Code

CS544 project

1. Introduction

1. Dataset US mass shootings dataset from 1966-2019 (partial)
2. Size: 339 rows and 24 columns
3. Columns:
4. Source: https://www.kaggle.com/myho63/us-mass-shooting-1966-2019?select=US+Mass+Shooting+1966-2019+%28cleaned%29.csv (https://www.kaggle.com/myho63/us-mass-shooting-1966-2019?select=US+Mass+Shooting+1966-2019+%28cleaned%29.csv)
5. Description:

This data covers raw data of mass shooting cases in US from 1966-2019. In fact, it does not cover all cases but it gives analyst/ viewer a general viewpoint of gun violence situation in US. I want to find some connections among those cases and certain conditions what causes mass shooting, either classification or prediction.

6. Goal:

Do some analysis to find pattern of US mass shooting cases.

Do all tasks required in the project requirements.

2. Preprocessing data

1. Read mass_shooting dataset and replace -999 to NA
2. Fix column names which are not well-formated
3. Sort the data by date and remove date input error.

Hide

```
df <- read_csv('mass_shooting.csv', na = c(-999, "NA") )
# check column names and remove unneeded ones
# fix Date
df <- df %>% rename(
  Incideent_are = `Incident Area`,
  Location_type = `Open/Close Location`,
  Shooter_status = `Shooter status`,
  Shooter_number = `No. of shooter/suspect`,
  Total_vicitimes = `Total victims`,
  Policeman_killed = `Policeman Killed`,
  Suspect_age = Age,
  Employeed = `Employeed (Y/N)`,
  Employed_at = `Employed at`,
  Mental_problem = `Mental Health Issues`
) %>% select(4, 5, 3, 7:24, -10) %>% mutate(
  Date = as.Date(df$Date, "%m/%d/%y"),
  Mental_problem = ifelse(Mental_problem == 'yes', 'Yes', Mental_problem),
  Mental_problem = replace_na(Mental_problem, 'Unknown'),
  Cause = replace_na(Cause, 'Unknown'),
  Target = replace_na(Target, 'Unknown')
) %>% arrange(desc(Date) ) %>% filter(Date < '2020-12-31')
# check again
glimpse(df)
```

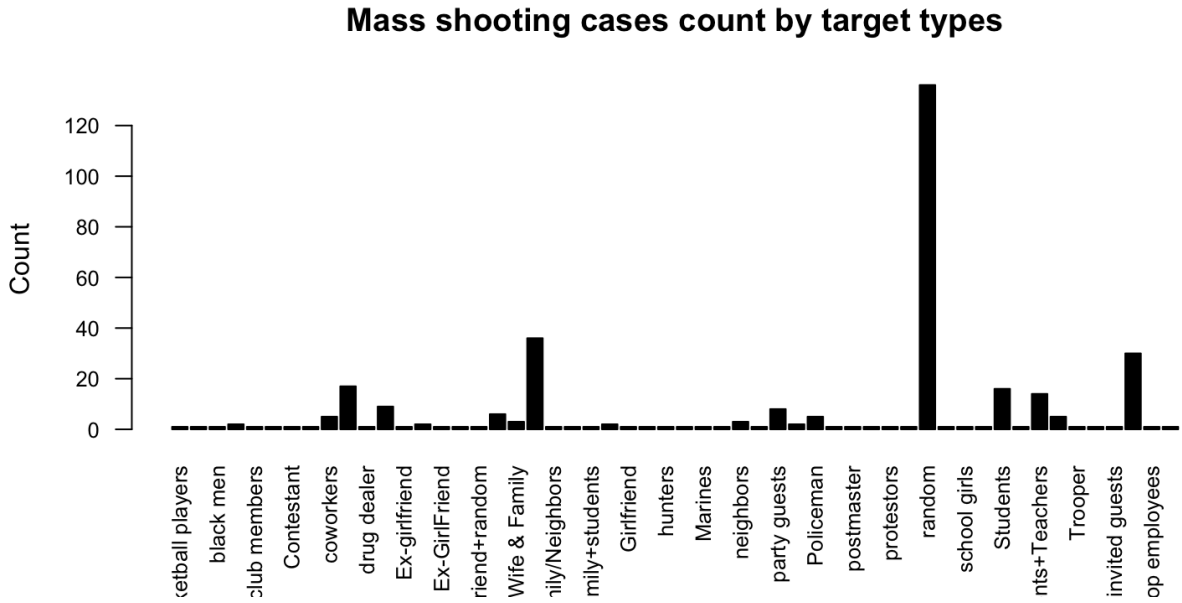
```
## Rows: 337
## Columns: 20
## $ Date             <date> 2019-12-10, 2019-12-06, 2019-08-31, 2019-08-04, 20
1…
## $ Area             <chr> "Public area", "Military classroom", "Public area",
…
## $ Location         <chr> "Jersey City, New Jersey", "Pensacola, Florida", "O
d…
## $ Location_type    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
…
## $ Target           <chr> "Unknown", "Unknown", "Unknown", "Unknown", "Unknow
n…
## $ Cause            <chr> "Unknown", "Unknown", "Unknown", "Unknown", "Unknow
n…
## $ Shooter_status   <chr> "killed", "killed", "killed", "arrested", "arreste
d"…
## $ Shooter_number   <chr> "two", "one", "one", "one", "one", "one", "one", "o
n…
## $ Fatalities       <dbl> 4, 3, 7, 9, 22, 3, 12, 5, 3, 5, 3, 12, 11, 3, 5, 3,
…
## $ Injured          <dbl> 3, 8, 25, 27, 26, 12, 4, 6, 1, 0, 0, 22, 6, 3, 0,
2,…
## $ Total_vicitimes  <dbl> 7, 11, 32, 36, 48, 15, 16, 11, 4, 5, 3, 34, 17, 6,
5…
## $ Policeman_killed <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
…
## $ Suspect_age      <chr> "47", "-", "36", "24", "21", "19", "40", "45", "2
1",…
## $ Employeed        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
…
## $ Employed_at      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
…
## $ Mental_problem   <chr> "Unknown", "Unknown", "Yes", "Unknown", "Unknown",
"…
## $ Race             <chr> "Black, Black American or African American", NA, "W
h…
## $ Gender           <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Ma
l…
## $ Latitude         <dbl> 40.71, 30.36, 31.93, 39.76, 31.77, 37.00, 36.75, 4
1.…
## $ Longitude        <dbl> -74.08, -87.29, -102.28, -84.18, -106.38, -121.58,
-…
```

