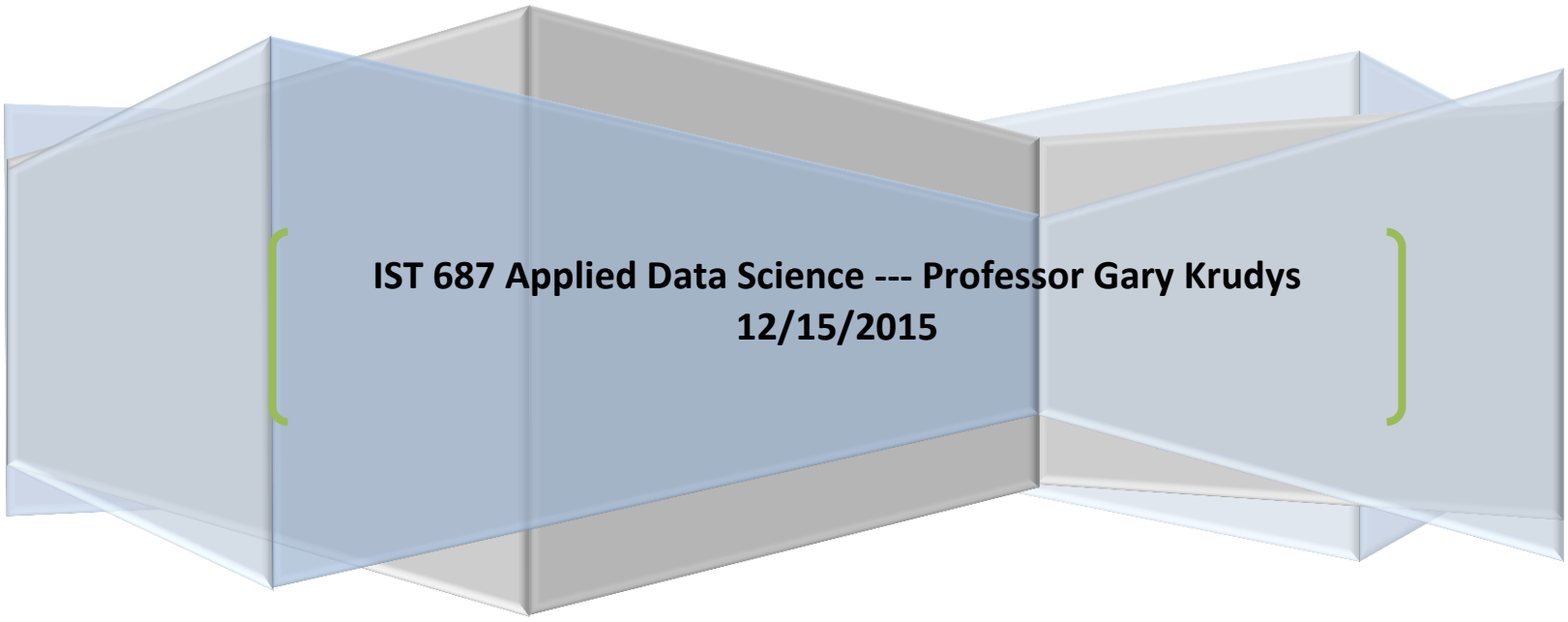
Analysis on Consumer Complaints about Financial Services

Final Group Project by

Matt Zhang, Lawrence Hamblin, Daniel Hanks Jr

School of Information Studies
SYRACUSE UNIVERSITY



IST 687 Applied Data Science --- Professor Gary Krudys
12/15/2015

Table of Contents

1. Business Questions

The Consumer Financial Protection Bureau (CFPB) started to accept and manage consumer complaints regarding financial products in 2011 due to the numerous issues related to financial products. It's the first federal agency that mainly focuses on consumer financial projection (CFPB, 2012).

Consumers can file complaints about certain financial products in a variety of ways. This includes the CFPB website, email, fax, phone, postal mail and referral channels. The workflow that the CFPB uses once it receives a complaint is fairly straightforward. The CFPB received the complaint, the bureau then reviews the complaint and forwards the complaints to the corresponding financial products companies. The bureau then lets the company review and respond to the consumers. Once the consumer received the company response, the consumer and the bureau both review the resolution process and result of the consumer complaints. Once the consumer complaint has been properly handled and reviewed by the bureau, the complaints will be stored and published in the Consumer Complaint database for the future and further analysis. This database includes the following information for each consumer complaint:

Complaint ID, Product, Sub-Product, Issue, Sub-issue, State, Zip Code, Submitted via, Date Received, Date sent to company, Company, Company response, Timely response? (Yes or No), Consumer dispute (Yes or No).

With the above collected data, we want explore the following questions about consumer complaints about financial products:

1. What are the Top 10 consumer complaints about (percentage wise)?
2. Based on the 3 most interesting rules you can find from this Consumer Complaint Database, what are the recommendations you would provide to companies or consumers to improve the customer relationship?
3. What is the complaint ratio about financial service across United States?
4. Do the characteristics of complaints in this database make it useful for predicting complaints over time?

2. Process Flow Components

Data screening, linking, cleansing

Data acquisition

We found the Consumer Complaint Database called Consumer_Complaints.csv file comes from --- http://catalog.data.gov/dataset/consumer-complaint-database. This database covers consumer complaints about financial products and services that the Consumer Financial Protection Bureau (CFPB) have received from consumers from year 2011 to 2015.
The consumer complaint database is in a comma-separated values file. This can be imported into R with the read.csv() command as follows:

complaintData <- read.csv("~/Consumer_Complaints_original_20151128.csv", header = TRUE, sep = ",")

The header parameter accounts for the column labels in the first row of the file, and the sep parameter indicates that values are separated by commas. The result is a data frame with 16 variables (columns) and 486,343 observations (rows).

Data architecture

Linking/structuring: All of the data to be used is contained in a single file, so there is no need to link it to any other data sets.

Quality/completeness: Some of the columns will not be used for analysis due to missing or incomplete data.

1. The "**consumer complaint narrative**" column is very sparsely populated relative to the other columns, suggesting that it was an optional item in a web form (or that it did not apply, in the case of complaints submitted via alternative methods, such as by phone). Some of the narratives are also duplicates, indicating that a consumer may have submitted the same complaint verbatim several times. Furthermore, as the narrative is free-form text rather than one of a number of discrete values, analyzing its contents would be more suited to human eyes than to the capabilities of R.
2. The ZIP codes are in an inconsistent format due to having been anonymized. Along with usable five-digit ZIP codes, some of the codes only contain four digits, and others have three with the last two digits replaced with a capital X. Since all

five digits are needed to plot a precise geographic location, the **ZIP code** data will be ignored in favor of state data.

3. **Dataset 'hierarchy' or sequence:** for easy data mining and data analysis, a dataset hierarchy will be created with many sub-datasets and data frames. For example:

   complaintData <- read.csv("~/Consumer_Complaints_original_20151128.csv", header = TRUE, sep = ",") # this is the original dataset.

   Issue <- complaintData$Issue # this is a new sub-dataset for the "Issue" variable.

   Sub.issue <- complaintData$Sub.issue # this is a new variable for all the sub issues.

4. **The final data architecture** will look like an hierarchical structure consisting both summary data sets and other dataset details, such as vectors, dataframes, data structures, functions and R packages.

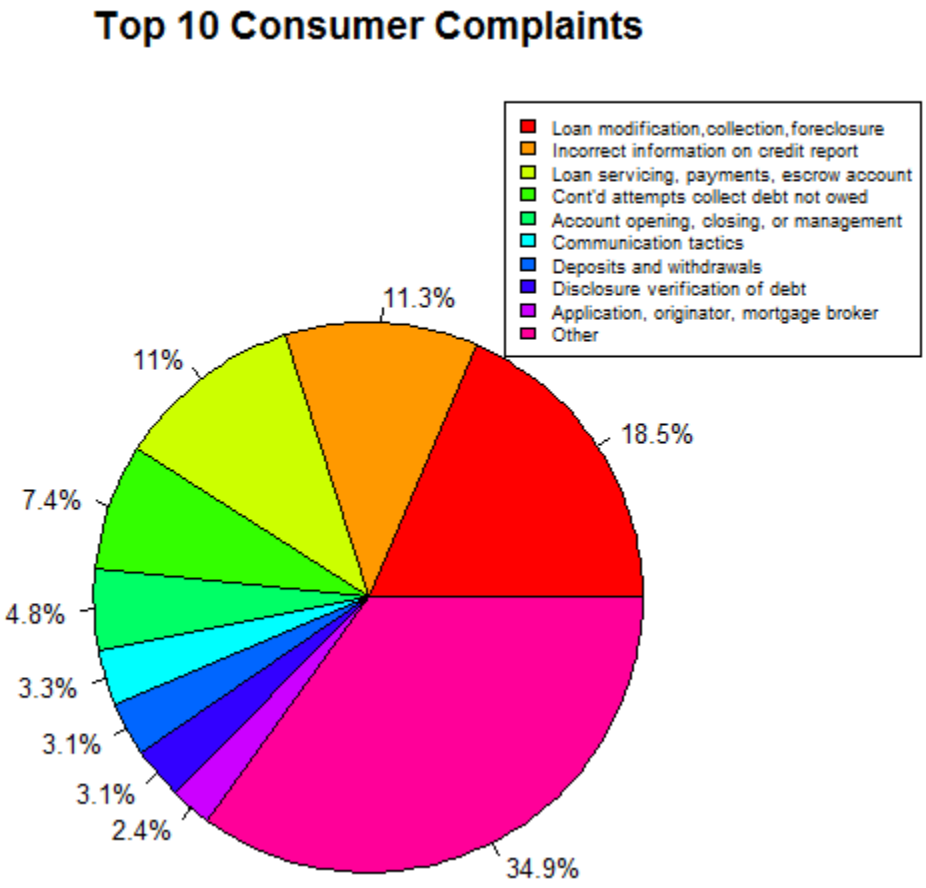
3. Presentation/Visualization & Assessment/Interpretation

Q1. What are the Top 10 consumer complaints about (percentage wise)?

summary(consumerdata$Issue, maxsum=10)

