

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 columns:

```
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 complain Company
##   <fct>   <fct>         <fct> <fct>      <chr>          <fct>
## 1 Debt co~ Other (i.e. ph~ Discl~ Not given e~ This company refu~ The CBE~
## 2 Debt co~ Credit card     Impro~ Talked to a~ "This complaint i~ SQUARET~
## 3 Debt co~ Credit card     Takin~ Sued w/o pr~ "I am writing to ~ Selip &~
## 4 Debt co~ Other (i.e. ph~ Cont'~ Debt result~ My identity was s~ Southwe~
## 5 Student~ Federal studen~ Can't~ Can't get f~ "I was dropped fr~ AES/PHE~
## 6 Debt co~ Credit card     Discl~ Not given e~ The first communi~ Blatt, ~
## 7 Debt co~ Other (i.e. ph~ Commu~ Frequent or~ "My complaint is ~ AR Reso~
## 8 Debt co~ I do not know   False~ Attempted t~ In a clearance in~ SANTAND~
## 9 Student~ Non-federal st~ Can't~ Can't tempo~ XXXX University, ~ Navient~
## 10 Student~ Non-federal st~ Deal~ Received ba~ I had attended XX~ CITIZEN~
## # ... with 81,930 more rows
```

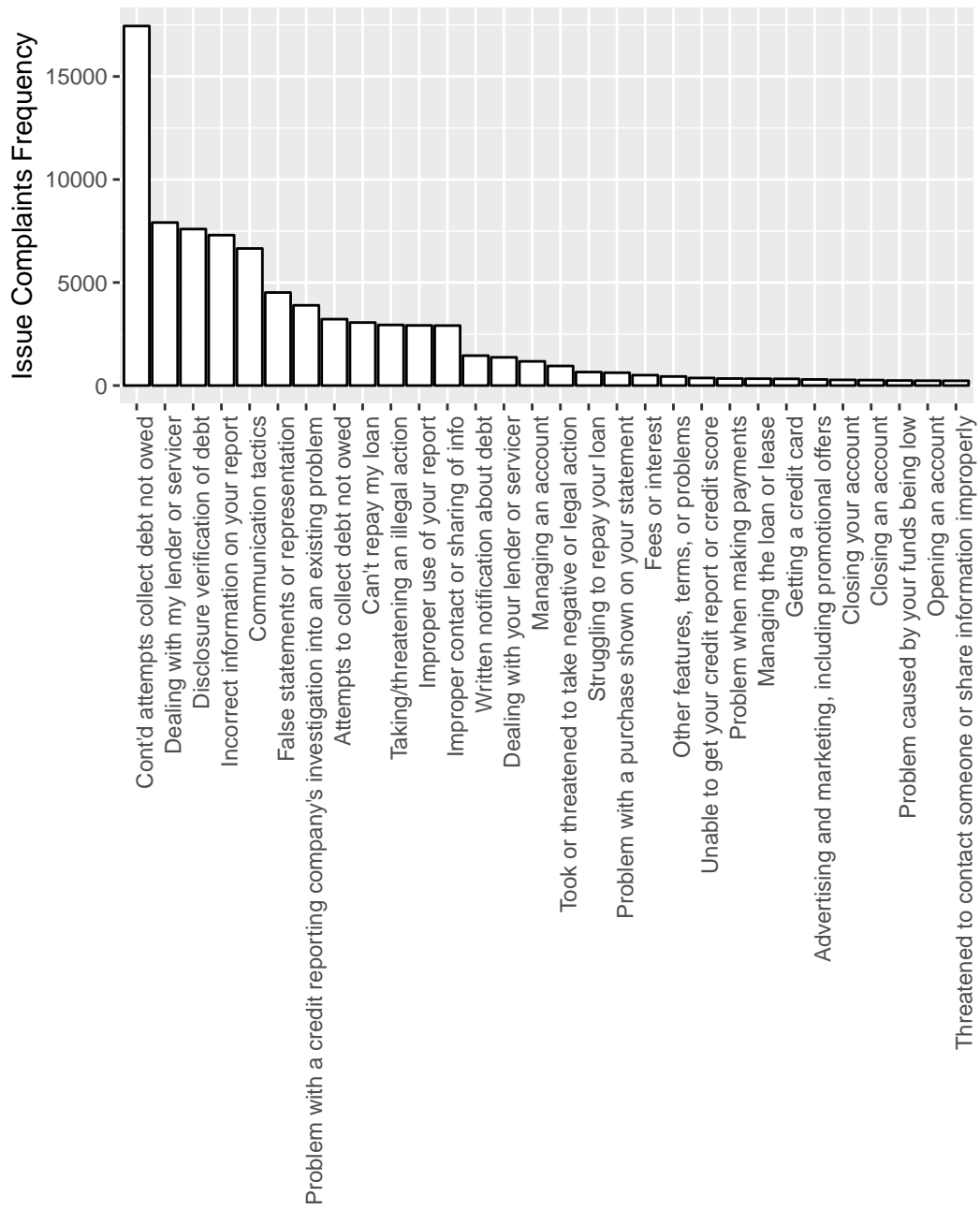
Converting all but narrative columns to factors:

```
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)
```