3. Analysis categorical variables of Target type and Cause

Target

Hide

```
barplot(table(df$Target), las = 2, cex.names = 0.8, cex.axis = 0.8,
        col = 'black', main = 'Mass shooting cases count by target types',
        ylab = 'Count')
```

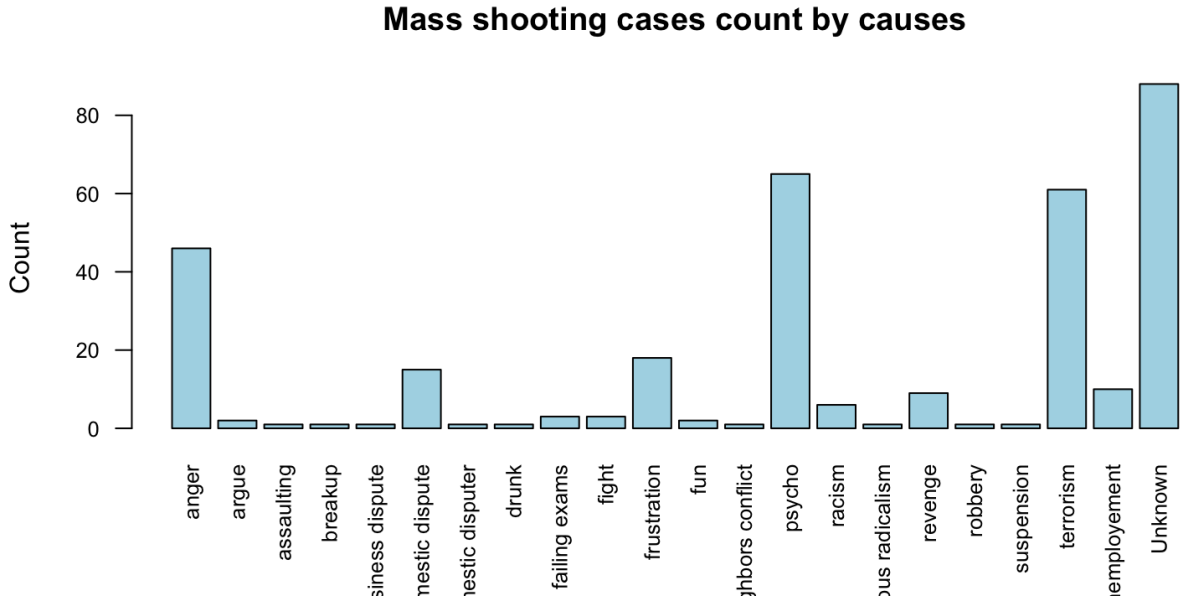


Most mass shooting cases are random targets

Causes

Hide

```
barplot(table(df$Cause), las = 2, cex.names = 0.8, cex.axis = 0.8,
        col = 'lightblue', main = 'Mass shooting cases count by causes',
        ylab = 'Count')
```

