Fields:

####S #: data number

####Location: City where the case occurred, state

####Date: Time

####Summary: Case summary

####Fatalities

####Injured: Number of injured

####Total victims

####Mental Health Issues:

####Race: Race

####Gender

####Latitude: Latitude

####Longitude: Longitude

1. Import package

library(tidyverse)

library(stringr)

library(data.table)

library(maps)

library(lubridate)

library(leaflet)

2. Import and View Datasets

shooting <- read. csv('Mass Shootings Dataset Ver 2.csv',stringsAsFactors = F,header = T)

summary(shooting)

glimpse(shooting)

Conclusion: There are 320 rows of data in total, and the 13 variables are not large, but the data needs to be reconstructed

3. Data reconstruction

#Convert the Date field and create a new variable year

shooting <- shooting %>% select(1:13) %>% mutate(Date=mdy(shooting$Date),year=year(Date))

summary(shooting$year)

#Extract gender

shooting$Gender<-if\_ else(shooting$Gender=="M","Male",shooting$Gender)

#Extract the race field

shooting$Race<-if\_ else(str\_detect(shooting$Race,"Black American or African American"),"Black",shooting$Race)

shooting$Race<-if\_ else(str\_detect(shooting$Race,"White American or European American"),"White",shooting$Race)

shooting$Race<-if\_ else(str\_detect(shooting$Race,"Asian American"),"Asian",shooting$Race)

shooting$Race<-if\_ else(str\_detect(shooting$Race,"Some other race"),"Other",shooting$Race)

shooting$Race<-if\_ else(str\_detect(shooting$Race,"Native American or Alaska Native"),"Native American",shooting$Race)

#Split the time data

shooting$yearcut<-cut(shooting$year,breaks = 10)

#Dispose of mental illness

shooting$Mental.Health.Issues<-if\_ else(str\_detect(shooting$Mental.Health.Issues,"Un"),"Unknown",shooting$Mental.Health.Issues)

shooting$Race<-str\_ to\_ upper(shooting$Race)

shooting$Mental.Health.Issues<-str\_ to\_ upper(shooting$Mental.Health.Issues)

#Decompose location into city and state variables

shooting$city <- sapply(shooting$Location,function(x){

return(unlist(str\_split(x,','))[1] %>% str\_ trim())

})

shooting$state <- sapply(shooting$Location,function(x){

return(unlist(str\_split(x,','))[2] %>% str\_ trim())

})

4. EDA analysis

4.1 Annual changes in the number of gunshot deaths

#Number of people killed by gunfire every year

shooting %>%

group\_ by(year) %>%

summarise(total=sum(Total.victims)) %>%

ggplot(aes(x=year,y=total)) +

geom\_ bar(stat = 'identity',fill='blue') +

geom\_ text(aes(label=total),vjust=-0.2) +

xlim(1969,2020) +

geom\_ line(color='red') +

ylab('Total victims every year') +

ggtitle('People died because of gun shoot every year')

Conclusion: After 2015, shootings occurred frequently in the United States, and the number of deaths due to shootings increased significantly in 2017

4.2 Location of shooting

#Geographical distribution of the number of injured

shooting %>%

select(Total.victims,Fatalities,Longitude,Latitude,Summary) %>%

na. omit() %>%

leaflet() %>%

addProviderTiles(providers$OpenStreetMap) %>%

fitBounds(-124,30,-66,43) %>%

addCircles(color='#8A0707',lng = ~Longitude,lat = ~Latitude,weight = 1,

radius = ~sqrt(Total.victims) \* 20000,popup = ~Summary)

#Geographical distribution of deaths

shooting %>%

select(Total.victims,Fatalities,Longitude,Latitude,Summary) %>%

na. omit() %>%

leaflet() %>%

addProviderTiles(providers$OpenStreetMap) %>%

fitBounds(-124,30,-66,43) %>%

addCircles(color='blue',lng = ~Longitude,lat = ~Latitude,weight = 1,

radius = ~sqrt(Fatalities) \* 20000,popup = ~Summary)

Distribution of injuries and deaths

Conclusion: From the perspective of geographic information and population information, the probability of shooting in the eastern United States is higher than that in the western United States

4.3 Gender distribution of gunmen

shooting %>%

ggplot(aes(x=factor(Gender),fill=factor(Gender)))+

geom\_ bar()+

xlab('Gender')+

ylab('Number of each Gender')+

ggtitle('The distribution of gender')

Conclusion: Men are far more likely to commit crimes than women

4.4 Ethnic distribution of shooting cases

shooting %>%

na. omit() %>%

group\_ by(Race) %>%

summarise(num=sum(Total.victims)) %>%

ggplot(aes(x=factor(Race),y=num,fill=factor(Race)))+

geom\_ bar(stat = 'identity')+

coord\_ polar(theta = 'y')+

labs(x='Race',y='Number of killed people',fill='Race')+

ggtitle('People killed by different race')

Conclusion: There are many white criminals, but the number of black criminals is also increasing

4.5 Monthly distribution of shooting cases

shooting %>%

mutate(month=month(Date)) %>%

group\_ by(month) %>%

summarise(n=sum(Total.victims)) %>%

ggplot(aes(x=factor(month),y=n)) +

geom\_ bar(stat = 'identity')+

labs(x='month',y='Number of killed people')+

ggtitle('The distribution of killed people every month')+

geom\_ text(aes(label=n),vjust=-0.2,color='red')+

theme\_ bw()

Conclusion: The number of shootings in October is the highest and the most dangerous