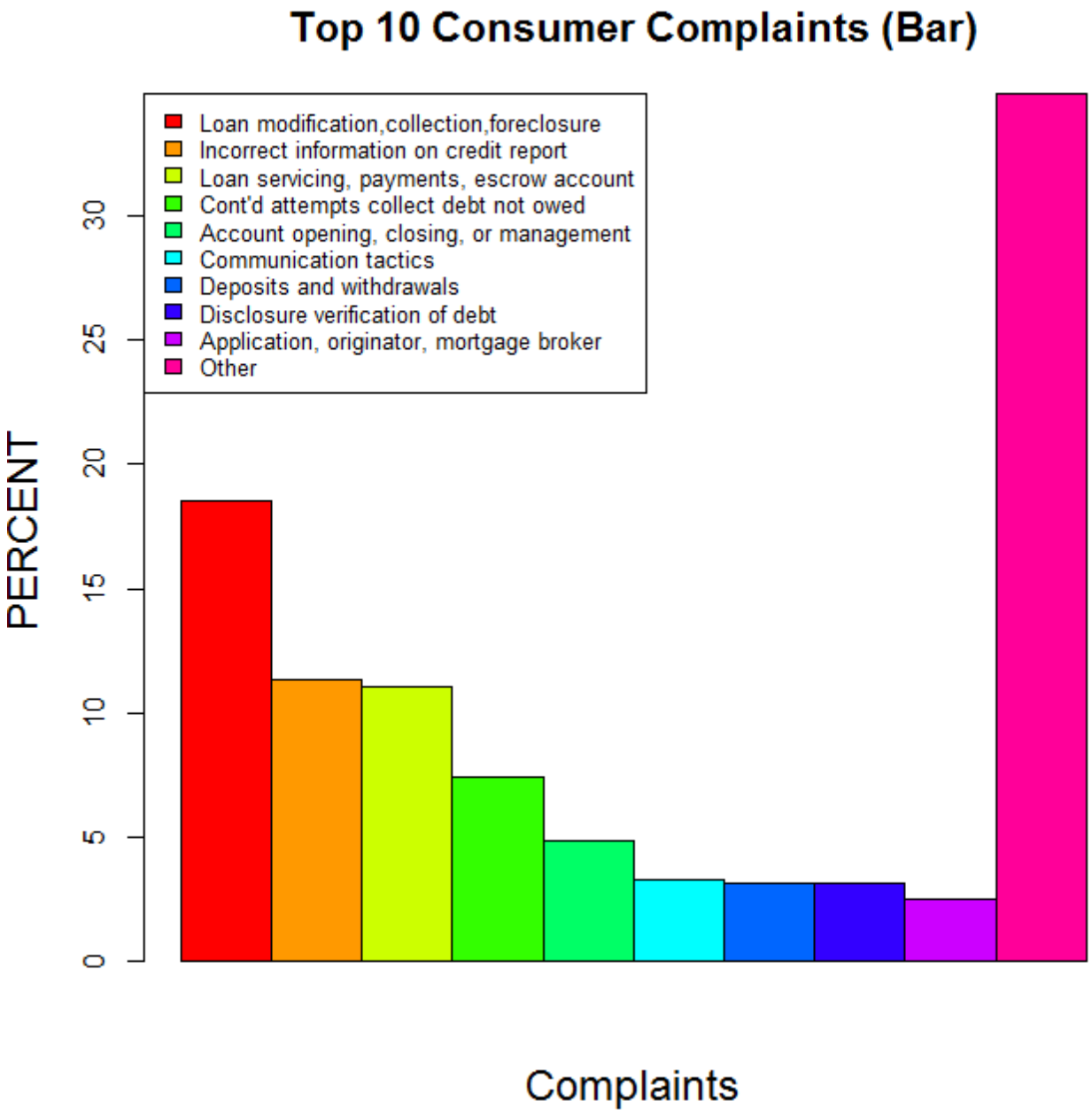
```
> summary(consumerdata$Issue, maxsum=10)
Loan modification,collection,foreclosure
                                   90208
  Incorrect information on credit report
                                   55130
Loan servicing, payments, escrow account
                                   53597
   Cont'd attempts collect debt not owed
                                   35910
 Account opening, closing, or management
                                   23477
                  Communication tactics
                                   15889
              Deposits and withdrawals
                                   15175
        Disclosure verification of debt
                                   15158
Application, originator, mortgage broker
                                   11907
                                (Other)
                                  169892
```

```
#convert count to percent
percentlabels <- round(100*topcomplaints/sum(topcomplaints), 1)
#create Percent labels
pielabels <- paste(percentlabels, "%", sep="")
#Create pie chart
pie(topcomplaints, main="Top 10 Consumer Complaints", labels=pielabels,
col=rainbow(length(topcomplaints)), cex=0.8, radius=.6)
#create legend
legend("topright",c("Loan modification,collection,foreclosure","Incorrect information on credit
report","Loan servicing, payments, escrow account","Cont'd attempts collect debt not owed","Account
opening, closing, or management","Communication tactics","Deposits and withdrawals","Disclosure
verification of debt","Application, originator, mortgage
broker","Other"),cex=0.6,fill=rainbow(length(topcomplaints)))
```



#Create bar graph and convert count into percentage

```
barplot(prop.table(as.matrix(topcomplaints))*100,2,main="Top 10 Consumer Complaints (Bar)",
ylab="PERCENT",xlab="COMPLAINTS", cex.lab=1.5, cex.main=1.4, beside=TRUE, col=colors)
#create legend
legend("topleft",c("Loan modification,collection,foreclosure","Incorrect information on credit
report","Loan servicing, payments, escrow account","Cont'd attempts collect debt not owed","Account
opening, closing, or management","Communication tactics","Deposits and withdrawals","Disclosure
verification of debt","Application, originator, mortgage
broker","Other"),cex=0.8,fill=rainbow(length(topcomplaints)))
```



As shown by the charts above, Loan modification, collection and foreclosure was the top ranked, identified issue with a little over 18% of the complaints.

Q2. Based on the 3 most interesting rules you can find from this Consumer Complaint Database, what are the recommendations you would provide to companies or consumers to improve the customer relationship?

To answer this question we must first define what makes a rule "interesting" and understand its importance. An interesting rule is considered a more useful rule because it's unexpected. The lift is one way to measure this. Lift takes into account the support for a rule while giving more weight to rules where the LHA and/or the RHS occur less frequently. So the larger the value of the lift, the more "interesting" the rule is considered.

```
#ccData is simply the consumer_complaints spreadsheet.  I did delete column 16, which was simply
#Complaint ID. ( Not needed)
rules <- apriori(ccData, parameter=list(support=0.005, confidence=0.8))

Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE   0.005      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2431

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[70053 item(s), 486343 transaction(s)] done [0.67s].
sorting and recoding items ... [161 item(s)] done [0.07s].
creating transaction tree ... done [0.46s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [1.87s].
writing ... [176630 rule(s)] done [0.03s].
creating S4 object  ... done [0.15s].
set of 176630 rules

#obviously too many rules to sort through so we make a subset of the larger set of rules by choosing
#rules with a high lift value.
goodrules <- rules[quality(rules)$lift > 80.0]
inspect(goodrules)
```

The highest lifts (Top 16) are shown below in Figure 1. As you can see, they all have to deal with improper use of credit report. Finding the lift has found a useful answer to this question since the sub-issues go into detail on what consumer's complaints were with the use of their credit reports. Our first recommendation would be to put some form of policy in place that prohibits the sharing of a customer's credit report without the customer's authorization and communicating this policy to the customer.

```
 1  {Sub.issue=Report improperly shared by CRC} => {Issue=Improper use of my credit report}    0.005228820  1.0000000 177.1096
 2  {Issue=Improper use of my credit report}    => {Sub.issue=Report improperly shared by CRC} 0.005228820  0.9260743 177.1096
 3  {Product=Credit reporting,
     Sub.issue=Report improperly shared by CRC} => {Issue=Improper use of my credit report}    0.005228820  1.0000000 177.1096
 4  {Product=Credit reporting,
     Issue=Improper use of my credit report}    => {Sub.issue=Report improperly shared by CRC} 0.005228820  0.9260743 177.1096
 5  {Sub.product=,
     Sub.issue=Report improperly shared by CRC} => {Issue=Improper use of my credit report}    0.005228820  1.0000000 177.1096
 6  {Sub.product=,
     Issue=Improper use of my credit report}    => {Sub.issue=Report improperly shared by CRC} 0.005228820  0.9260743 177.1096
 7  {Sub.issue=Report improperly shared by CRC,
     Timely.response.=Yes}                      => {Issue=Improper use of my credit report}    0.005193865  1.0000000 177.1096
 8  {Issue=Improper use of my credit report,
     Timely.response.=Yes}                      => {Sub.issue=Report improperly shared by CRC} 0.005193865  0.9259531 177.0864
 9  {Product=Credit reporting,
     Sub.product=,
     Sub.issue=Report improperly shared by CRC} => {Issue=Improper use of my credit report}    0.005228820  1.0000000 177.1096
10 {Product=Credit reporting,
     Sub.product=,
     Issue=Improper use of my credit report}    => {Sub.issue=Report improperly shared by CRC} 0.005228820  0.9260743 177.1096
11 {Product=Credit reporting,
     Sub.issue=Report improperly shared by CRC,
     Timely.response.=Yes}                      => {Issue=Improper use of my credit report}    0.005193865  1.0000000 177.1096
12 {Product=Credit reporting,
     Issue=Improper use of my credit report,
     Timely.response.=Yes}                      => {Sub.issue=Report improperly shared by CRC} 0.005193865  0.9259531 177.0864
13 {Sub.product=,
     Sub.issue=Report improperly shared by CRC,
     Timely.response.=Yes}                      => {Issue=Improper use of my credit report}    0.005193865  1.0000000 177.1096
14 {Sub.product=,
     Issue=Improper use of my credit report,
     Timely.response.=Yes}                      => {Sub.issue=Report improperly shared by CRC} 0.005193865  0.9259531 177.0864
15 {Product=Credit reporting,
     Sub.product=,
     Sub.issue=Report improperly shared by CRC,
     Timely.response.=Yes}                      => {Issue=Improper use of my credit report}    0.005193865  1.0000000 177.1096
16 {Product=Credit reporting,
     Sub.product=,
     Issue=Improper use of my credit report,
     Timely.response.=Yes}                      => {Sub.issue=Report improperly shared by CRC} 0.005193865  0.9259531 177.0864
```

