MA615_GroupAssignment

Becky Yu , Mira Tang , Kecheng Liang , Chaoqun Yin

September 30, 2018

We got the dataset from the Suspicious Activities Report statistics tool. After filtering specific industries (Insurance Company, Loan or Finance Company, Money Services BUsiness and Securities/Futures) and suspicious activities(Fraud), we got a dataset with 43454 observations of 8 variables. Looking into the data, we found that observations with "[Total]" entries were summarized data that should be seprated from raw data, so we removed those entires and ended with 23890 observations.

Data Cleaning

```
#Import data
Finance <- read.csv("SARStats.csv")
#View(Finance)
summary(Finance)
```

```
##    Year.Month             State
##  2017   :9310   California   : 3425
##  2016   :8994   New York     : 3214
##  2018   :7729   Texas        : 2713
##  2015   :7273   Florida      : 1755
##  2014   :5609   Massachusetts: 1684
##  2013   :3665   Illinois     : 1499
##  (Other): 874   (Other)      :29164
##                            Industry             Suspicious.Activity
##  [Total]                       :  384   Other Fraud (Type):9137
##  Insurance Company             : 3138   Wire              :9130
##  Loan or Finance Company       : 1129   Check             :6905
##  Money Services Business (MSB):17832   Credit/Debit Card :6046
##  Securities/Futures            :20971   ACH               :5875
##                                         Mail              :2035
##                                         (Other)           :4326
##           Regulator               Product
##  IRS           :17538   [Total]       :9296
##  SEC           :12240   Debit Card    :8270
##  [Total]       : 4789   Other         :7041
##  OCC           : 4537   Credit Card   :4352
##  FRB           : 2655   Prepaid Access:3511
##  Not Applicable:  707   Mutual Fund   :2585
##  (Other)       :  988   (Other)       :8399
##                    Instrument        Count
##  [Total]                :19564   1      :16698
##  Funds Transfer         : 7924   2      : 5974
##  U.S. Currency          : 4262   3      : 3208
##  Personal/Business Check: 4186   4      : 2073
##  Foreign Currency       : 1994   5      : 1241
##  Other                  : 1914   6      : 1140
##  (Other)                : 3610   (Other):13120
```

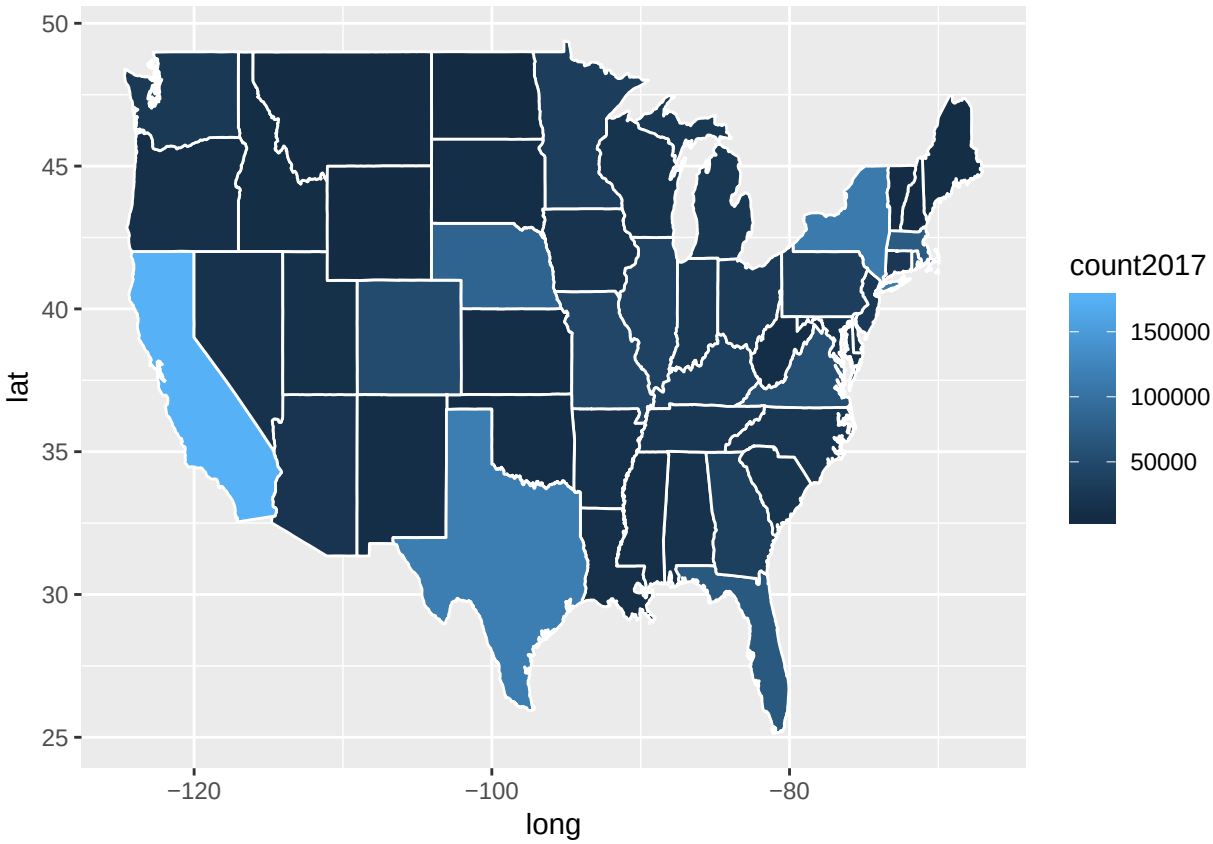
```r
#Eliminate meaningless data
Finance <- filter(Finance, !(str_detect(string = Finance$State,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Industry,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Product,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Instrument,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Suspicious.Activity,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Regulator,pattern = "\\[Total\\]")))
Finance <- filter(Finance, !(str_detect(string = Finance$Year.Month,pattern = "\\[Total\\]")))

#rename the inappropriate column name
colnames(Finance)[1]<-"Year"

#Transform the type of data
Finance$Count <- as.numeric(Finance$Count)
```

