FIT3152 Assignment 1

Name: Ng Wei Hong

Student ID: 28055322

Table of contents

Table of Contents

No table of contents entries found.

Pre-processing

In the pre-processing, the following columns of the original dataset have been excluded, these proofs from the codes are as below:

```
> unique(bankcomplaints$Product)
[1] Bank account or service
Levels: Bank account or service
> unique(bankcomplaints$Sub.issue)
[1] NA
```

The columns excluded are:

- Product
    - Because this column has only 1 category as shown in the code block below, hence we can assume all complaints are of the same product category.
- Sub.Issue
    - Because all values for this column are Null as shown above therefore it is not useful for our purposes.

In pre-processing, no rows are removed from data because all rows are unique, meaning there are no duplicate data. Furthermore, column [Date.received] and [Date.sent.to.company] are set to be stored as date data type for our analysis' purposes with the code below.

```
#ensure date data are stored in date format

bankcomplaints$Date.received = as.Date(bankcomplaints$Date.received,format = "%d/%m/%Y")
bankcomplaints$Date.sent.to.company = as.Date(bankcomplaints$Date.received,format =
"%d/%m/%Y")
```

Complaint Parameter Analysis

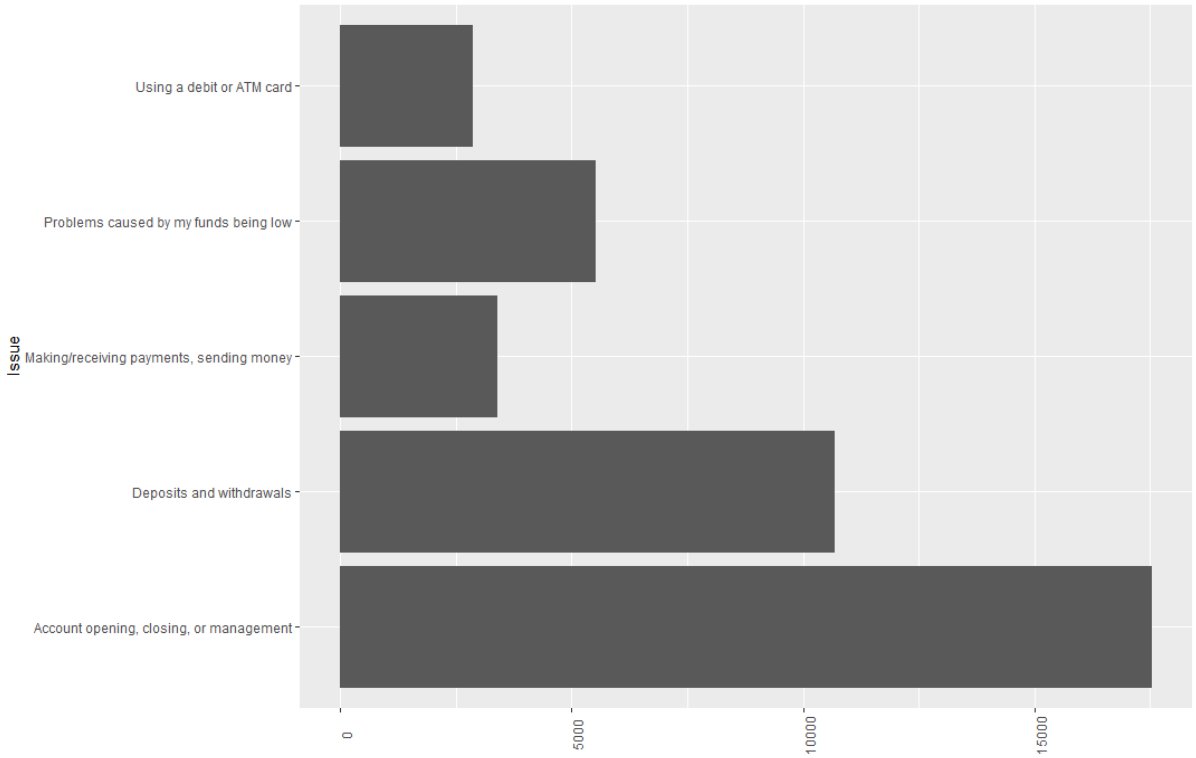
Notable analysis:

1. **Most common complaints with all companies included**

   a. **R-code:**

```
> #plot histogram for most common issues
> qplot(Issue, data = bankcomplaints)+ theme(axis.text.x = e
lement_text(angle = 90)) + coord_flip()
```

   b. **Graph:**



   c. **Analysis:**

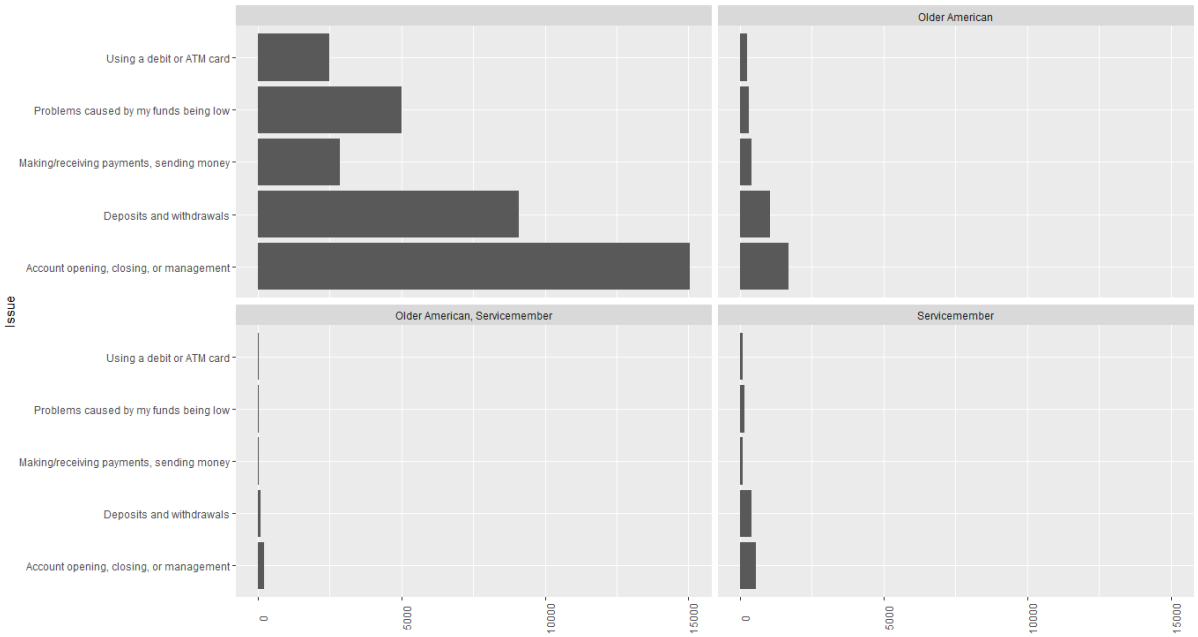
      i. As shown in the graph above, the most common customer complaints with all companies included are of “Account opening, closing and management”. With “Deposits and withdrawals” as the second most common issues, and with “problems caused by my funds being low” coming in third.

2. Histogram of issues, separated by customer tags

a. R-code

```
> #histogram separated by customer tags
> qplot(Issue, data = bankcomplaints, facets = Tags~.)+ face
t_wrap(~Tags) +theme(axis.text.x = element_text(angle = 90))
 + coord_flip()
```

b. Graph:



c. Analysis:

i. Older Americans customers seems to have less problems with "problems caused by my funds being low" compared to other customers. This is derived from R-code below:

```
> #perform chi squared test for being older ameri
can is associated with having less problems with
"funds being low", compared to non older americas
s
> chisq.test(chitestmatrix, correct=FALSE)

        Pearson's Chi-squared test

data:  chitestmatrix
X-squared = 59.935, df = 1, p-value = 9.806e-15
```

As seen from the p-value being less than 0.05, we reject the null hypothesis, and concluding that there is an association between being an Older American and having association with

