Project Report

Wenyi Zhong

1. Preparation for using R

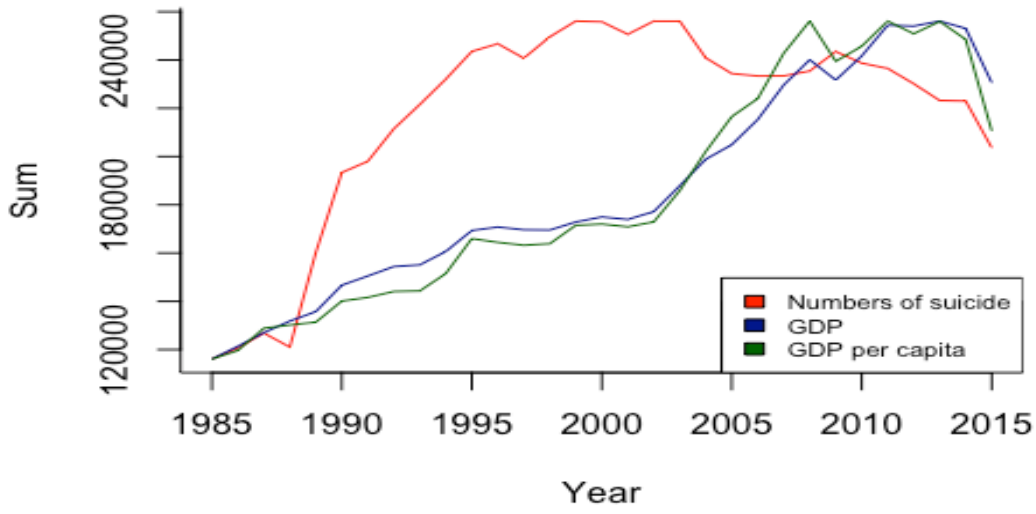
```
setwd("~/Desktop/MET CS544/Final Project")
library(foreign)
library(plyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(plotrix)
library(sampling)
```

2. Importing dataset and cleaning dataset

```
par(mfrow = c(1,1))
data = read.csv("SuicideRates.csv") # Import dataset of suicide rate
data = data[data$year <= "2015", ] # Remove the data of year 2016 (data
 is incomplete in 2016) to get rid of possible outliers
names(data)[7] = "rate" # Change the name of a variable in order to mak
e it easier to use
names(data)[1] = "region"


# List all countries that appear in the dataset
country = as.data.frame(unique(data$region))

# Summary of the total number of people who committed suicide by year.
```

3. Numeric data analysis

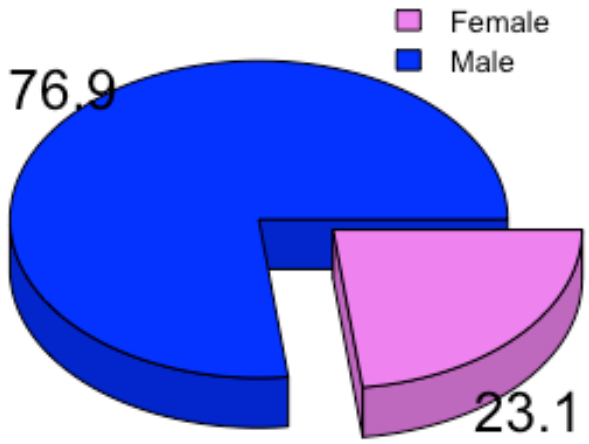
```
suicide_sum_year = aggregate(data$suicides_no, by = list(data$year), FU
N = sum)
colnames(suicide_sum_year) = c("Year", "Sum")
plot(suicide_sum_year, type = "l", col = "red", bty = "l") # Plot of th
e summary above
par(new = TRUE)
data$gdp = as.numeric(gsub(",","", data$gdp_for_year....))
gdp_year = unique(data[c("year", "gdp")])
gdp_year_sum = aggregate(gdp_year$gdp, by = list(gdp_year$year), FUN =
sum)
colnames(gdp_year_sum) = c("Year", "GDP")
plot(gdp_year_sum, type = "l", axes = FALSE, xlab = "", ylab = "", col
= "darkblue")
legend("bottomright", c("Numbers of suicide", "GDP"), cex = 0.7, ncol =
 1, fill = c("red", "darkblue"))
par(new = TRUE)
gdppc_year = unique(data[c("year", "gdp_per_capita....")])
gdppc_year_sum = aggregate(gdppc_year$gdp_per_capita...., by = list(gdp
pc_year$year), FUN = sum)
colnames(gdppc_year_sum) = c("Year", "GDPPC")
plot(gdppc_year_sum, type = "l", axes = FALSE, xlab = "", ylab = "", co
l = "darkgreen")
legend("bottomright", c("Numbers of suicide", "GDP", "GDP per capita"),
 cex = 0.7, ncol = 1, fill = c("red", "darkblue", "darkgreen"))
```



By plotting the sum of people who committed suicide, GDP and GDP per capita t ogether, we can see that the relationship between suicided number of people and G DP or GDP per capita may be a negative related relationship, which can also be expla ined by commonsense that one is less likely to commit suicide if the economic situati on gets better.