Figure 1

We have to dig a little further to find more interesting rules. We can do this by reducing the requirements of the lift.

```
goodrules <- rules[quality(rules)$lift > 75.0]
inspect(goodrules)
```

As shown below in Figure 2, just dropping the lift slightly has lead us to another answer.

```
21 {Product=Debt collection,
    Sub.issue=Talked to a third party about my debt,
    Submitted.via=Web}                              => {Issue=Improper contact or sharing of info} 0.005611266  1.0000000  76.20542
22 {Product=Debt collection,
    Sub.issue=Talked to a third party about my debt,
    Company.public.response=}                       => {Issue=Improper contact or sharing of info} 0.005442661  1.0000000  76.20542
23 {Product=Debt collection,
    Sub.issue=Talked to a third party about my debt,
    Consumer.complaint.narrative=}                  => {Issue=Improper contact or sharing of info} 0.005417987  1.0000000  76.20542
24 {Product=Debt collection,
    Sub.issue=Talked to a third party about my debt,
    Timely.response.=Yes}                           => {Issue=Improper contact or sharing of info} 0.005712018  1.0000000  76.20542
25 {Sub.issue=Talked to a third party about my debt,
    Submitted.via=Web,
    Timely.response.=Yes}                           => {Issue=Improper contact or sharing of info} 0.005183584  1.0000000  76.20542
26 {Sub.issue=Talked to a third party about my debt,
    Consumer.complaint.narrative=,
    Company.public.response=}                       => {Issue=Improper contact or sharing of info} 0.005019091  1.0000000  76.20542
27 {Sub.issue=Talked to a third party about my debt,
    Company.public.response=,
    Timely.response.=Yes}                           => {Issue=Improper contact or sharing of info} 0.005002642  1.0000000  76.20542
```

Figure 2

This complaint doesn't come as a surprise. The improper sharing of people's information has always made people hesitant to give out information to companies. Similar to the findings above, prohibiting the selling or distribution of consumer's information and communicating it to the consumer would reduce complaints. This would also build trust between the consumer and company.

Once again, we lower the lift number to find more information.

```
goodrules <- rules[quality(rules)$lift > 70.0]
inspect(goodrules)
```

```
375 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Consumer.complaint.narrative=,
     Company.public.response=,
     Submitted.via=Web,
     Company.response.to.consumer=Closed with explanation} => {Issue=False statements or representation}       0.005506402  1.0000000  69.10244
376 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Company.public.response=,
     Submitted.via=Web,
     Company.response.to.consumer=Closed with explanation,
     Timely.response.=Yes}                          => {Issue=False statements or representation}       0.005858006  1.0000000  69.10244
377 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Consumer.complaint.narrative=,
     Submitted.via=Web,
     Company.response.to.consumer=Closed with explanation,
     Timely.response.=Yes}                          => {Issue=False statements or representation}       0.005559862  1.0000000  69.10244
378 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Company.public.response=,
     Submitted.via=Web,
     Timely.response.=Yes,
     Consumer.disputed.=No}                         => {Issue=False statements or representation}       0.005276112  1.0000000  69.10244
379 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Consumer.complaint.narrative=,
     Submitted.via=Web,
     Timely.response.=Yes,
     Consumer.disputed.=No}                         => {Issue=False statements or representation}       0.005002642  1.0000000  69.10244
380 {Product=Debt collection,
     Sub.issue=Attempted to collect wrong amount,
     Consumer.complaint.narrative=,
     Company.public.response=,
     Submitted.via=Web,
     Timely.response.=Yes}                          => {Issue=False statements or representation}       0.006655796  1.0000000  69.10244
```

Figure 3

As shown above in Figure 3, the next major issue with high lift value is “False statements or representation”. The sub-issue is debt collectors trying collect the wrong amount. This is an issue among not only debt collectors, but banks and student loan companies. They need to put more effort into verifying the debt before going after it. Prove the debt to the consumer, communicate the debt through letters, email and finally phone calls so the consumer understands it’s a valid debt they are responsible for. This is an issue that gives debt collectors a bad reputation. Validating this debt and reducing the complaints on this issue would go a long way improving relationships with customers.

Conclusion

Using apriori algorithm to mine this dataset we identified 3 issues that must be addressed for companies to improve relations with its’ customers. They were as follows:

1. Improper use of credit report (Sharing credit report information)
2. Improper sharing or use of information (Customers personal information)
3. False statements or representation (Collecting wrong amount due)

Solutions to the first two issues could be as simple as policies that restrict a company from sharing credit report information or customer’s personal information (email, phone number, address, etc.). Communicating these policies to the customer is critical to gain the trust of them and reduce complaints. The final issue is a matter of verifying the debt consumers have and communicating it to them.

Q3. What is the complaint ratio about financial service across the United States?

Mash- up

The R code used (1):

```
> summary(consumerdata.map2)
    region              value
 Length:51          Min.   :  399
 Class :character   1st Qu.: 2013
 Mode  :character   Median : 5325
                    Mean   : 9381
                    3rd Qu.:11825
                    Max.   :71497

> View(consumerdata.map2)
```

Min = North Dakota (399); Max = California (71497)

Input

```
# Consumer complaint ratio
min.issue=min(consumerdata.map2$value)
max.isse=max(consumerdata.map2$value)
min(consumerdata.map2$value);max(consumerdata.map2$value)
sum.issue=sum(consumerdata.map2$value)
issue.ratio <- round((consumerdata.map2$value)/sum(consumerdata.map2$value),4)
issue.ratio
issue.ration.percent <- issue.ratio*100
issue.ration.percent
summary(issue.ration.percent)
> min(consumerdata.map2$value);max(consumerdata.map2$value)
[1] 399
[1] 71497

> issue.ration.percent
 [1]  1.02  0.43  0.12  2.25 14.94  1.73  1.19  0.59  0.54  9.76  4.44  0.41
 0.35
[14]  3.58  1.11  0.48  0.72  0.97  2.04  3.22  0.36  2.69  1.17  1.33  0.47
 0.16
[27]  2.82  0.08  0.32  0.53  4.10  0.49  1.22  6.94  3.16  0.66  0.41  1.20
 3.64
[40]  0.35  1.26  0.14  1.52  7.48  0.54  3.28  0.18  2.10  1.13  0.26  0.10

> summary(issue.ration.percent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.08    0.42    1.11    1.96    2.47   14.94
```

Analysis:

To answer the complaint ratio question about the consumer complaints across the United States, we used the above code to find out the consumer complaint ratio for each state --- from 0.08% to 14.94%. North Dakota has the lowest complaint rate with 399 complaints; California has the highest rate with about 71497 consumer complaints. To illustrate the consumer complaint ratio, we used the following code from both choroplethr and choroplethrMaps R packages.

Please note, it could be very time consuming to get the consumer dataset work with those two R packages and produce maps to illustrate the consumer complaint ratio. Make sure to compare the structure of data frame with the built-in data frame such as df_pop_state. If the data set you are using for plotting the map is not the exact the same format with the built-in data frame , including the headers, go back and reformat your .csv file based on the choroplethrAdmin1 package.

R code input – mapping

```
# Group-08: Matt Zhang, Lawrence Hamblin, Daniel Hanks Jr
# Q3. What is the complaint ratio about financial service across United States?
# Mashup(1) --- mapping
# To read this file into a data frame called "consumerdata.map"
consumerdata.map2 <- read.csv(file="C:/Data-Apps/Download-All/isssue-count-by-region-value-fff.csv",header=TRUE,sep=",",stringsAsFactors = FALSE)
consumerdata.map2
str(consumerdata.map2)
consumerdata.map2$value <- as.numeric(consumerdata.map2$value)
str(consumerdata.map2)
# compare the structure, data format with built-in df_pop_state  data frame;
## if not the exact the same format, go back and reformat your .csv file based on the choroplethrAdmin1 package
# install the R packages
# install.packages("choroplethr")
library(choroplethr)
# install.packages("choroplethrMaps")
library(choroplethrMaps)
summary(consumerdata.map2)
View(consumerdata.map2)
class(consumerdata.map2)
head(consumerdata.map2)
# to make the map
?state_choropleth
state_choropleth(consumerdata.map2)
# to add a legend
state_choropleth(consumerdata.map2,
          title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
          legend = "Issues")
# to use a continous scale to see outliers by specifying the num_color=1
state_choropleth(consumerdata.map2,
          title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
          legend = "Issues",
          num_colors = 1)

# to change the quantile by setting the num_colors=2 to represent Consumer Complaints that above and below the median
state_choropleth(consumerdata.map2,
          title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
```

```
        legend = "Issues",
        num_colors = 2)
# to set a continuous scale and zoom for certain states, such as "california", "texas", "florida",
"georgia", "new york", "pennsylvania".
state_choropleth(consumerdata.map2,
        title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
        legend = "Issues",
        num_colors = 1,
        zoom = c("california", "texas", "florida", "georgia", "new york", "pennsylvania"))
```