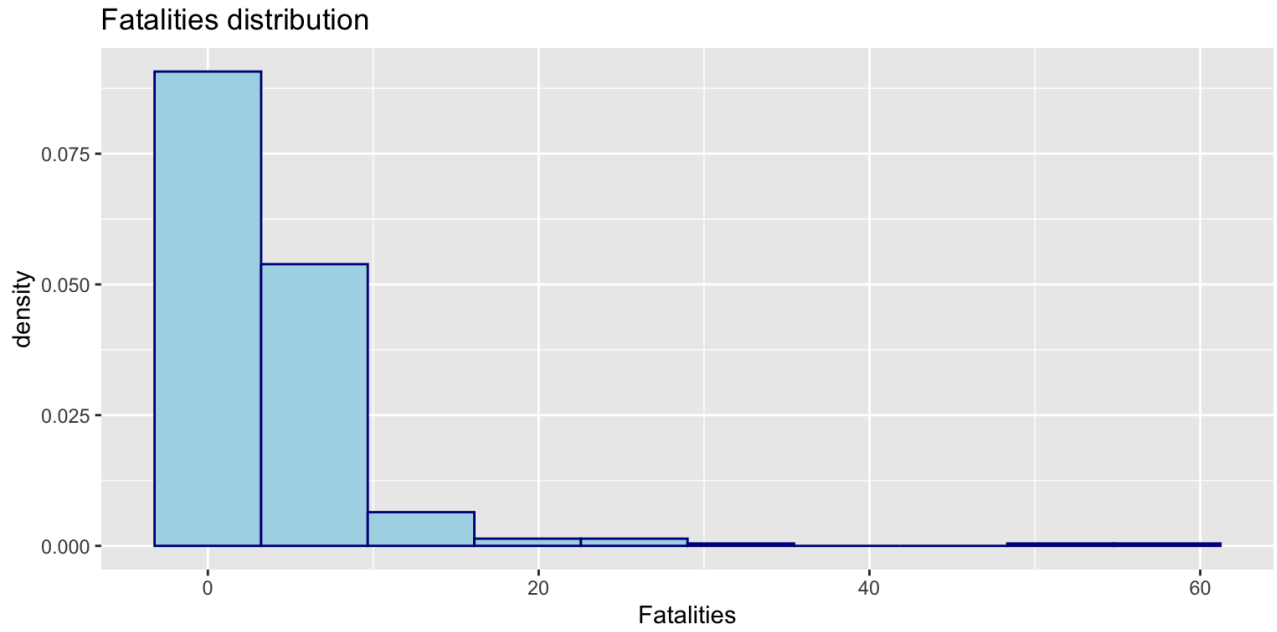


We can see psycho causes the most mass shooting cases.

Analysis numeric variable of Fatality

Hide

```
ggplot(df) + geom_histogram(aes(x = Fatalities, y = ..density..), bins = 10,
color="darkblue", fill="lightblue") + labs(title = 'Fatalities distribution')
```



Most mass shooting cases have fatalities in the range of 2 to 10.

4. Check if total victim number and fatalities are related to cause or suspect mental problem

Hide

```
# first to check correlation between deaths and victims, they should strongly p
        ositive associated

cor(df$Total_vicitimes, df$Fatalities)
```

```
## [1] 0.7084
```