less complaints related to the issue of “problems caused by my funds being low

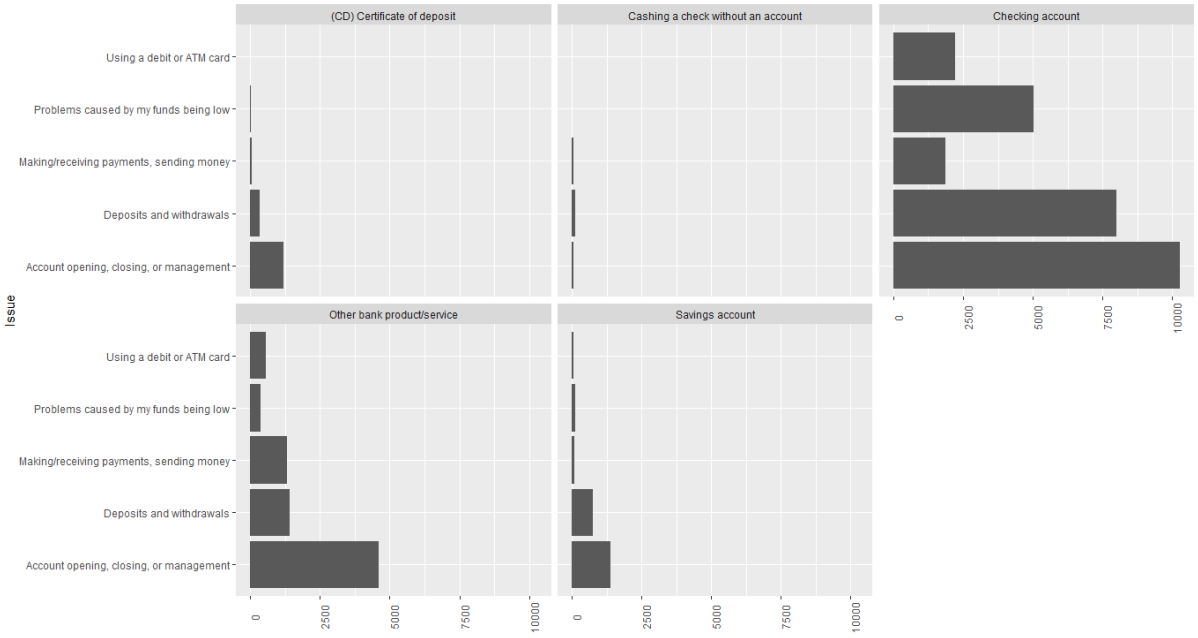
i. As shown in the histogram above, when compared with the histogram for all issues for all companies with all their customers, there is no apparent irregularity besides the histogram for Older Americans, customers with other tags all follow patterns for customer issues with all companies.
ii. No irregularities can be seen in the graph for “Older Americans, Servicemember” and “Servicemember”, may be due to because the sample size being too small, making patterns in the data to be not noticeable. A solution to this would be to increase sample size for making patterns more detectable.

3. Histogram of issues, separated by sub-product identified by complaint

a. R-code

```
> #histogram seperated by sub.product
> qplot(Issue, data = bankcomplaints, facets = Sub.product~
.)+ facet_wrap(~Sub.product) +theme(axis.text.x = element_t
ext(angle = 90)) + coord_flip()
```

b. Graph:



c. Analysis:

i. Complaints for "Other bank product/service" category of sub-product has particular low frequency with the issue "problems caused by funds being low", compared to all other complaints. This may be due to bank product and services costing customers little to no money.

ii. Other customer complaints separated by sub product seems to have no irregularity when compared to the histogram for all issues for all companies with all their products.

iii. Sample size for complaints with sub-product type "certificate of deposit" as well as complaints with sub-product type "Cashing a check without an account" seems to be small, which may be the cause of no particularly visible patterns being discoverable in these samples. Solution would be to gather more sample for complaints of these categories.

Activity and Complaint Over Time Analysis

Notable analysis:

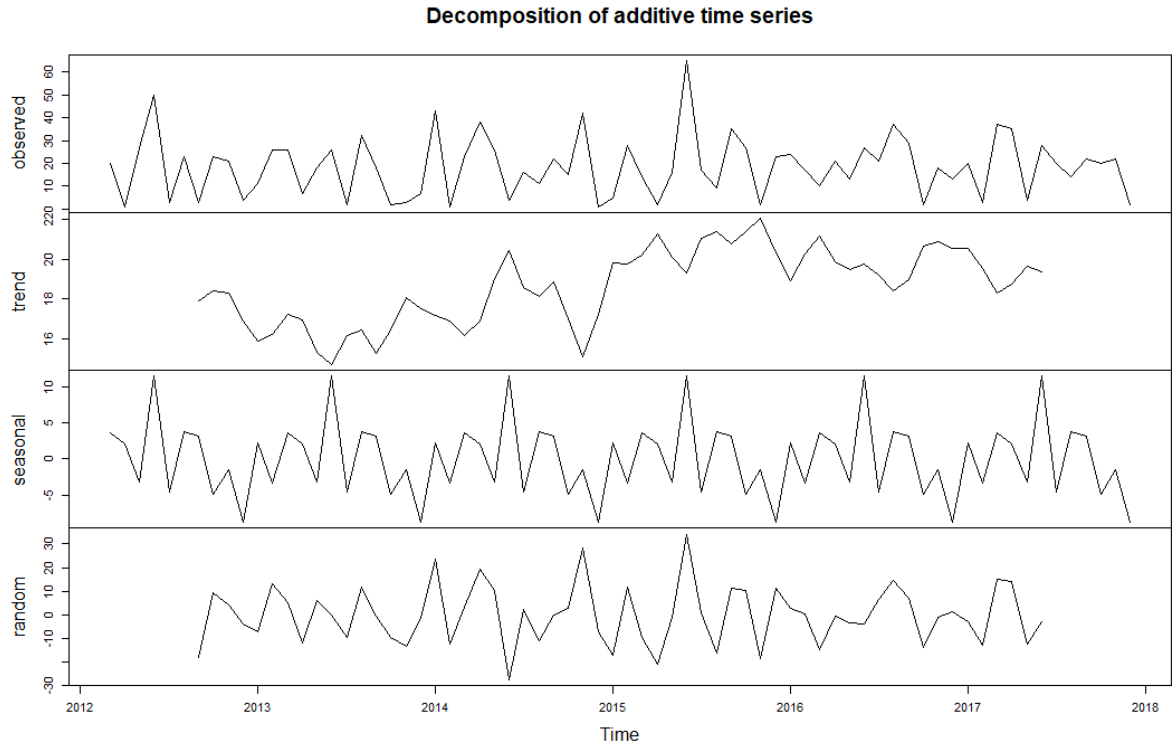
(Note: for more detailed R code to product the graphs below this, see R-codes in Appendix)