EDA

State

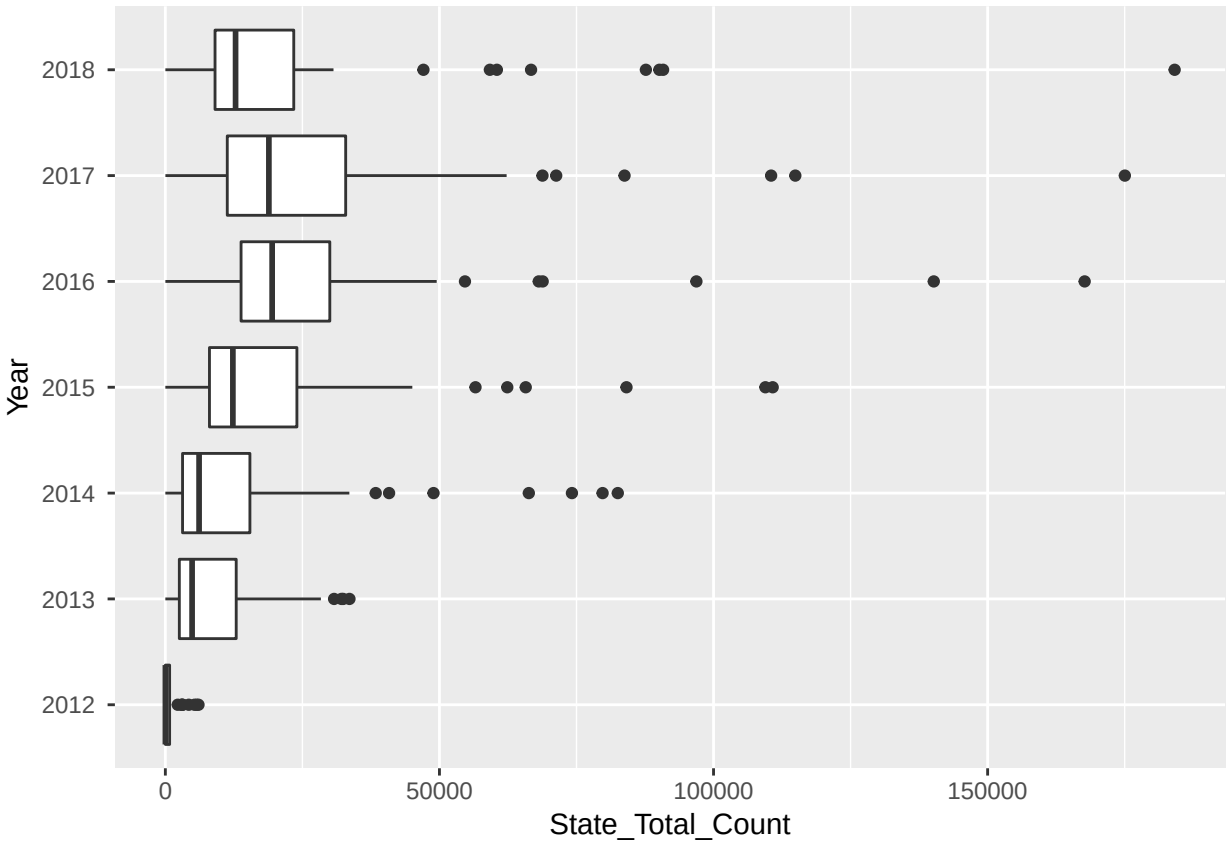
Firstly, let us take a quick look at the total count of frauds reported in 2017 for each state:

```r
Finance %>%
  group_by(Year,State) %>%
  summarise(State_Total_Count = sum(as.numeric(Count)))-> State1

#load us map data
all_states <- map_data("state")

#mutate the count data into map data
State2017<-State1[which(as.character(State1$Year)=='2017 '),]
State2017$State<-tolower(State2017$State)
s<-State2017$State
for (i in 1:length(s)){
  if (s[i] %in% unique(all_states$region)){
    all_states[which(all_states$region==s[i]),"count2017"]<-State2017$State_Total_Count[i]
  }
}

#plot all states with ggplot
ggplot(all_states)+
  geom_polygon(aes(x=long, y=lat, group = group,fill=count2017),colour="white" )
```