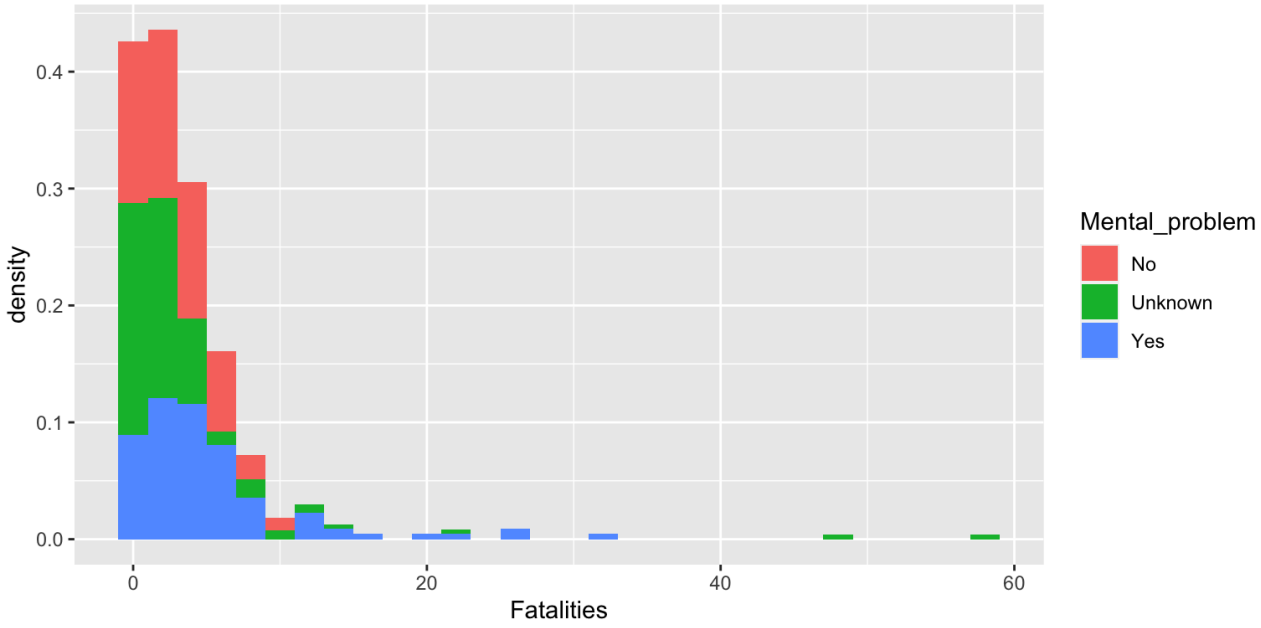
Hide

```
df %>% group_by(Mental_problem) %>% summarise(
  death = mean(Fatalities),
  victim = mean(Total_vicitimes)
) %>% arrange( desc(death, victim))
```

```
## # A tibble: 3 x 3
##   Mental_problem death victim
##   <chr>          <dbl>  <dbl>
## 1 Yes             5.60  12.1
## 2 Unknown         3.60  11.9
## 3 No              3.39   6.53
```

Hide

```
ggplot(df) + geom_histogram(aes(Fatalities, y = ..density..,
                                fill = Mental_problem))
```



Hide

```
df %>% group_by(Cause) %>% summarise(
  death = mean(Fatalities),
  victim = mean(Total_vicitimes)
) %>% arrange( desc(death, victim))
```

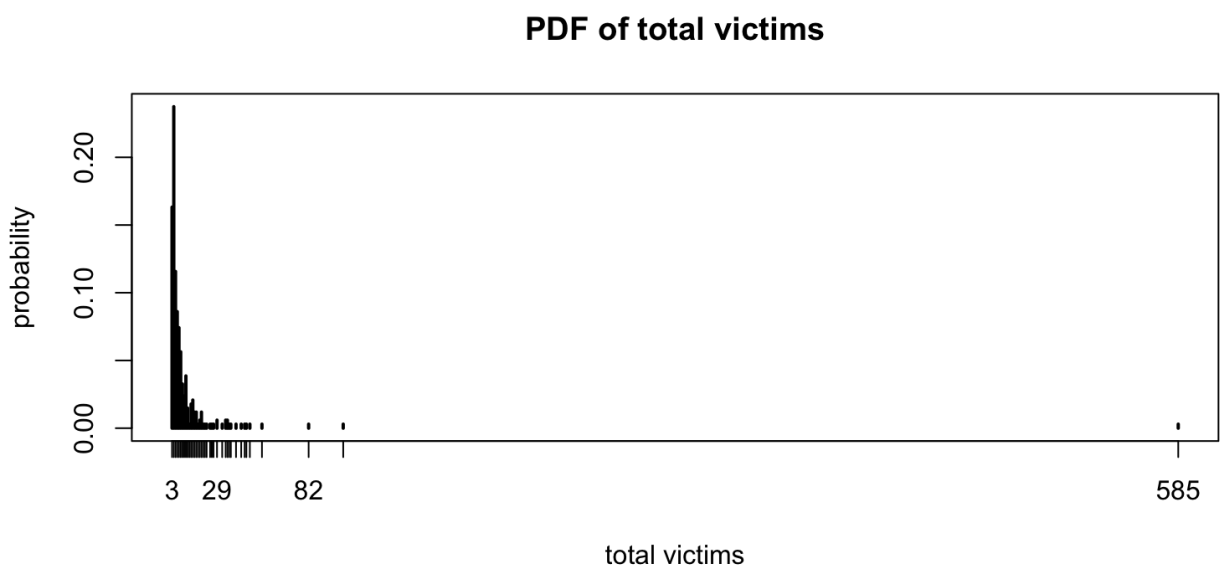
```
## # A tibble: 22 x 3
##    Cause                death victim
##    <chr>                <dbl>  <dbl>
##  1 robbery               9      9
##  2 drunk                 7      9
##  3 unemployement         6.6   10.4
##  4 domestic disputer     6      6
##  5 terrorism             5.33  12.8
##  6 neighbors conflict    5      5
##  7 religious radicalism  5      9
##  8 Unknown               4.94  16.2
##  9 psycho                4.4    8.32
## 10 business dispute      4      4
## # … with 12 more rows
```