Feature engineering

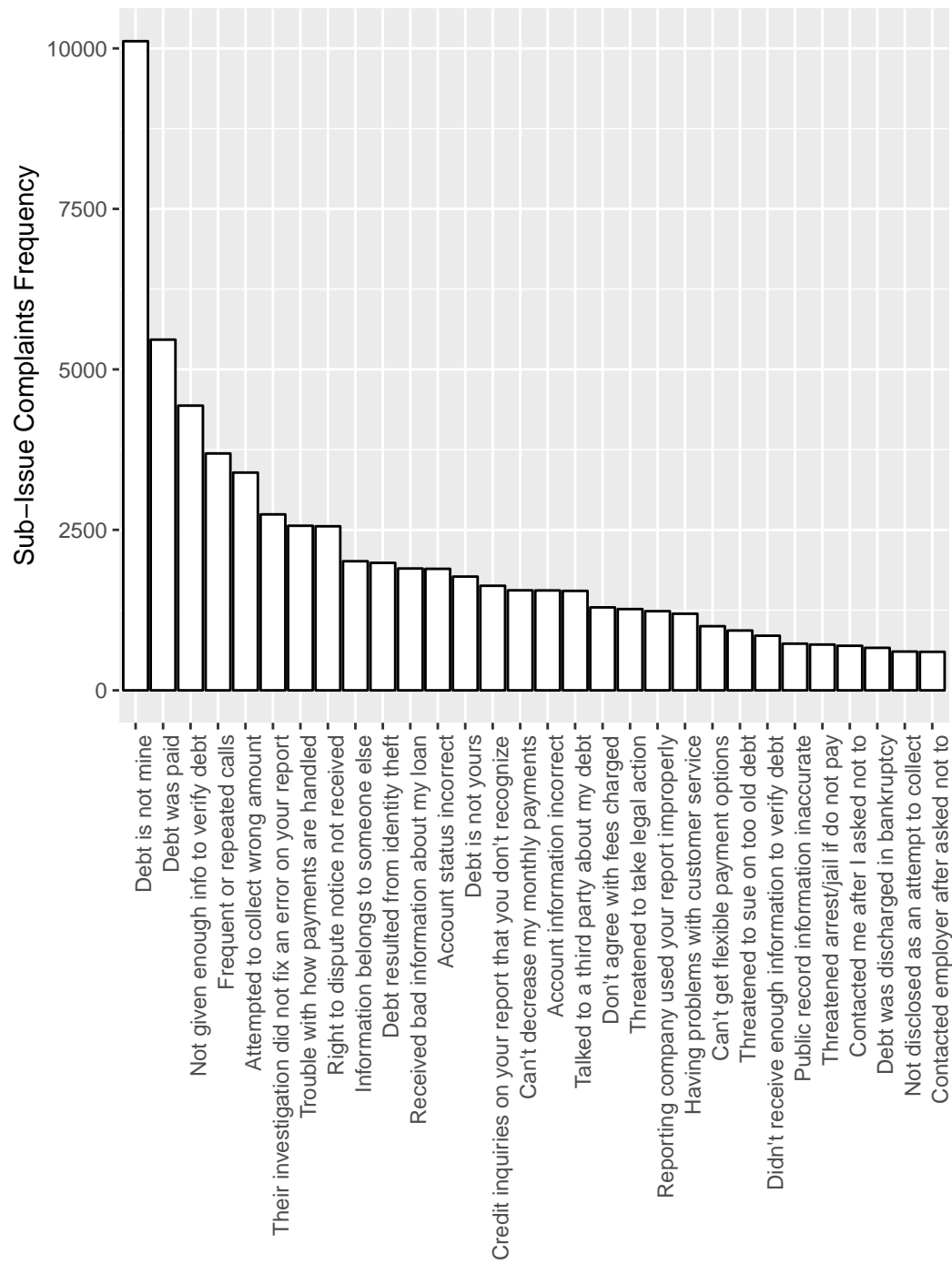
Distribution of the most frequent “Issue” complaints

```
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))
```