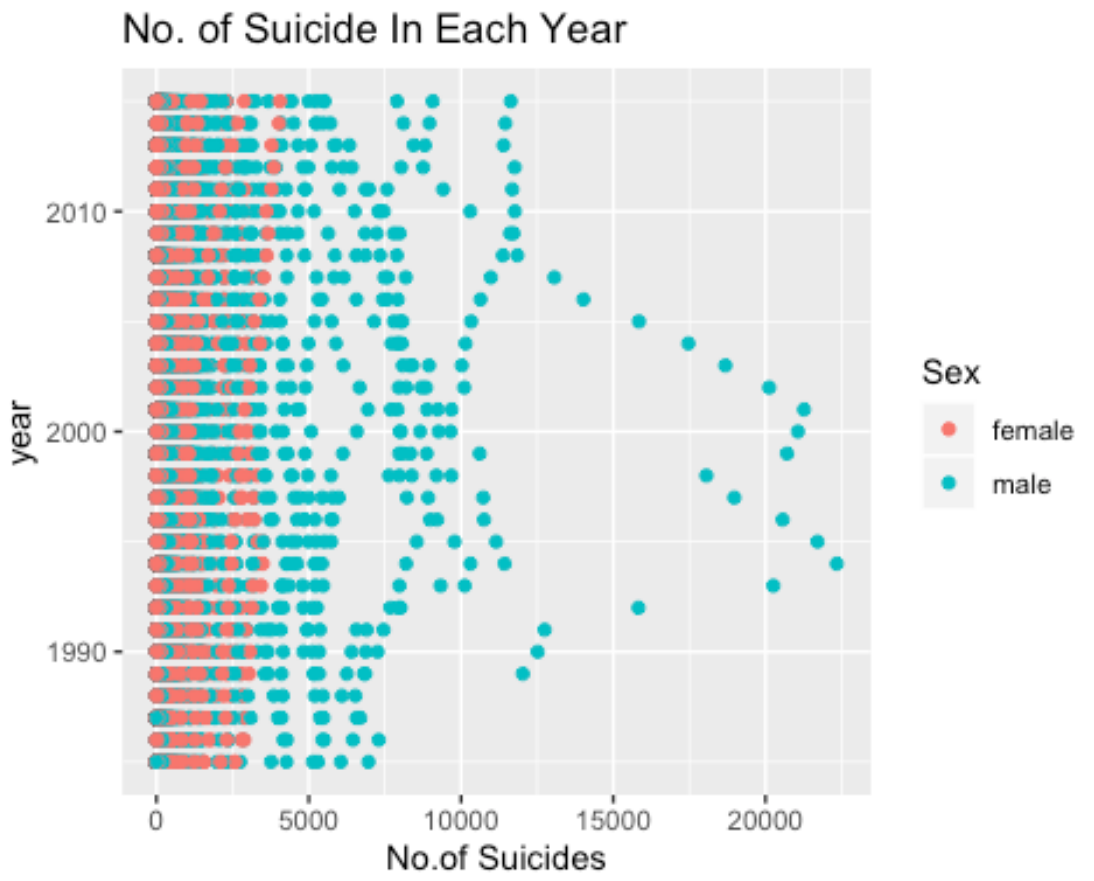
4. Categorical data analysis

```
male = sum(subset(data$suicides_no, data$sex == "male"))
female = sum(subset(data$suicides_no, data$sex == "female"))
gender_sum = c(male, female)
piepercent = round(100 * gender_sum / sum(gender_sum), 1)
pie3D(gender_sum, labels = piepercent ,col = c("blue", "violet"), explo
de = 0.2, theta = 1)
legend("topright", c("Female","Male"), cex = 0.8,ncol=1,fill = c("viole
t", "blue"), bty = "n")
```




```
NO.Of_Suicidies = data$suicides_no
Year = data$year
Sex = factor(data$sex)
```

```
ggplot(data, aes(x=NO.Of_Suicidies, y=Year, color=Sex))+geom_point(shape=19)+labs(x="No.of Suicides",y="year",title="No. of Suicide In Each Year")
```