Next, here are the boxplots of total counts of all states per year:

```
ggplot(State1, aes(x=Year, y=State_Total_Count))+
  geom_boxplot()+
  coord_flip()
```



From the boxplot, we can compare the numbers of fraud cases from 2012 to 2018 easily. From 2012 to 2017, the total number of fraud cases has a obvious trend of increasing.

Make a table

```
#get a summarized dataframe with top five states for each year
Finance %>%
  group_by(Year , State) %>%
  summarise(Total_Count = sum(as.numeric(Count))) %>%
  arrange(Year,desc(Total_Count)) %>%
  slice(1:5) -> State2

#tidy the long table using string concating skills
table1<-aggregate(State~Year,data =State2,paste,collapse=",")
table2<-aggregate(Total_Count~Year,data =State2,paste,collapse=",")

#join the sub tables
table<-left_join(table1,table2,by="Year")
kable(table,caption = "Top five States each year", "html" ) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

Top five States each year

Year

State

Total_Count

2012

California,Indiana,North Carolina,Florida,New York

6045,5727,5316,4272,3146

2013

Massachusetts,California,New York,Kentucky,Rhode Island

33583,32431,32125,30782,28373

2014

Massachusetts,New York,Rhode Island,California,Nebraska

82535,79759,74172,66306,48920

2015

New York,California,Texas,Nebraska,Massachusetts

110787,109461,84118,65739,62372

2016

California,New York,Texas,Nebraska,Massachusetts

167700,140177,96872,68830,68116

2017

California,Texas,New York,Nebraska,Massachusetts

175031,114916,110504,83762,71301

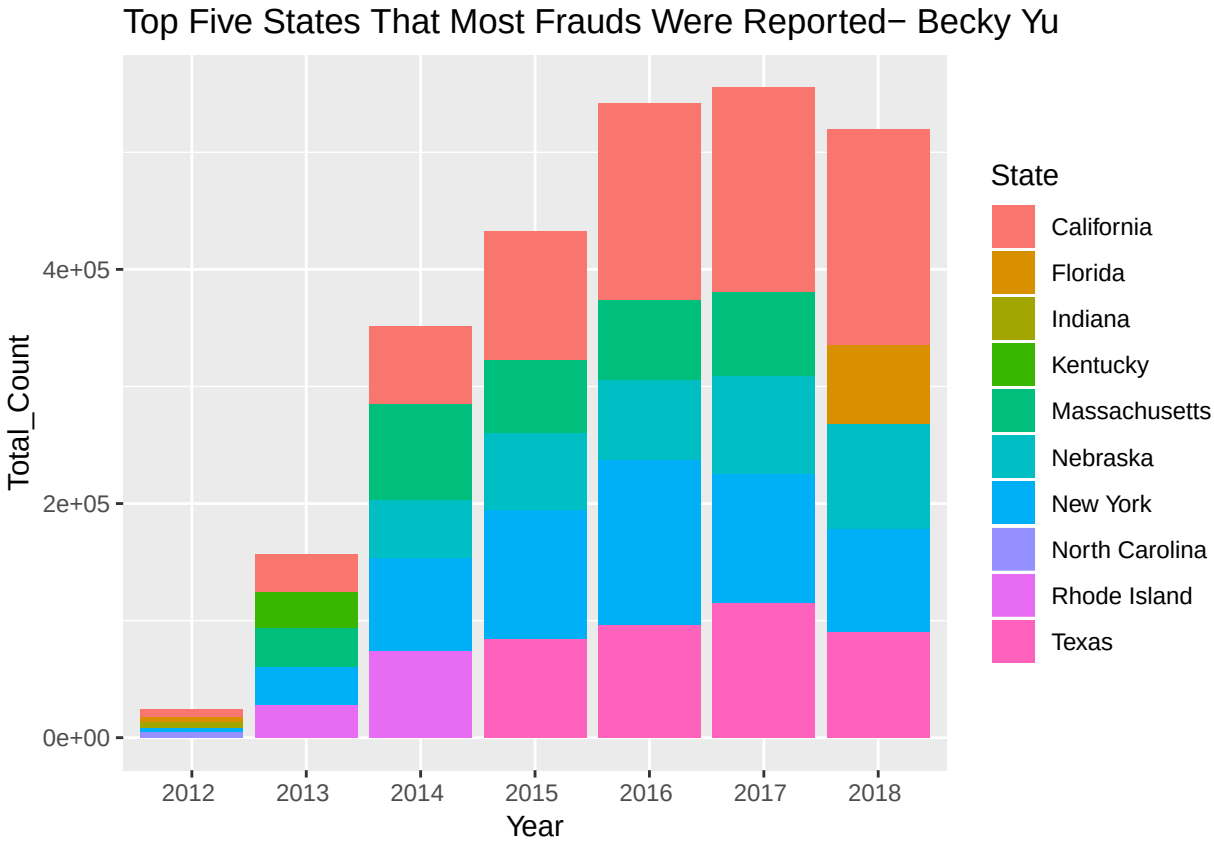
2018

California,Texas,Nebraska,New York,Florida

184114,90797,90119,87658,66714

Make a plot

```
ggplot(State2,aes(x = Year,y = Total_Count,fill = State))+
  geom_bar(stat = 'identity',position = 'stack') +
  labs(title = "Top Five States That Most Frauds Were Reported- Becky Yu")
```



The count of frauds reported increased largely through the last few years. Since the data of 2018 has not been complete yet, we may still predict a trend of growth. California and New York were typically among the top five States that most frauds were reported. Massachusetts ranked first in 2013 and 2014, but ended with fifth in 2015, 2016 and 2017.

Industry

```
Finance %>% group_by(Year,Industry)  %>%
  summarize(Count = sum(Count)) %>%
  arrange(Year,desc(Count)) -> Finance.ind
```

First of all, we can see how count change with year and industry through the contour plot.