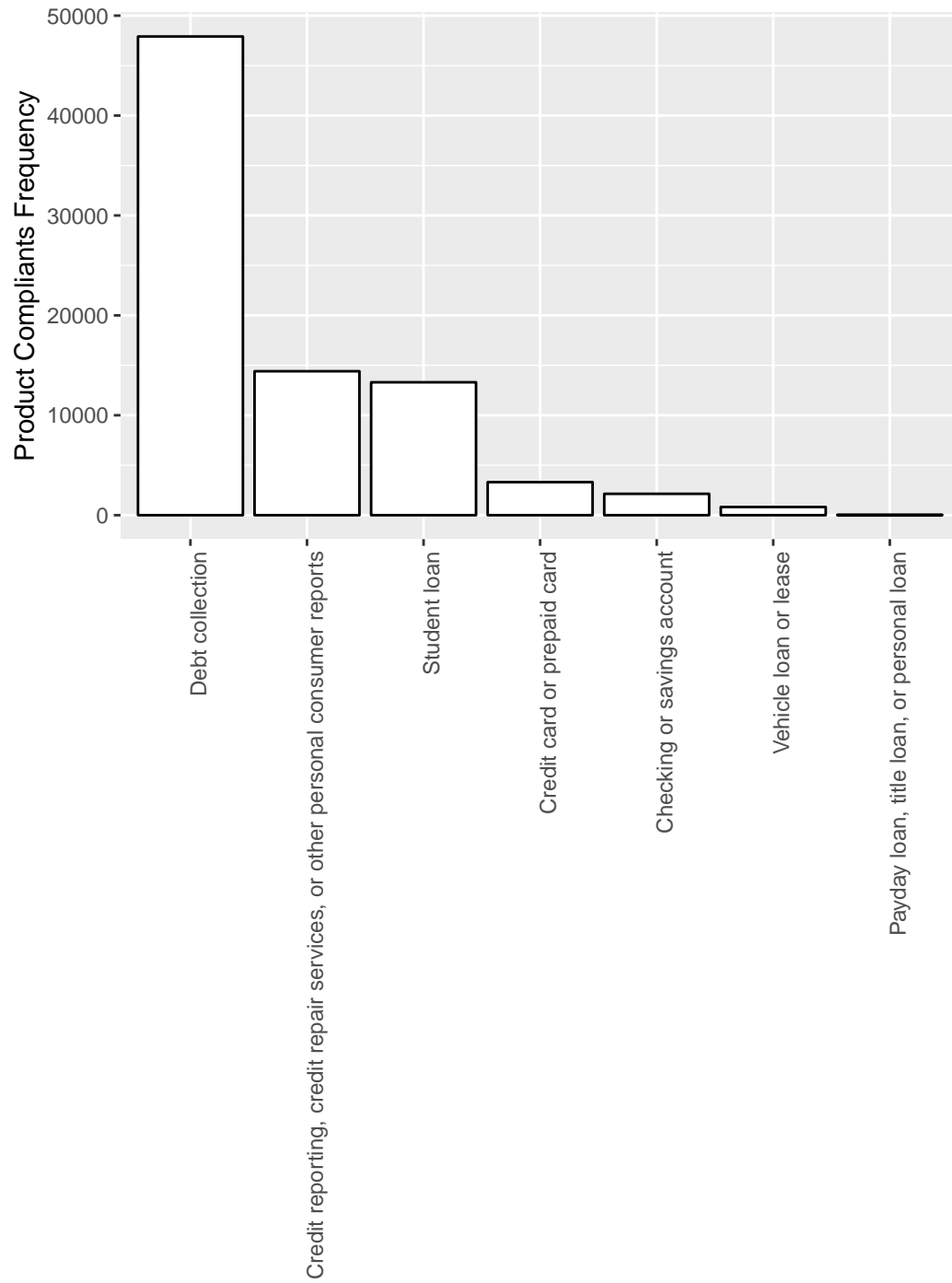
Distribution of the most frequent “Sub-issue” complaints

```
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))
```



