Dataset can be found at https://data.world/dataquest/bank-and-credit-card-complaints under Bank_Account_or_Service_Complaints.csv

Pre-processing:

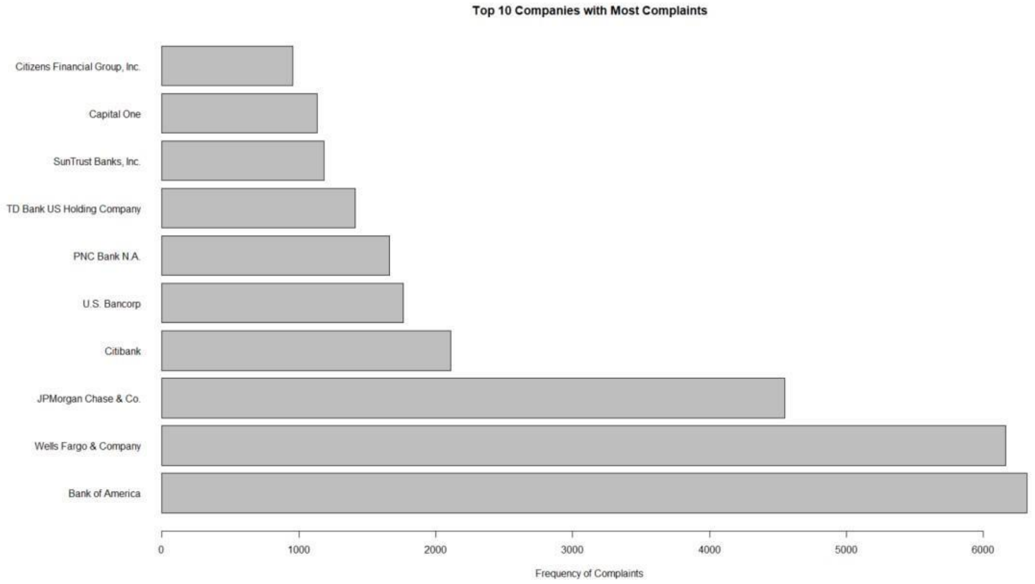
The dataset is generated through the given code. Looking through the dataset, I have identified some data columns that can be removed. These data columns are Sub.issue, Consumer.complaint.narrative, Company.public.response, Consumer.consent.provided., and Date.received. Sub.issue and Consumer.complaint.narrative are removed due to the majority of data not having any data on it, and we can't assume what data the empty dataspace infers. Consumer.consent.provided. is removed as the data has no significance in helping the analysis of the complaints. Date.received is removed as there is another Date data that is similar, namely Date.sent.to.company. Since Date.sent.to.company signifies the date customers wrote the complaint, this data was used over Date.received as it more accurately shows the date said complaint was written.

Code:

```
library(ggplot2)
rm(list = ls())
set.seed(29800463)
bankcomplaints <- read.csv("bankcomplaints.csv") bankcomplaints <-
bankcomplaints[sample(nrow(bankcomplaints),40000),]
a <- c("Sub.issue", "Consumer.complaint.narrative", "Company.public.response",
"Consumer.consent.provided.", "Date.received")
cleanedBC <- bankcomplaints[,!names(bankcomplaints) %in% a, drop = F]
```