1. **Complaints Frequency regarding All Issue Over Time**

   a. **R-code**

```
> #overall time series for the whole data, for number of com
plaints over time
> vvv <- aggregate(bankcomplaints[c(17)],bankcomplaints[3],s
um)
> whole_issue_frequency <- ts(vvv$count, frequency = 12,star
t = c(2012,3) , end = c(2017,12))
> plot(decompose(whole_issue_frequency))
```

   b. **Graph:**



   c. **Analysis:**

i. The data sample starts on the March 2012, and ends December 2017, as shown in the graph above.
ii. Seasonally, number of customer complaints peak near the middle of the year, and there are 5 peaks every year.
iii. For trend, numbers of complaints steadily grow from 2012, until the third quarter of 2014, where it sharply drops. This sharp drop rises until the number of complaints reaches equilibrium on the first quarter of 2015 and does not significantly decrease or increase from first quarter of 2015 till December 2017, where our dataset's time span ends.

2. Analysis of frequency over time for complaints regarding each issue

Note: In this section, issues, also known as complaints, are as follows:

Issue:

1. Account opening, closing, or management
2. Deposits and withdrawals
3. Problems caused by my funds being low
4. Using a debit or ATM card
5. Making/receiving payments, sending money

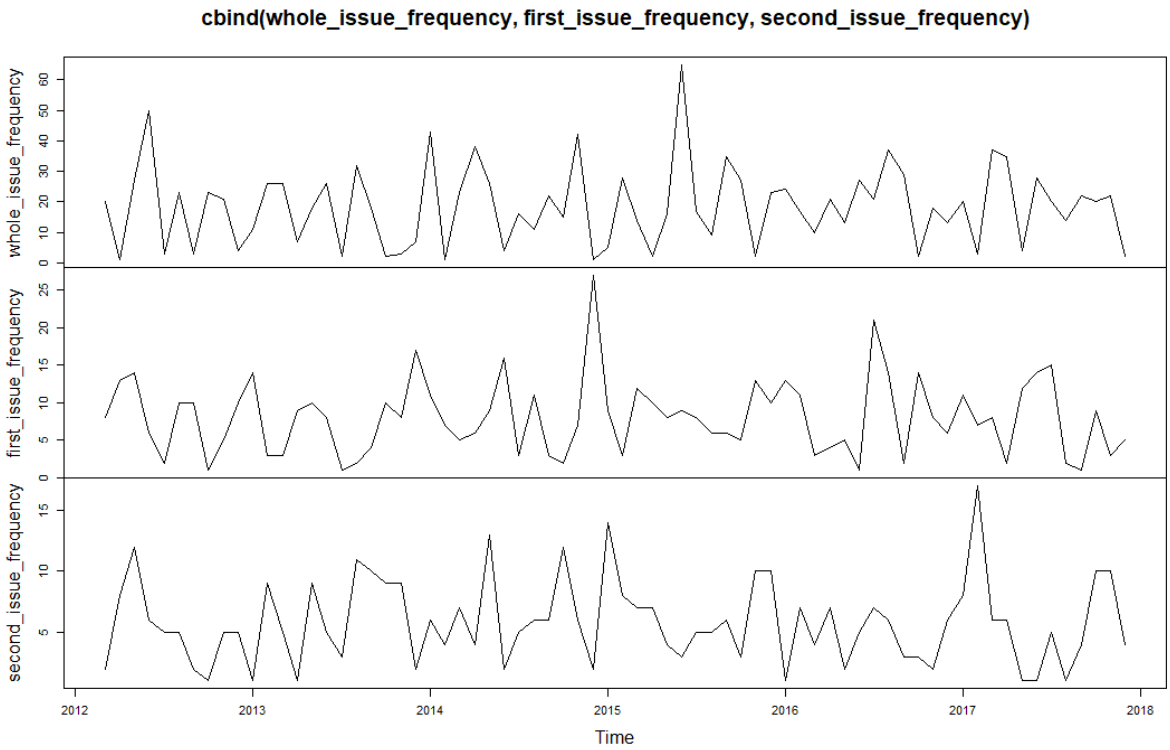
For example, issue one means it is a complaint regarding “Account opening, closing, or management”. Issue two means a complaint regarding “deposits and withdrawal”.