R code output – mapping

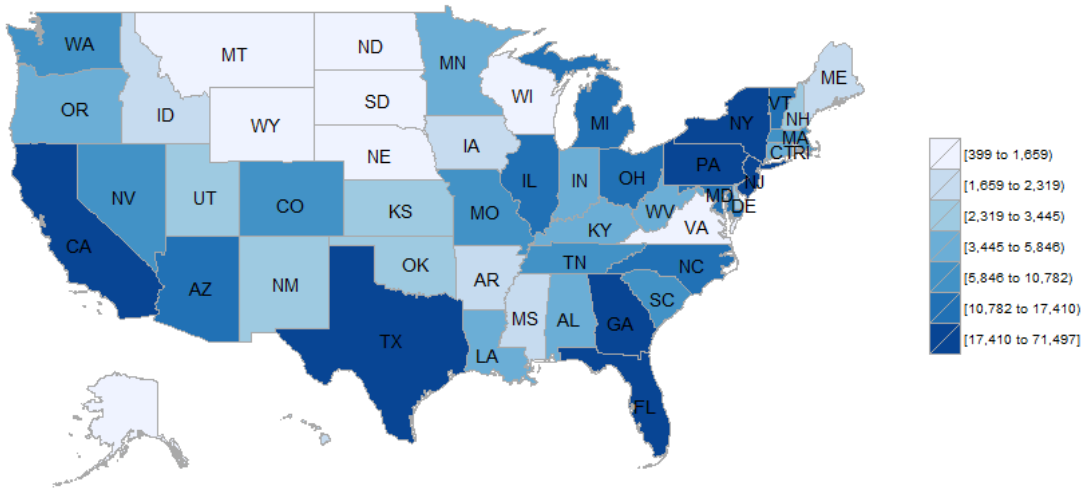
Code Output: ratio

```
> # Group-08: Matt Zhang, Lawrence Hamblin, Daniel Hanks Jr
> Q3. What is the complaint ratio about financial service across United States?
> # To read this file into a  data frame called "consumerdata.map"
> consumerdata.map2 <- read.csv(file="C:/Data-Apps/Download-All/isssue-count-by-region-value-
fff.csv",header=TRUE,sep=",",stringsAsFactors = FALSE)
> str(consumerdata.map2)
'data.frame':     51 obs. of  2 variables:
 $ region: chr  "alabama" "arkansas" "alaska" "arizona" ...
 $ value : int  4894 2060 564 10782 71497 8261 5696 2828 2585 46677 ...
> consumerdata.map2$value <- as.numeric(consumerdata.map2$value)
> str(consumerdata.map2)
'data.frame':     51 obs. of  2 variables:
 $ region: chr  "alabama" "arkansas" "alaska" "arizona" ...
 $ value : num  4894 2060 564 10782 71497 ...
> # compare the structure, data format with built-in df_pop_state  data frame;
> ## if not the exact the same format, go back and reformat your .csv file based on the
choroplethrAdmin1 package
> # install the R packages
> # install.packages("choroplethr")
> library(choroplethr)
> # install.packages("choroplethrMaps")
> library(choroplethrMaps)
> str(df_pop_state)
'data.frame':     51 obs. of  2 variables:
 $ region: chr  "alabama" "alaska" "arizona" "arkansas" ...
 $ value : num  4777326 711139 6410979 2916372 37325068 ...
> # head(df_pop_state)   # notice the exact format
> # double verify the data and functions
> str(consumerdata.map2)
'data.frame':     51 obs. of  2 variables:
 $ region: chr  "alabama" "arkansas" "alaska" "arizona" ...
 $ value : num  4894 2060 564 10782 71497 ...
> summary(consumerdata.map2)
   region             value
 Length:51          Min.   :  399
 Class :character   1st Qu.: 2013
```

```
 Mode  :character   Median : 5325
                Mean   : 9381
                3rd Qu.:11825
                Max.   :71497
> View(consumerdata.map2)
> class(consumerdata.map2)
[1] "data.frame"
> head(consumerdata.map2)
     region value
1   alabama  4894
2  arkansas  2060
3    alaska   564
4   arizona 10782
5 california 71497
6  colorado  8261
>
```

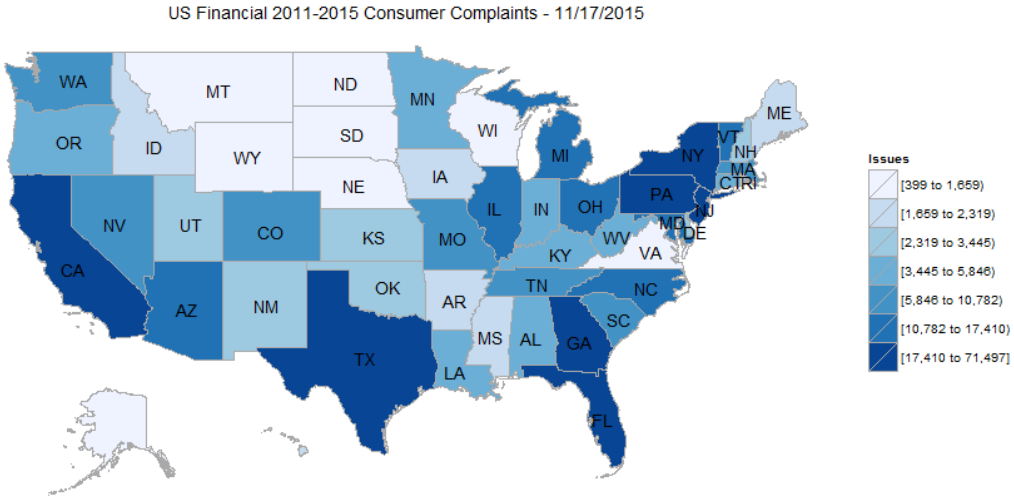
> # to make the map

```
> ?state_choropleth
> state_choropleth(consumerdata.map2)
```



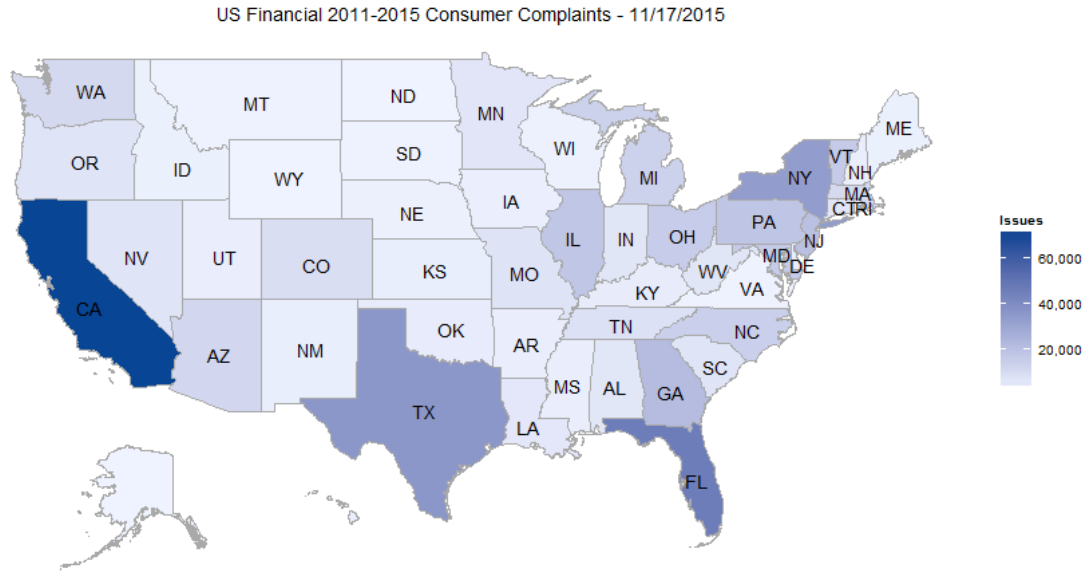
> # to add a lengend

```
state_choropleth(consumerdata.map2,
          title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
          legend = "Issues")
```



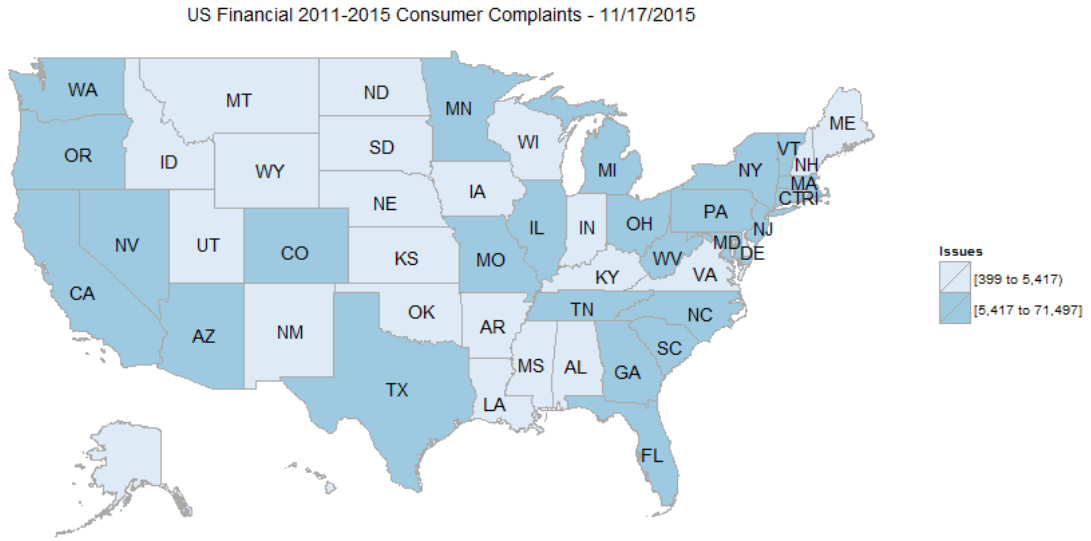
 # to use a continuous scale to see outliers by specifying the num_color=1

```
state_choropleth(consumerdata.map2,
        title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
        legend = "Issues",
        num_colors = 1)
```



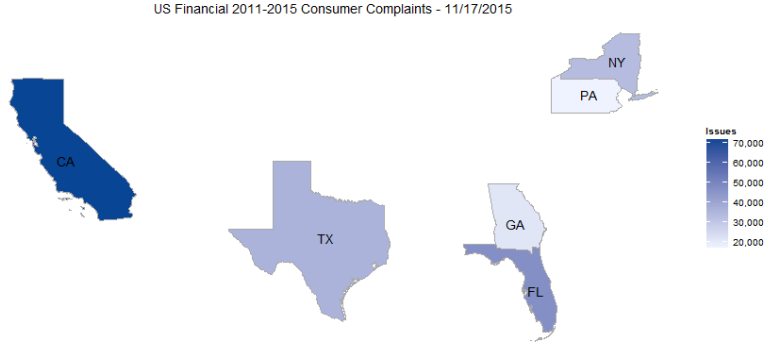
> # to change the quantile by setting the num_colors=2 to represent Consumer Complaints that above and below the median

```
state_choropleth(consumerdata.map2,
        title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
        legend = "Issues",
        num_colors = 2)
```