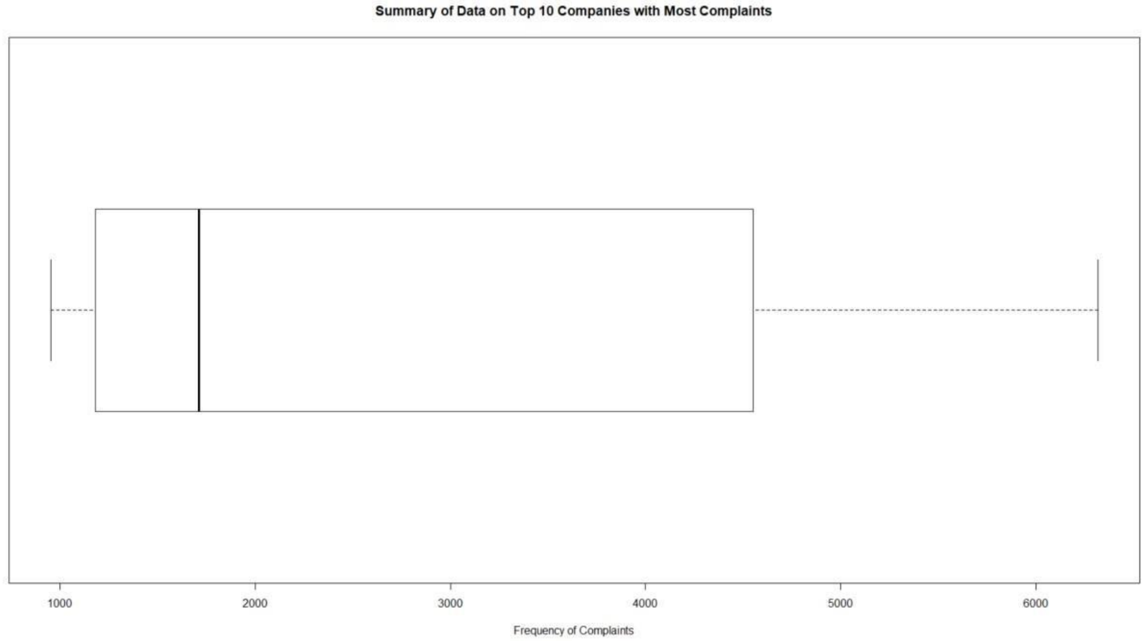
Analysis:

There are a total of 325 companies within this dataset. When looking at the data for number of complaints grouped by company, we can see that the minimum number of complaints a company got is 1, and the maximum number of complaints a company got is 6318. The median number of complaints a company got is 2. On average, a company listed within this dataset got 123.1 complaints. There is no accompanying visualisation to this data as the visualisations would not provide any meaningful understanding to the data, and the reason for this will be discussed later.



The above is a bar chart of the top 10 companies with most complaints. This subset is used as to gain insight on what companies got the most complaints and the amount they got. The company that got the greatest number of complaints is Bank of America, with 6318 complaints. This is followed by Wells Fargo & company which had 6162 complaints, JPMorgan Chase & Co. with 4550 complaints, Citibank with 2109 complaints, U.S. Bancorp with 1761 complaints, PNC Bank N.A with 1665 complaints, TD Bank US Holding Company with 1414 complaints, SunTrust Banks, Inc. with 1182 complaints, Capital One with 1136 complaints, and Citizens Financial Group, Inc. with 954 complaints.

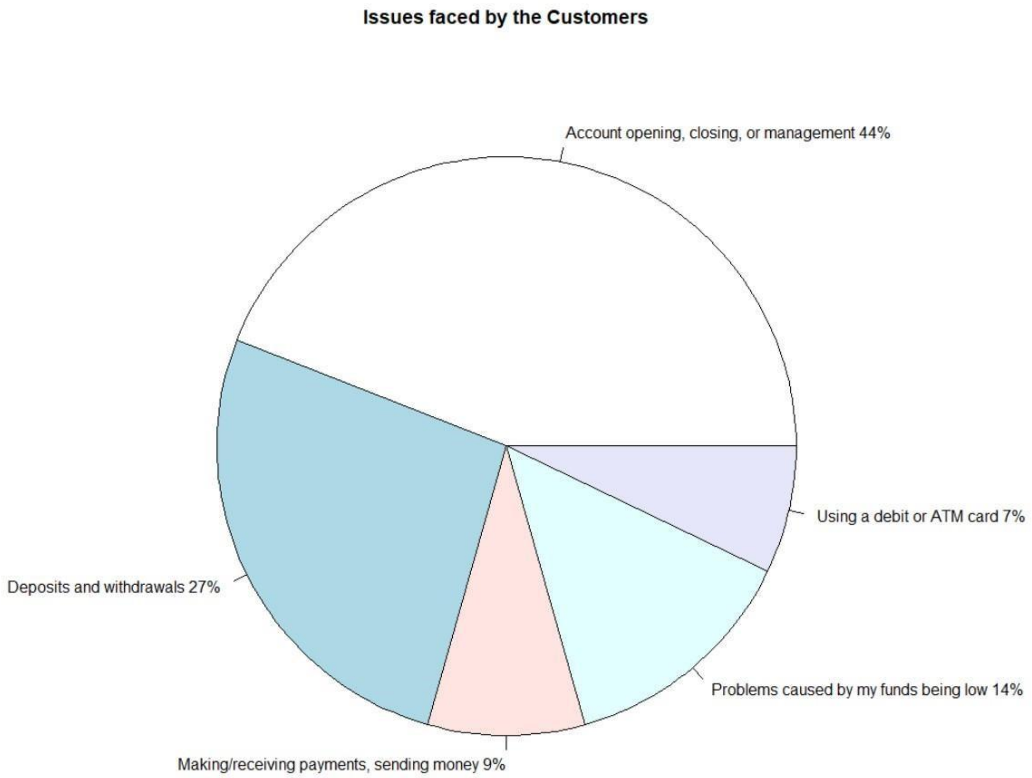
Based on the data of these top 10 companies, the minimum number of complaints is 954, and the maximum number of complaints is 6318. The median number of complaints was 1713 and these 10 companies got 2725 complaints on average. Below shows the summary of data on the top 10 companies with most complaints.



