Practicum

Jiaqing Ge

2/19/2020

1. Packages needed

```r
library(dplyr)
library(ggplot2)
library(knitr)
library(kableExtra)
library(wordcloud)
library(plotly)
library(ggthemes)
library(gapminder)
library(scales)
library(usmap)
library(tidycensus)
library(lubridate)
library(TTR)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(tidytext)
library(stringr)
```

2. Import data

```r
complaints <- read.csv('complaints.csv',stringsAsFactors = FALSE)
saveRDS(complaints, file = "complaints_raw.rds") # This line change complaints.csv to dataset that works
complaints_raw <- readRDS("complaints_raw.rds")
```

3. Randomly draw 5000 rows of data to do the analysis.

```r
subset <- complaints_raw[sample(nrow(complaints_raw), size = 5000, replace = FALSE),]
saveRDS(subset, file = "complaints_sub.rds")
```

4. take a look at it.

```r
complaints_test <- readRDS('complaints_sub.rds')
head(complaints_test,100) %>%
  View()
glimpse(complaints_test)
```

```
## Observations: 5,000
## Variables: 18
## $ Date.received                <chr> "2019-08-23", "2019-10-17", "2018-03-2...
## $ Product                      <chr> "Credit reporting, credit repair servi...
## $ Sub.product                  <chr> "Credit reporting", "Credit reporting"...
## $ Issue                        <chr> "Incorrect information on your report"...
## $ Sub.issue                    <chr> "Information belongs to someone else",...
## $ Consumer.complaint.narrative <chr> "", "Equifax is reporting incorrectly ...
## $ Company.public.response      <chr> "Company has responded to the consumer...
## $ Company                      <chr> "TRANSUNION INTERMEDIATE HOLDINGS, INC...
## $ State                        <chr> "GA", "CA", "FL", "CA", "IL", "NY", "T...
## $ ZIP.code                     <chr> "30349", "952XX", "347XX", "95835", "6...
## $ Tags                         <chr> "", "", "", "", "", "", "", "", "", ""...
## $ Consumer.consent.provided.   <chr> "Consent not provided", "Consent provi...
## $ Submitted.via                <chr> "Web", "Web", "Web", "Web", "Web", "We...
## $ Date.sent.to.company         <chr> "2019-08-23", "2019-10-17", "2018-03-2...
## $ Company.response.to.consumer <chr> "Closed with explanation", "Closed wit...
## $ Timely.response.             <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Ye...
## $ Consumer.disputed.           <chr> "N/A", "N/A", "N/A", "No", "No", "No",...
## $ Complaint.ID                 <int> 3351802, 3409076, 2851469, 2373071, 11...
```

5. Change date received to year-month-day

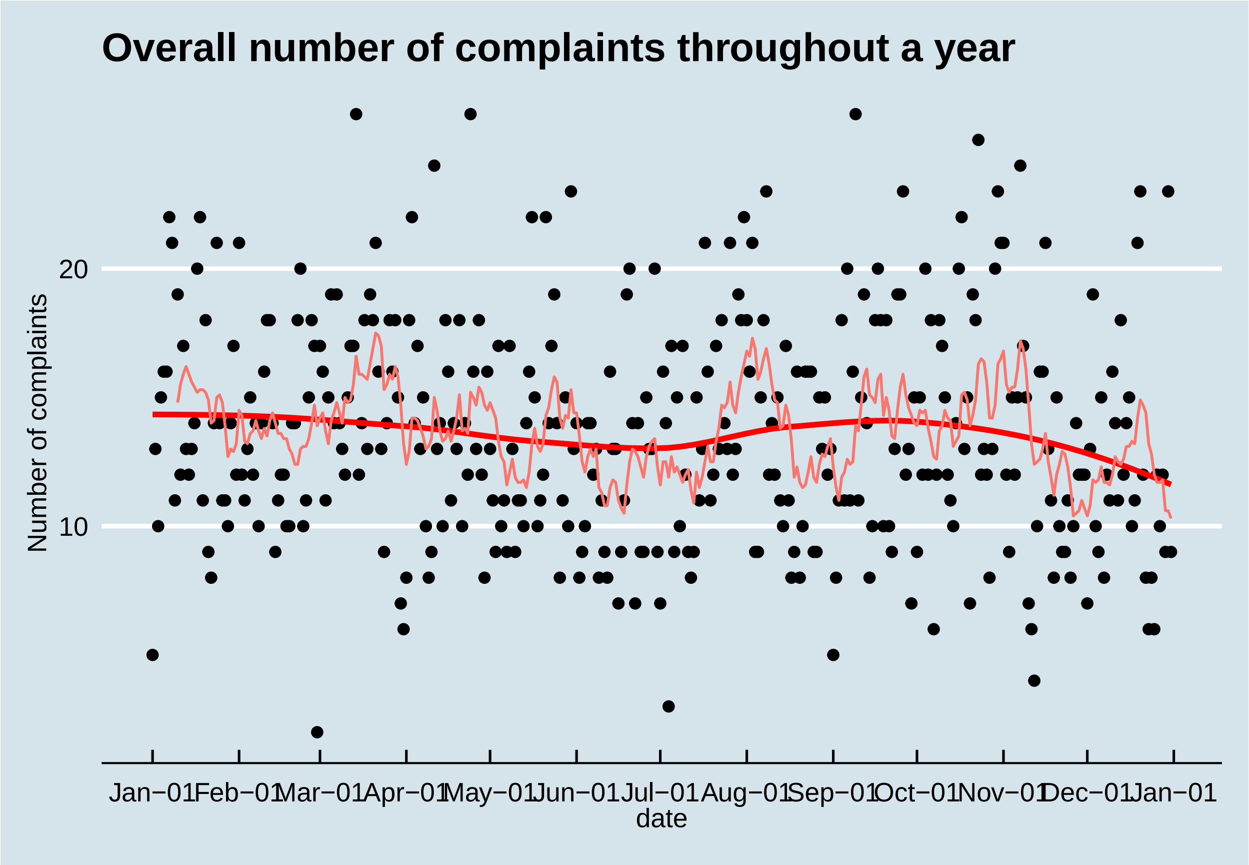
```r
complaints_test <- complaints_test %>%
  mutate(year = as.integer(substr(Date.received, start = 1, stop = 4))) %>%
  mutate(month = as.integer(substr(Date.received, start = 6, stop = 7))) %>%
  mutate(day = as.integer(substr(Date.received, start =9 , stop = 10)))
complaints_test$Date.received <- ymd(complaints_test$Date.received)
str(complaints_test)
```

```
## 'data.frame':    5000 obs. of  21 variables:
##  $ Date.received               : Date, format: "2019-08-23" "2019-10-17" ...
##  $ Product                     : chr  "Credit reporting, credit repair services, or other personal co
##  $ Sub.product                 : chr  "Credit reporting" "Credit reporting" "Credit reporting" "Conve
##  $ Issue                       : chr  "Incorrect information on your report" "Problem with a credit r
##  $ Sub.issue                   : chr  "Information belongs to someone else" "Their investigation did
##  $ Consumer.complaint.narrative: chr  "" "Equifax is reporting incorrectly collections in the amount
##  $ Company.public.response     : chr  "Company has responded to the consumer and the CFPB and chooses
##  $ Company                     : chr  "TRANSUNION INTERMEDIATE HOLDINGS, INC." "EQUIFAX, INC." "Exper
##  $ State                       : chr  "GA" "CA" "FL" "CA" ...
##  $ ZIP.code                    : chr  "30349" "952XX" "347XX" "95835" ...
##  $ Tags                        : chr  "" "" "" "" ...
##  $ Consumer.consent.provided.  : chr  "Consent not provided" "Consent provided" "Consent provided" "C
##  $ Submitted.via               : chr  "Web" "Web" "Web" "Web" ...
##  $ Date.sent.to.company        : chr  "2019-08-23" "2019-10-17" "2018-03-22" "2017-03-07" ...
##  $ Company.response.to.consumer: chr  "Closed with explanation" "Closed with explanation" "Closed wi
##  $ Timely.response.            : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Consumer.disputed.          : chr  "N/A" "N/A" "N/A" "No" ...
##  $ Complaint.ID                : int  3351802 3409076 2851469 2373071 1118688 1874286 3371286 313324
##  $ year                        : int  2019 2019 2018 2017 2014 2016 2019 2019 2016 2015 ...
##  $ month                       : int  8 10 3 3 11 4 9 1 2 7 ...
##  $ day                         : int  23 17 22 7 16 11 11 25 22 23 ...
```