1. We can criminals with mention conditions committed the majority of mass shooting cases with big numbers of victims and deaths. Thus, we could say that taking care of people who have mention issues could reduce the damage of mass shooting cases.
2. Robbery is the most dangerous reason for shooting cases. In general robbery cases, criminals primarily wanted money but the situations always escalated to gun violence.

5. See the distribution of total victims

Hide

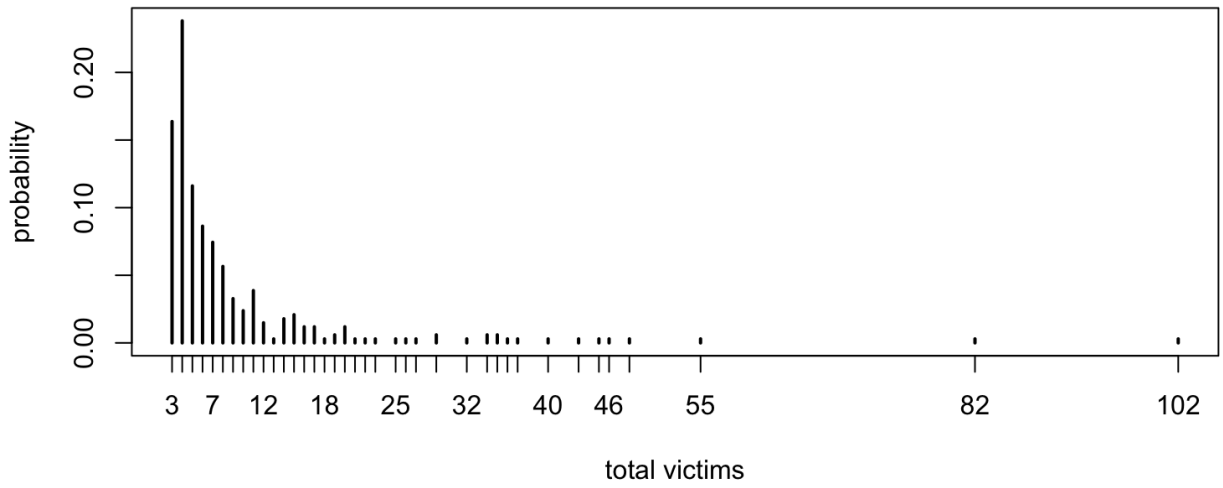
```
plot(prop.table(table(df$Total_vicitimes)), type = 'h',
     ylab = 'probability', xlab = 'total victims', main = 'PDF of total victim
        s')
```



Hide

```
# remove the outlier 585 and plot the pdf again

victim2 = subset(df$Total_vicitimes, df$Total_vicitimes != max(df$Total_vicitim
        es))
plot(prop.table(table(victim2)), type = 'h',
     ylab = 'probability', xlab = 'total victims', main = 'PDF of total victim
        s')
```



Total victims are right skewed, a outlier of 585 pulled the distribution to the right side. After removing the outlier, the plot is still right skewed. It means a few mass shooting cases have more than 50 victims. Those cases are a larger-scale compared to other cases.