There were 139 companies that got only 1 complaint within the dataset. This means that about 42.8% of companies only had 1 complaint, while 289 companies had less or equal to 100 complaints. This would mean that roughly 88.9% of companies had less or equal to 100 complaints. This is a strong contributor to the low median and mean, even though the company with the most complaints has 6318 complaints. This also made the box plot unreadable due to the large gap between the min, 1st quartile, median and 3rd quartile when compared with the maximum.

Code:

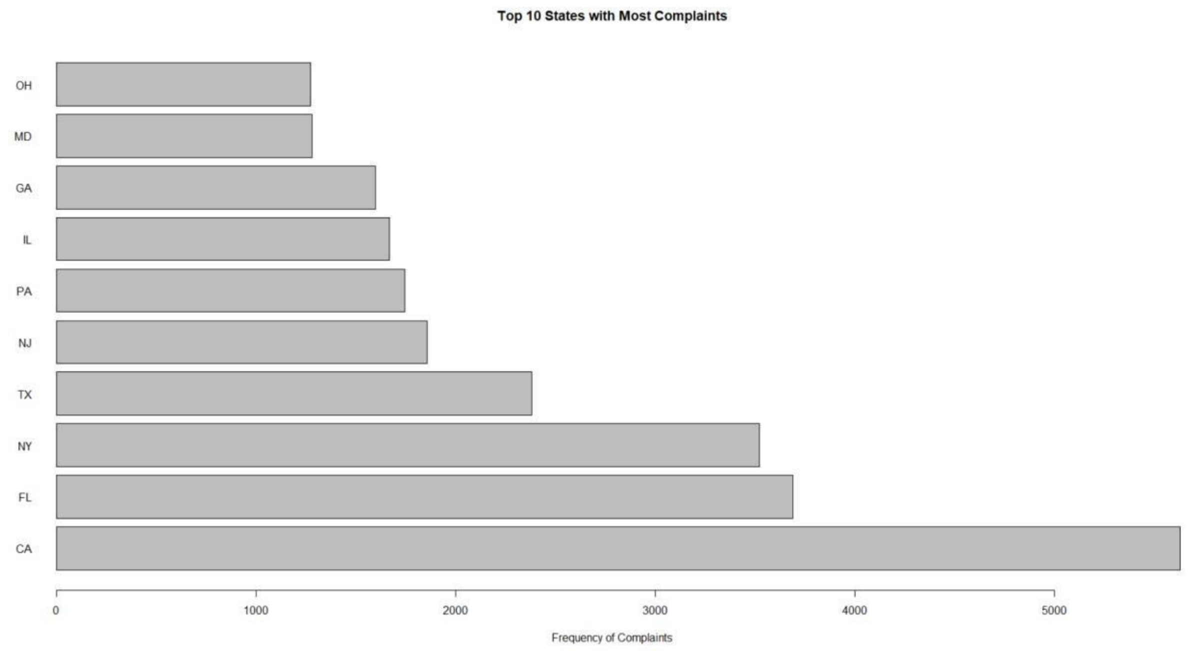
```
test <-data.frame(table(cleanedBC$Company))
summary(test$Freq) b <- test[order(test$Freq),]
c <- head(b,n = 10) par(mar=c(5,15,3,3))
barplot(c$Freq, main = "Top 10 Companies with Most Complaints", xlab = "Frequency of
Complaints", names.arg = c$Var1, horiz = TRUE,las = 1)
boxplot(c$Freq,main = "Summary of Data on Top 10 Companies with Most Complaints",
xlab = "Frequency of Complaints",col = "white",border = "black", horizontal=
TRUE) summary(c$Freq) length(which(test$Freq == 1))
length(which(test$Freq <= 100))
length(which(test$Freq == 1))/nrow(test)
length(which(test$Freq <= 100))/nrow(test)
```