6. Ploting of the number of complaints throughout the year

Ploting overall number of complaints throughout the year

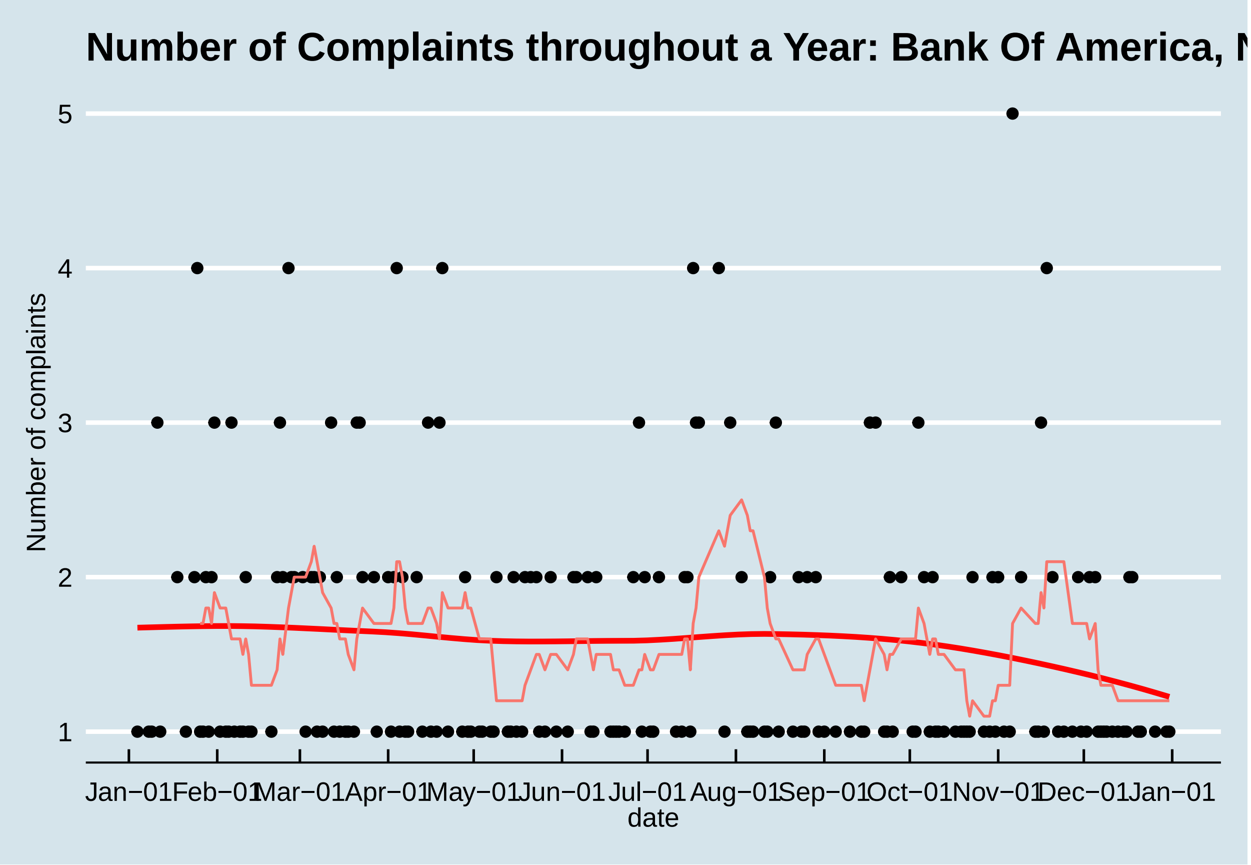
```
p <- complaints_test %>%
  mutate(date = as.POSIXct(paste(month , day , sep = "." )  , format = "%m.%d" )) %>%
  group_by(date) %>%
  summarise(number_of_complaints = n()) %>%
  ggplot(aes(x = date, y = number_of_complaints))+ geom_point()+ylab("Number of complaints")+
  theme_economist()+
  scale_x_datetime(labels=  date_format("%b-%d"),date_breaks = '1 month')+
  geom_smooth(lwd=1, se=FALSE,color = 'red')+
  geom_line(aes(x=date, y=SMA(number_of_complaints,10), color = 'red'))+
  theme(legend.position="none")+
  ggtitle("Overall number of complaints throughout a year")
p
```



In this graph, I plot a graph showing the number of complaints through out a year to see whether there are more complaints in certain months or day. We can see from the graph that there are a few outliars that are very low appears at the beginning of the month. On Jan-01, Mar-01, Jul-01 and Sep-01, the number of complaints drops significantly.