```
plot_ly(
  x = Finance.ind$Year,
  y = Finance.ind$Industry,
  z = Finance.ind$Count,
  type = "contour"
)
```

Make a table

```
#table grouped by year
kable(Finance.ind[,c(2,3)], caption = "Securities/Futures Suspicious Activity Reports by Industry" , "h
  kable_styling(bootstrap_options = c("striped", "hover", "condensed")) %>%
  group_rows("2012", 1, 3) %>%
```

```
  group_rows("2013", 4, 6) %>%
  group_rows("2014", 7, 10) %>%
  group_rows("2015", 11, 14) %>%
  group_rows("2016", 15, 18) %>%
  group_rows("2017", 19, 22) %>%
  group_rows("2018", 23, 26)
```

Securities/Futures Suspicious Activity Reports by Industry

Industry

Count

2012

Money Services Business (MSB)

27182

Insurance Company

9797

Securities/Futures

2579

2013

Money Services Business (MSB)

247631

Securities/Futures

198989

Insurance Company

29279

2014

Securities/Futures

432317

Money Services Business (MSB)

364868

Insurance Company

22727

Loan or Finance Company

3

2015

Money Services Business (MSB)

685468

Securities/Futures

486001

Insurance Company

23120

Loan or Finance Company

11969

2016

Money Services Business (MSB)

1066324

Securities/Futures

536431

Insurance Company

27123

Loan or Finance Company

14131

2017

Money Services Business (MSB)

969278

Securities/Futures

580305

Insurance Company

47178

Loan or Finance Company

24568

2018

Money Services Business (MSB)

775374

Securities/Futures

539040

Insurance Company

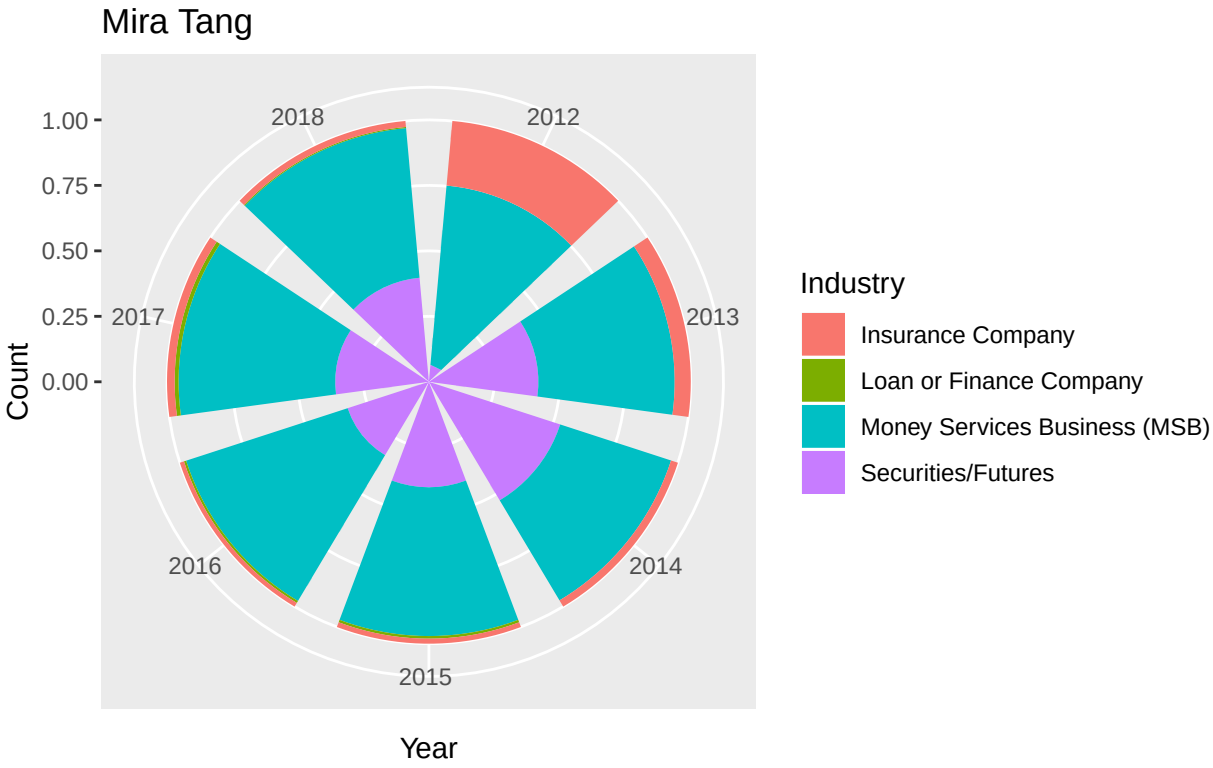
30987

Loan or Finance Company

5732

Make a plot