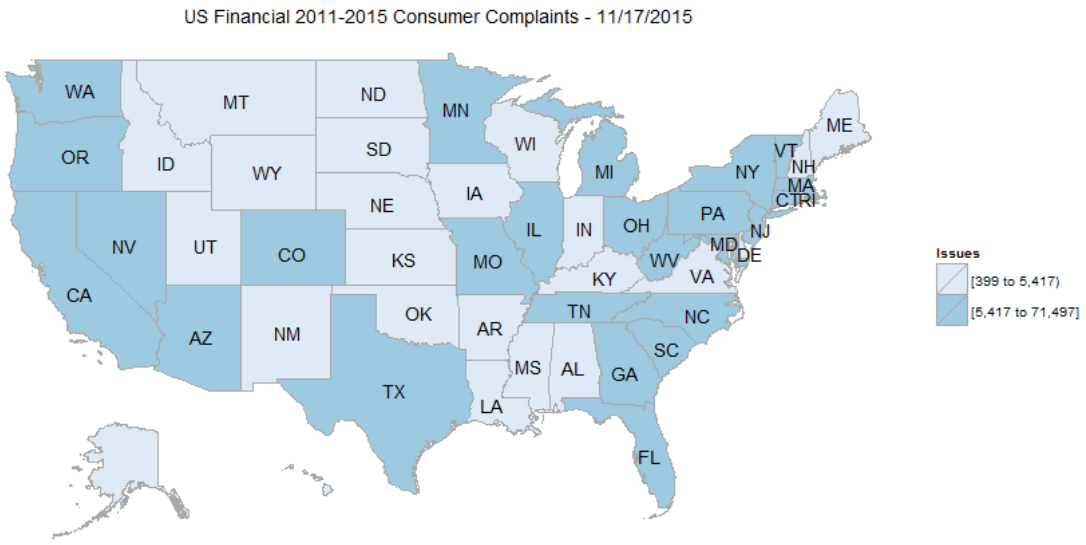
> # to set a continuous scale and zoom for certain states, e.g.

```
> state_choropleth(consumerdata.map2,
+         title  = "US Financial 2011-2015 Consumer Complaints - 11/17/2015",
+         legend = "Issues",
+         num_colors = 1,
+         zoom = c("california", "texas", "florida", "georgia", "new york", "pennsylvania"))
```

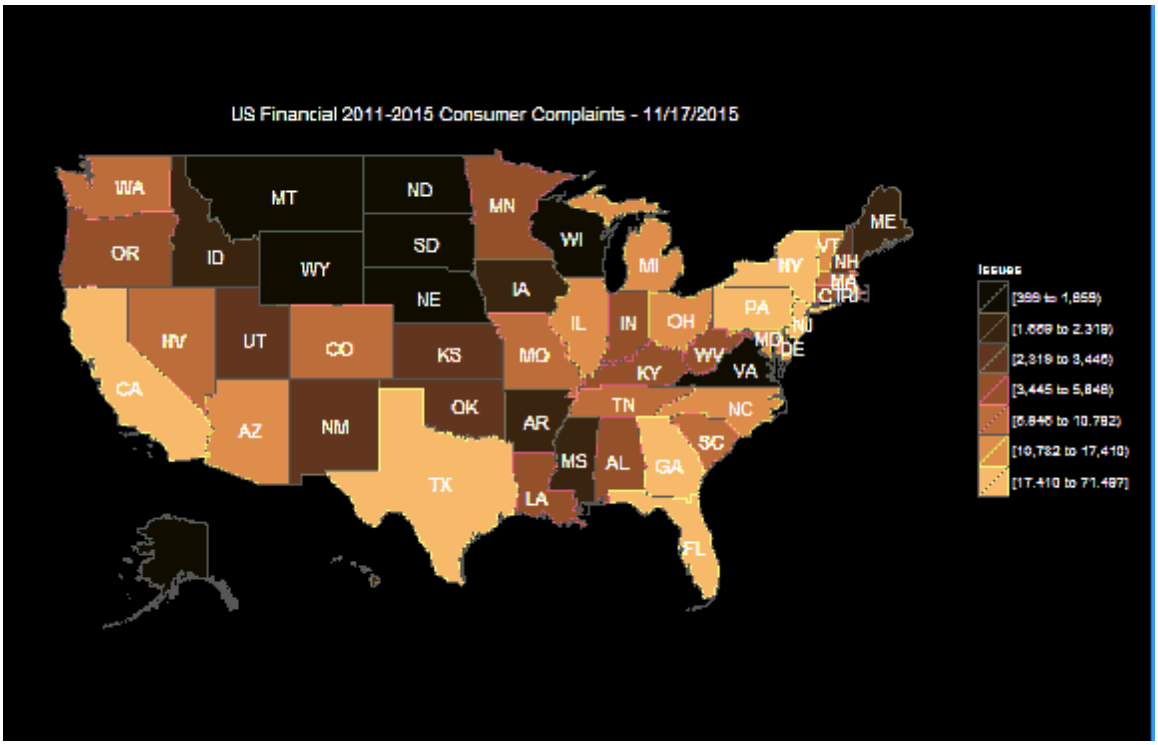


Analysis:

This map represents Consumer Complaints that above and below the median. The dark blue color represents States that have complaints above the average (5325); the light blue color represents States that have complaints below the average.



From the map below, we can see, from the brightness of color, the five top states that have the most consumer complaints --- California, Florida, Texas, New York, and Georgia.



Code input – plotting

```
# Group-08: Matt Zhang, Lawrence Hamblin, Daniel Hanks Jr
# Q3. Mashup(2) - plotting to graphics
# To use the qplot, read this file into a "factor" data frame with labels
complaindata.plot <- read.csv(file="C:/Data-Apps/Download-All/isssue-count-by-region-value-fff.csv",header=TRUE,sep=",",stringsAsFactors = TRUE)
complaindata.plot
str(complaindata.plot)
complaindata.plot$value <- as.numeric(complaindata.plot$value)
###  Plotting via qplot() function
# install.packages("ggplot2")
# install.packages("ggmap")
library(ggmap)
library(ggplot2)
# to make a hstogram
qplot(complaindata.plot$value, data=complaindata.plot, color="red", fill=region,
    xlab="Comsumer Complaints across 51 States")
qplot(log(complaindata.plot$value), data=complaindata.plot, color="red", fill=region, gemo="desity",
    xlab="Comsumer Complaints across 51 States")
###  Plotting via ggplot() function
# Transform the data format
# install.packages("reshape2") first
library(sp)
library(reshape2)
library(ggplot2)
View(complaindata.plot)
```

```
str(complaindata.plot)
# to convert or melt the original data, use metl() function
complaindata.plot.m = melt(complaindata.plot)
View(complaindata.plot.m)
str(complaindata.plot.m)
ls(complaindata.plot.m)
# Change the default "Var1, Var2, value" to match the real column name
colnames(complaindata.plot.m) = c("State", "variable", "Issue")
View(complaindata.plot.m)
# now let's make  some basic plots
plot(complaindata.plot.m)

# now let's make a scatterplot, add geom_point() function. Make sure type in "ggplot"
ggplot(complaindata.plot.m, aes(x=State, y=Issue, color=variable)) + geom_point()
```