6. Sampling

simple random sampling

Hide

```
victim = df$Total_vicitimes
N = nrow(df)
sample_size = 50
n_samples = 30
xbar = numeric(n_samples)
for (i in 1:n_samples) {
  p = sample(victim, sample_size)
  xbar[i] = mean(p)
}
xbar
```

```
##  [1]  8.10  7.74 11.10  8.32  6.86  9.92  7.72 12.66 19.52 14.68  7.20 19.58
## [13]  8.28  9.62  8.72  9.62 10.30  8.60  8.58 19.40  9.54  9.78  8.12 23.98
## [25]  7.82  8.00  7.70  8.00  9.04  7.08
```

Hide

```
mean(xbar)
```

```
## [1] 10.52
```

Hide

```
mean(victim)
```

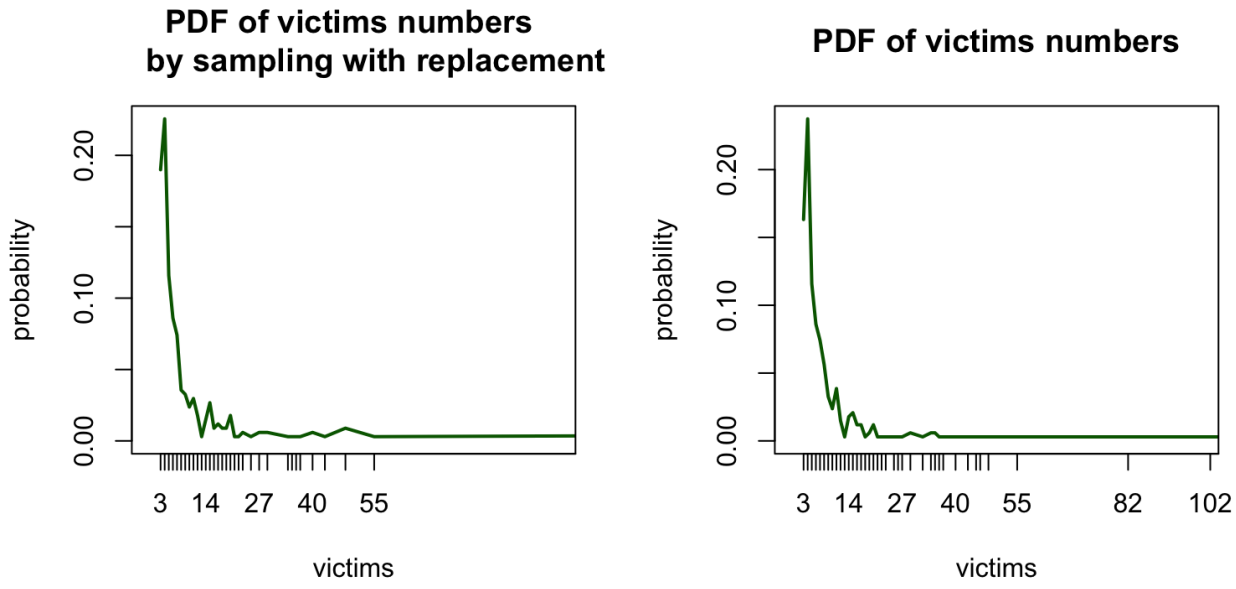
```
## [1] 10.46
```

We can see the mean of total victims of random samples is close to the mean of population. So, Central limit theory is applicable to total victimes.

bootstrapping

Hide

```
# total number of samples from the population with replacement
bootstrapping = sample(victim, N, replace = T)
par(mfrow = c(1, 2))
plot(prop.table(table(bootstrapping)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims numbers
     by sampling with replacement',
     ylab = 'probability', xlab = 'victims', xlim = c(0, 100))
plot(prop.table(table(df$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims numbers',
     ylab = 'probability', xlab = 'victims', xlim = c(0, 100))
```

