone or share information improperly

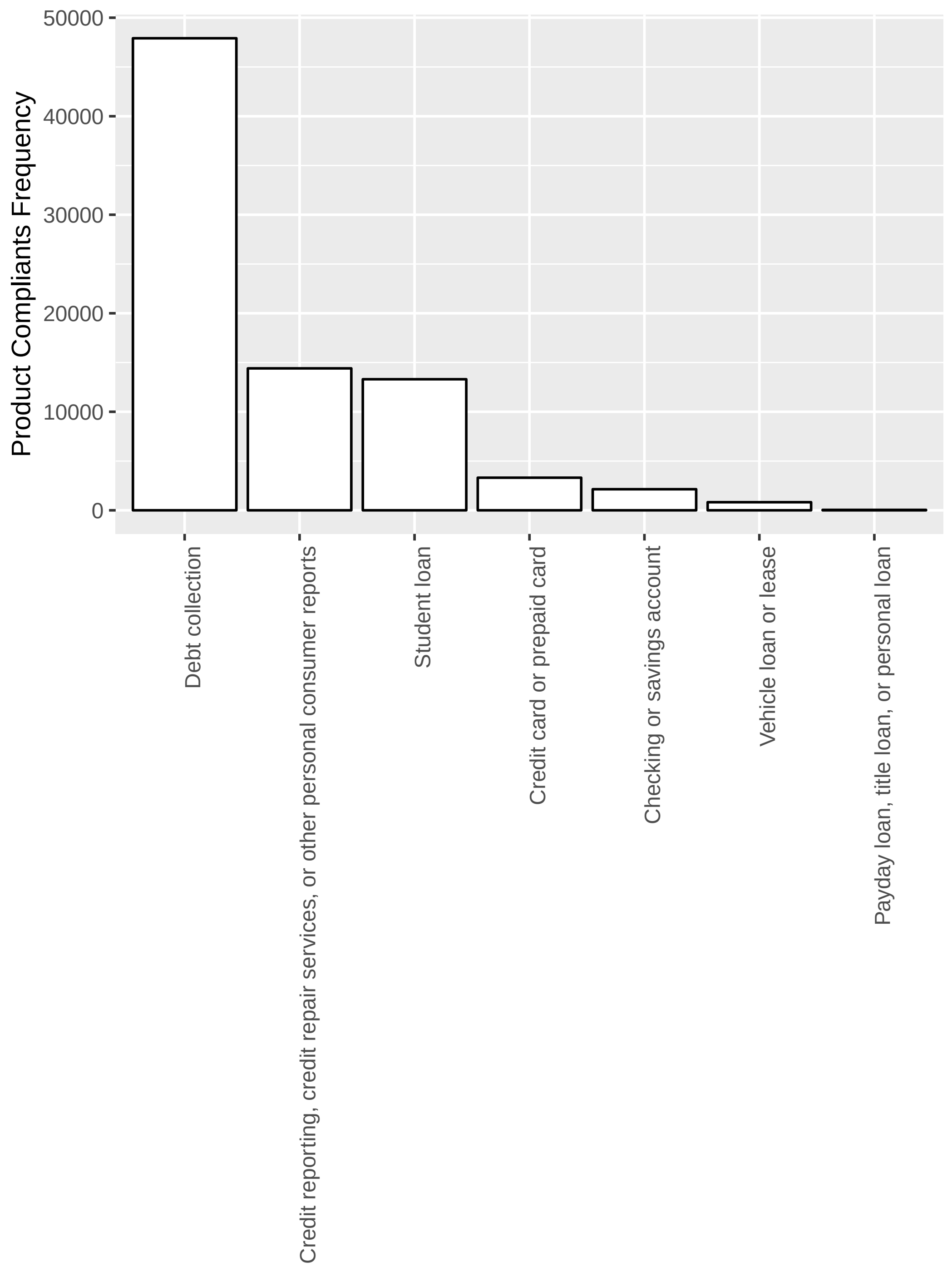
Distribution of the most frequent "Sub-issue" complaints

```
most_freq_subissues_list <- levels(fct_infreq(df$`Sub-issue`))[1:30]
ggplot() + aes(fct_infreq(df[df$`Sub-issue` %in% most_freq_subissues_list,]$`Sub-issue`))+
  geom_histogram(colour="black", fill="white", stat = "count")+
  ylab("Sub-Issue Complaints Frequency") + xlab("")+
  theme(axis.text.x = element_text(angle =90, hjust = 1))
```