Ploting number of complaints of certain complany through out the year through out the year

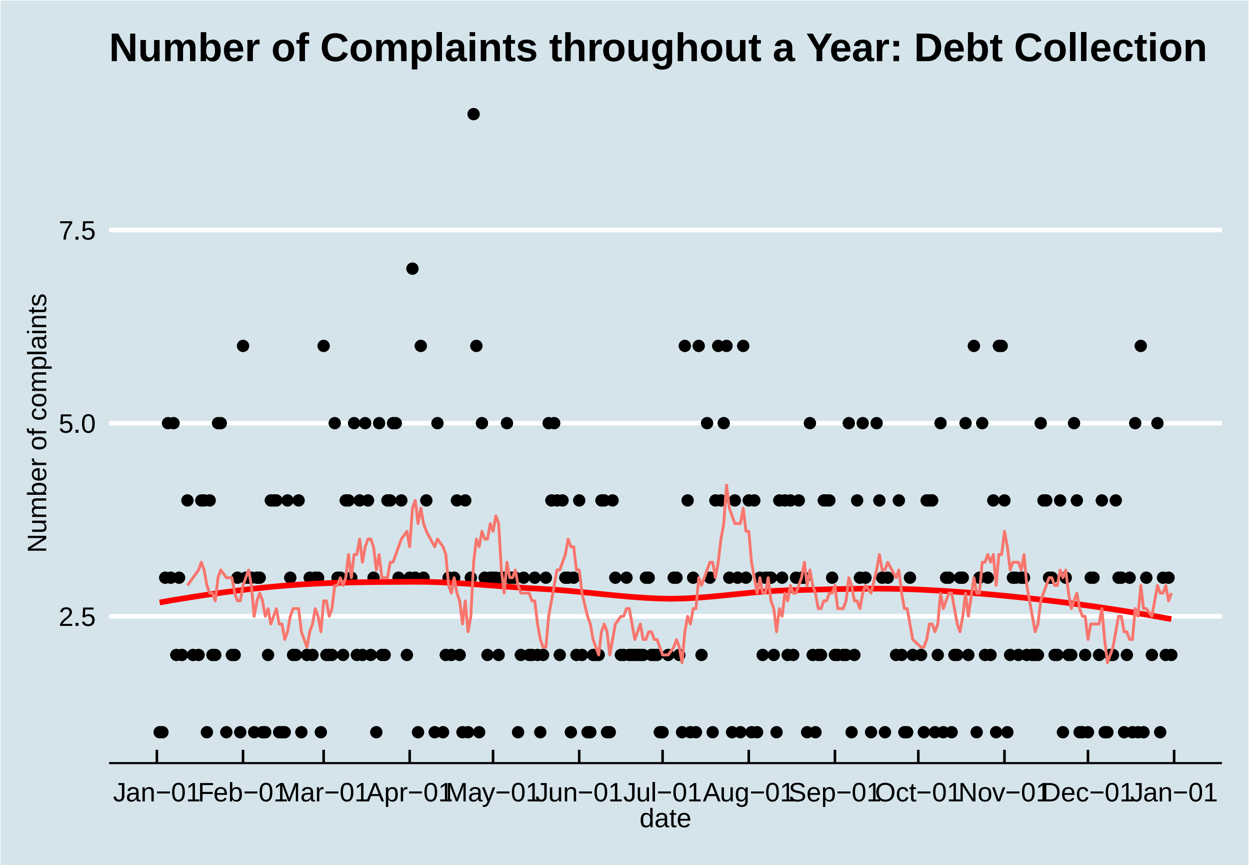
```
simpleCap <- function(x) {
  s <- strsplit(x, " ")[[1]]
  paste(toupper(substring(s, 1,1)), substring(s, 2),
      sep="", collapse=" ")
} # This line change the format of company names
complaints_pattern_company <- function(x){
p_ <-  complaints_test %>%
  filter(Company == x)  %>%
  mutate(date = as.POSIXct(paste(month , day , sep = "." )  , format = "%m.%d" )) %>%
  group_by(date) %>%
  summarise(number_of_complaints = n()) %>%
  ggplot(aes(x = date, y = number_of_complaints))+ geom_point()+ylab("Number of complaints")+
  theme_economist()+
  scale_x_datetime(labels=  date_format("%b-%d"),date_breaks = '1 month')+
  geom_smooth(lwd=1, se=FALSE,color = 'red')+
  geom_line(aes(x=date, y=SMA(number_of_complaints,10), color = 'red'))+
  theme(legend.position="none")+
  ggtitle(paste("Number of Complaints throughout a Year:",simpleCap(tolower(x)), sep = ' '))
p_}
complaints_pattern_company('BANK OF AMERICA, NATIONAL ASSOCIATION')
```



Ploting number of complaints of certain product through out the year through out the year

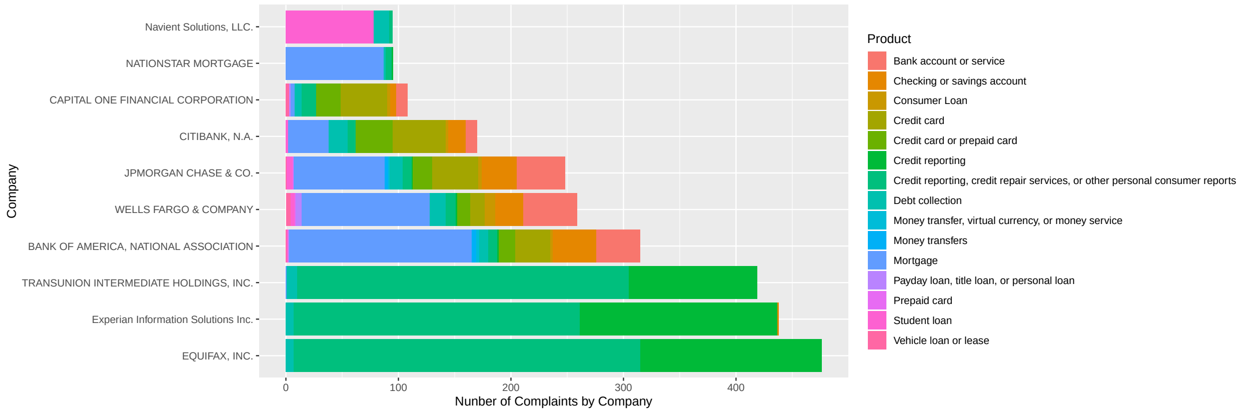
```
complaints_pattern_product <- function(x){
p_ <-  complaints_test %>%
  filter(Product == x)  %>%
  mutate(date = as.POSIXct(paste(month , day , sep = "." )  , format = "%m.%d" )) %>%
  group_by(date) %>%
  summarise(number_of_complaints = n()) %>%
  ggplot(aes(x = date, y = number_of_complaints))+ geom_point()+ylab("Number of complaints")+
  theme_economist()+
  scale_x_datetime(labels=  date_format("%b-%d"),date_breaks = '1 month')+
  geom_smooth(lwd=1, se=FALSE,color = 'red')+
  geom_line(aes(x=date, y=SMA(number_of_complaints,10), color = 'red'))+
  theme(legend.position="none")+
  ggtitle(paste("Number of Complaints throughout a Year:",simpleCap(tolower(x)), sep = ' '))
p_}
complaints_pattern_product('Debt collection')
```

```
## Warning: Removed 9 rows containing missing values (geom_path).
```