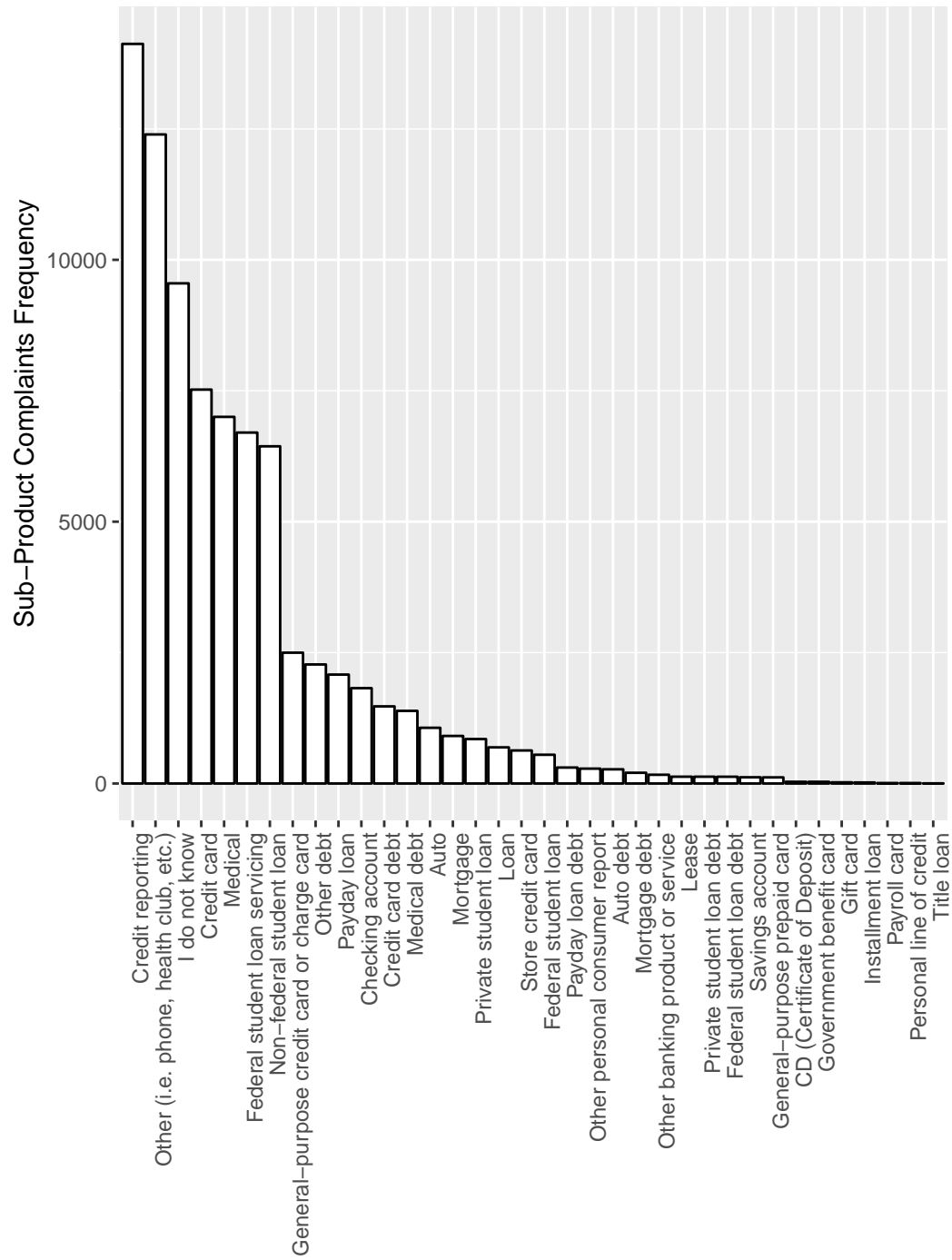
Distribution of the most frequent “Product” complaints

```
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))
```