```
ggplot(Finance.ind,aes(x = Year , Count ,fill = Industry)) +
  geom_bar(stat = "identity",position = "fill",width = 0.8) +
  coord_polar(theta = "x") +
  labs(title = "Mira Tang")
```



We can know from the plot that MSB is the most frequent industry reported securities/futures suspicious activity in each year. Although insurance company's suspicious activities happend a lot in 2012, it decreased by year.

Suspicious Activities

```
#Select some useful columns from the database
suspicion <- select(Finance, Year, Suspicious.Activity, Count)

# Change the column's name
colnames(suspicion)[1] <- "Year"
suspicion %>%
  group_by(Year, Suspicious.Activity) %>%
  summarize(Count = sum(Count)) %>%
  arrange(Year, desc(Count)) %>%
  slice(1:3) -> sus.table
```

Make a table

```
kable(sus.table, caption = "Top 5 Reported Susicious Activities by Years") %>%
  kable_styling(bootstrap_options = c("striped") )
```

Make a plot

```
ggplot(suspicion, aes(x = Year, y = Count, fill = Suspicious.Activity)) +
  geom_col(position = "stack") +
  labs(title = "Kecheng Liang")
```

Table 1: Top 5 Reported Susicious Activities by Years

Year	Suspicious.Activity	Count
2012	Other Fraud (Type)	13938
2012	ACH	9900
2012	Wire	6859
2013	Wire	187095
2013	Other Fraud (Type)	72266
2013	Check	66244
2014	Wire	275259
2014	Other Fraud (Type)	140279
2014	ACH	127467
2015	Wire	312110
2015	Other Fraud (Type)	278494
2015	Credit/Debit Card	192922
2016	Other Fraud (Type)	519922
2016	Wire	513993
2016	Check	195172
2017	Other Fraud (Type)	486738
2017	Wire	412141
2017	Credit/Debit Card	239798
2018	Other Fraud (Type)	485287
2018	Wire	313448
2018	ACH	188977

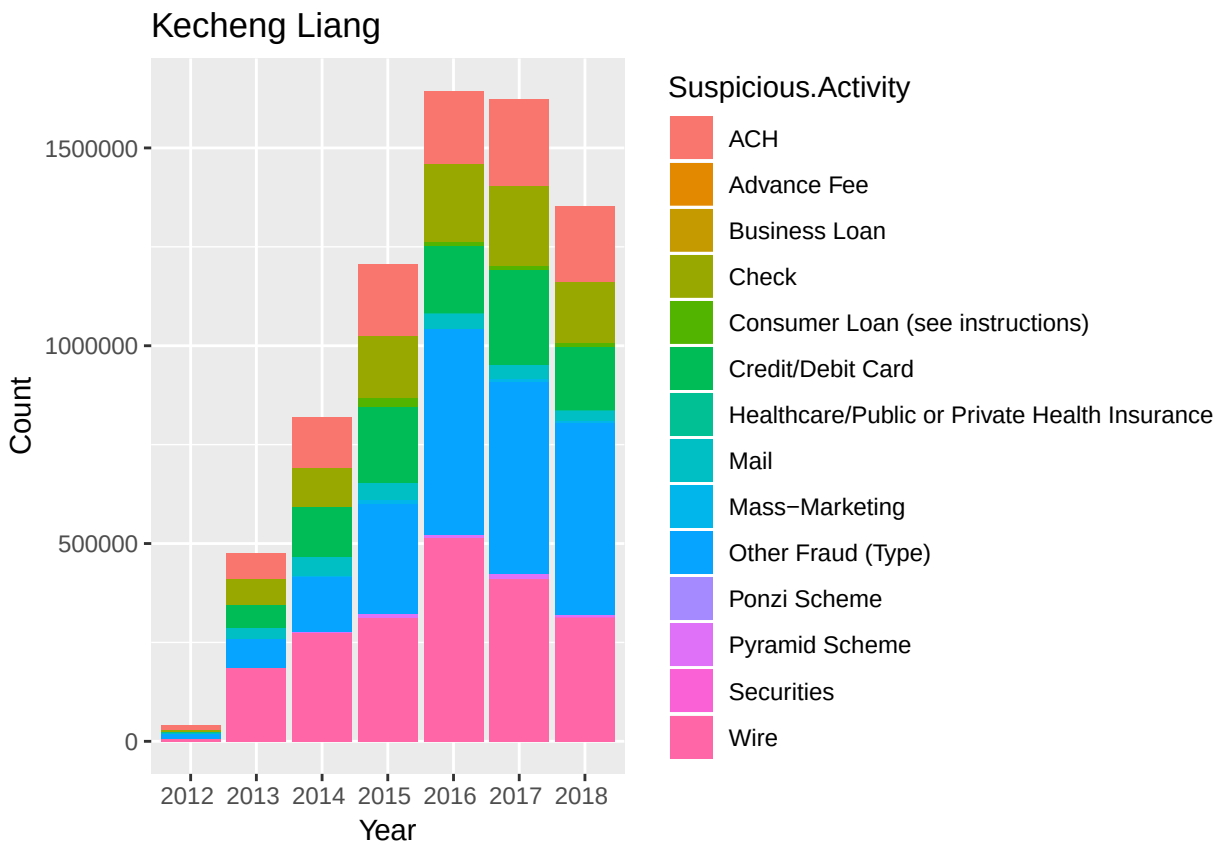


Table 2: Top 3 Regulators against Criminal Each Year

Year	Regulator	Count
2012	IRS	36979
2012	SEC	2577
2012	CFTC	2
2013	IRS	276910
2013	SEC	195515
2013	FRB	1544
2014	IRS	356676
2014	SEC	326661
2014	OCC	44769
2015	IRS	678236
2015	SEC	366601
2015	OCC	96939
2016	IRS	1075441
2016	SEC	359824
2016	OCC	142726
2017	IRS	992367
2017	SEC	404706
2017	OCC	102136
2018	IRS	781004
2018	SEC	353141
2018	FRB	127091

The graph shows that we should pay more attetion on wire, credit/debit card, and check.

Regulator