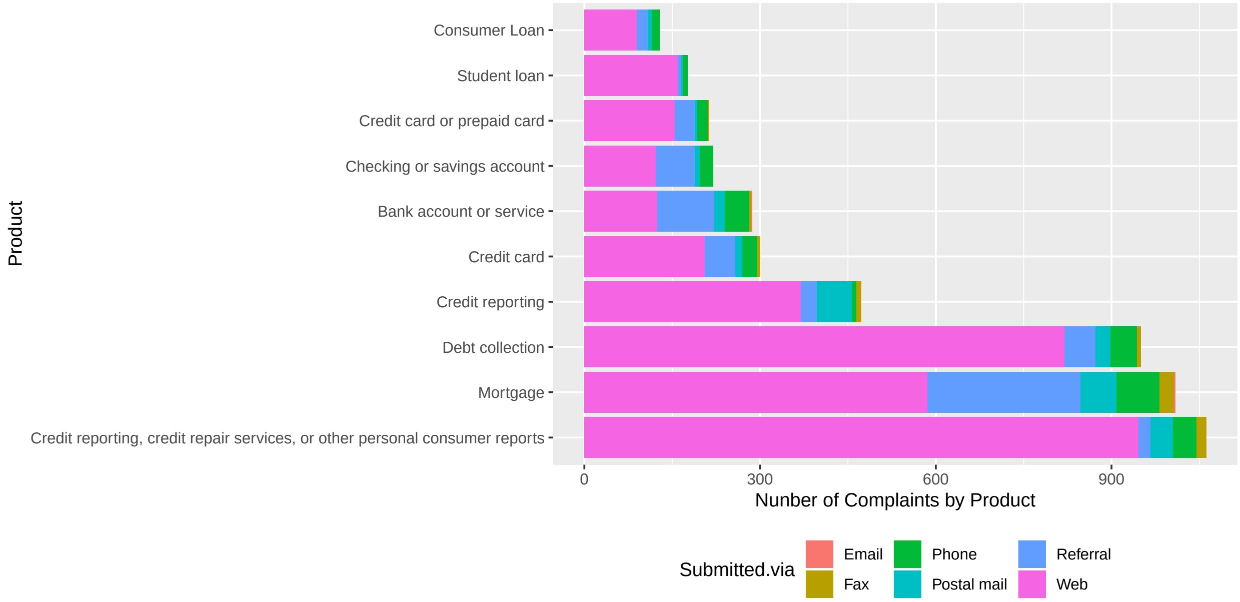
Ploting companies with highest number of complaints

```
Top_10_comanies <- complaints_test %>%
  group_by(Company) %>%
  summarise(number_of_complaints = n())%>%
  arrange(desc(number_of_complaints))%>%
  head(10)%>%
  select(Company) # This line select the companies with most complaints
p_1 <- complaints_test %>%
  filter(Company %in% Top_10_comanies$Company) %>%
  group_by(Company,Product)%>%
  summarise(number_of_complaints = n())%>%
  ungroup()%>%
  mutate(Company = factor(Company, levels=Top_10_comanies$Company))%>%
  ggplot(aes(fill=Product, y=number_of_complaints, x=Company)) +
  geom_bar(position="stack", stat="identity")+ylab('Nunber of Complaints by Company')+coord_flip()+theme
p_1
```



Ploting products with highest number of complaints

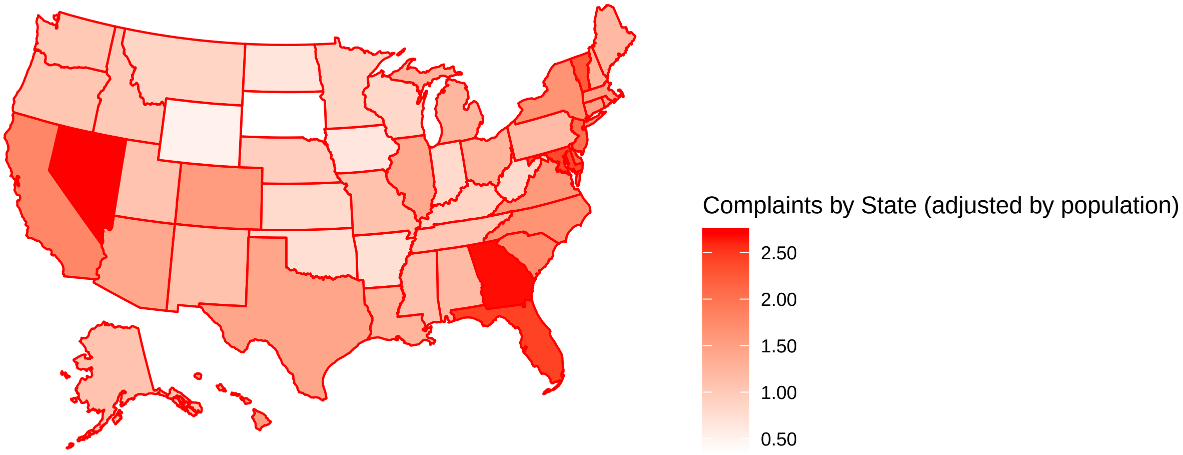
```
Top_10_product <- complaints_test %>%
  group_by(Product) %>%
  summarise(number_of_complaints = n())%>%
  arrange(desc(number_of_complaints))%>%
  head(10)%>%
  select(Product) # This line select the companies with most complaints
p_2 <- complaints_test %>%
  filter(Product %in% Top_10_product$Product) %>%
  group_by(Product,Submitted.via)%>%
  summarise(number_of_complaints = n())%>%
  ungroup()%>%
  mutate(Product = factor(Product, levels=Top_10_product$Product))%>%
  ggplot(aes(fill=Submitted.via, y=number_of_complaints, x=Product)) +
  geom_bar(position="stack", stat="identity")+ylab('Nunber of Complaints by Product')+coord_flip()+theme
p_2
```