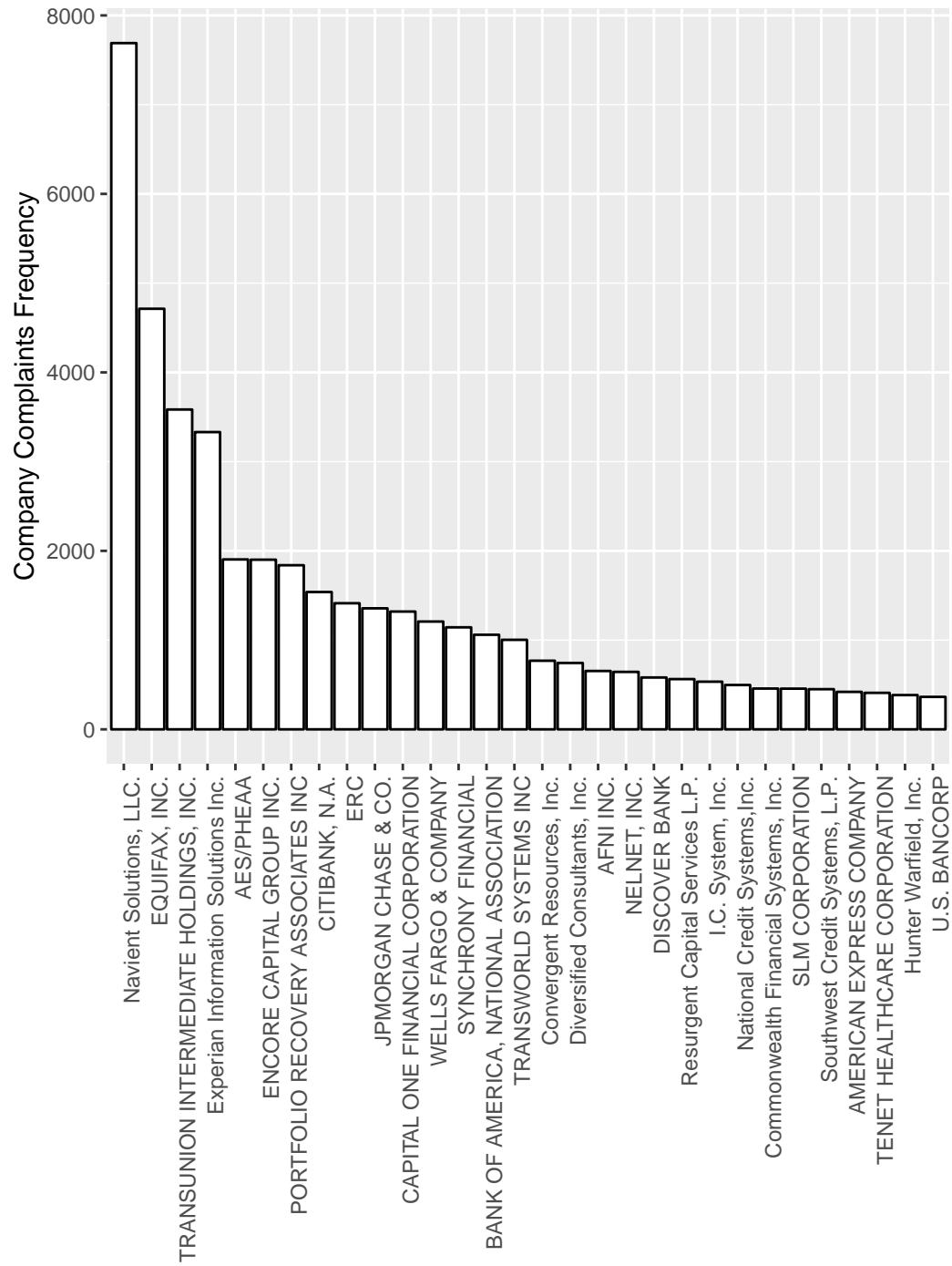
Distribution of the most frequent “Sub-product” complaints

```
most_freq_subproduct_list <- levels(fct_infreq(df$`Sub-product`))
ggplot() + aes(fct_infreq(df[df$`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

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

```
## [1] " company refuses provide verification validation debt fdcpa debt mine"
## [2] " writing request assistance deceptive practices collection lawfirm appears
tactics violating consumer protection law debt collection practices depriving consumers
rights dispute received notice company day contacted offices instructed memo dated
instruct contact plaintiff attorney court instructions provided contacted plaintiff
attorney phone faxed letter disputing debt letter company responded letter dated sending
bill due date requested bill showing balance back made payment back disputing amount owed
disputing charges wrote back company faxed dispute letter continue dispute amount owed
company response furnished information disputing owe disputing balance inaccurate needed
proof charges activity account paragraph letter disputing amount send letter letter
company disputing balance requesting documents company responded letter communication
received letter copy default judgment filed court clerk office indicating failed respond
judgment facts judgment firm served original judgment attaching dispute letters showing
responded instructed office occasions plaintiff failed respond dispute furnish
information providedand unable obtain proof original debt credible legal procedure settle
debts utilized unfaithful dirty tactics violated rights court committed perjury law
filing false documents court defaulted judgment failed respond fact responded failed
furnish proof court house clerk office told company notify offices contact company told
court clerk office respond summons clerks office granted default judgement based false
information respond summons filed false affirmation clerk office"
## [3] " 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"

[4] " attended forced loan attend school loan interest rate make payments attended school enrolled told classmates received reduction balance curious criteria receive reduction"

[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"

Steming

```
dictCorpus <- myCorpus
myCorpus <- tm_map(myCorpus, stemDocument)
print(myCorpus[1:5]$content)
```

[1] "compani refus provid verif valid debt fdcpa debt mine"

