

Group#9

Xuan, Megha, Jianhao, Guangyan

September 29, 2018

Readme

We are using the data from:

```
head(read.csv("SARStats.csv"))
```

##	Year.Month	State	Countym	Industry
## 1	2018	California	Alameda County, CA	Depository Institution
## 2	2018	California	Alameda County, CA	Depository Institution
## 3	2018	California	Alameda County, CA	Depository Institution
## 4	2018	California	Alameda County, CA	Depository Institution
## 5	2018	California	Alameda County, CA	Depository Institution
## 6	2018	California	Alameda County, CA	Depository Institution

##	Suspicious.Activity	Product	Instrument	Count
## 1		ACH Credit Card	U.S. Currency	6
## 2		Check Credit Card	U.S. Currency	10
## 3	Consumer Loan (see instructions)	Credit Card	U.S. Currency	2
## 4		Credit/Debit Card	Credit Card U.S. Currency	27
## 5		Mail Credit Card	U.S. Currency	2
## 6		Mass-Marketing Credit Card	U.S. Currency	6

This is how we generated the dataset:

Suspicious Activity Report Statistics (SAR Stats) FAQ/Glossary

Industry Type
× Securities/Futures × Other

Year & Month ☐ Display Months
× July × August × September

Suspicious Activity Category / Type
× Securities × Wire × Other Fraud (Type)

States / Territories
× California

County / Metro & Micro Area ☒ County ☐ Metro
× Yuba County, CA

Instrument Type(s) / Payment Mechanism(s)
× U.S. Currency

Product Type
× Credit Card

Relationship
Filter is unavailable in combination with County/Metro

Regulator
[Select items to filter]

Xuan's plot

As we get a lot of regions here, it is hard to put all the information on the same plot. So I randomly picked 4 counties: ("Los Angeles County, CA", "Santa Clara County, CA", "Orange County, CA", "Santa Barbara County, CA").

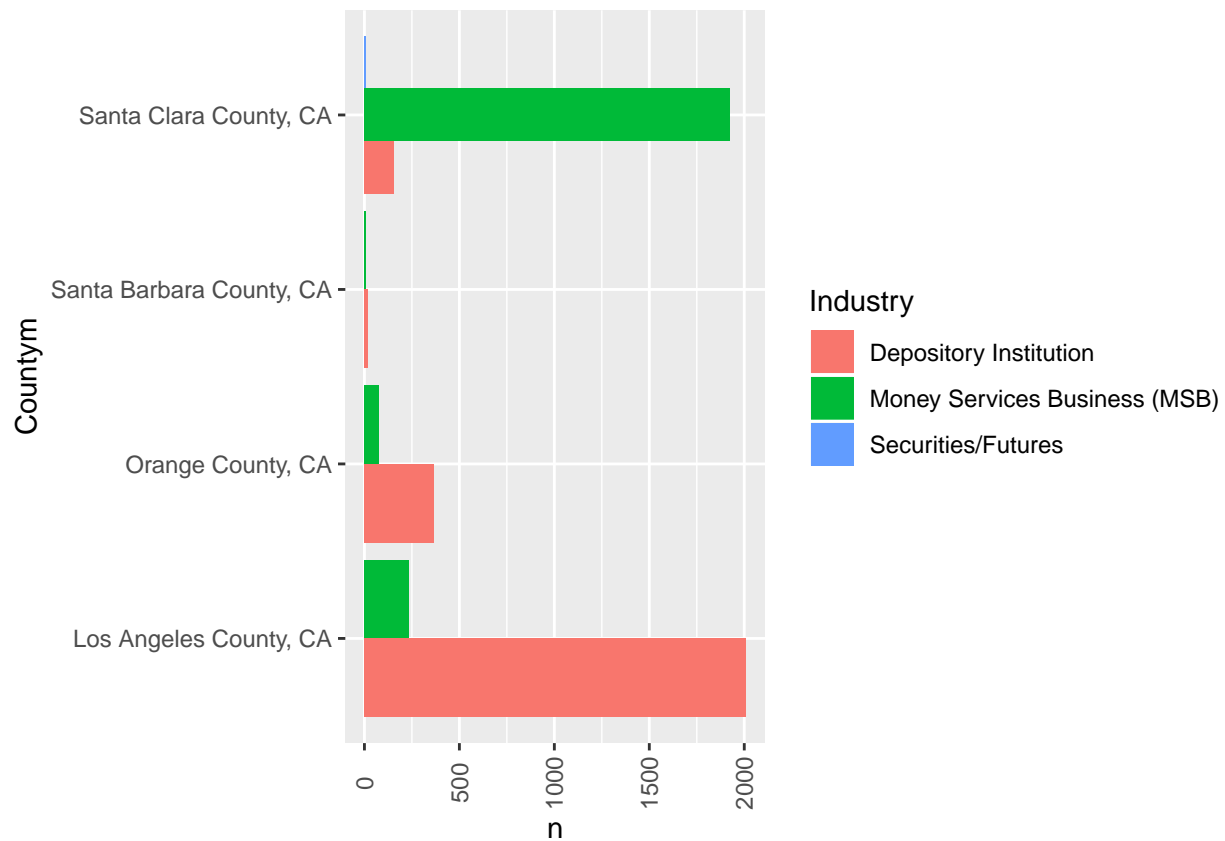
I don't want the text to be overlapped with each other, so I switch the direction of x-axis and y-axis. Now the x-axis becomes vertical and the y-axis becomes horizontal.