```
#Seclect specific olumns for analysis
regulator <- select(Finance, Year, Regulator, Count)

#More data cleaning eliminating "Not Applicable"
regulator$Regulator <- str_replace_all(regulator$Regulator, fixed(" "), "")
regulator <- filter(regulator, !str_detect(string = regulator$Regulator,"NotApplicable"))
```

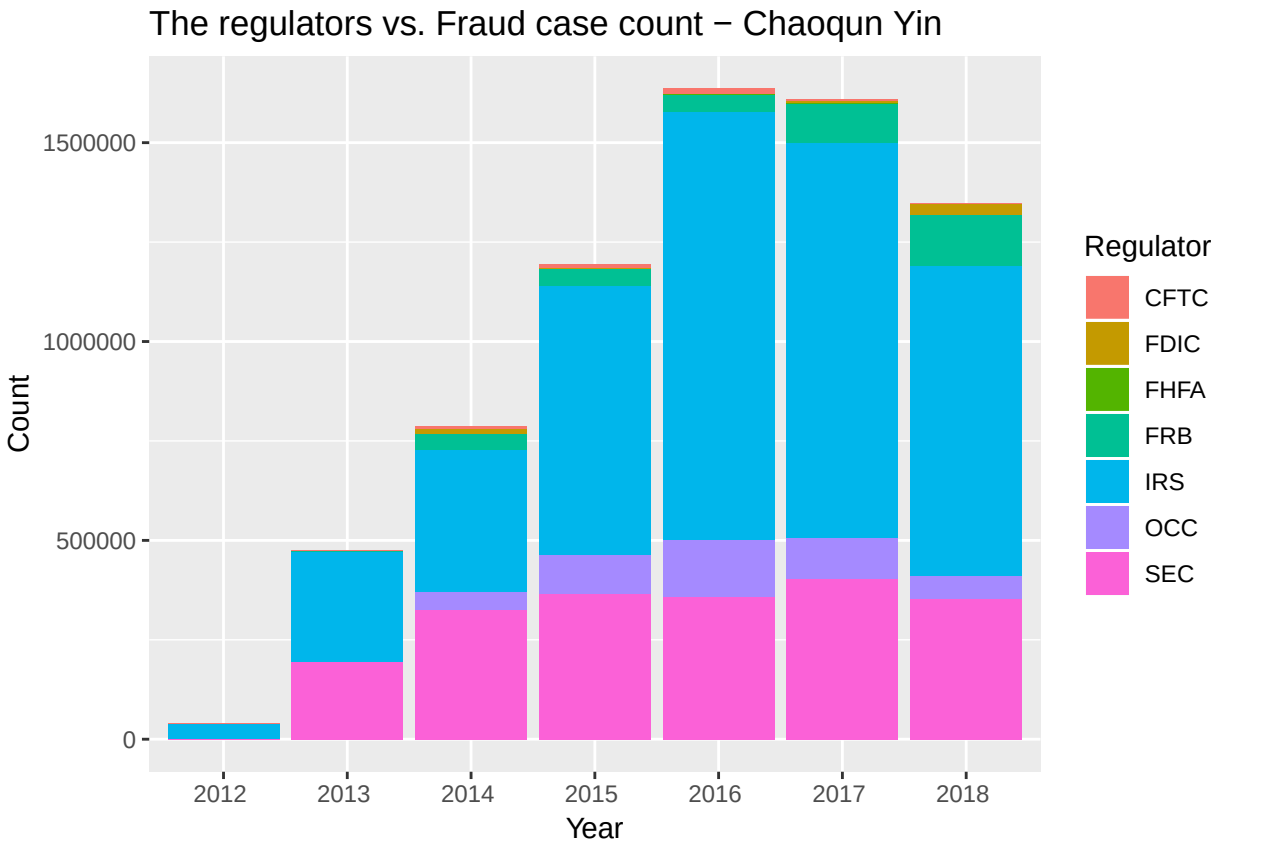
Make a table