[2] "write request assist decept practic collect lawfirm appear tactic violat consum protect law debt collect practic depriv consum right disput receiv notic compani day contact offic instruct memo date instruct contact plaintiff attorney court instruct provid contact plaintiff attorney phone fax letter disput debt letter compani respond letter date send bill due date request bill show balanc back made payment back disput amount owe disput charg wrote back compani fax disput letter continu disput amount owe compani respons furnish inform disput owe disput balanc inaccur need proof charg activ account paragraph letter disput amount send letter letter compani disput balanc request document compani respond letter communic receiv letter copi default judgment file court clerk offic indic fail respond judgment fact judgment firm serv origin judgment attach disput letter show respond instruct offic occas plaintiff fail respond disput furnish inform providedand unabl obtain proof origin debt credibl legal procedur settl debt util unfaith dirti tactic violat right court commit perjuri law file fals document court default judgment fail respond fact respond fail furnish proof court hous clerk offic told compani notifi offic contact compani told court clerk offic respond summon clerk offic grant default judgement base fals inform respond summon file fals affirm clerk offic"

[3] "communic receiv debt collector court summon deliv mother law home receiv summon hand week summon state alledg owe money debt collector advis disput debt day demand debt collector valid debt attempt demand valid debt onlin research certifi letter day ago post offic told today approxim certifi letter attempt deliv day mail accept debt collector today withoyut knowledg inform suffici form opinion truth accuraci claim base deni general specif claim debt collector"

[4] "attend forc loan attend school loan interest rate make payment attend school enrol told classmat receiv reduct balanc curious criteria receiv reduct"

[5] "year ago harass issu ask send applic form sign understood charg extra pet damag paid addit month pet cover damag want photo damag properti manag told walk left apart thing scam money contact request inform"

Building a Document-Term Matrix

This operation requires 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.

```
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: 27440, documents: 40970)>>  
## Non-/sparse entries: 1585496/1122631304  
## Sparsity          : 100%  
## Maximal term length: 124  
## Weighting          : term frequency (tf)  
## Sample            :  
##  
##      Docs  
## Terms      11425 11904 31399 31930 33039 34137 35502 35599 36932 9698  
## account    26    21    60    62    18    64     5     8     0     6  
## call        3    10     2     0    18     0     3     0     2     7  
## collect     0     0     5     0     0     4     0     6     0    21  
## credit      3     1     0     0     0     2    13     2     2     0  
## debt        0     0     1     1     0     1     0     0     3    68  
## inform     11     5     8     8     6     8    10    73     5     9  
## loan        5     0     2     0    11     0     2     0     3     0  
## payment     2    12    27    27    24    27     7     0     0     1  
## receiv      4     2     0     0    25     0     1     0     8    10  
## report      1     0    25    27     1    28    26     8     0     0
```