7. Mapping the complaints in different states

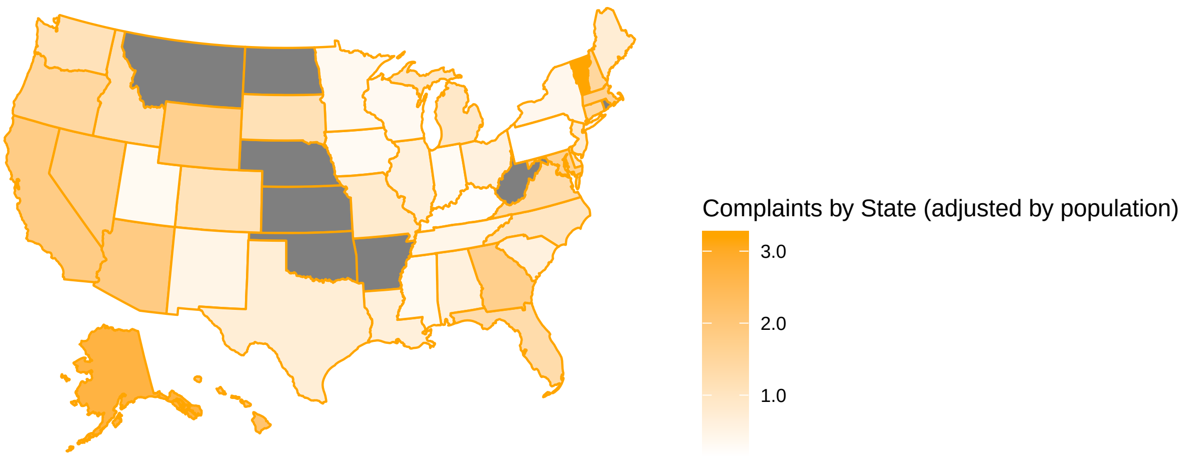
The overall complaints number accross U.S.

```
census_api_key(Sys.getenv('CENSUS_API_KEY'))
state_pop <- get_acs(
  geography = "state",
  variables = "B01003_001",
  year = 2018,
  survey = "acs5"
  )%>%
  mutate(fips = fips(NAME))%>%
  mutate(State = state.abb[match(NAME,state.name)])%>%
  select(NAME,estimate,fips,State)%>%
  mutate(state = NAME)%>%
  mutate(state_population = estimate)%>%
  select(-NAME,-estimate)
state_pop <- as.data.frame(state_pop)
state_complaint <- complaints_test %>%
  group_by(State) %>%
  summarise(number_of_complaints = n())%>%
  ungroup()
state_complaint <- left_join(state_complaint,state_pop)
state_complaint <- state_complaint%>%
  mutate(state_adjusted_complaints = number_of_complaints/state_population*100000)
state_complaint <- as.data.frame(state_complaint)%>%
  filter(!is.na(state))
plot_usmap(data = state_complaint, values = "state_adjusted_complaints", color = "red")+scale_fill_conti
    low = "white", high = "red", name = "Complaints by State (adjusted by population)", label = scales:
```



complaints number accross U.S.

```
company_complaints_mapping <- function(x){
state_complaint <- complaints_test %>%
  filter(Company == x) %>%
  group_by(State) %>%
  summarise(number_of_complaints = n())%>%
  ungroup()
state_complaint <- left_join(state_complaint,state_pop)
state_complaint <- state_complaint%>%
  mutate(state_adjusted_complaints = number_of_complaints/state_population*1000000)
state_complaint <- as.data.frame(state_complaint)%>%
  filter(!is.na(state))
plot_usmap(data = state_complaint, values = "state_adjusted_complaints", color = "orange")+scale_fill_c
    low = "white", high = "orange", name = "Complaints by State (adjusted by population)", label = scal
company_complaints_mapping('BANK OF AMERICA, NATIONAL ASSOCIATION')
```



8. Text mining and word cloud for complaints involving monetary relief or not.

set up corpus for narrative complaints

```
complaints_narrative_corp <- complaints_test %>%
  select(Company.response.to.consumer,Consumer.complaint.narrative)%>%
  filter(!is.na(Consumer.complaint.narrative))%>%
  filter(Consumer.complaint.narrative != '')%>%
  group_by(Company.response.to.consumer)%>%
  summarise(narrative = paste0(Consumer.complaint.narrative,collapse = " "))%>%
  ungroup()%>%
  unnest_tokens(word, narrative) %>%
  count(Company.response.to.consumer,word, sort = TRUE)
```

data cleaning

```
complaints_narrative_corp <- complaints_narrative_corp %>%
  filter(word %in% stopwords("english") == FALSE)%>%
  filter(word %in% c("xxxx", "xxxxxxxx", "xx") == FALSE)%>%
  mutate(word = tolower(word))%>%
  filter(str_detect(word, "^[0-9]")==FALSE)%>%
  filter(str_detect(word,'[[:punct:] ]+')==FALSE)%>%
  filter(str_detect(word,' ')==FALSE)
```

calculate number of words of each issue

```
total_words <- complaints_narrative_corp %>%
  group_by(Company.response.to.consumer) %>%
  summarize(total = sum(n))
complaints_narrative_corp <- left_join(complaints_narrative_corp, total_words)
```

```
complaints_narrative_corp <- complaints_narrative_corp %>%
  bind_tf_idf(word, Company.response.to.consumer, n)
```

