Visualization for IBM Survey

Nikky Xiong

Feb 6, 2020

This is the analysis for the first week of this project, aiming to map out a guide line for the future data exploratory work, like data exploration for certain type of complaints or for one company, text mining, predictive models, and maps. Things that we can dive into for EDA include data summary by companies, products, issues, states. Also, visualization that we can show include world cloud, line charts, scatter plots, and bar charts. We can try to integrate other data resources, such as the total population of each states to make further analysis.

```
library(readr)
library(dplyr)
library(tidytext)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
```

data summary

```
df <- readr::read_csv('complaints.csv')
df %>% summary()
```

```
##  Date received          Product          Sub-product
##  Min.   :2011-12-01   Length:1495408     Length:1495408
##  1st Qu.:2015-03-26   Class :character   Class :character
##  Median :2017-03-21   Mode  :character   Mode  :character
##  Mean   :2016-11-17
##  3rd Qu.:2018-09-07
##  Max.   :2020-02-05
##     Issue            Sub-issue         Consumer complaint narrative
##  Length:1495408     Length:1495408     Length:1495408
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Company public response   Company             State
##  Length:1495408          Length:1495408     Length:1495408
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##    ZIP code             Tags           Consumer consent provided?
##  Length:1495408     Length:1495408     Length:1495408
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Submitted via      Date sent to company Company response to consumer
##  Length:1495408     Min.   :2011-12-01   Length:1495408
##  Class :character   1st Qu.:2015-03-31   Class :character
##  Mode  :character   Median :2017-03-22   Mode  :character
##                     Mean   :2016-11-19
##                     3rd Qu.:2018-09-08
##                     Max.   :2020-02-05
##  Timely response?   Consumer disputed?  Complaint ID
##  Length:1495408     Length:1495408     Min.   :      1
##  Class :character   Class :character   1st Qu.:1303398
##  Mode  :character   Mode  :character   Median :2396664
##                                        Mean   :2135220
##                                        3rd Qu.:3012962
##                                        Max.   :3522181
```

```
df <- df %>% head(500)
```

text cleaning function

```
clean_text <- function(variable){

    text = tibble(txt = variable)

    text <- text %>%
    unnest_tokens(word, txt) %>%
    anti_join(stop_words) %>%
    na.omit() %>%
    count(word, sort = TRUE)

    return(text)
}
```

text cleaning by variables

```
product <- clean_text(df$Product)
issue <- clean_text(df$Issue)
complaints <- clean_text(df$`Consumer complaint narrative`)
response <- clean_text(df$`Company public response`)
```

visualization

```
wordcloud(words = product$word, freq = product$n,
          colors=brewer.pal(6, "Dark2"), min.freq = 1)
```

`Credit` is the most important and the most frequently mentioned key words in the products. `Consumer`, `reports`, `perosnl`, `repair`, and `services` are also worth paying attention to.

```
wordcloud(words = issue$word, freq = issue$n,
          colors=brewer.pal(6, "Dark2"), min.freq = 1)
```

`Report[ing] incorrect information` seems to be the most important and the most frequently mentioned key words in the issues.

In [1]:
```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

pd.options.display.max_rows = 10
```