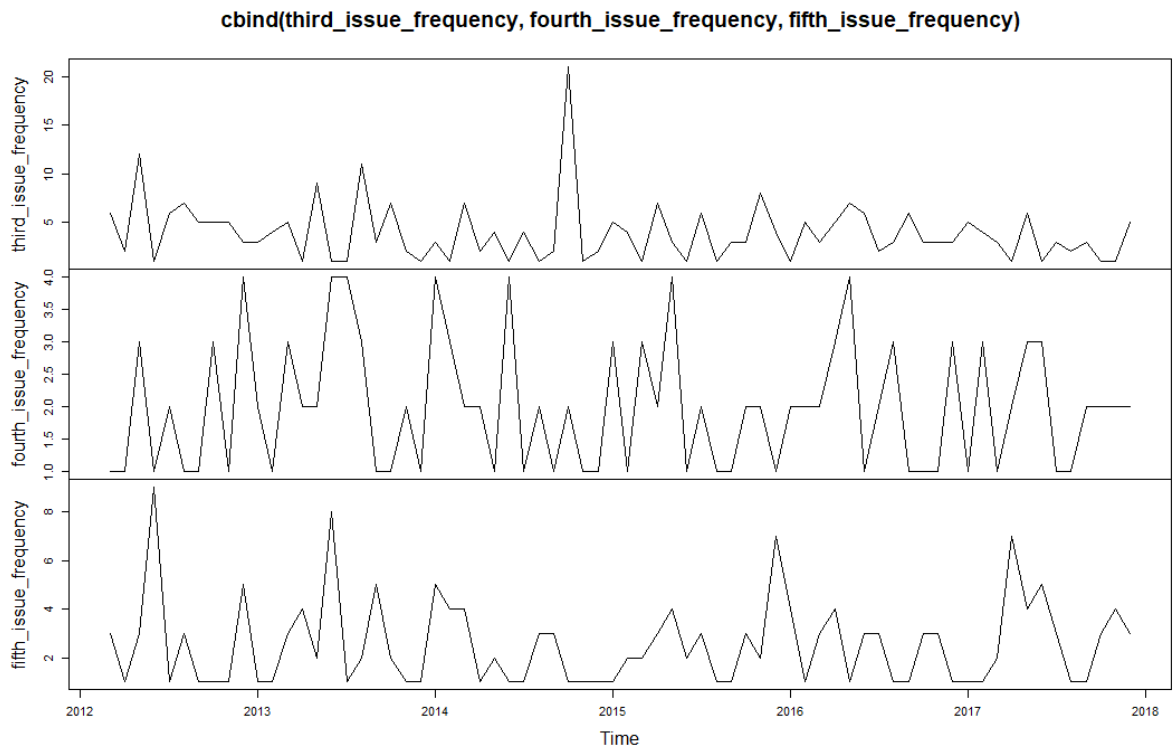
a. R-code

```
> #seprated in 3 and 3 plots, because 6 plots together is no
t very readable

> plot.ts(cbind(whole_issue_frequency, first_issue_frequency
, second_issue_frequency))
> plot.ts(cbind(third_issue_frequency, fourth_issue_frequenc
y,fifth_issue_frequency))
```

b. Graph:





c. Analysis:

iv. The number of customer complaints regarding the first issue and third issue seems to be steady throughout 2012 to 2017, but has a sharp increase, on the third quarter of 2014, which has then dropped back to normal levels by the first quarter of 2015. This may suggest a correlation between complaints regarding the first and third issue. Further data analysis regarding this possible correlation has not been done in this report, as the report is not meant to be too long, so it is suggested that further analysis be done should this relate to our main shareholder's interests.

v. There seems to be nothing notable about the number of complaints regarding the second issue as well as for the fourth issue throughout 2012 to 2017, the number of complaints for both seems to be steady with no significant spikes or drops throughout the dataset's time span.

vi. For the fifth issue, it is steady through 2012 to 2017, excluding the latter half of 2014, where the average number of complaints per month regarding this issue sharply drops, and this rises to normal levels again by the first quarter of 2015.