To understand the type of activities that lead to the most complaints, we will analyse the issues faced by the customers. The leading issue faced by customers is Account opening, closing or management with 17618 complaints. This is followed by deposits and withdrawals with 10626 complaints, making/receiving payments, sending money with 3501 complaints, problems caused by my funds being low with 5417 complaints, and using a debit or ATM card with 2838 complaints. The pie chart above shows a breakdown of the issues faced by customers along with the percentage (rounded to nearest percentage. Total not equal to 100% due to rounding.) of complaints containing the issue.

Code:

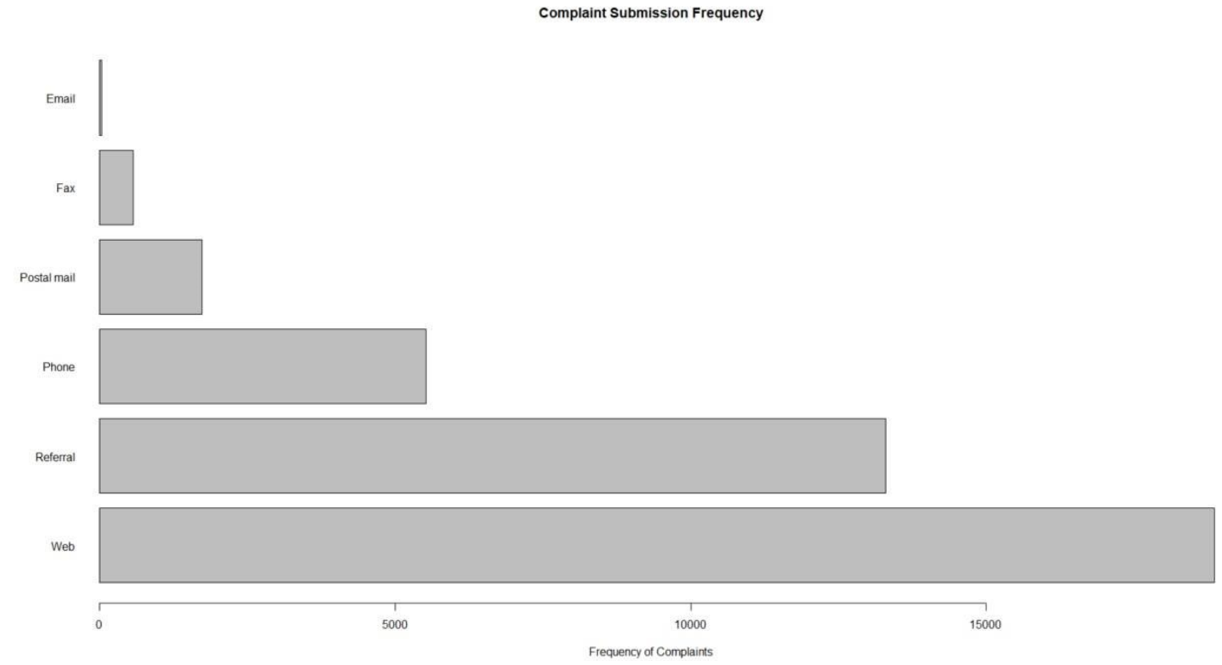
```
test <-data.frame(table(cleanedBC$Issue))
percentage <-round(test$Freq/sum(test$Freq)*100)
label <- paste(test$Var1, percentage) label <- paste(label,"%", sep = "")
pie(test$Freq, labels = label, main = "Issues faced by the Customers")
```



The state with the most complaints sent is California with 5631 complaints, followed by Florida with 3689 complaints, New York with 3520 complaints, Texas with 2381 complaints, New Jersey with 1856 complaints, Pennsylvania with 1743 complaints, Illinois with 1666 complaints, Georgia with 1596 complaints, Maryland with 1280 complaints, and Ohio with 1271 complaints. Above is a bar chart showing the top 10 states by frequency of complaints along with the number of complaints. This is done as such in order to ensure the visualisation is readable and easy to view and understand.