In [2]:
```
df = pd.read_csv("complaints.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	Date received	Product	Sub-product	Issue	Sub-issue	Consumer complaint narrative	Company public response	Compa
0	2019-09-24	Debt collection	I do not know	Attempts to collect debt not owed	Debt is not yours	transworld systems inc. \nis trying to collect...	NaN	TRANSWOR SYSTEMS I
1	2019-09-19	Credit reporting, credit repair services, or o...	Credit reporting	Incorrect information on your report	Information belongs to someone else	NaN	Company has responded to the consumer and the ...	Experi Informati Solutions I
2	2019-10-25	Credit reporting, credit repair services, or o...	Credit reporting	Incorrect information on your report	Information belongs to someone else	I would like to request the suppression of the...	Company has responded to the consumer and the ...	TRANSUNI INTERMEDIA HOLDING IN
3	2019-11-08	Debt collection	I do not know	Communication tactics	Frequent or repeated calls	Over the past 2 weeks, I have been receiving e...	NaN	Diversifi Consultan I
4	2019-02-08	Vehicle loan or lease	Lease	Problem with a credit reporting company's inve...	Their investigation did not fix an error on yo...	NaN	NaN	HYUND CAPIT AMERI

In [4]:
```
df.describe(include='all')
```

Out[4]:

	Date received	Product	Sub-product	Issue	Sub-issue	Consumer complaint narrative	Company public response	Company	1
count	1495408	1495408	1260243	1495408	937919	491046	560987	1495408	
unique	2988	18	76	166	218	463288	10	5526	
top	2017-09-08	Credit reporting, credit repair services, or o...	Credit reporting	Incorrect information on your report	Information belongs to someone else	There are many mistakes appear in my report wi...	Company has responded to the consumer and the ...	EQUIFAX, INC.	
freq	3553	336054	330136	207899	106939	1035	410681	148685	
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

11 rows × 18 columns

In [7]:
```
# subset only the first 500 entries of the dataset
subset = df[0:500]
```

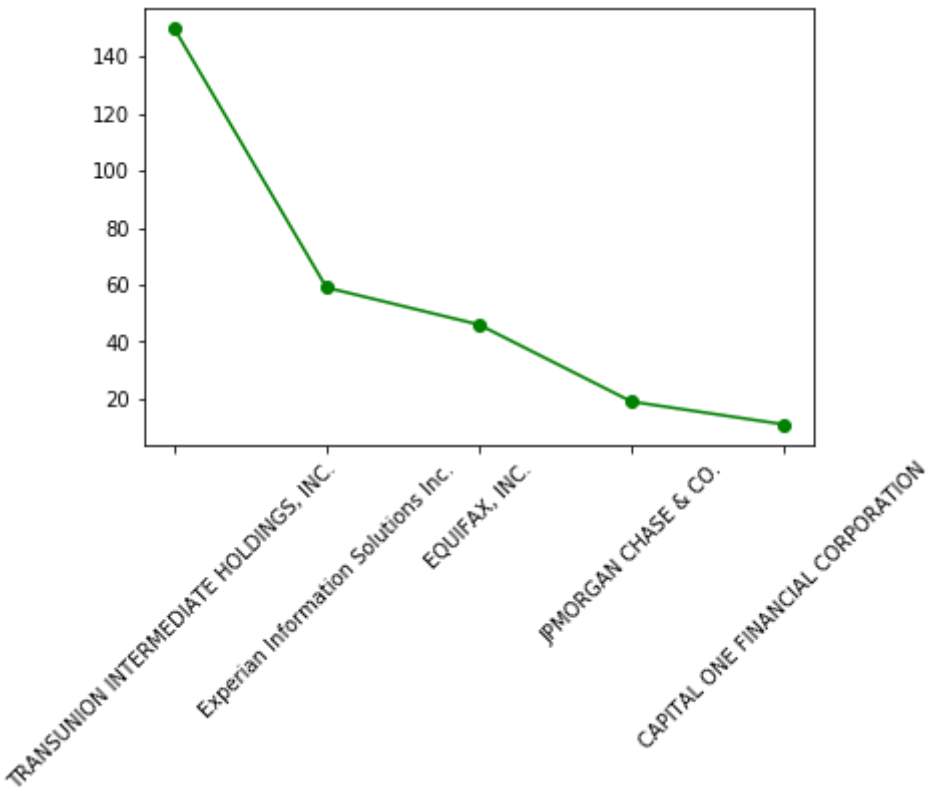
In [46]:
```
# how many products from each company have complaints
product = subset.groupby('Company')['Product'].count().sort_values(ascen
ding= False).head(10)
product = product.to_frame()
print(product)
# TRANSUNION INTERMEDIATE HOLDINGS, INC. is the company in the subset ha
s the most products that received complaints
```