3. Heat map for the 5 companies with most complaints over time

Justification for analysing only top 5 companies with most customer complaints in heat map:

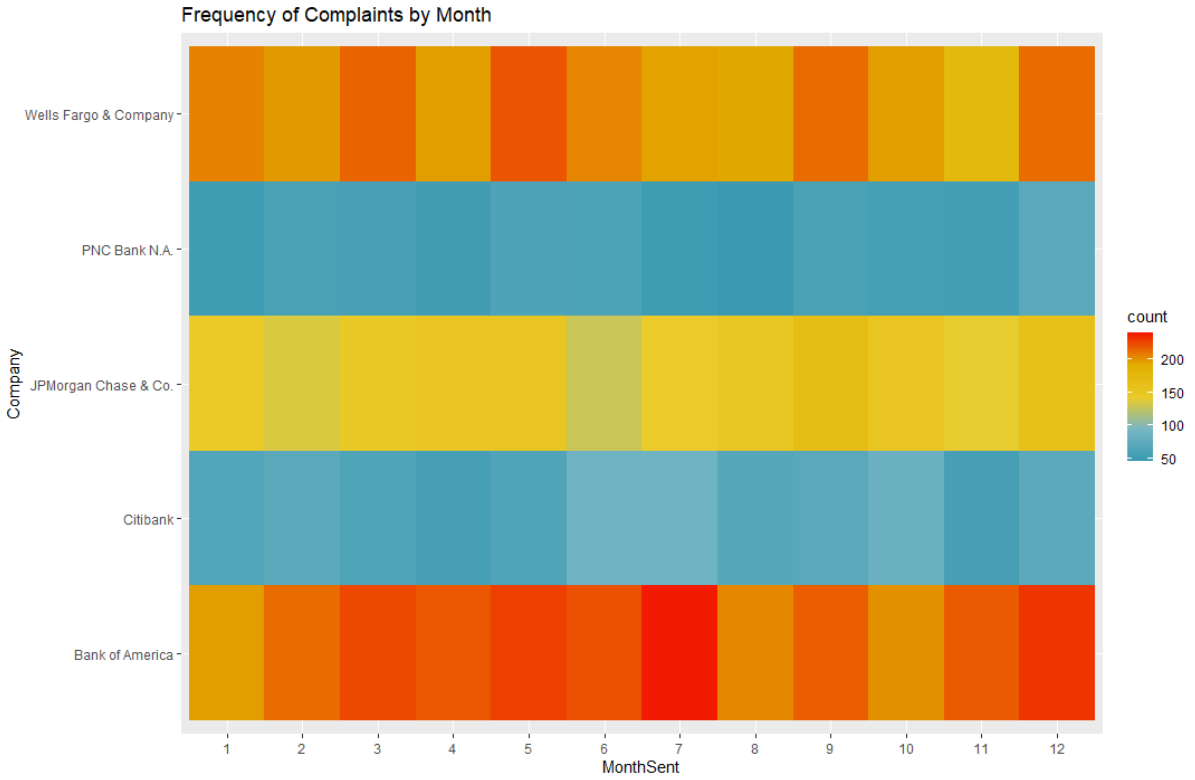
- There is hundreds of companies, including all of them would make the heat map unreadable.
- Top 10 companies heat map is not used, because it has been analysed and there is nothing notable in the 6th~10th ranked companies. All analysis and graphs included here includes only what is notable.

a. R-code

```
> #plot heat map
> g = ggplot (data = companyaggreteddata , aes (x = MonthSe
nt, y = Company))
> g = g + geom_tile(aes(fill = count))
> g = g + ggtitle("Frequency of Complaints by Month")
>
> pal <- wes_palette("Zissou1", 100, type = "continuous")
> g = g + scale_fill_gradientn(colours = pal)
> g
```

b. Graph

Note: The heat map shows the average number of complaints, in each month, for top 5 companies with the most complaints in our dataset. The average for each month, is the average of the respective month, for the years 2012 to 2017 which our dataset spans.



c. Analysis:

vii. Bank of America has the highest average number of customer complaints, during July.

viii. Citibank and PNC Bank NA has comparatively lower number of complaints throughout the year, compared to Wells Fargo & Company, JPMorgan Chase & Co, and Bank of America.

ix. Wells Fargo & Company seems to have larger number of complaints on average during March and May, compared to its other months. For its other months, the number of complaints they receives seems to be on similar levels.

x. PNC Bank NA, JPMorgan Chase & Co, and Citibank seems to have steady levels of average customer complaints throughout the year.

xi. Bank of America seems to have peaks for average number of customer complaints during March, May, July and December, with the highest peak during July. The average number of customer complaints for other months seems to be on normal levels.