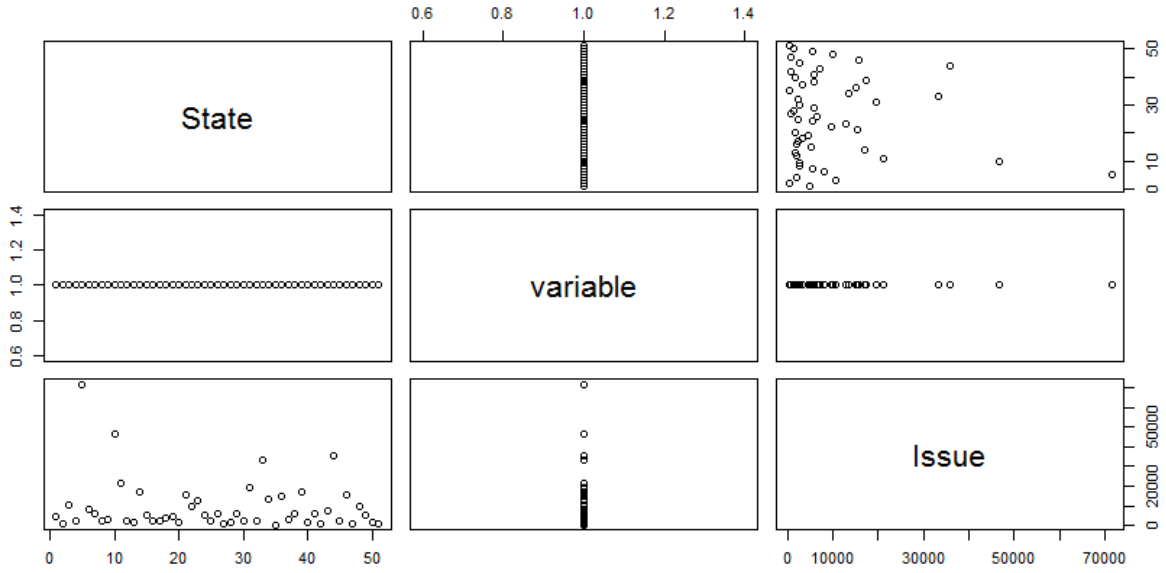
Code output – plotting

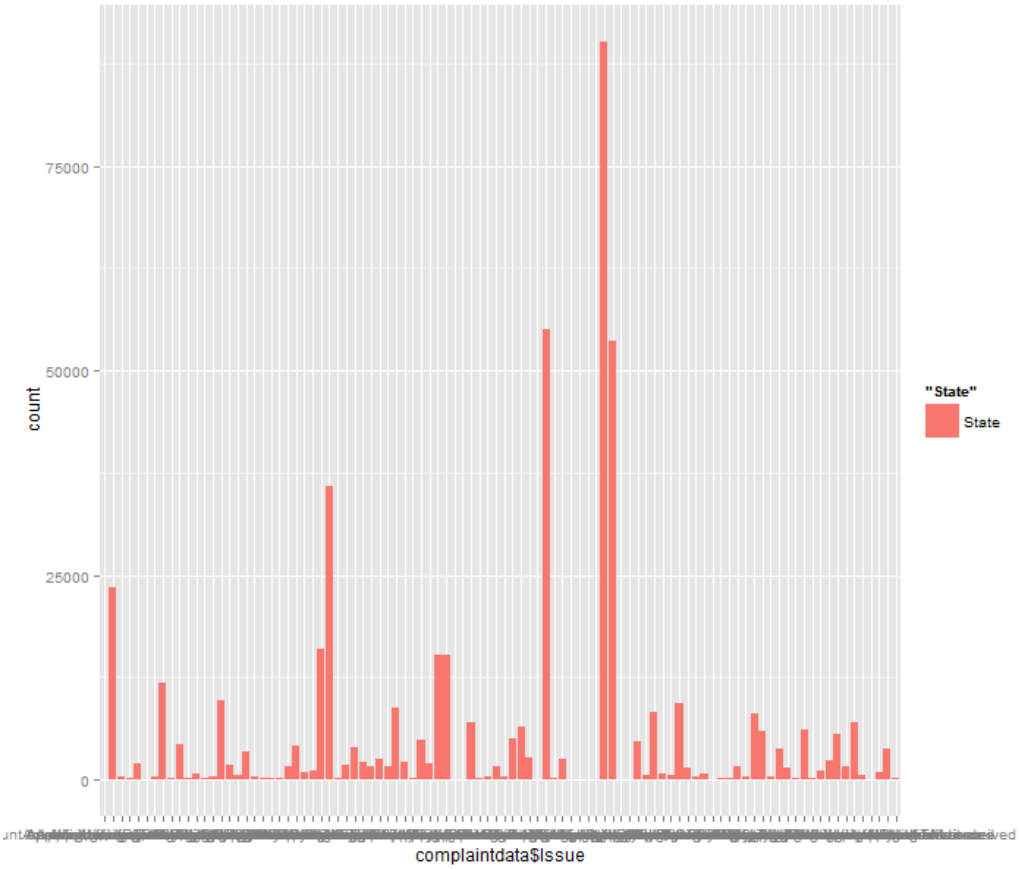
```
> # Group-08: Matt Zhang, Lawrence Hamblin, Daniel Hanks Jr
> # Q3. Mashup(2) - plotting to graphics
> # To use the qplot, read this file into a  "factor" data frame with labels
> complaindata.plot <- read.csv(file="C:/Data-Apps/Download-All/isssue-count-by-r
egion-value-fff.csv",header=TRUE,sep=",",stringsAsFactors = TRUE)
> str(complaindata.plot)
'data.frame':  51 obs. of  2 variables:
 $ region: Factor w/ 51 levels "alabama","alaska",..: 1 4 2 3 5 6 7 9 8 10 ...
 $ value : int  4894 2060 564 10782 71497 8261 5696 2828 2585 46677 ...
> complaindata.plot$value <- as.numeric(complaindata.plot$value)
> ###  Plotting via qplot() function
> # install.packages("ggplot2")
> # install.packages("ggmap")
> library(ggmap)
> library(ggplot2)
> # to make a hstogram
> qplot(complaindata.plot$value, data=complaindata.plot, color="red", fill=region
,
+ xlab="Comsumer Complaints across 51 States")
> qplot(log(complaindata.plot$value), data=complaindata.plot, color="red", fill=r
egion, gemo="desity",
+ xlab="Comsumer Complaints across 51 States")
> ###  Plotting via ggplot() function
> # Transform the data format
> # install.packages("reshape2") first
> library(sp)
> library(reshape2)
> library(ggplot2)
> View(complaindata.plot)
> str(complaindata.plot)
'data.frame':  51 obs. of  2 variables:
 $ region: Factor w/ 51 levels "alabama","alaska",..: 1 4 2 3 5 6 7 9 8 10 ...
 $ value : num  4894 2060 564 10782 71497 ...
> # to convert or melt the original data, use metl() function
> complaindata.plot.m = melt(complaindata.plot)
Using region as id variables
> View(complaindata.plot.m)
> str(complaindata.plot.m)
'data.frame':  51 obs. of  3 variables:
 $ region  : Factor w/ 51 levels "alabama","alaska",..: 1 4 2 3 5 6 7 9 8 10 ...
 $ variable: Factor w/ 1 level "value": 1 1 1 1 1 1 1 1 1 1 ...
```

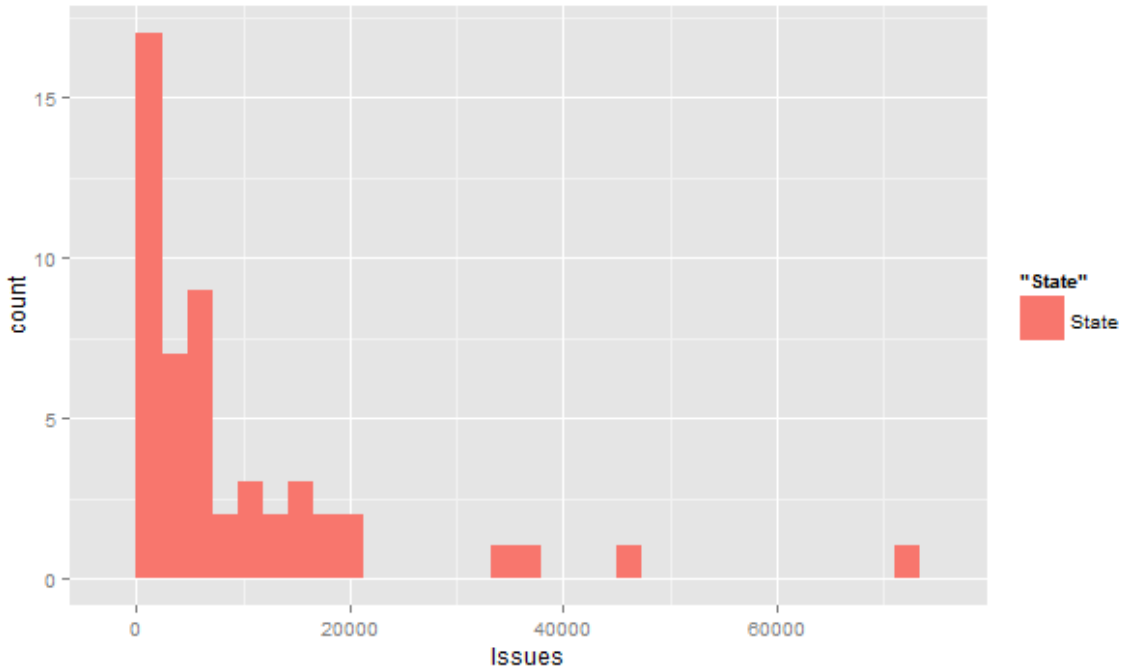
```
 $ value    : num  4894 2060 564 10782 71497 ...
> ls(complaindata.plot.m)
[1] "region"   "value"    "variable"
> # Change the default "Var1, Var2, value" to match the real column name
> colnames(complaindata.plot.m) = c("State", "variable", "Issue")
> View(complaindata.plot.m)
> # now let's make  some basic plots
> plot(complaindata.plot.m)
> # now let's make a scatterplot, add geom_point() function. Make sure type in "g
gplot"
> ggplot(complaindata.plot.m, aes(x=State, y=Issue, color=variable)) + geom_point
()
```

Analysis with Plots Outputs

Here is the plotting overview of the consumer complaints across 51 states:

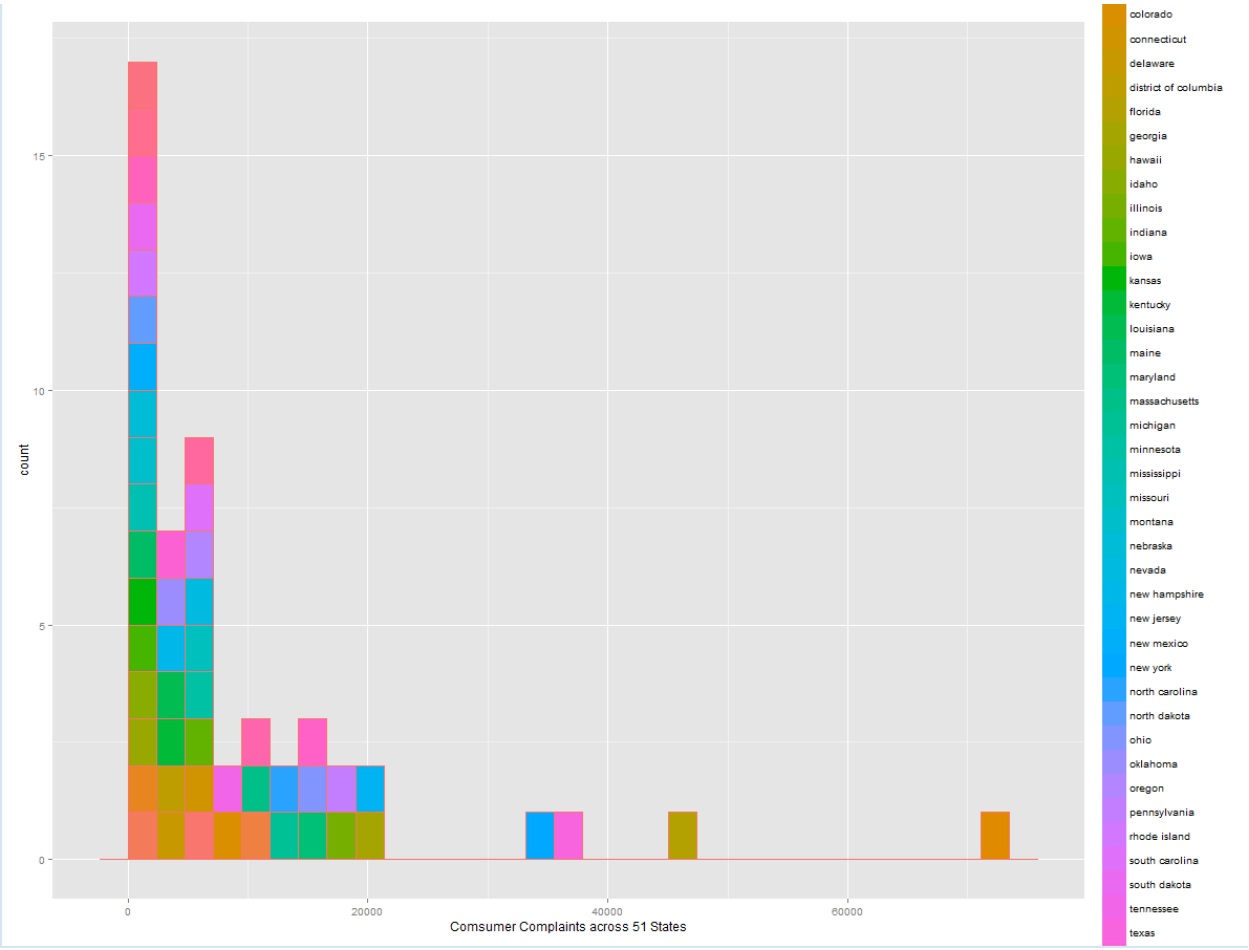




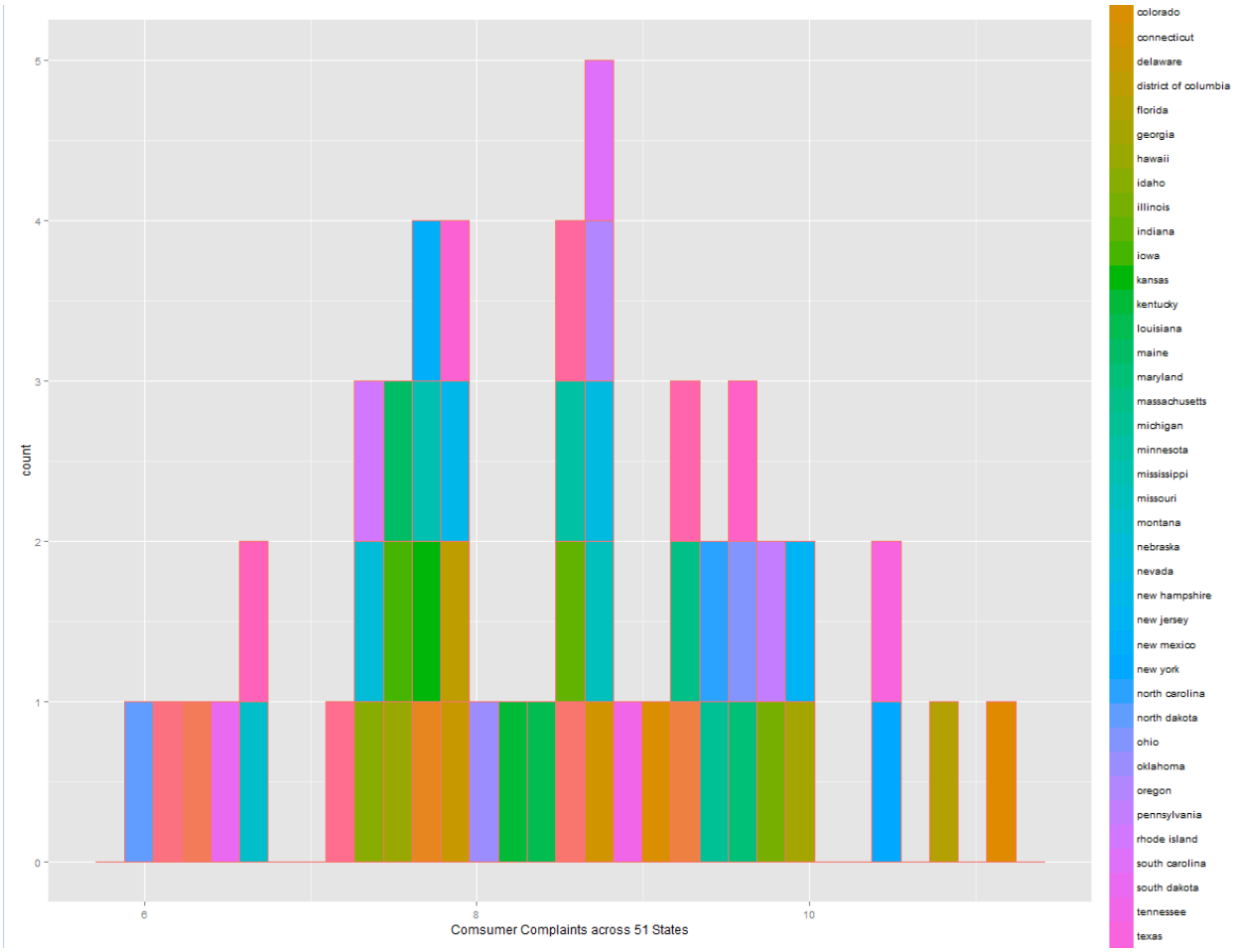


From the above histograms, we can see that most of the complaint counts are under the 12,500 scale across the United states. Some of the issue counts are way above the

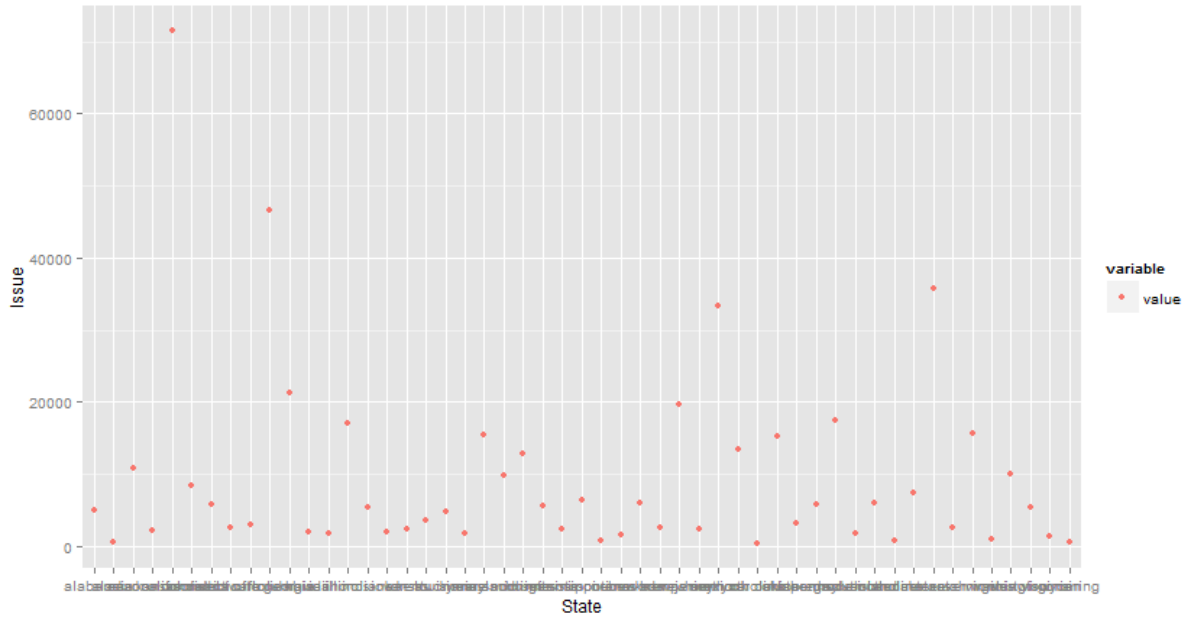
50,000 scale.

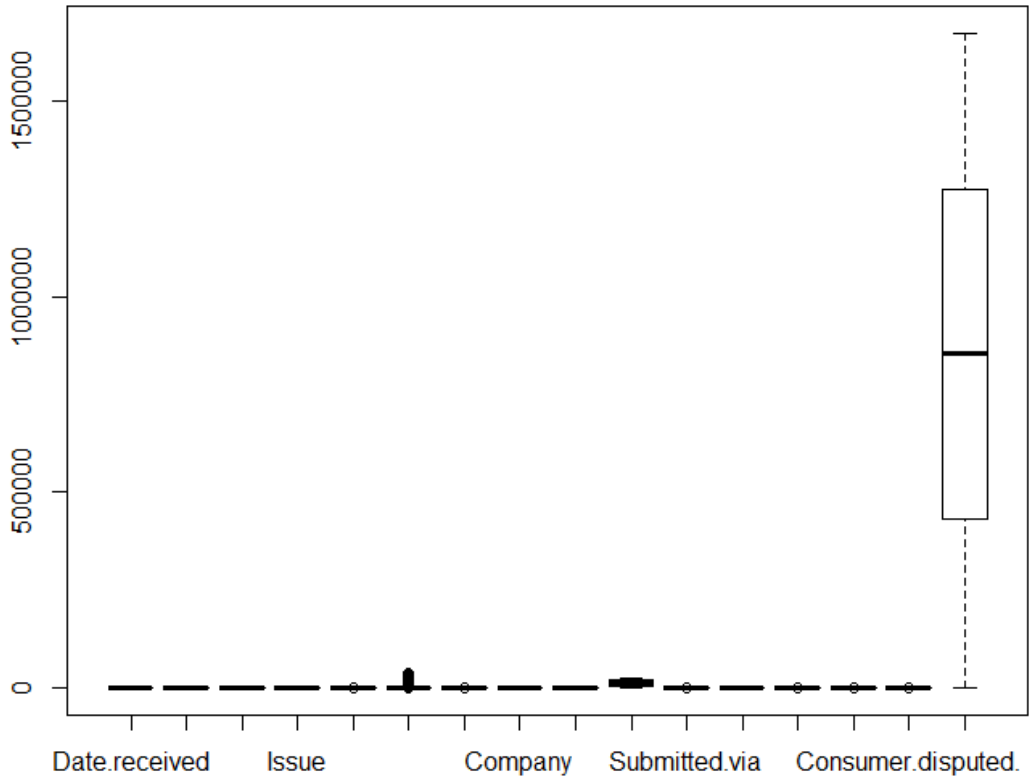


The above histogram shows the complain counts across 51 States.



The above histogram shows the consumer complain counts across 51 States when using the log(complaindata.plot$value) function.





The above scatterplot and boxplot also show the outliers in consumer complaint counts across the United States.

Q4. Do the characteristics of complaints in this database make it useful for predicting complaints over time?

The inclusion of dates in the database allows for the tracking of other variables in the data over time: volume of complaints received, which states were the most common origins of complaints, how often companies responded to consumers in a particular way, and so on. Combined with the predictive capability of a support vector machine, it becomes possible to use a portion of data from an earlier time period to predict the distribution of data in a later time period. Thus, the following question can be answered: Do the characteristics of complaints received by the CFPB over a given time period

exhibit enough uniformity to reliably predict the kinds of complaints received in a later time period?

A more specific question of this type to attempt to answer is whether it is possible to predict the issue appearing in a given complaint based on its location (i.e., state) and product category as reliably in a later period as in an earlier period.