```
regulator %>% group_by(Year,Regulator) %>%
  summarize(Count=sum(Count)) %>%
  arrange(Year,desc(Count)) %>%
  slice(1:3)-> sus.table

kable(sus.table, caption = "Top 3 Regulators against Criminal Each Year") %>%
  kable_styling(bootstrap_options = c("striped") )
```

Make a plot of the regulators vs. fraud case count

```
ggplot(regulator , aes(x = Year,y = Count,fill = Regulator)) +
  geom_col(position = "stack") +
```

```
labs(title = "The regulators vs. Fraud case count - Chaoqun Yin")
```



From the plot, the fraud cases charged by IRS increase greatly from 2012 to 2016. Then in 2016, the fraud cases charged by IRS get to the peek then the numbers started to decrease. Among all the data points in the plot, the most financial criminal cases are handled by IRS and SEC during the 7 years.

Conclusion

- As for the numbers of fraud cases in different states:

From the map plot and boxplot, we can compare the numbers of fraud cases in different states from 2012 to 2018 easily. From 2012 to 2017, the total number of fraud cases has a obvious trend of increasing.Then, the count of frauds reported increased largely through the last few years. Since the data of 2018 has not been complete yet, we may still predict a trend of growth. California and New York were typically among the top five States that most frauds were reported. Massachusetts ranked first in 2013 and 2014, but ended with fifth in 2015, 2016 and 2017.

- As for the fraud cases reported in different industries:

We can know from the plot that MSB is the most frequent industry reported securities/futures suspicious activity in each year. Although insurance company's suspicious activities happend a lot in 2012, it decreased by year. But the fraud cases reported in MSB have a trend of increasing, so we should be extremly cautious about the industry.

- As for the suspicious activity type of fraud cases:

The wire, mail and debit/credit card are often used by criminals to make fraud. And from the plot we can see that ACH type are used more frequently recent years, so we can predict that it will continue increasing in 2018. We should be caucious about this new kind of financial derivatives.

- As the regulators against the fraud actions:

The fraud cases charged by IRS increase greatly from 2012 to 2016. Then in 2016, the fraud cases charged by IRS get to the peek then the numbers started to decrease. It is suggested by the plot that FRB plays more and more important role in the recent years against fraud criminals. Among all the data points in the plot, the most financial criminal cases are handled by IRS, SEC and FRB during the 7 years from 2012 to 2018.