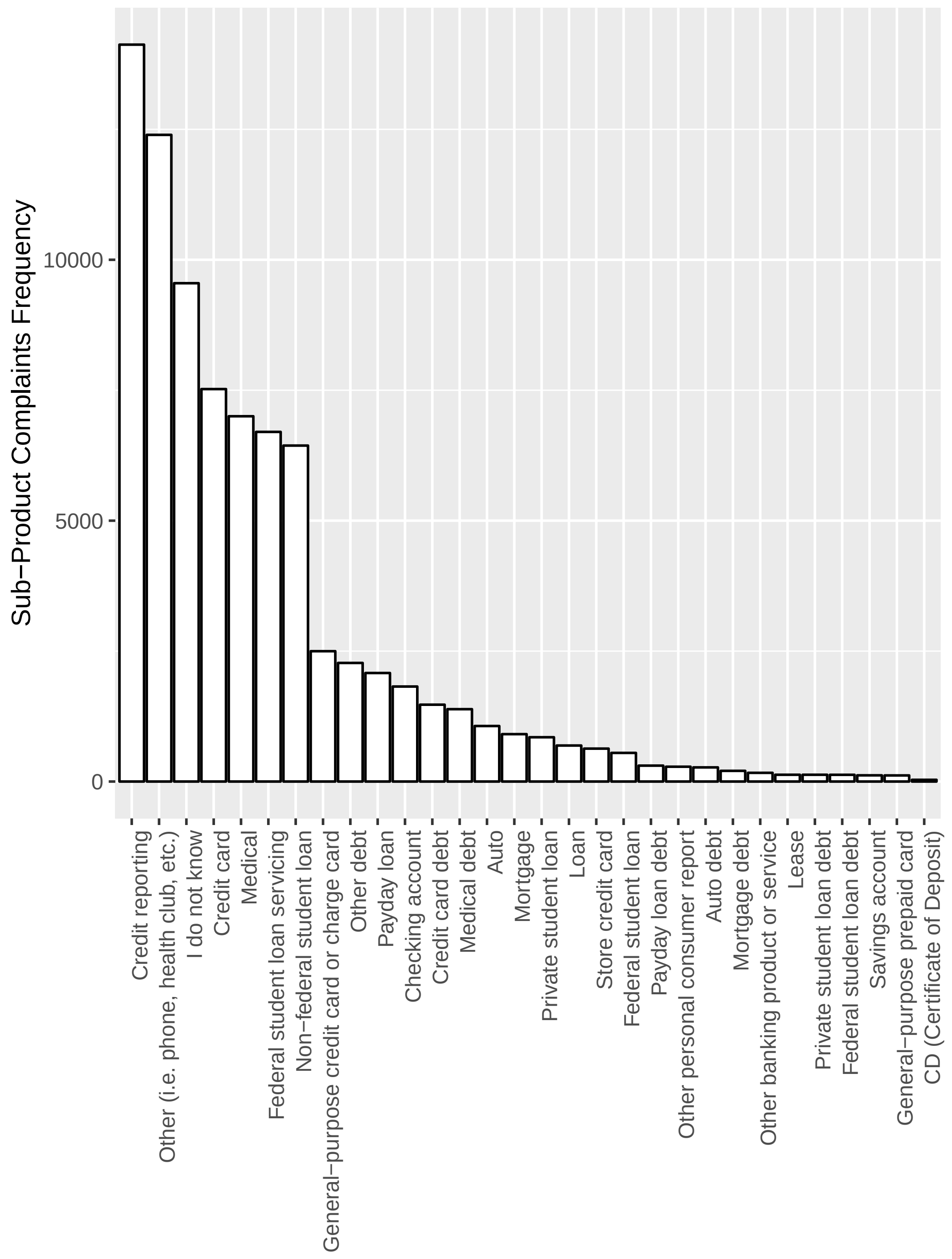
Distribution of the most frequent "Product" complaints

```r
most_freq_product_list <- levels(fct_infreq(df$Product))[1:30]
ggplot() + aes(fct_infreq(df[df$Product %in% most_freq_product_list,]$Product))+
  geom_histogram(colour="black", fill="white", stat = "count")+
  ylab("Product Compliants Frequency") + xlab("")+
  theme(axis.text.x = element_text(angle =90, hjust = 1))
```