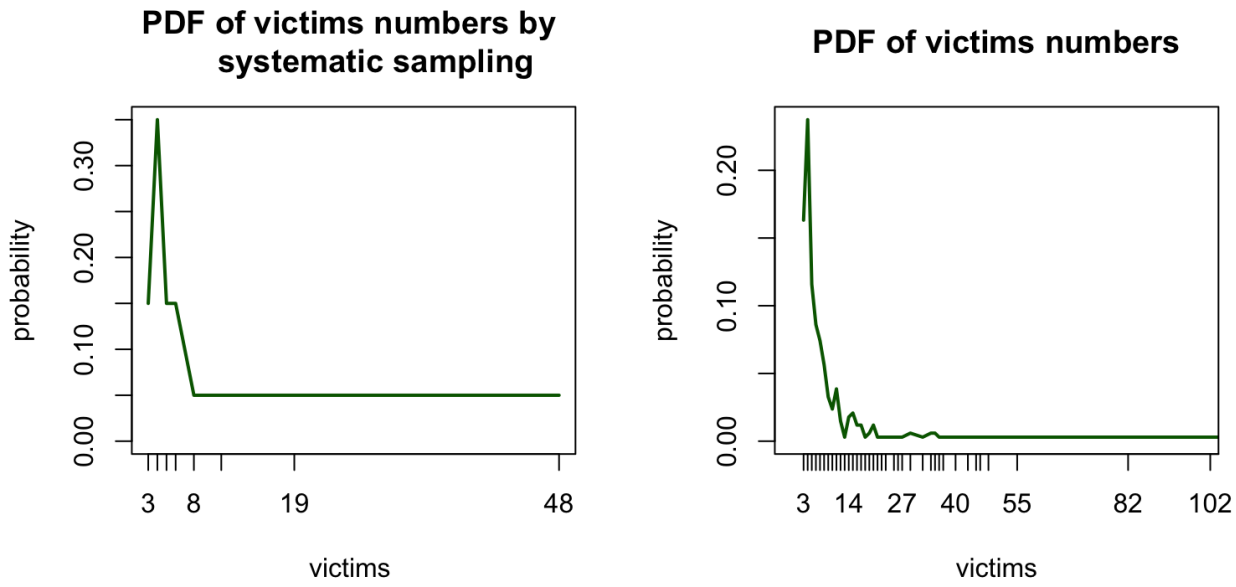


systematic sampling with step size 20

Hide

```
par(mfrow = c(1, 2))
k = ceiling(N / 20)
sys_samples = df[seq(from = sample(k, 1), by = k, to= N), ]
plot(prop.table(table(sys_samples$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims numbers by
     systematic sampling',
     ylab = 'probability', xlab = 'victims ')

plot(prop.table(table(df$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims numbers',
     ylab = 'probability', xlab = 'victims', xlim = c(0, 100))
```



Hide

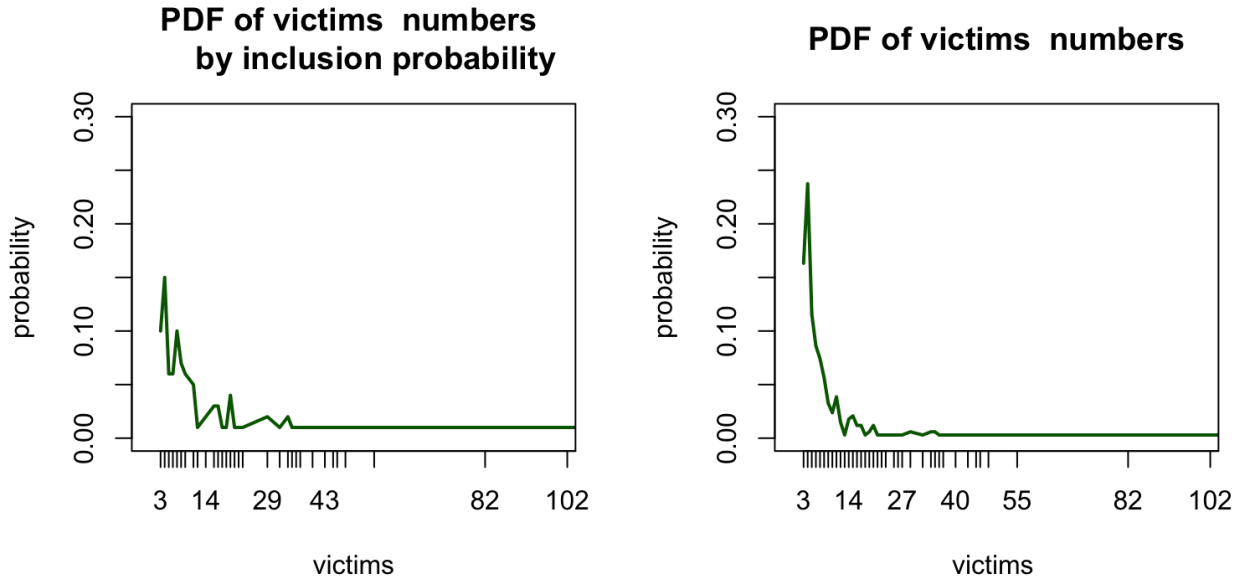
```
par(mfrow = c(1, 1))
```

inclusion probability

Hide

```
pik <- inclusionprobabilities(df$Fatalities, 100)
sp <- UPsystematic(pik)
in_samples = df[sp != 0, ]
par(mfrow = c(1, 2))
plot(prop.table(table(in_samples$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims  numbers
     by inclusion probability',
     ylab = 'probability', xlab = 'victims', ylim = c(0, 0.3),
     xlim = c(0, 100))

plot(prop.table(table(df$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims  numbers',
     ylab = 'probability', xlab = 'victims ',
     ylim = c(0, 0.3), xlim = c(0, 100))
```