Position = "dodge" places overlapping objects directly beside one another. This makes it easier to compare individual values.

```
SARStats <- read_csv("SARStats.csv")

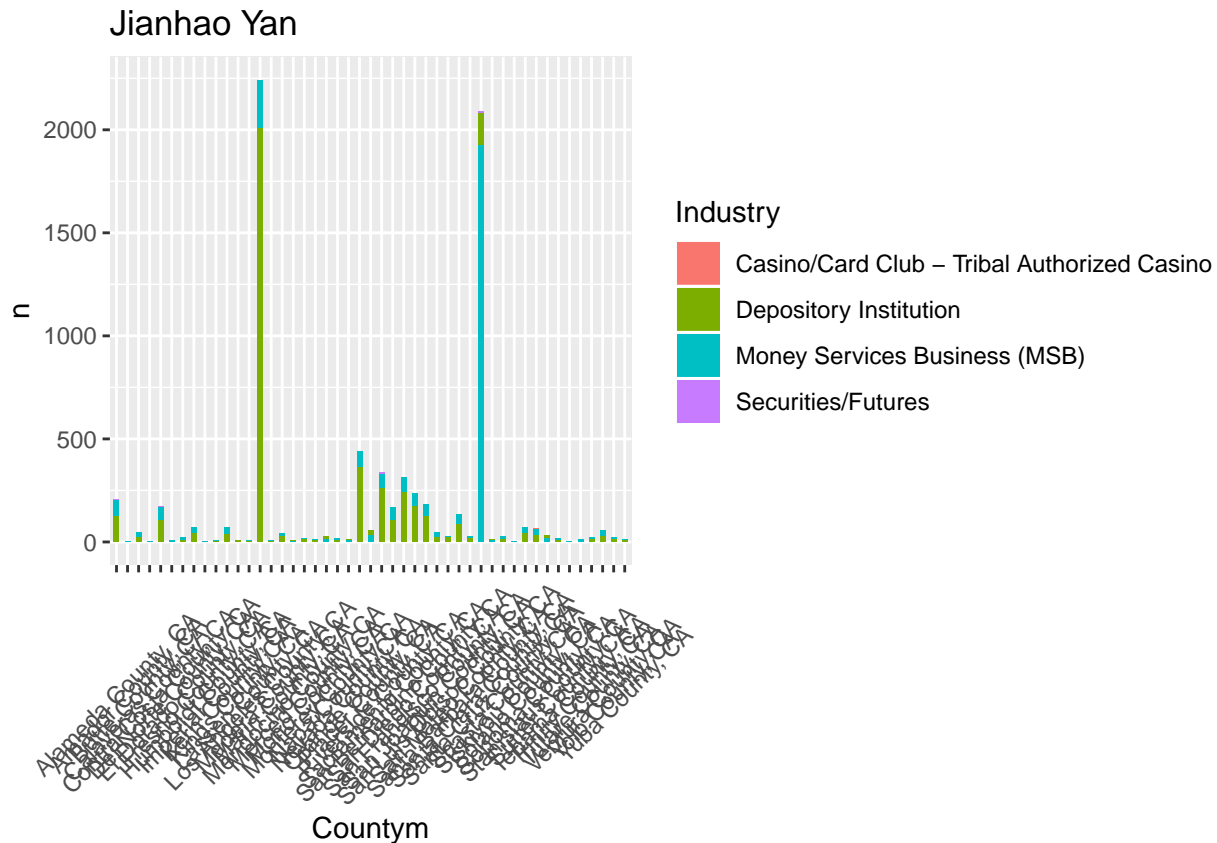
## Parsed with column specification:
## cols(
##   `Year Month` = col_character(),
##   State = col_character(),
##   Countym = col_character(),
##   Industry = col_character(),
##   `Suspicious Activity` = col_character(),
##   Product = col_character(),
##   Instrument = col_character(),
##   Count = col_number()
## )

SARStats$Count <- as.numeric(SARStats$Count)
xuan <- SARStats %>%
  group_by(Industry, Countym) %>%
  summarise(n = sum(Count)) %>%
  filter(Industry != '[Total]' & Countym %in% c("Los Angeles County, CA", "Santa Clara County, CA", "Orange County, CA", "Santa Barbara County, CA"))
  arrange(desc(n))
ggplot(data=xuan, mapping=aes(x=Countym, y=n), group=factor(1), xlab(Countym)) +
  geom_bar(position = "dodge", aes(fill=Industry), stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 0.5, vjust = 0.5)) +
  coord_flip()
```