```
                                                Product
Company
TRANSUNION INTERMEDIATE HOLDINGS, INC.              150
Experian Information Solutions Inc.                  59
EQUIFAX, INC.                                        46
JPMORGAN CHASE & CO.                                 19
CAPITAL ONE FINANCIAL CORPORATION                    11
CITIBANK, N.A.                                       10
SYNCHRONY FINANCIAL                                   8
WELLS FARGO & COMPANY                                 7
AMERICAN EXPRESS COMPANY                              6
Navient Solutions, LLC.                               5
```

Out[46]:
```
pandas.core.frame.DataFrame
```

In [72]:
```
plt.plot(product['Product'], color='green', marker='o', linestyle='soli
d')
plt.xticks(rotation=45)
# There is a sharp decrease in the number of products from companies tha
t got complaints.
# Top three companies had the highest number of products with complaints
that took a large share of the complaints.
```

Out[72]: ([0, 1, 2, 3, 4], <a list of 5 Text xticklabel objects>)

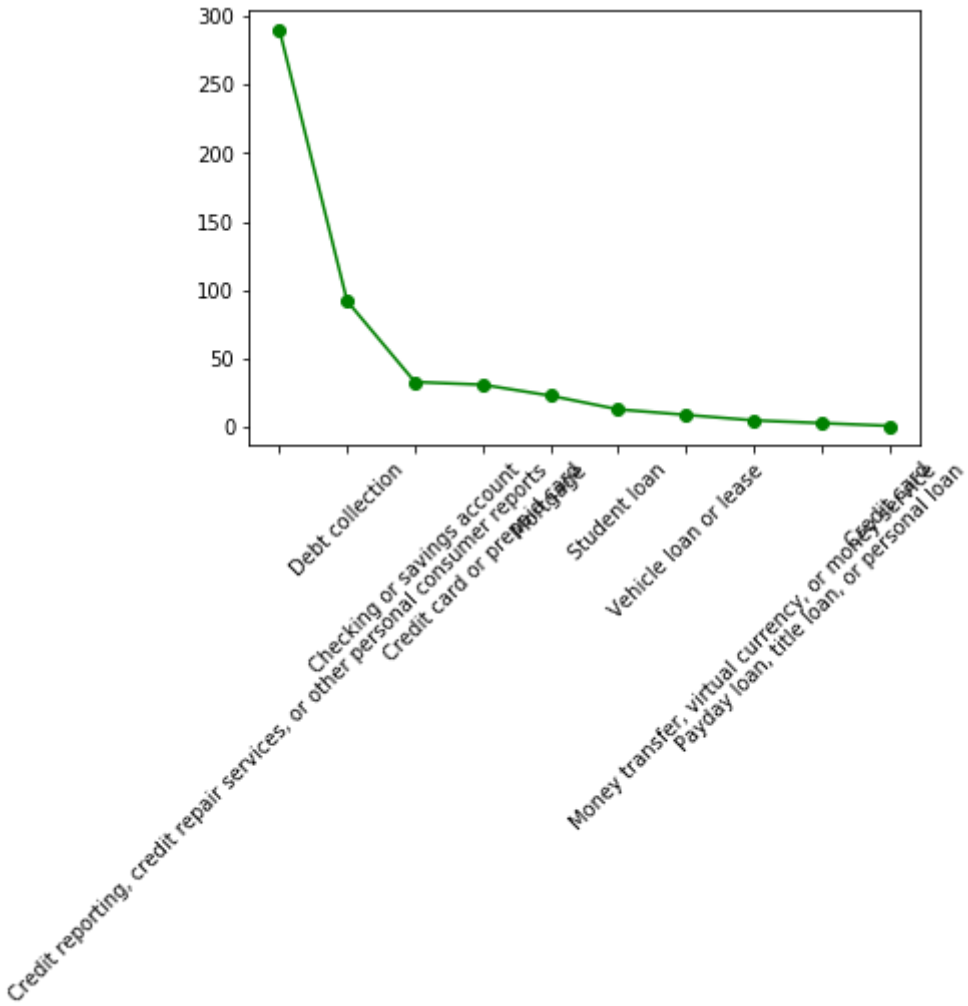


In [48]:
```
# how many issues from each product have complaints
issue = subset.groupby('Product')['Issue'].count().sort_values(ascending
= False).head(10)
issue = issue.to_frame()
print(issue)
# Credit reporting, credit repair services, or other personal consumer r
eports is the product
# in the subset has the most issues that received complaints
```

Product	Issue
Credit reporting, credit repair services, or ot...	290
Debt collection	92
Checking or savings account	33
Credit card or prepaid card	31
Mortgage	23
Student loan	13
Vehicle loan or lease	9
Money transfer, virtual currency, or money service	5
Payday loan, title loan, or personal loan	3
Credit card	1

In [73]:
```
plt.plot(issue['Issue'], color='green', marker='o', linestyle='solid')
plt.xticks(rotation=45)
# There is a sharp decrease in the number of issues from products that g
ot complaints.
# The top issue is about a product focused on Debt collection.
```

Out[73]:
```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9], <a list of 10 Text xticklabel objects
>)
```

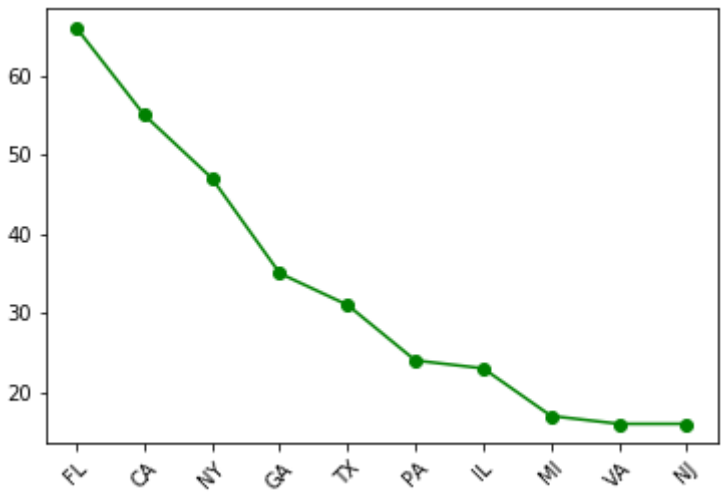


In [49]:
```
# how many issues from each state
state = subset.groupby('State')['Issue'].count().sort_values(ascending=
False).head(10)
state = state.to_frame()
print(state)
# Florida is the state that fires the most complaints
```

```
       Issue
State
FL        66
CA        55
NY        47
GA        35
TX        31
PA        24
IL        23
MI        17
VA        16
NJ        16
```

In [75]:
```
plt.plot(state['Issue'], color='green', marker='o', linestyle='solid')
plt.xticks(rotation=45)
# Top 7 states made the mass majority of the complaints.
```

Out[75]:
```
([0, 1, 2, 3, 4, 5, 6, 7, 8, 9], <a list of 10 Text xticklabel objects
>)
```



In [ ]:
```
In conclusion, it seems that we can investigate into companies with the
most issues as the top companies receive the mass majority of the compla
ints. The same applies to . We can dig further into these areas to find
more specific patterns.
```