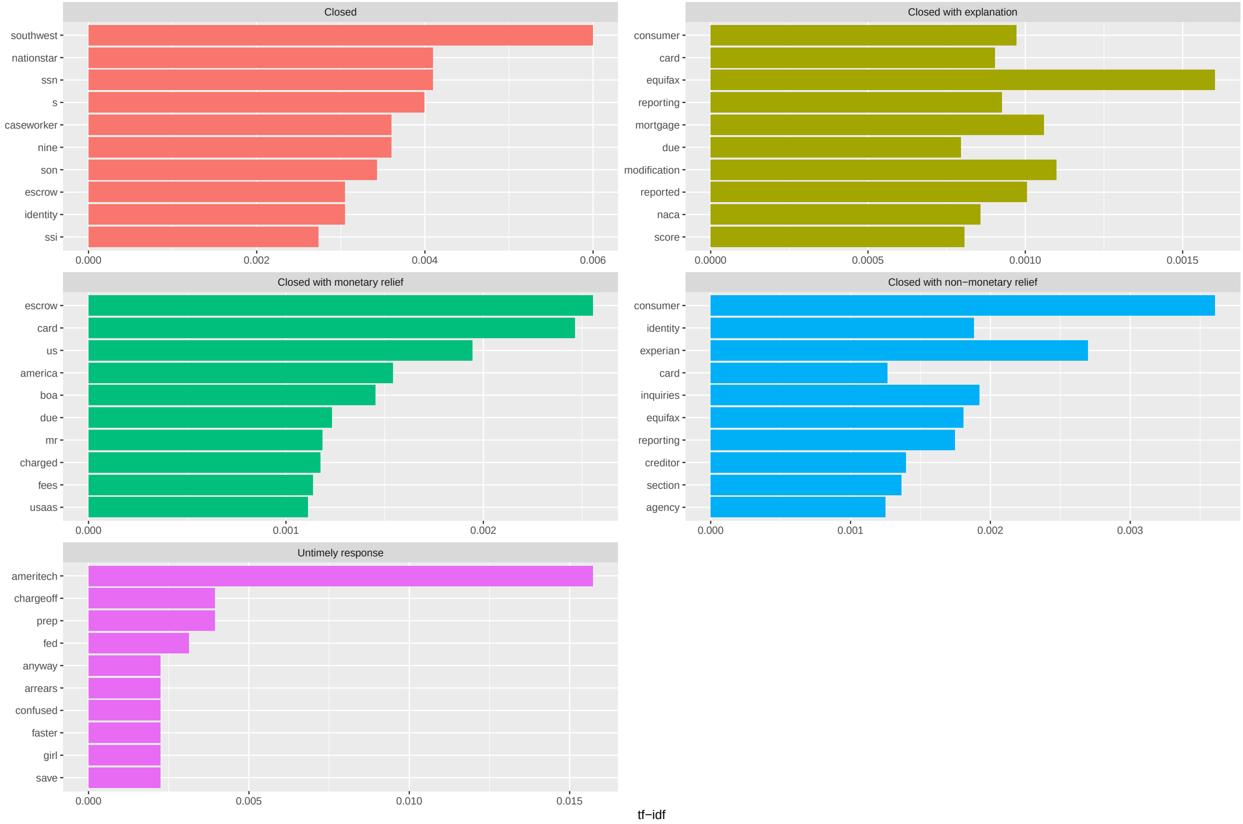
```
complaints_narrative_corp <- complaints_narrative_corp %>%
  select(-total) %>%
  arrange(desc(tf_idf))
str(complaints_narrative_corp)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    13979 obs. of  6 variables:
##  $ Company.response.to.consumer: chr  "Untimely response" "Closed" "Closed" "Closed" ...
##  $ word                        : chr  "ameritech" "southwest" "nationstar" "ssn" ...
##  $ n                           : int  8 5 6 6 24 2 2 111 3 3 ...
##  $ tf                          : num  0.00977 0.00373 0.00447 0.00447 0.01788 ...
##  $ idf                         : num  1.609 1.609 0.916 0.916 0.223 ...
##  $ tf_idf                      : num  0.01572 0.006 0.0041 0.0041 0.00399 ...
```

```
p_5 <- complaints_narrative_corp %>%
  arrange(desc(tf_idf)) %>%
  mutate(word = factor(word, levels = rev(unique(word)))) %>%
  group_by(Company.response.to.consumer) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(word, tf_idf, fill = Company.response.to.consumer)) +
  geom_col(show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~Company.response.to.consumer, ncol = 2, scales = "free") +
  coord_flip()
```

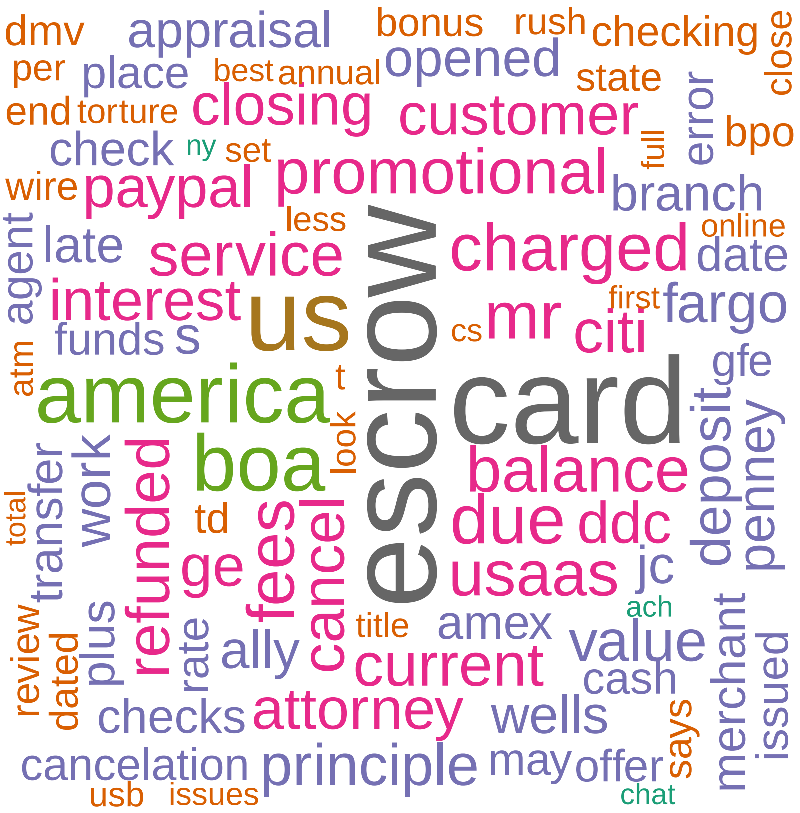
```
## Selecting by tf_idf
```

```
p_5
```



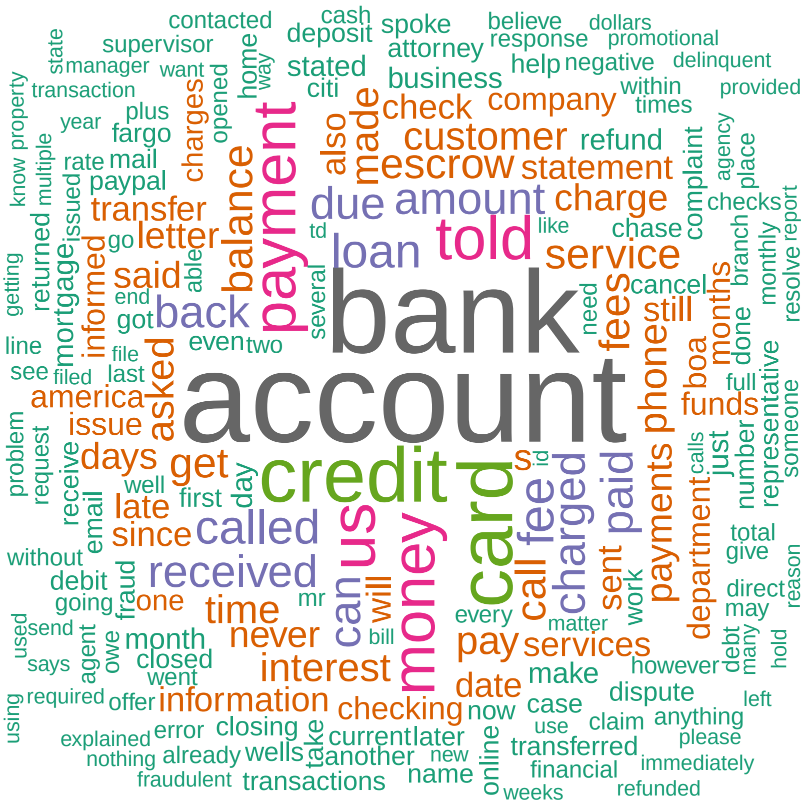
word cloud with tf_idf

```
complaints_narrative_corp <- complaints_narrative_corp %>%
    filter(Company.response.to.consumer == 'Closed with monetary relief')
wordcloud(words = complaints_narrative_corp$word, freq = complaints_narrative_corp$tf_idf,
    max.words=200, random.order=FALSE, rot.per=0.35,
    colors=brewer.pal(8, "Dark2"))
```