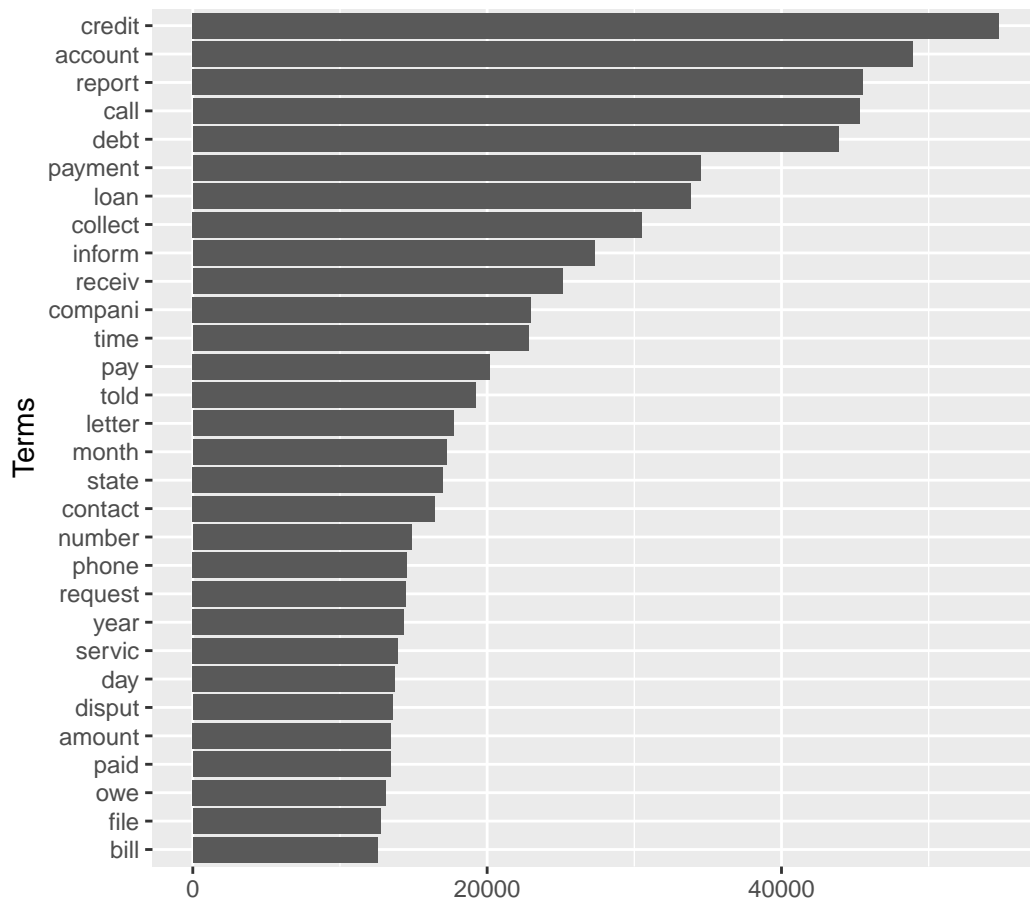


Figure 1: 30 Most Frequent Terms

Frequent Terms and Association

```
freq.terms <- findFreqTerms(myDtm, lowfreq=5)
term.freq <- rowSums(as.matrix(myDtm))
term.freq <- subset(term.freq, term.freq >= 5)

dfTerms <- data.frame(term = names(term.freq), freq = term.freq)
ggplot(dfTerms[order(-dfTerms$freq),][1:30,], aes(x = reorder(term, freq), y = freq)) +
  geom_bar(stat = "identity") + xlab("Terms") + ylab("") + coord_flip()
```

Which words are associated with term “loan”?

```
findAssocs(myDtm, c('loan'), 0.35)
```

```
## $loan
## student payment privat navient interest consolid
##      0.60    0.40    0.39    0.39    0.38    0.36
```

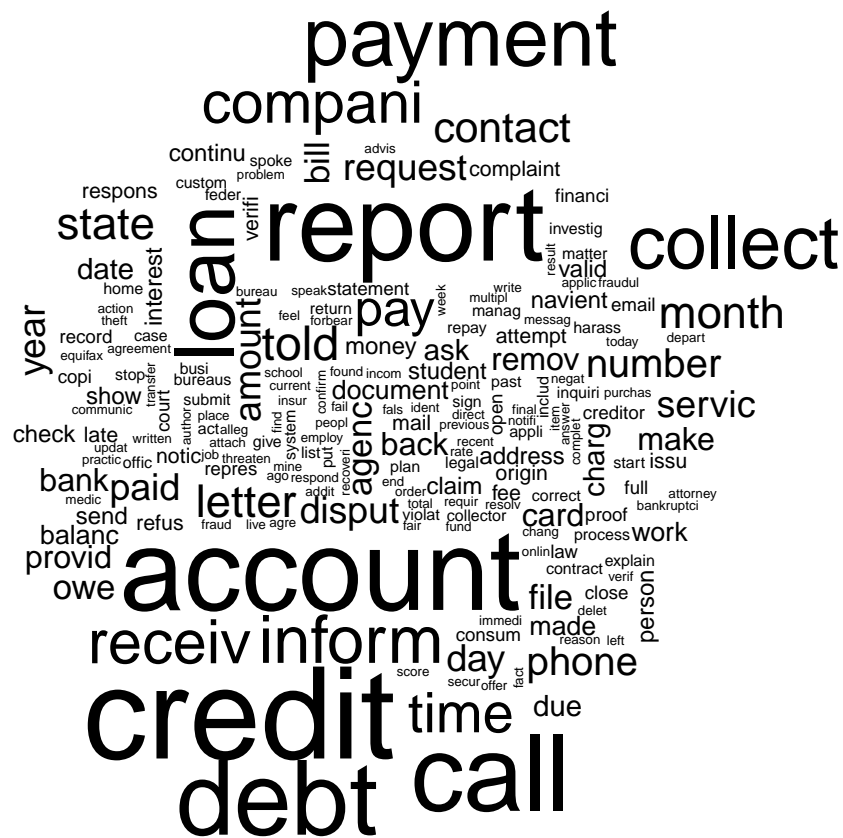


Figure 2: Words Cloud of Complaints

Building word cloud:

```
library(wordcloud)
m <- as.matrix(myDtm)
v <- sort(rowSums(m), decreasing=TRUE)
myNames <- names(v)
d <- data.frame(word=myNames, freq=v)
wordcloud(d$word, d$freq, min.freq=20, scale=c(4,.2), max.words = 200)
```