Code:

```
test <-data.frame(table(cleanedBC$Submitted.via))
b <- test[order(-test$Freq),]
c <- head(b,n = 10)
barplot(c$Freq, main = "Top 10 States with Most Complaints", xlab = "Frequency of
Complaints", names.arg = c$Var1, horiz = TRUE,las = 1)
```



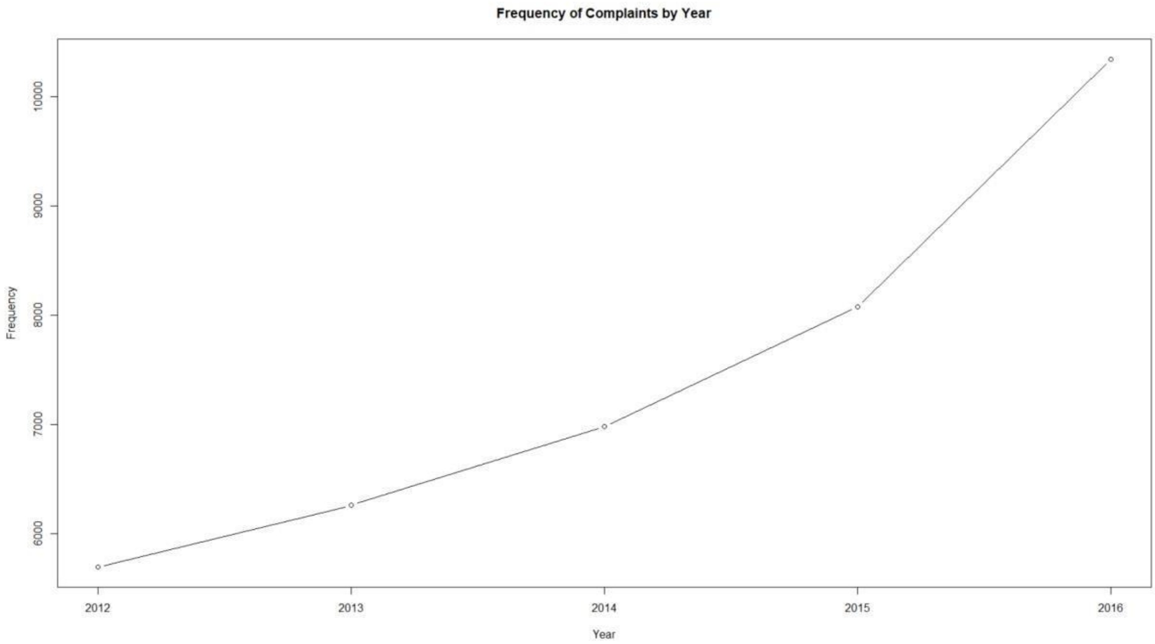
Attached above is a bar chart showing all the complaint submissions and their frequency. Most complaints within this dataset are sent in via web, with 18863 complaints sent in via this method. This is followed by referral with 13298 complaints, Phone with 5517 complaints, Postal mail with 1729 complaints, fax with 564 complaints, and email with 29 complaints. For every complaint sent in via email, there is roughly 650 complaints sent in through the web.

Code:

```
test <-data.frame(table(cleanedBC$Submitted.via))
b <- test[order(-test$Freq),]
par(mar=c(5,7,4,2))
barplot(b$Freq, main = "Complaint Submission Frequency", xlab = "Frequency of Complaints", names.arg = b$Var1, horiz = TRUE,las = 1)
```

To better view the trends in complaints, we analyse the data based on year. There were 5691 complaints within the dataset that was sent during 2012. In 2013, 6258 complaints were sent. 6978 complaints were sent during 2014 while 8078 complaints were sent in during 2015. 2016 saw the most complaints, with 10339 complaints sent in during that year.