word cloud with frequency of words

```
wordcloud(words = complaints_narrative_corp$word, freq = complaints_narrative_corp$n,
    max.words=200, random.order=FALSE, rot.per=0.35,
    colors=brewer.pal(8, "Dark2"))
```



repeat all of that using bigram

```r
complaints_narrative_corp_2 <- complaints_test %>%
  select(Company.response.to.consumer,Consumer.complaint.narrative)%>%
  filter(!is.na(Consumer.complaint.narrative))%>%
  filter(Consumer.complaint.narrative != '')%>%
  group_by(Company.response.to.consumer)%>%
  summarise(narrative = paste0(Consumer.complaint.narrative,collapse = " "))%>%
  ungroup()%>%
  unnest_tokens(word, narrative)%>%
  filter(word %in% stopwords("english") == FALSE)%>%
  filter(word %in% c("xxxx", "xxxxxxxx", "xx") == FALSE)%>%
  mutate(word = tolower(word))%>%
  filter(str_detect(word, "^[0-9]")==FALSE)%>%
  filter(str_detect(word,'[[:punct:] ]+')==FALSE)%>%
  filter(str_detect(word,' ')==FALSE)%>%
  group_by(Company.response.to.consumer)%>%
  summarise(narrative = paste0(word,collapse = " "))%>%
  unnest_tokens(bigrams, narrative, token = "ngrams", n = 2)%>%
  count(Company.response.to.consumer,bigrams, sort = TRUE)
```

```r
total_words_2 <- complaints_narrative_corp_2 %>%
  group_by(Company.response.to.consumer) %>%
  summarize(total = sum(n))
complaints_narrative_corp_2 <- left_join(complaints_narrative_corp_2, total_words_2)
```

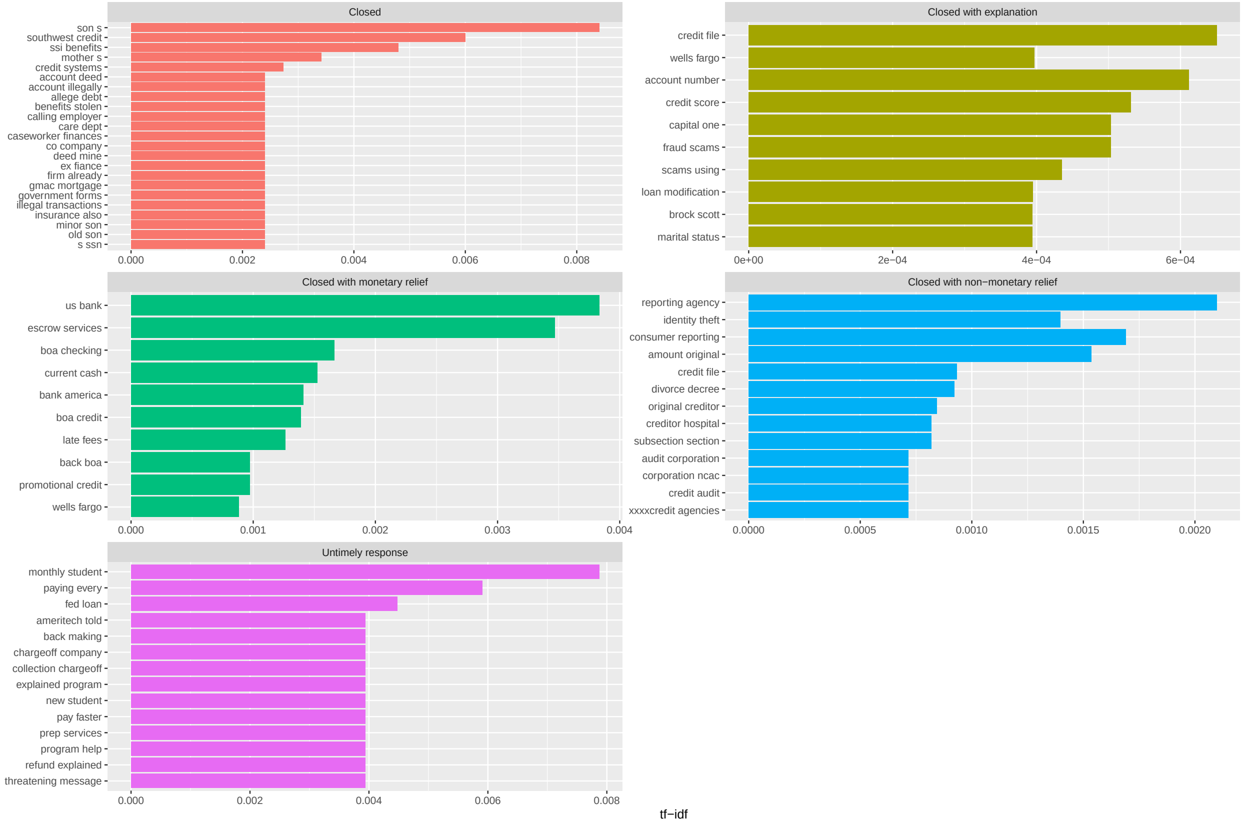
```
complaints_narrative_corp_2 <- complaints_narrative_corp_2 %>%
  bind_tf_idf(bigrams, Company.response.to.consumer, n)
```

```
complaints_narrative_corp_2 <- complaints_narrative_corp_2 %>%
  select(-total) %>%
  arrange(desc(tf_idf))
```

```
p_6 <- complaints_narrative_corp_2 %>%
  arrange(desc(tf_idf)) %>%
  mutate(bigrams = factor(bigrams, levels = rev(unique(bigrams)))) %>%
  group_by(Company.response.to.consumer) %>%
  top_n(10) %>%
  ungroup() %>%
  ggplot(aes(bigrams, tf_idf, fill = Company.response.to.consumer)) +
  geom_col(show.legend = FALSE) +
  labs(x = NULL, y = "tf-idf") +
  facet_wrap(~Company.response.to.consumer, ncol = 2, scales = "free") +
  coord_flip()
```