Jianhao's plot

```
data_new<-SARStats %>%
  group_by(Countym,Industry) %>%
  summarise(n=sum(Count))%>%
  filter(Countym!='[Total]')%>%
  filter(Industry!='[Total]')%>%
  arrange(desc(n))
ggplot(data_new, aes(x = Countym,y =n, group = factor(1))) +
  geom_bar(stat = "identity", width = 0.5,aes(fill=Industry))+theme(axis.text.x = element_text(angle = 45))
```



Discussion & Conclusion

From this graph, we can find that the Los Angeles County, CA has the most financial frauds, and these frauds mostly happened in depository industry.

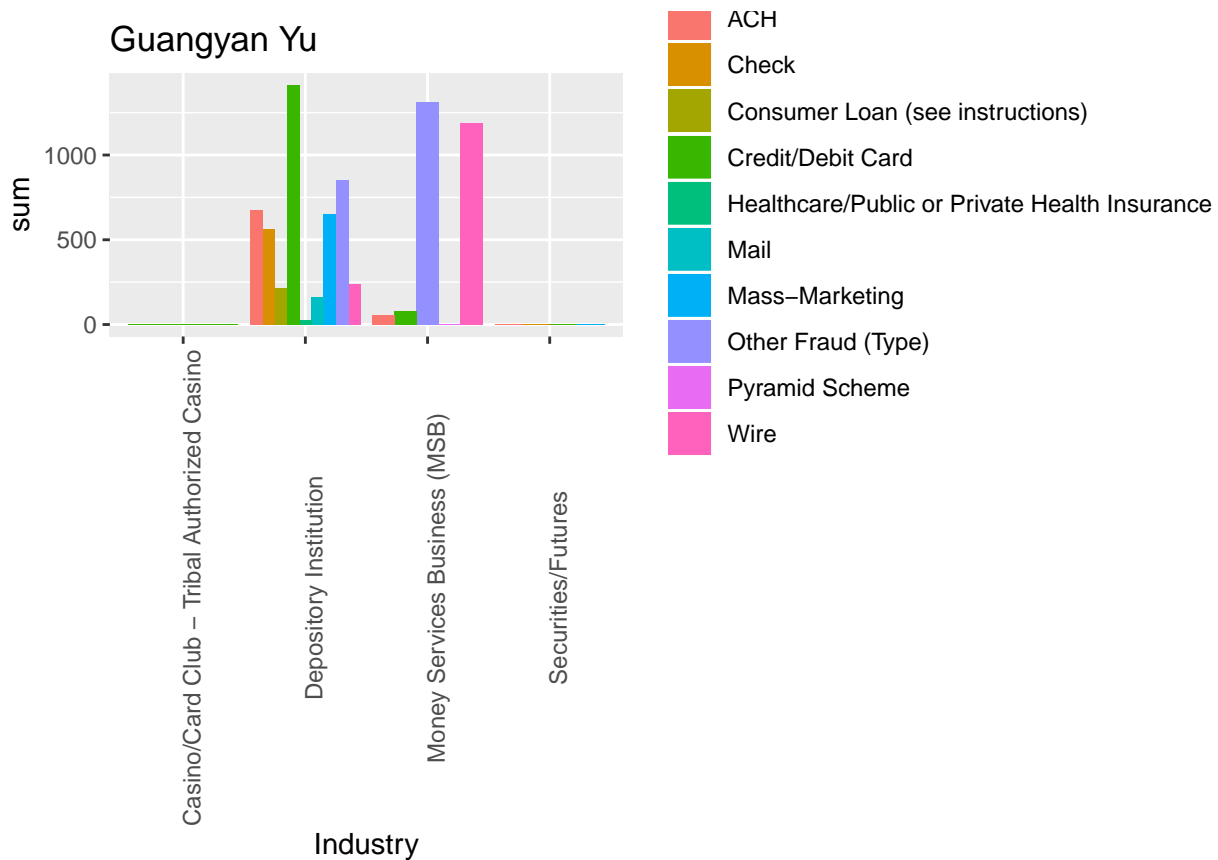
Guangyan's Plot

I want to figure out the relationship between Industry and Suspicious Activity, and find what is the most happening suspicious activity in every Industry so that we could intentionally decrease the suspicious activities.

```
library(knitr)
library(tidyverse)
data<-read.csv("SARStats.csv")
#summary(data)
data$Count <- as.numeric(data$Count)
data <- filter(data, !(str_detect(string = data$Industry,pattern = "\\[Total\\]")))
data <- filter(data, !(str_detect(string = data$Suspicious.Activity,pattern = "\\[Total\\]")))
data <- filter(data, !(str_detect(string = data$Count,pattern = "\\[Total\\]")))
data1<-data %>%
  group_by(Industry,Suspicious.Activity) %>%
  summarise(sum=sum(Count)) %>%
  group_by(Industry) %>%
  arrange(desc(sum))
kable(data1)
```

Industry	Suspicious.Activity	sum
Depository Institution	Credit/Debit Card	1413
Money Services Business (MSB)	Other Fraud (Type)	1315