Building LDA model for the ‘train’ set of complaints

```
library(topicmodels)
dtm <- as.DocumentTermMatrix(myDtm)
ui = unique(dtm$i) #array of unique row ids in 'train' set
if(!file.exists("./user-complaints-mining/lda10.Rds")) {
  df.new = df[ui,]
  lda <- LDA(dtm.new, k = 7) # identify topics
  saveRDS(lda, file = "./user-complaints-mining/lda10.Rds")
  gc()
} else {
  lda <- readRDS("./user-complaints-mining/lda10.Rds")
}
(term <- terms(lda, 10)) # first terms of every topic
```

```
##      Topic 1  Topic 2  Topic 3      Topic 4  Topic 5  Topic 6
## [1,] "bank"    "payment" "inform"    "payment" "report"  "card"
## [2,] "account" "account" "address"  "month"   "credit"  "credit"
## [3,] "check"   "balanc"  "file"     "interest" "account" "charg"
## [4,] "money"   "paid"    "letter"   "pay"     "remov"   "purchas"
## [5,] "call"    "late"    "number"   "loan"    "disput"  "servic"
## [6,] "told"    "amount"  "receiv"   "make"    "inform"  "car"
## [7,] "close"   "due"     "secur"    "plan"    "compani" "month"
## [8,] "fund"    "date"    "complaint" "time"    "bureaus" "chase"
## [9,] "well"    "charg"   "bankruptci" "year"    "inquiri" "time"
## [10,] "back"   "month"   "state"    "rate"    "file"    "move"

##      Topic 7  Topic 8  Topic 9  Topic 10
## [1,] "call"    "debt"    "loan"    "debt"
## [2,] "phone"   "collect" "student"  "valid"
## [3,] "told"    "owe"     "navient"  "request"
## [4,] "number"  "bill"    "school"   "provid"
## [5,] "ask"     "compani" "year"     "law"
## [6,] "time"    "agenc"   "privat"   "violat"
## [7,] "contact" "receiv"  "time"     "collect"
## [8,] "inform"  "letter"  "pay"      "document"
## [9,] "compani" "paid"    "feder"    "court"
## [10,] "receiv" "pay"     "consolid" "act"
```

Show complaints correlation to topics in 'train' set

```
topic <- topics(lda, 5, threshold=.15)
for (i in c(1,55,500,333)) {
  print (paste(" "))
  print (paste("Complaint# ", i))
  print(paste("Topics found: ", topic[i]))
  print (train$`Consumer complaint narrative`[ui[i]])
}

## [1] " "
## [1] "Complaint# 1"
## [1] "Topics found: c(8, 10)"
## [1] "This company refuses to provide me verification and validation of debt per my
right under the FDCPA. I do not believe this debt is mine."
## [1] " "
## [1] "Complaint# 55"
## [1] "Topics found: 7"
## [1] "I have received calls from Midland Credit company since XX/XX/2015. They do not
ask for me by name but say they are trying to collect a debt. The same number of this
company has called me every day for the last several weeks. The number they used before
in XX/XX/XXXX was XXXX. They are now calling from XXXX. I spoke to a manager XXXX XXXX
and requested that they remove my number from their call list. This is the third time I
have ask them to quit calling my number. They have never complied. I HAVE NEVER GIVEN
THEM MY NAME AS THEY DONT ADDRESS ME BY NAME! All these people have is my phone number to
continue to harass me. They all sound like they are from XXXX."
## [1] " "
## [1] "Complaint# 500"
## [1] "Topics found: c(4, 9)"
## [1] "I am seeking information about my Genesis loans through XXXX in XXXX, PA. I have
heard about all of the loan forgiveness because of the XXXX XXXX settlement. I was one of
the many students who were talked into the lending scheme brought on by these schools and
Genesis lending company. I was charged horrible interest payments that I could n't
afford, while still in school. As well as, pulled out of class because I could not afford
payments. Horrible interest rates still haunt me, along with unreasonable monthly
payments. I am already on deferment, on which they still require me to make minimum
payments. I have a loan balance of {$21000.00} and have no idea how I will be able to
keep up with this. I feel as though I was a victim as well as other students who have
already had their loans forgiven. The way the company went about handing these loans out
to young students who know nothing about them, was wrong. My loan servicer XXXX told me
that Genesis has already handled the persons eligible for forgiveness and passed me along
to their company. So my question is, why was I considered not eligible? I attended school
in XX/XX/XXXX-XX/XX/XXXX and suffered just like everyone else who attended these scheming
schools. Please help!"
## [1] " "
## [1] "Complaint# 333"
## [1] "Topics found: c(1, 10)"
## [1] "Talked to someone not related to me about debt and left information about debt
with my name on front door where everyone could see it and read it. They are also listing
someone else in the document I do n't know, does n't exist and/Or has nothing to do with
this debt. The documents were so cigarette smoke infiltrated they made me sick."
```