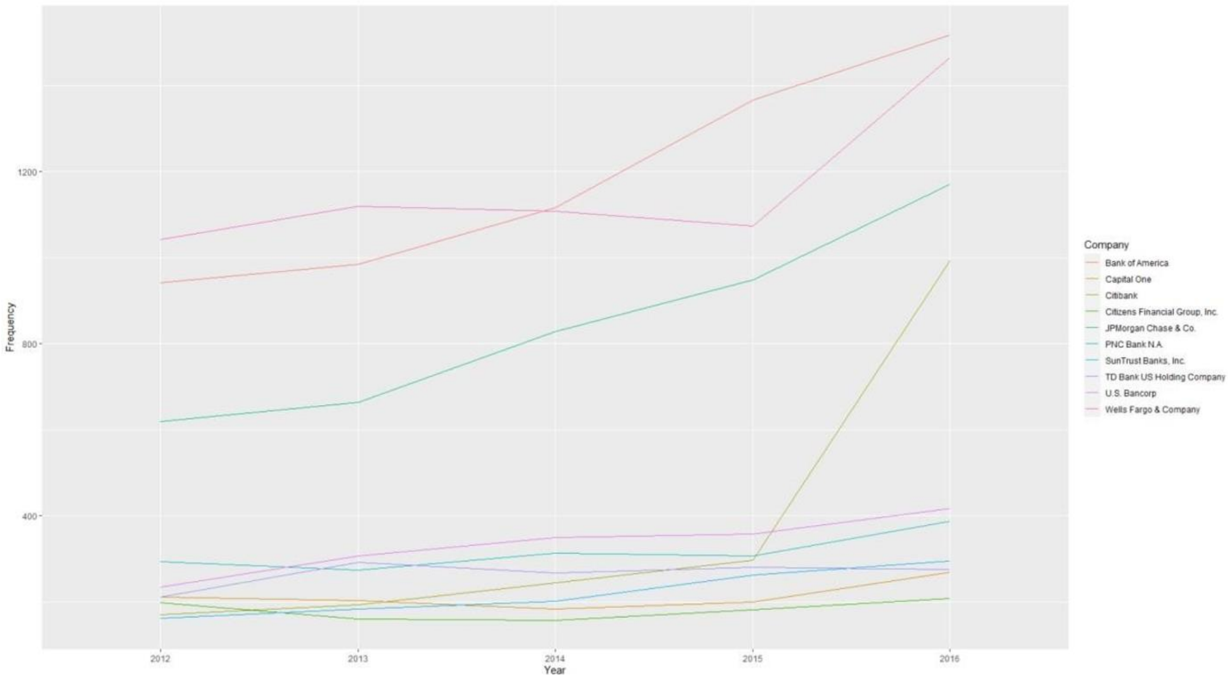
Year 2017 was not included within this part of the analysis as there was no sufficient data to allow proper analysis of 2017. As the latest data that was from 2017 was at 10/04/2017, the inclusion of 2017 makes it look like the amount of people sending in complaints during 2017 was low. Such a conclusion is untrue due to the lack of data to back it up, and as such was not considered in this part of the analysis. The years used are based on the date customers sent complaint to company, rather than the date received by company, as I believe the date customers sent complaint to company better reflects the time of complaint.



As we can see from the line chart above, there is an increase in the number of complaints received as the years increase. The amount in which the frequency of complaints increase is also increasing, as seen with the curve upwards within the line graph.

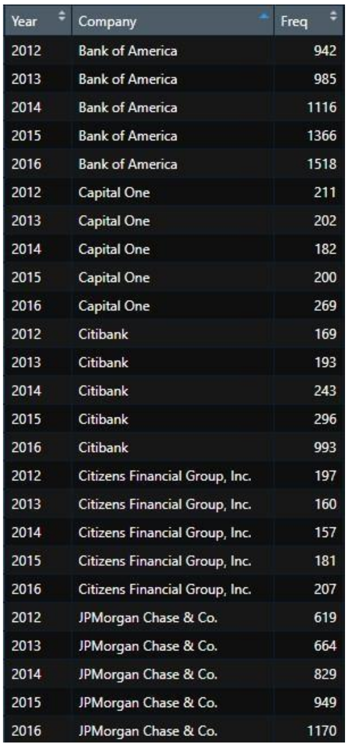
Code:

```
test2 <- as.Date(cleanedBC$Date.sent.to.company, "%m/%d/%Y")
test2 <- format(test2, "%Y")
test2 <- data.frame(table(test2))
test2 <- format.data.frame(test2)
test2 <- test2[-c(6),]
plot(test2, xlab = “Year”, ylab = “Frequency”, main = “Frequency of complaints by
year”, type = “b”)
```