Appendix

1. Assumptions:

- All data are under the product category of “Bank account or service, because

```
> unique(bankcomplaints$Product)
[1] Bank account or service
Levels: Bank account or service
> #Result: only 1 unique category shown hence can be excluded
```

2. R-code

```
rm(list = ls())



library(ggplot2)
library(lubridate)
library(wesanderson)
#setwd('C:/Users/User/Downloads/FIT3152/Assignment 1')



#draw custom sample of 40,000 rows from data set for analysis
set.seed(28055322) # 28055322 = your student ID
bankcomplaints <-read.csv("bankcomplaints.csv")
bankcomplaints <-bankcomplaints [sample(nrow(bankcomplaints), 40000), ] # 40000 rows



################      Part a - Pre-processing      ######################

### DATA QUALITY CHECK
#*Note that only notable checks are shown here,
#*Note checks that show no irregularity, or where the column may be useful and so are included are not shown here.

#check for whether it can be excluded
```

```
unique(bankcomplaints$Product)

#Result: only 1 unique category shown hence can be excluded


unique(bankcomplaints$Sub.issue)

#Result: only Null value is shown, excluding this row



#check for redundant//dupliate data, to ensure data quality

unique(bankcomplaints)

#Result: all rows are unique, data quality in this aspect is good and hence no data transformation are needed



### DATA TRANSFORMATION, AND EXTRACTION



#exclude column 2 and 5, which are colums for [product] and [sub.issue]

#column  3,6,7,11,12,15, may or may not be useful and hece are placed at the back of the dataset

bankcomplaints = bankcomplaints[c(18, 1, 14, 4, 8, 9, 10, 13, 16, 17, 3, 6, 7,11, 12, 15)]


#ensure date data are stored in date format

bankcomplaints$Date.received = as.Date(bankcomplaints$Date.received,format = "%d/%m/%Y")

bankcomplaints$Date.sent.to.company = as.Date(bankcomplaints$Date.received,format = "%d/%m/%Y")

#################################################################
```

```
################      Part b - Complaint Parameter Analysis      ######################

#list of unique issues

issuesunique <- data.frame(unique(bankcomplaints$Issue))



#plot histogram for most common issues with all companies included

qplot(Issue, data = bankcomplaints)+ theme(axis.text.x = element_text(angle = 90)) + coord_flip()

#histogram separated by customer tags

qplot(Issue, data = bankcomplaints, facets = Tags~.)+ facet_wrap(~Tags) +theme(axis.text.x =
element_text(angle = 90)) + coord_flip()



#statistical test: [Older Americans] have less issue with "problems with my funds being low"

olderamericans = bankcomplaints[(bankcomplaints$Tag == 'Older American'),]

olderamericansfundslow= olderamericans[(olderamericans$Issue == "Problems caused by my funds
being low"),]

countolderamericans = nrow(olderamericans)

countolderamericansfundslow = nrow(olderamericansfundslow)

countolderamericanswithnolowfundsissue = countolderamericans - countolderamericansfundslow
```

```
countnonoldamericans = nrow(bankcomplaints) - row(olderamericans)

datawithissueoflowfund= bankcomplaints[(bankcomplaints$Issue == "Problems caused by my funds being low"),]

countnonoldamericanswithlowfunds = nrow(datawithissueoflowfund) - countolderamericansfundslow

countnonoldamericanswithnolowfundsissue = countnonoldamericans - countnonoldamericanswithlowfunds


chitestmatrix = matrix(c(countolderamericansfundslow, countnonoldamericanswithlowfunds,countolderamericanswithnolowfundsissue,countnonoldamericanswithnolowfundsissue),nrow =2, ncol =2)

#perform chi squared test for being older american is associated with having less problems with "funds being low", compared to non-older americams

chisq.test(chitestmatrix, correct=FALSE)



#histogram seperated by sub.product

qplot(Issue, data = bankcomplaints, facets = Sub.product~.)+ facet_wrap(~Sub.product) +theme(axis.text.x = element_text(angle = 90)) + coord_flip()



#####################################################################


################       Part c - Activity and Complaint Over Time Analysis
######################


#initialize, for counting frequency of the issue, each row has a count of 1 complaint, hence each row are set as 1

bankcomplaints$count = 1


#get unique issues from dataset

issueunique <-data.frame(unique(bankcomplaints$Issue))
```

```
#seperate dataset into different variables, grouped by issues