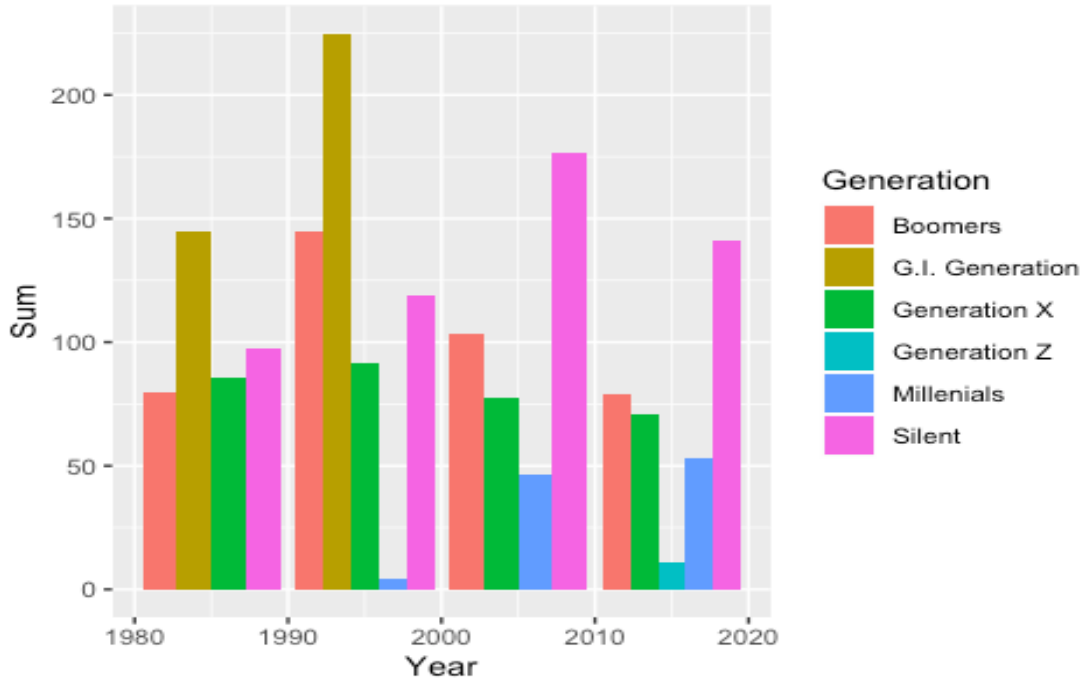


After plotting the suicided number of people by gender, we found that more than 3/4 of people who committed suicide are male, and only less than 1/4 are female. Moreover, from the second plot, male seem to be more sensitive to the real-world situation, the number of males suicided experienced a major increase in year 1995(probably the Asian Financial Crisis) and 2002 (probably the Terrorists Attack). However, the number of suicided females is relatively stable overtime. This may not fit the knowledge of commonsense, but can explained by science that more Estrogen leads to a more stable psychological condition.

5. Multivariable Ananlysis

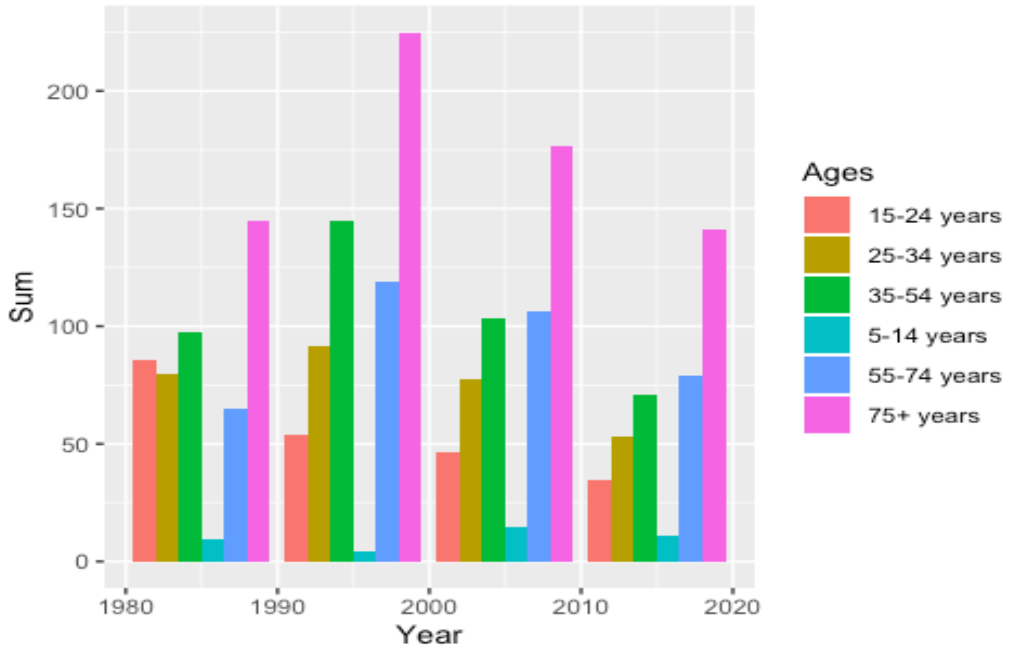
```
data1985 = data[which(data$year == 1985), ]
data1995 = data[which(data$year == 1995), ]
data2005 = data[which(data$year == 2005), ]
```

```
data2015 = data[which(data$year == 2015), ]
datasub = rbind(data1985, data1995, data2005, data2015)
p1 = ggplot(datasub, aes(x = datasub$year, y = datasub$rate,
fill=datasub$generation)) + geom_bar(position = "dodge", stat =
"identity")
p1 + xlab("Year") + ylab("Sum") + labs(fill = "Generation")
```