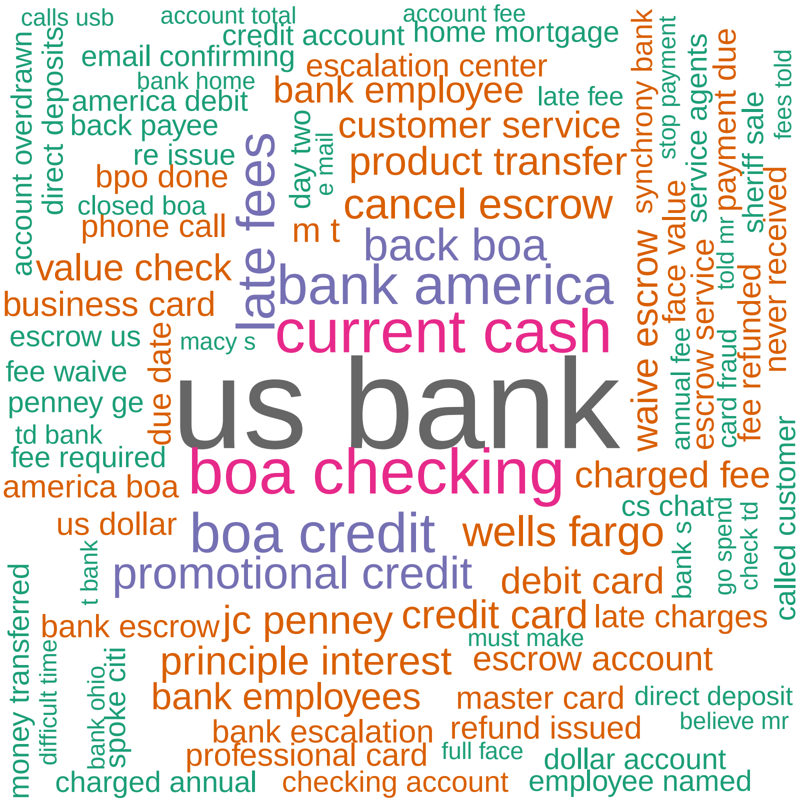
```
## Selecting by tf_idf
```

```
p_6
```



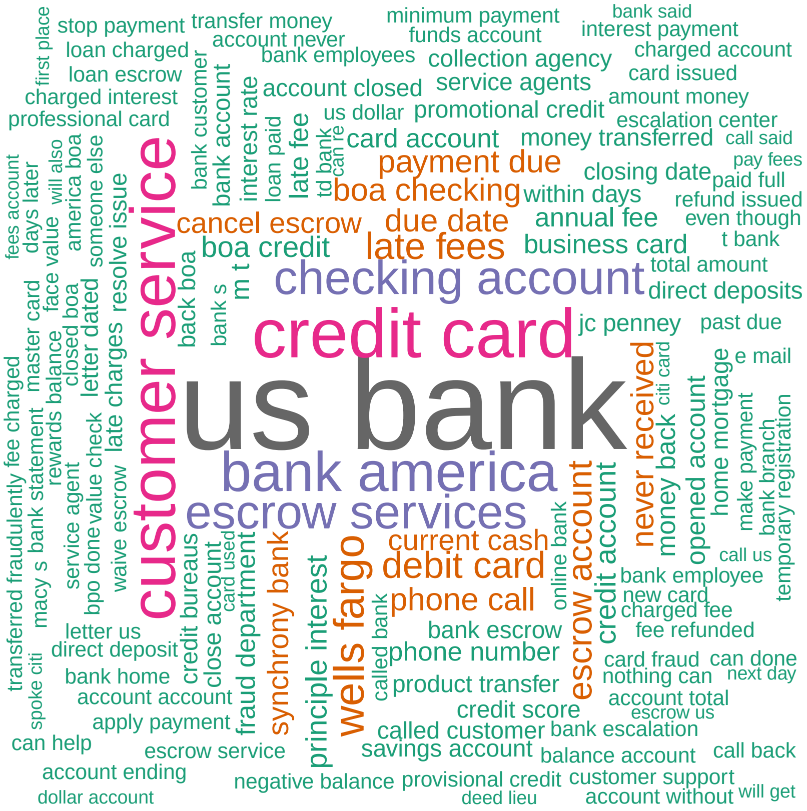
word cloud with tf_idf

```r
complaints_narrative_corp_2 <- complaints_narrative_corp_2 %>%
    filter(Company.response.to.consumer == 'Closed with monetary relief')
wordcloud(words = complaints_narrative_corp_2$bigrams, freq = complaints_narrative_corp_2$tf_idf,
    max.words=200, random.order=FALSE, rot.per=0.35,
    colors=brewer.pal(8, "Dark2"))
```



word cloud with frequency of words

```r
wordcloud(words = complaints_narrative_corp_2$bigrams, freq = complaints_narrative_corp_2$n,
    max.words=200, random.order=FALSE, rot.per=0.35,
    colors=brewer.pal(8, "Dark2"))
```


us bank
credit card
customer service
checking account
bank america
escrow services
wells fargo
late fees
due date
boa checking
payment due
cancel escrow
debit card
phone call
current cash
escrow account
never received
synchrony bank
principle interest
boa credit
stop payment
transfer money
minimum payment
bank said
interest payment
funds account
account never
loan charged
charged account
bank employees
collection agency
card issued
loan escrow
first place
service agents
account closed
charged interest
amount money
promotional credit
professional card
us dollar
escalation center
card account
money transferred
call said
pay fees
paid full
closing date
within days
refund issued
even though
annual fee
t bank
business card
total amount
direct deposits
jc penney
past due
e mail
bank customer
bank account
interest rate
loan paid
late fee
td bank
can re
fees account
days later
will also
america boa
someone else
resolve issue
face value
closed boa
master card
rewards balance
letter dated
late charges
back boa
bank s
m t
citi card
money back
opened account
home mortgage
make payment
bank branch
temporary registration
credit account
online bank
call us
bank employee
new card
charged fee
fee refunded
transferred fraudulently fee charged
macy s bank statement
service agent
value check
bpo done
waive escrow
credit bureaus
close account
card used
fraud department
called bank
bank escrow
phone number
product transfer
credit score
card fraud
can done
nothing can
next day
account total
escrow us
bank escalation
letter us
direct deposit
bank home
account account
apply payment
spoke citi
can help
escrow service
called customer
savings account
balance account
call back
account ending
negative balance
provisional credit
customer support
dollar account
deed lieu
account without
will get