4.5 Does the shooter have mental illness

shooting %>%

na. omit() %>%

ggplot(aes(x=Mental.Health.Issues)) +

geom\_ bar()+

scale\_ x\_ discrete(limits=c("NO","YES"))+

theme\_ bw()

Conclusion: Whether the killer has mental illness is not a main reason

4.6 Whether the number of people with mental illness and those without mental illness is different

shooting %>%

na. omit() %>%

group\_ by(Mental.Health.Issues) %>%

summarise(n=sum(Total.victims)) %>%

ggplot(aes(x=factor(Mental.Health.Issues),y=n,group=1)) +

geom\_ bar(stat = 'identity',fill='pink')+

scale\_ x\_ discrete(limits=c('NO','YES'))+

geom\_ text(aes(label=n),vjust=-0.2)+

geom\_ line(color='red')

Conclusion: The number of murderers with mental illness is twice that of those without mental illness, and the harm of psychiatric gunmen is greater

4.7 Statistics of shooter race in different time periods

shooting %>%

na. omit() %>%

group\_ by(yearcut) %>%

ggplot(aes(x=yearcut,fill=Race))+

geom\_ bar(position = 'dodge')

Conclusion: It can be seen that although the shooting cases are mainly white, the number of black overturned cases has also been increasing in recent years

4.8 Age distribution of gunmen

#Extract age from summary by regular expression

tmp <- mutate(shooting,age=str\_extract\_all(shooting$Summary,pattern = '(,\\s)\\d{2}(,)'),

age2 = str\_ extract\_ all(shooting$Summary,pattern = '(a\\s)\\d{2}(-year)'))

tmp$age <- str\_ sub(tmp$age,3,4)

tmp$age2 <- str\_ sub(tmp$age2,3,4)

#Remove fields with unknown age

te <- subset(tmp,tmp$age != 'ar')

te2 <- subset(tmp,tmp$age2 != 'ar')

te <- rbind(te,te2)

for(i in 1:nrow(te)){

if(te$age[i] == 'ar'){

te$age[i] = te$age2[i]

}

}

te <- arrange(te,age)

te <- te[-c(1:4),]

te <- arrange(te,S.)

te$age <- as.integer(te$age)

te3 <- te %>%

select(S.,age) %>%

mutate(agecut=cut(te$age,breaks = 10\*(1:7)))

shoot\_ age <- left\_ join(te3,shooting)

ggplot(data=shoot\_age,aes(x=agecut))+

geom\_ bar(fill='blue')+

theme\_ bw()

Conclusion: From the perspective of age distribution, young people are more likely to commit crimes, and impulsivity is the devil

4.9 Distribution of mental disorders in different age groups

ggplot(data=shoot\_age,aes(x=agecut,fill=Mental.Health.Issues))+

geom\_ bar()

Conclusion: Gunmen aged 10-20 and 30-40 years old are the high risk group of mental diseases

4.10 Urban distribution and state distribution of shooting cases

#Urban distribution

shooting %>%

group\_ by(city) %>%

summarise(count=n()) %>%

filter(city != '' & count >= 2) %>%

ggplot(aes(x=reorder(city,count),y=count))+

geom\_ bar(stat = 'identity',fill='lightblue')+

coord\_ flip()+

labs(x='City',y='Number of gun-shot happended')+

ggtitle('The number of case happened in each city')

#State distribution

shooting %>%

group\_ by(state) %>%

summarise(count=n()) %>%

filter(state != '' & count >= 2) %>%

ggplot(aes(reorder(state,count),y=count))+

geom\_ bar(stat='identity',fill='lightblue')+

coord\_ flip()+

labs(x='State',y='Number of gun-shot happended')+

ggtitle('The number of case happened in each state')

1. Judging from the gender of the gunmen, the majority of men commit crimes

2. From the perspective of the shooter's race, white people are the main perpetrators, but the number of black criminals is also increasing year by year

3. Judging from the age distribution of gunmen, young and middle-aged people between the ages of 10 and 50 account for the vast majority

4. From the perspective of the shooter's mental illness, although the number of shooters suffering from mental illness and not suffering from mental illness is not significant, the shooters suffering from mental illness will cause greater harm, so we must focus on controlling

5. In terms of the time of shooting cases, gun crimes increased the most in 2015, but in 2017, there was an extreme increase, which shows the importance of gun control

6. From the geographical information of gun cases, on the whole, the number of shooting cases in the east is greater than that in the west

7. In terms of the number of shootings, California has the largest number of shootings in recent years

Work3

library(Matching)

library(tableone)

bc<-read.csv("E:/r/test/zaochan.csv",sep=',',header=TRUE)

Dput (names (bc)) # # Output variable name

allVars <-c("age", "lwt", "race", "smoke", "ptl", "ui",

"Ftv") # # # All variable names

Fvars<- c ("race", "smoke", "ui") # Classification variables are defined as fvars

tab2 <- CreateTableOne(vars = allVars, strata = "ht" , data = bc, factorVars=fvars,

AddOverall=TRUE) # # # Draw baseline table

Print (tab2) # Output table

fit1<- glm(ht ~age + lwt + race + smoke + ptl + ui + ftv, data=bc,

family=binomial(link = "logit"))

rr <- Match(Y=Y, Tr=Tr, X=X, M=1)

bcMatched <- bc[unlist(rr [c("index.treated","index.control")]), ]

图表, 散点图

描述已自动生成 图片包含 照片, 厨房, 桌子, 侧面

描述已自动生成