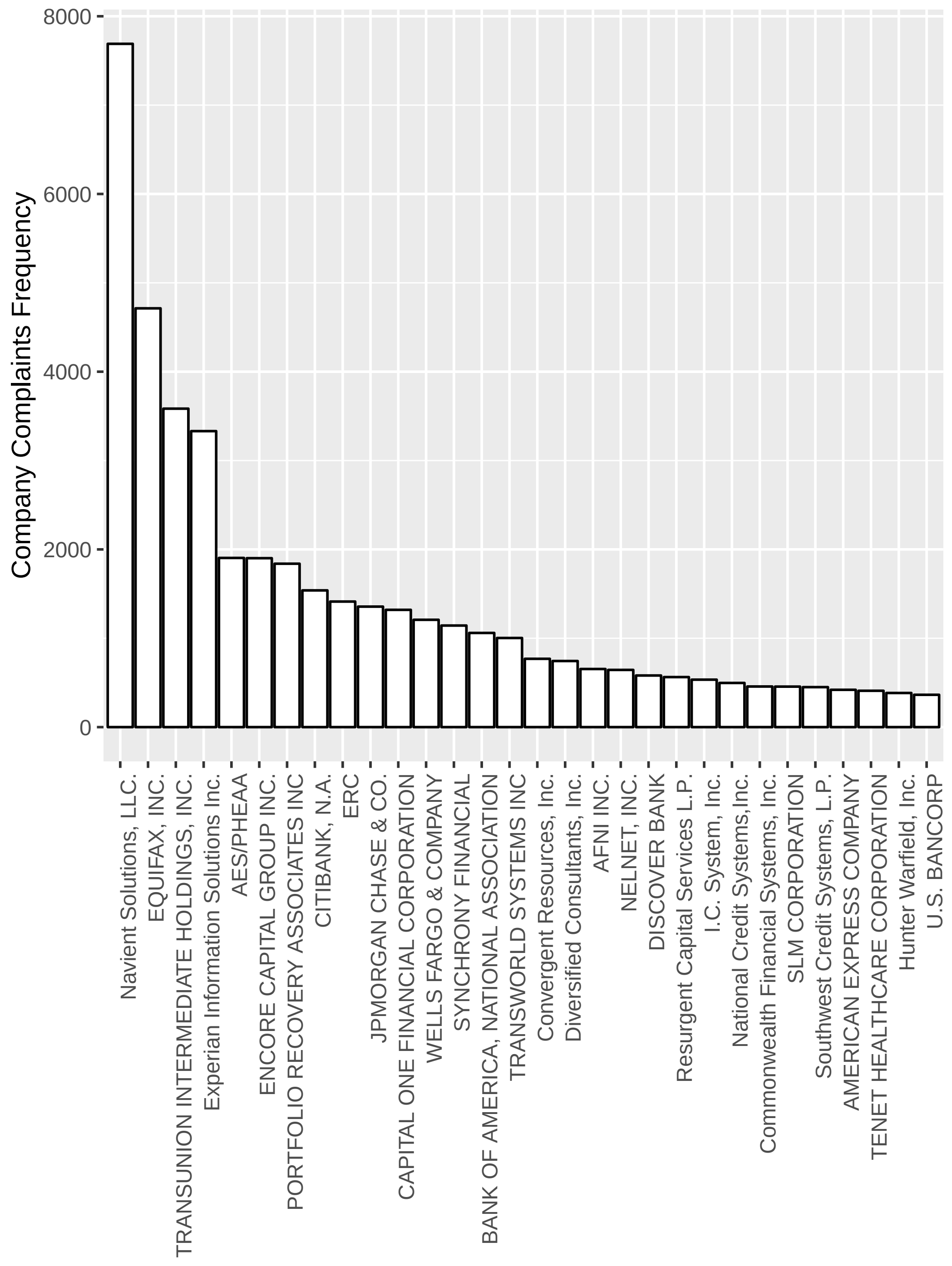
Distribution of the most frequent "Sub-product" complaints

```r
most_freq_subproduct_list <- levels(fct_infreq(df$`Sub-product`))[1:30]
ggplot() + aes(fct_infreq(df[df$`Sub-product` %in% most_freq_subproduct_list,]$`Sub-product`))+
  geom_histogram(colour="black", fill="white", stat = "count")+
  ylab("Sub-Product Complaints Frequency") + xlab("")+
  theme(axis.text.x = element_text(angle =90, hjust = 1))
```



Distribution of the most frequent "Company" complaints

```r
most_freq_company_list <- levels(fct_infreq(df$Company))[1:30]
ggplot() + aes(fct_infreq(df[df$Company %in% most_freq_company_list,]$Company))+
  geom_histogram(colour="black", fill="white", stat = "count")+
  ylab("Company Complaints Frequency") + xlab("")+
  theme(axis.text.x = element_text(angle =90, hjust = 1))
```



