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:

```
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
df
```

```
## # A tibble: 81,940 x 6
##    Product  `Sub-product` Issue `Sub-issue` `Consumer complaint n~ Company
##    <chr>    <chr>         <chr> <chr>       <chr>                 <chr>
##  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
```

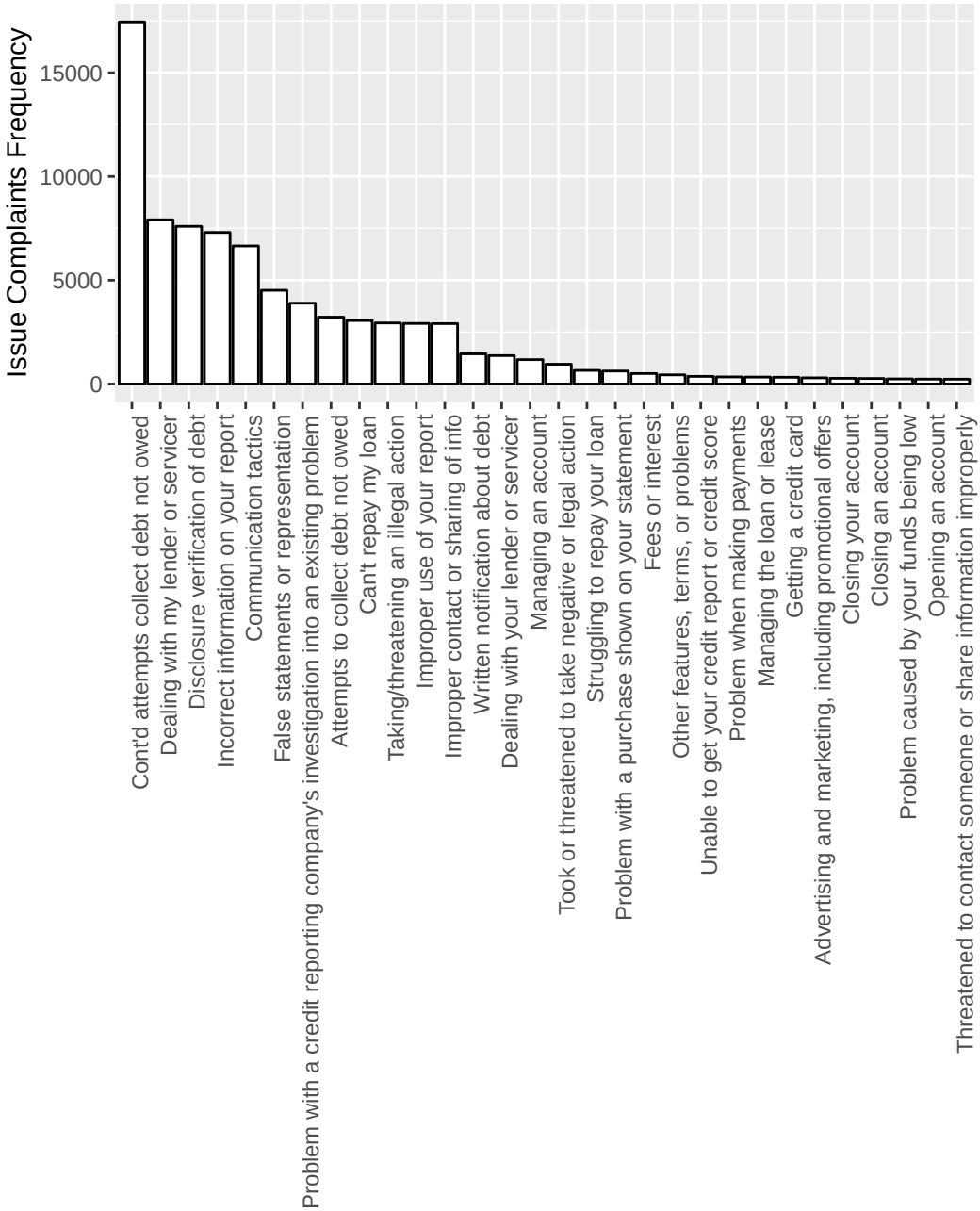
Converting all but narrative columns to factors:

```
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

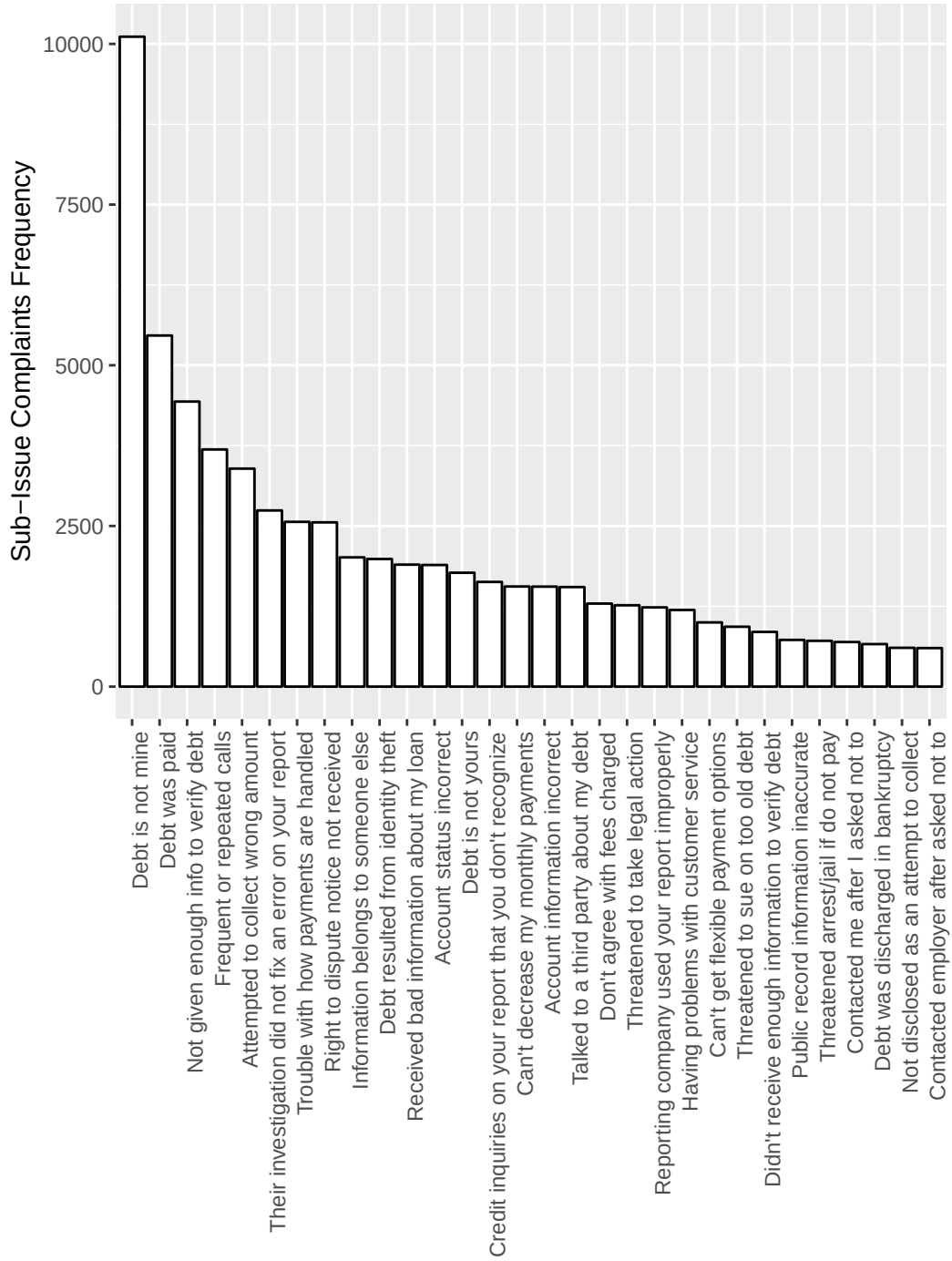
Distribution of the most frequent "Issue" complaints