issueone <- bankcomplaints[(bankcomplaints$Issue==issuesunique[1,1]),]
issuetwo <- bankcomplaints[(bankcomplaints$Issue==issuesunique[2,1]),]
issuethree <- bankcomplaints[(bankcomplaints$Issue==issuesunique[3,1]),]
issuefour <- bankcomplaints[(bankcomplaints$Issue==issuesunique[4,1]),]
issuefive <- bankcomplaints[(bankcomplaints$Issue==issuesunique[5,1]),]



###get earliest and latest start date for plotting time series

timereceviedmax <- bankcomplaints[which.max(bankcomplaints$Date.received),]
timereceviedmax <- timereceviedmax$Date.received
timereceviedmax


timereceviedmin <- bankcomplaints[which.min(bankcomplaints$Date.received),]
timereceviedmin <- timereceviedmin$Date.received
timereceviedmin



###plot time series graph for all issues
#NOTE: because we are not using only using time series with "ts()" for the whole complaint data, the other 5 issue's plot commands are commented out.
#NOTE: time used here is based on the date complaint is received

#overall time series for the whole data, for number of complaints over time
vvv <- aggregate(bankcomplaints[c(17)],bankcomplaints[3],sum)
```

```
whole_issue_frequency <- ts(vvv$count, frequency = 12,start = c(2012,3) , end = c(2017,12))
plot(decompose(whole_issue_frequency))



#issue 1
v <- aggregate(issueone[c(17)],issueone[3],sum)
first_issue_frequency <- ts(v$count, frequency = 12, start = c(2012,3), end = c(2017,12))
#plot(decompose(first_issue_frequency))


#issue 2
vtwo <- aggregate(issuetwo[c(17)],issuetwo[3],sum)
second_issue_frequency <- ts(vtwo$count, frequency = 12,start = c(2012,3), end = c(2017,12))
#plot(decompose(second_issue_frequency))


#issue 3
vthree <- aggregate(issuethree[c(17)],issuethree[3],sum)
third_issue_frequency <- ts(vthree$count, frequency = 12,start = c(2012,3), end = c(2017,12))
#plot(decompose(third_issue_frequency))


#issue 4
vfour <- aggregate(issuefour[c(17)],issuefour[3],sum)
fourth_issue_frequency <- ts(vfour$count, frequency = 12,start = c(2012,3), end = c(2017,12))
#plot(decompose(fourth_issue_frequency))
```

```
#issue 5

vfive <- aggregate(issuefive[c(17)],issuefive[3],sum)

fifth_issue_frequency <- ts(vfive$count, frequency = 12,start = c(2012,3), end = c(2017,12))

#plot(decompose(fifth_issue_frequency))









#plot.ts(cbind(whole_issue_frequency, first_issue_frequency, second_issue_frequency,
third_issue_frequency, fourth_issue_frequency,fifth_issue_frequency))

#seprated in 3 and 3 plots, because 6 plots together is not very readable

plot.ts(cbind(whole_issue_frequency, first_issue_frequency, second_issue_frequency))

plot.ts(cbind(third_issue_frequency, fourth_issue_frequency,fifth_issue_frequency))



#################################################################



################  Part c latter part- 1.  Heat map for the 5 companies with most complaints
over time  ######################



bankcomplaints$MonthSent = month(bankcomplaints$Date.sent.to.company)



#choose top 5 companies with most complaints for heat map

#top 5 is not chosen because top 10 is not notable

numberoftopchosen <- 5

choosingcompaniesforheatmap = aggregate(bankcomplaints[c(17)],bankcomplaints[5],sum)

choosingcompaniesforheatmap$Rank = rank(-choosingcompaniesforheatmap$count)
```

```
chosencompanies =
head(choosingcompaniesforheatmap[order(choosingcompaniesforheatmap$Rank),],numberoftopch
osen)


#select only the top 10 companies with most complaints, from original dataset, for analysis

analysingcompanies = bankcomplaints[(bankcomplaints$Company %in%
chosencompanies$Company ),]


#extract company, month, and count

extracteddata = analysingcompanies[c(5, 18, 17)]




#count frequency for each month for each company, regardless of year

attach(extracteddata)

companyaggreteddata = as.data.frame(as.table(by(count, list(Company, MonthSent), sum)))

detach(extracteddata)


colnames(companyaggreteddata) = c("Company", "MonthSent", "count")


#remove null rows

#companyaggreteddata <- companyaggreteddata[(!is.null(companyaggreteddata$count))]

companyaggreteddata <- companyaggreteddata[(companyaggreteddata$Company %in%
chosencompanies$Company),]



#plot heat map

g = ggplot (data = companyaggreteddata , aes (x = MonthSent, y = Company))

g = g + geom_tile(aes(fill = count))

g = g + ggtitle("Frequency of Complaints by Month")
```

```
pal <- wes_palette("Zissou1", 100, type = "continuous")

g = g + scale_fill_gradientn(colours = pal)

g
```

##