Text Minig

Split data into test and train sets

```
set.seed(123)
sample = sample.split(df$`Consumer complaint narrative`, SplitRatio = .5)
train = subset(df, sample == TRUE)
test  = subset(df, sample == FALSE)
```

Word analysis

Building a corpus, which is a collection of text documents VectorSource specifies that the source is character vectors.After that, the corpus needs a couple of transformations, including changing letters to lower case, removing punctuations/numbers and removing stop words. The general English stop-word list is tailored by adding some words specific to the documents in question.

```
if(!file.exists("./user-complaints-mining/myCorpus.Rds")) {
  myCorpus <- Corpus(VectorSource(train$`Consumer complaint narrative`))
  myCorpus <- tm_map(myCorpus, removePunctuation)
  myCorpus <- tm_map(myCorpus, removeNumbers)
  myCorpus <- tm_map(myCorpus, tolower)
  myStopwords <- c(stopwords(language="en", source="smart"), "xxxx", "xxxxxxxxxxxx")
  myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
  myCorpus <- tm_map(myCorpus, stripWhitespace)
  saveRDS(myCorpus, file = "./user-complaints-mining/myCorpus.Rds")
  gc()
} else {
  myCorpus <- readRDS("./user-complaints-mining/myCorpus.Rds")
}
print(myCorpus[1:5]$content)
```

```
## [1] " identity stolen filed complaint police report affidavit contacted original
company collection agency resolve debt sending resolving issue spoken supervisors company
answer contact company debt removed credit fraudulent collection agency offered settle
paying dime company account "
## [2] " communication received debt collector court summons delivered mother laws home
received summons hand week summons stated alledgedly owed money debt collector advisement
dispute debt days demand debt collector validate debt attempted demand validation debt
online research certified letter days ago post office told today approximately certified
letter attempted delivered days mailed accepted debt collector today withoyut knowledge
information sufficient form opinion truth accuracy claim based deny generally
specifically claim debt collector"
## [3] " university pamy monthly payment afford nt anymore forbearance deferment time
left back school months school deferment left navient suggested lean cosigner afford days
default mortgage make payments"
## [4] " gave vehicle signed bill sale turned release liability dmv years report dollar
debt vehicle impoundedi contacted reporting persons told party responsible couldnt find
record owning vehiclei working company providing info dmv wasnt find attached vehicle
refused investagate "
## [5] "years ago harassed issue asked send application form signed understood charged
extra pet damages paid additional month pet covered damages wanted photos damages
property manager told walk left apartment thing scam money contacted requested
information"
```

Building a Document-Term Matrix

This operation requiers 64GB of RAM. To avoid calculation, the pre-build myDtm object will be loaded from the file system. To recalculate it needs to be removed from the file system first.

```r
if(!file.exists("./user-complaints-mining/myDtm.Rds")) {
  myDtm <- TermDocumentMatrix(myCorpus, control = list(minWordLength = 1))
  rowTotals <- apply(myDtm , 1, sum) #Find the sum of words in each Document
  myDtm <- myDtm[rowTotals > 0, ] #remove all docs without words
  saveRDS(myDtm, file = "./user-complaints-mining/myDtm.Rds")
  gc()
} else {
  myDtm <- readRDS("./user-complaints-mining/myDtm.Rds")
}
inspect(myDtm)
```

```
## <<TermDocumentMatrix (terms: 38750, documents: 40970)>>
## Non-/sparse entries: 1810376/1585777124
## Sparsity           : 100%
## Maximal term length: 125
## Weighting          : term frequency (tf)
## Sample             :
##              Docs
## Terms          11425 30866 31034 31396 31931 32354 32614 34040 35267 35598
##   account         13    62    60    60    62     4     5     2     1     4
##   company          0     0     0     1     0     2     3     2     4     0
##   credit           3     0     0     0     0     6     6   113     5     2
##   debt             0     0     0     1     0     0    13     1     0     0
##   information     10     8     8     8     8    37     9    27     9    71
##   loan             5     0     0     2     0     0     0     1     1     0
##   payment          1    27    26    27    27     6     0     0     9     0
##   report           1     0     1     0     0     0     2     9     1     3
##   told             0     0     0     1     0     0     0     3    55     0
##   xxxxxxxx         0     0     0     0     0     0     3     0    79     1
```