Money Services Business (MSB)	Wire	1186
Depository Institution	Other Fraud (Type)	854
Depository Institution	ACH	678
Depository Institution	Mass-Marketing	649
Depository Institution	Check	564
Depository Institution	Wire	238
Depository Institution	Consumer Loan (see instructions)	215
Depository Institution	Mail	164
Money Services Business (MSB)	Credit/Debit Card	82
Money Services Business (MSB)	ACH	54
Depository Institution	Healthcare/Public or Private Health Insurance	28
Securities/Futures	ACH	4
Securities/Futures	Check	2
Securities/Futures	Credit/Debit Card	2
Securities/Futures	Mass-Marketing	2
Casino/Card Club - Tribal Authorized Casino	Credit/Debit Card	1
Money Services Business (MSB)	Pyramid Scheme	1

```
ggplot(data1,aes(x=Industry,y = sum,fill = Suspicious.Activity)) + geom_bar(stat="identity",position =
```



Discussion and Conclusion

Through this plot, we can know that, firstly, the two industries—Casino/Card Club - Tribal Authorized Casino and Securities/Futures, have low suspicious activity, while Depository Institution and Money Services Business (MSB) have relatively high number of suspicious activity. Secondly, in the two high suspicious activity industry, Depository Institution has more kinds of suspicious activity than MSB. Thirdly, it is obvious that suspicious activity in Credit/Debit Card sets the most proportion in Depository Institution, and for MSB, suspicious activity in Other Fraud sets the most proportion.