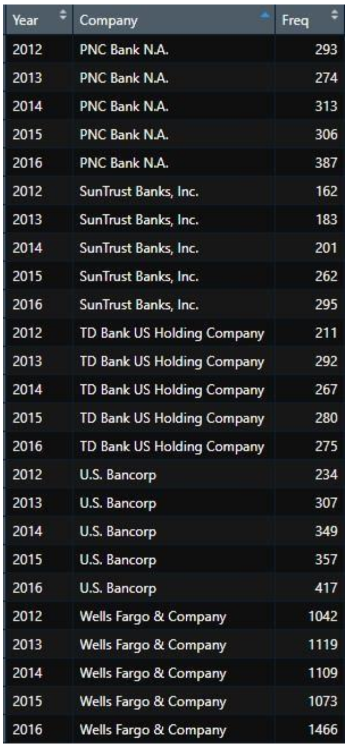


The above graph shows a multivariable line chart that shows the frequency of complaints against year for the top 10 companies with the most complaints. These 10 companies are used as they were used in our analysis previously. As we have better knowledge on the data of these companies, viewing the trends also allow us to better know how time affects these companies.

In general, the frequency of complaints increases as the years increase. However, there are some companies that show a reduction in frequency of complaints between 2012-2016, such as Capital One's complaint frequency decreased from 2012-2014, Citizens Financial Group, Inc.'s decrease in frequency from 2012-2014, PNC Bank N.A.'s reduction in frequency from 2012-2013 and from 20142015, and Wells Fargo & Company's decrease in complaints from 2013-2015, just to name a few. Added below are the specific frequency of complaints, grouped by company and year.



Year	Company	Freq
2012	Bank of America	942
2013	Bank of America	985
2014	Bank of America	1116
2015	Bank of America	1366
2016	Bank of America	1518
2012	Capital One	211
2013	Capital One	202
2014	Capital One	182
2015	Capital One	200
2016	Capital One	269
2012	Citibank	169
2013	Citibank	193
2014	Citibank	243
2015	Citibank	296
2016	Citibank	993
2012	Citizens Financial Group, Inc.	197
2013	Citizens Financial Group, Inc.	160
2014	Citizens Financial Group, Inc.	157
2015	Citizens Financial Group, Inc.	181
2016	Citizens Financial Group, Inc.	207
2012	JPMorgan Chase & Co.	619
2013	JPMorgan Chase & Co.	664
2014	JPMorgan Chase & Co.	829
2015	JPMorgan Chase & Co.	949
2016	JPMorgan Chase & Co.	1170



Year	Company	Freq
2012	PNC Bank N.A.	293
2013	PNC Bank N.A.	274
2014	PNC Bank N.A.	313
2015	PNC Bank N.A.	306
2016	PNC Bank N.A.	387
2012	SunTrust Banks, Inc.	162
2013	SunTrust Banks, Inc.	183
2014	SunTrust Banks, Inc.	201
2015	SunTrust Banks, Inc.	262
2016	SunTrust Banks, Inc.	295
2012	TD Bank US Holding Company	211
2013	TD Bank US Holding Company	292
2014	TD Bank US Holding Company	267
2015	TD Bank US Holding Company	280
2016	TD Bank US Holding Company	275
2012	U.S. Bancorp	234
2013	U.S. Bancorp	307
2014	U.S. Bancorp	349
2015	U.S. Bancorp	357
2016	U.S. Bancorp	417
2012	Wells Fargo & Company	1042
2013	Wells Fargo & Company	1119
2014	Wells Fargo & Company	1109
2015	Wells Fargo & Company	1073
2016	Wells Fargo & Company	1466

Code:

```
cleanedBC$Year <- as.Date(cleanedBC$Date.sent.to.company, "%m/%d/%Y")
cleanedBC$Year <- format(cleanedBC$Year, "%Y")
test <- data.frame(table(cleanedBC$Year, cleanedBC$Company))
test <- subset(test, Var2 == "Bank of America" |Var2 == "Wells Fargo & Company"|Var2
== "JPMorgan Chase & Co." |Var2 == "Citibank" |Var2 == "U.S. Bancorp" |Var2 == "PNC
Bank N.A." |Var2 == "TD Bank US Holding Company"|Var2 == "SunTrust Banks, Inc."
|Var2 == "Capital One"|Var2 == "Citizens Financial Group, Inc.")
test <- test[!(test$Var1==2017),]
ggplot(data=test, aes(x=Var1, y=Freq, group=Var2, color =
Var2)) + xlab("Year") + ylab("Frequency") + labs(color = "Company") + geom_line()
```