To do attempt to answer this question, the data first needs to be divided into earlier and later portions and randomized. The early data will be used to train the SVM, and the later data will serve as test data. To test for any change in the predictive power of the data over time, two SVMs will be trained and their test results compared. The first will use the earliest half of the data as training data, and the next most recent quarter as test data. The second SVM will then use the earliest three quarters of the data for training and the most recent quarter for testing. The following code will allow for the proper divisions.

```
randIndex <- sample(1:dim(complaintData)[1])
cutHalf <- floor(dim(complaintData)[1]/2)
cutFirstQuarter <- floor(dim(complaintData)[1]/4)
```

The dates in the table can be used to show exactly which time periods are covered by the portions of the data used. However, they are represented as integers and require modification before they can be manipulated as dates. This is accomplished with the as.Date() command, which accepts an input vector and the format of the dates it contains as parameters.

```
dates <- as.Date(complaintData$Date.received, "%m/%d/%Y")
complaintData <- data.frame(complaintData, dates)
```

Now it becomes simple to calculate time differences using difftime().

```
complaintData$dates[cutHalf]; complaintData$dates[dim(complaintData)[1]]
[1] "2014-05-15"
[1] "2011-12-01"
difftime(complaintData$dates[cutHalf], complaintData$dates[dim(complaintData)[1]], units="weeks")
Time difference of 128 weeks
```

```
complaintData$dates[1]; complaintData$dates[cutFirstQuarter-1]
[1] "2015-11-25"
[1] "2015-03-06"
difftime(complaintData$dates[cutFirstQuarter], complaintData$dates[cutHalf-1], units="weeks")
Time difference of 42.14286 weeks

complaintData$dates[1]; complaintData$dates[cutFirstQuarter-1]
[1] "2015-11-25"
[1] "2015-03-06"
difftime(complaintData$dates[1], complaintData$dates[cutFirstQuarter-1], units="weeks")
Time difference of 37.71429 weeks
```

This indicates that we would be using almost two and a half years of data to predict the following ten months, and then three and a half years of data to predict the next nine months.

Next, the data set is split and assigned to new data frames.

```
trainData <- complaintData[randIndex[cutHalf:dim(complaintData)[1]],]
testData <- complaintData[randIndex[cutFirstQuarter:cutHalf – 1],]
trainData2 <- complaintData[randIndex2[cutFirstQuarter:dim(complaintData)[1]],]
testData2 <- complaintData[randIndex2[1:cutFirstQuarter – 1],]
```

With the data properly divided, the first SVM can be run using ksvm(), followed by an overview and a histogram.

```
svmOutput <- ksvm(issueEval ~ Product + State, type="C-svc", data=trainData, kernel="rbfdot",
kpar="automatic", C=5, cross=3, prob.model=TRUE)
svmOutput
hist(alpha(svmOutput)[[1]])
```

Then the SVM is used to predict the test data, and a confusion matrix is printed to display the accuracy of the SVM's predictions.

```
svmPred <- predict(svmOutput, testData, type="votes")
compTable <- data.frame(testData[, 4], svmPred[1,])
```

This is followed by the second SVM with the same sequence of commands.

```
svmOutput2 <- ksvm(issueEval ~ Product + State, type="C-svc", data=trainData2, kernel="rbfdot",
kpar="automatic", C=5, cross=3, prob.model=TRUE)
svmOutput2
hist(alpha(svmOutput2)[[1]])
```

```
svmPred2 <- predict(svmOutput2, testData2, type="votes")
compTable2 <- data.frame(testData2[, 4], svmPred[1,])
```

Due to the size of the table, each SVM is likely to take hours, if not days, to run. Using smaller random samples from the same time periods will be far less time-consuming, though at the cost of accuracy.

4. Business Actions

Based on our analysis and visualization of the data, we recommend that the CFPB and providers of financial services such as the ones contained in the complaint database take the following actions.

1. Mortgage products consistently receive the most complaints from consumers. The CFPB and financial services providers should direct a significant part of their public relations efforts at informing consumers about mortgages, including the application process, payment, and associated risks.

2. As complaints point to satisfaction with a specific product, financial services companies should take complaints seriously and engage with consumers to identify the root causes of their displeasure. It is an essential step in identifying systemic problems that have the potential to affect more people than those who take the time to submit complaints.

3. Our observations revealed trends in the frequency of complaints received based on time of year. Companies and the CFPB should be proactive in reaching out to consumers at these times.

4. Though the number of complaints received from each state exhibits a relationship with its population, the types of complaints received most often also vary by state: Florida and New York have more complaints about mortgages, most complaints in Texas concern credit reporting, etc. Strategies for informing consumers and dealing with complaints should be tailored to each geographic region.

5. As with population, companies with the most complaints such as Bank of America, Wells Fargo, and JP Morgan Chase may simply have more customers, but that in itself is sufficient reason for them to be the most diligent in handling consumer complaints. They should direct their efforts at establishing themselves as receptive to consumer concerns.

6. Complaints are overwhelmingly submitted to the CFPB online. However, the CFPB should continue to promote other methods to submit complaints so as not to exclude consumers with limited access to the Web. Likewise, its outreach efforts should not be limited to its website.

5. Information Value Chain

Data: Our data set is the CFPB's database of consumer complaints about financial services. It is detailed, well-formatted and easy to understand.

Information: Each row contains data of various types that tell the story of a single complaint when linked together. Linking individual rows together to tell a broader story was only possible through the process of cleaning the data and converting it to a format that could be used with R.

Knowledge: Rigorous analysis of the entire data set and relevant subsets allowed for the observation of patterns and trends. Through analysis, we were able not only to observe a single complaint, but compare it and contrast it with other complaints in order to make generalized statements about them.

Intelligence: Understanding the patterns that arose from the data not only made it clear what consumers' problems were, but also how they should be solved. A majority of complaints about a particular financial product suggests that the companies providing that product should prioritize their dispute resolution efforts toward it.

Wisdom: This data set captures consumer concerns from the recent past. Those who are savvy and diligent enough can make use of it as a tool to handle current issues with consumer satisfaction, and to prepare for future issues. Even those companies that opt not to use this information may realize its value based on the decisions (and results) of their competitors, or any analysis the CFPB makes public.

6. Appendix:

Data Source

Collected by the Consumer Financial Protection Bureau (CFPB), this Consumer Complaint Database called Consumer_Complaints.csv file comes from --- http://catalog.data.gov/dataset/consumer-complaint-database.

We retrieved the dataset on Nov 28, 2015, which containing the data until Sep 26, 2015, 2015.

Meta Data

Attributes:

Column name	Column description	Data type	Number of possible values
Date received	The date the CFPB received the complaint	Date	N/A
Product	The type of product the consumer identified in the complaint	Text	11
Sub-product	The type of sub-product the consumer identified in the complaint	Text	45
Issue	The issue the consumer identified in the complaint	Text	95
Sub-issue	The sub-issue the consumer identified in the complaint	Text	47
Consumer complaint narrative	Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint. Consumers must opt-in to share their narrative. We will not publish the narrative unless the consumer consents, and consumers can opt-out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer.	Text	N/A
Company public response	The company's optional, public-facing response to a consumer's complaint. Companies provide a public response to the CFPB, for posting on the public database, by selecting a response from a set list of options.	Text	9
Company	The complaint is about this company	Text	3,382
State	The consumer's reported mailing state for the complaint	Text	63

ZIP code	Mailing ZIP code provided by the consumer. This field may: i) include the first five digits of a ZIP code; ii) include the first three digits of a ZIP code (if the consumer consented to publication of their complaint narrative); or iii) be blank (if ZIP codes have been submitted with non-numeric values, if there are less than 20,000 people in a given ZIP code, or if the complaint has an address outside of the United States).	Text	N/A
Submitted via	How the complaint was submitted to CFPB	Text	6
Date sent to company	The date the CFPB sent the complaint to the company	Date	N/A
Company response to consumer	This is how the company responded	Text	8
Timely response?	Whether the company gave a timely response	Text	2
Consumer disputed?	Whether the consumer disputed the company's response	Text	2
Complaint ID	The unique identification number for a complaint	Integer	N/A

R packages utilized

choroplethrAdmin1 package

choroplethrAdmin1 package

choroplethrMaps

choroplethrMaps

ggmap

ggplot2

reshape2

sp

apriori

kernlab

Summary of R Code used in this project

(see details in Q1, 2, 3, 4)

qplot(complaintdata$Issue, data=complaintdata, fill="State")

> qplot(complaindata.plot$value, data=complaindata.plot, color="red", fill=region,

+ xlab="Consumer Complaints across 51 States")

> qplot(log(complaindata.plot$value), data=complaindata.plot, color="red", fill=region, g

emo="desity",

+ xlab="Comsumer Complaints across 51 States")

plot(complaindata.plot.m)

boxplot(complaintdata)

names() - Functions to get or set the names of an object

View(complaindata.plot.m)

plot(complaindata.plot.m)

library(sp)

library(reshape2)

library(ggplot2)

head(consumerdata.map2)

log()

data.frame()- This function creates data frames, tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure by most of R's modeling software

head() - Returns the first or last parts of a vector, matrix, table, data frame or function

subset() - Return subsets of vectors, matrices or data frames which meet conditions

str() - Compactly display the internal structure of an R object, a diagnostic function and an alternative to summary (and to some extent, dput)

summary() - a generic function used to produce result summaries of the results of various model fitting functions.

pie() – Draw pie charts.

7. References:

Consumer Complaint Database, Sep 26, 2015, Retrieved from http://catalog.data.gov/dataset/consumer-complaint-database

CFPB (2015). 2014 Consumer Response Annual Report. *Consumer Financial Product Bureau.* Retrieved from: http://files.consumerfinance.gov/f/201503_cfpb_consumer-response-annual-report-2014.pdf

R Seek, http://rseek.org/ (for searching R related questions)

McCoy, K. (2014). Texas Auto Lender Fined $2.75M for Credit Errors. *USA Today.* Retrieved from: www.usatoday.com/story/money/business/2014/-8/20/cfpd-first-investor-financial-fined/14340119/

RStudio, Lecture: Visualizing Data in R. Retrieved from https://rpubs.com/astauffer/iceu3

Quick-R, Graphics with ggplot2. Retrieved from http://www.statmethods.net/advgraphs/ggplot2.html