```
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))
```



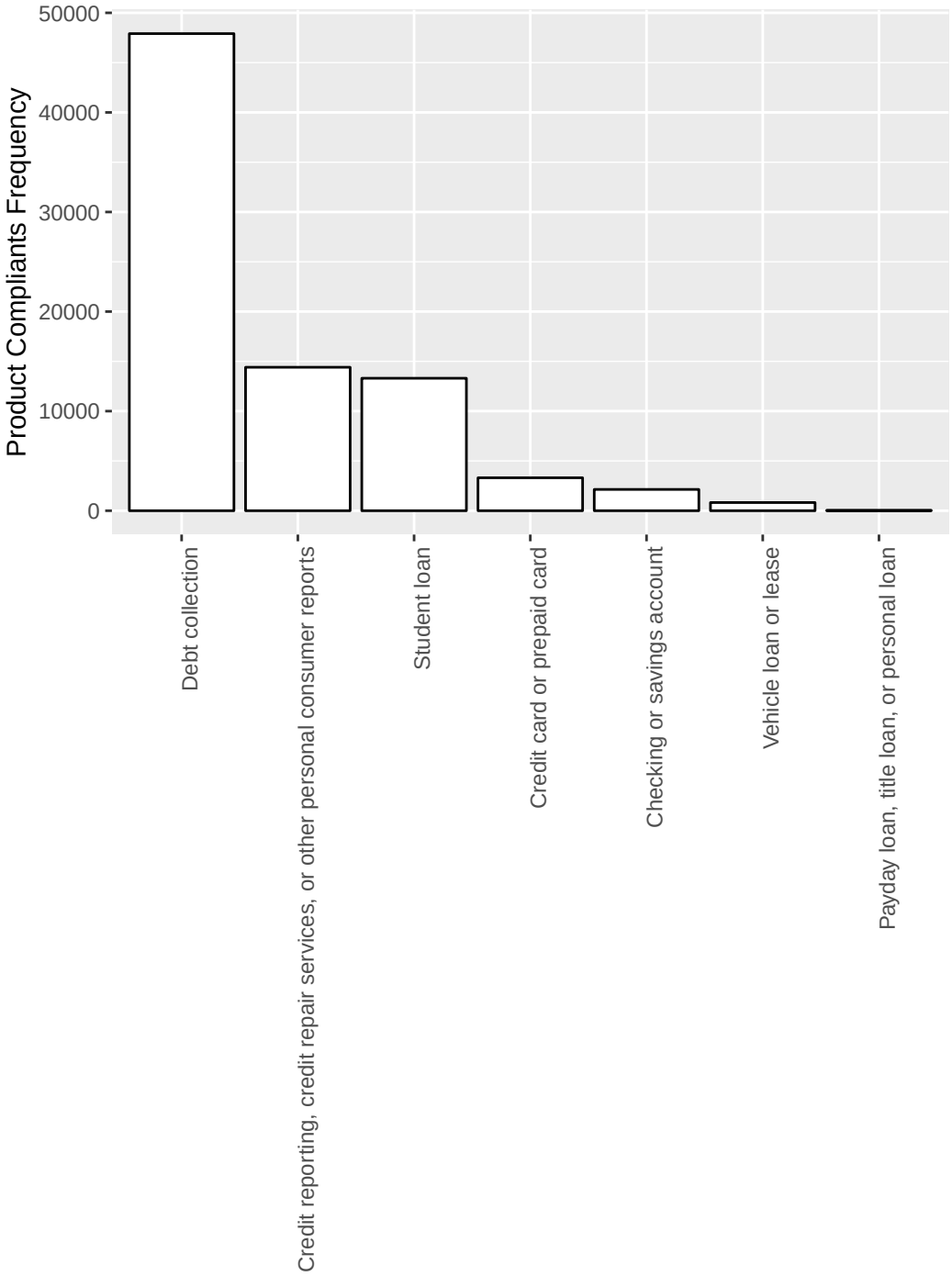
Distribution of the most frequent "Sub-issue" complaints

```
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))
```



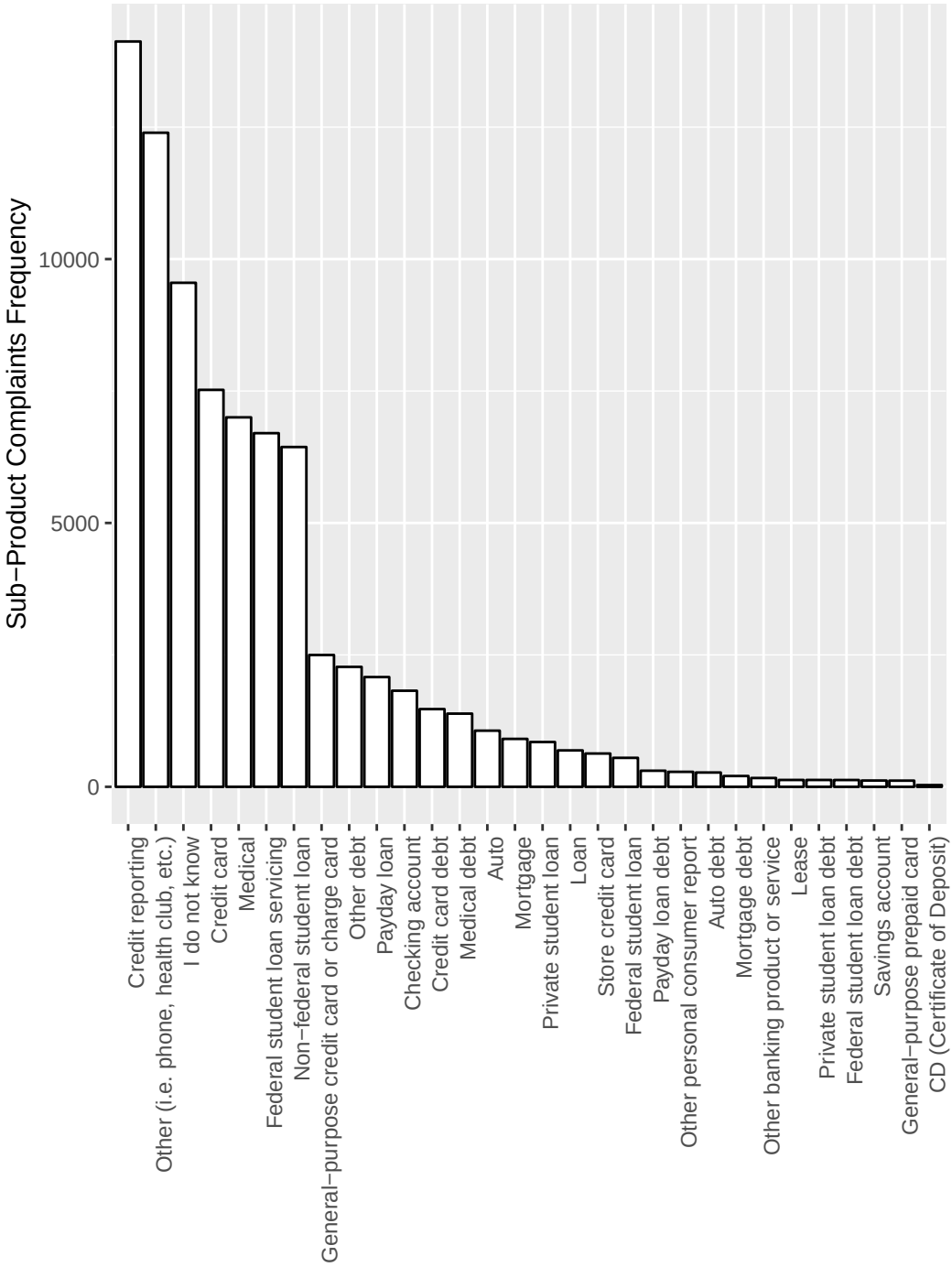
Distribution of the most frequent "Product" complaints

```
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$`Sub-product`))[1:30]
ggplot() + aes(fct_infreq(df[df$`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

```
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))
```