Hide

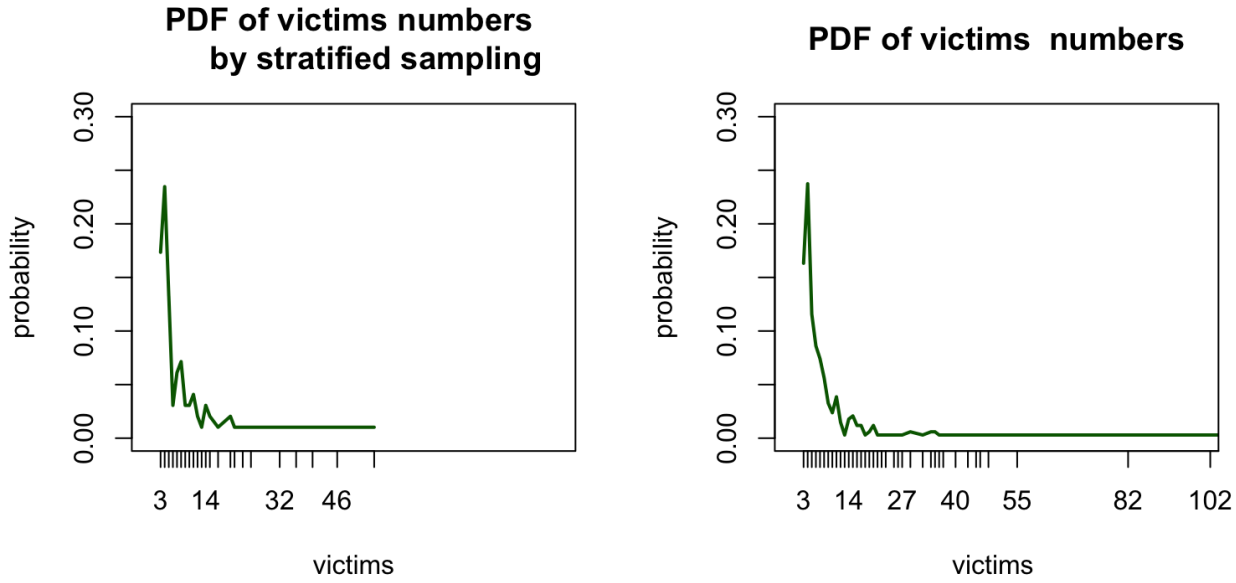
```
par(mfrow = c(1, 1))
```

stratified sampling

Hide

```
df_stra <- df[, c('Mental_problem', 'Total_vicitimes')]
freq <- table(df_stra$Mental_problem)
st_size = 100 * freq / sum(freq)
stra_reg <- strata(df_stra, stratanames = 'Mental_problem', size = st_size, met
        hod = 'srswor')
stra_simples = getdata(df_stra, stra_reg)$Total_vicitimes
par(mfrow = c(1, 2))
plot(prop.table(table(stra_simples)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims numbers
     by stratified sampling',
     ylab = 'probability', xlab = 'victims', ylim = c(0, 0.3),
     xlim = c(0, 100))

plot(prop.table(table(df$Total_vicitimes)), type = 'l',
     col = 'darkgreen', main = 'PDF of victims  numbers',
     ylab = 'probability', xlab = 'victims ',
     ylim = c(0, 0.3), xlim = c(0, 100))
```



Hide

```
par(mfrow = c(1, 1))
```

Simple sampling and stratified sampling with mental problem have similar distributions with population. In addition, stratified sampling are limited by label columns which is used to sample victims.

7. Conclusion

From this US mass shooting dataset, I explored the relationships between different variables. The result shows US mass shooting cases are caused by multiple social problems, such as psychopath, poverty and alcoholic.The majority of cases are not terrorism attacks. So, if the gun control could be tight, mass shooting cases could be reduced significantly. The other two measure could be improve mental health care and reduce poverty.

8. What I learned

1. Real world data are not as clean and neat as what we used in class. Cleaning data could be a big work.
2. Bootstrapping could mostly preserve sample distribution when the population is not available. If we want to do classifications, stratified sampling is better because labels are not always balanced. Dealing with imbalanced dataset could be a trick work. The most population approaches are applying different sampling technics(oversampling and undersampling) or using imbalanced mathmetical formulas.
3. Visualization interprets results derived from dataset, to make results being understood by people.