When we plotted the sum of suicide rate by generation for the selected year (1985, 1995, 2005, 2015), we found that the G.I. Generation (born between 1901 and 1927) is the generation with the highest suicide rate in both 1985 and 1990, after 1995, the suicide rate of the G.I. Generation became zero because the age of people of the generation, and the Silent Generation became the one with the highest suicide rate. What's more, we can see that the suicide rate of the Generation X keeps stable through 1985 to 2015 with a slightly decrease, which may concluded as the Generation X has the most stable mind among all the generation that listed above.

```
p2 = ggplot(datasub, aes(x = datasub$year, y = datasub$rate, fill = dat
asub$age)) + geom_bar(position = "dodge", stat = "identity")
p2 + xlab("Year") + ylab("Sum") + labs(fill = "Ages")
```

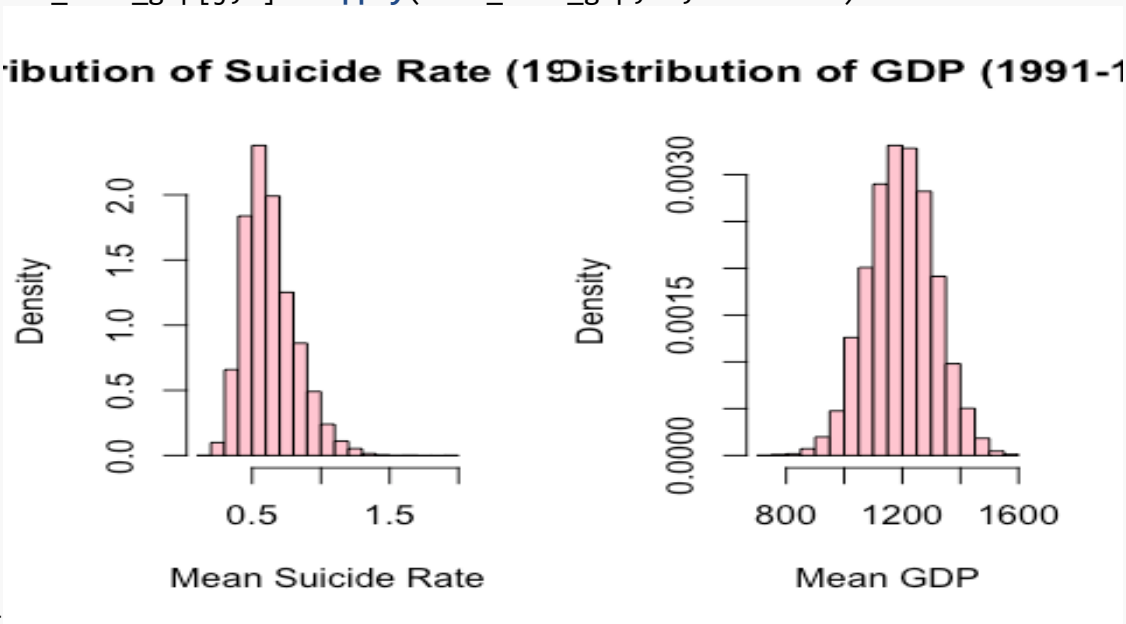


The plot above shows the how suicide rate varies from time to time among different groups of ages. From the plot we can see that people aged 75 and above are always the group with the highest suicide rate, which is so different from what I expected from the very beginning. People aged 35-54 are also with a relatively high suicide rate among the remaining groups, however, age 55-74 shows a trend that its suicide rate gradually became more than the mid-aged group. From the observation above, we can know that elderlies face a more severe problem than we thought it might be, it would be better if we look for some approaches to decrease the suicide rate of elderly people.

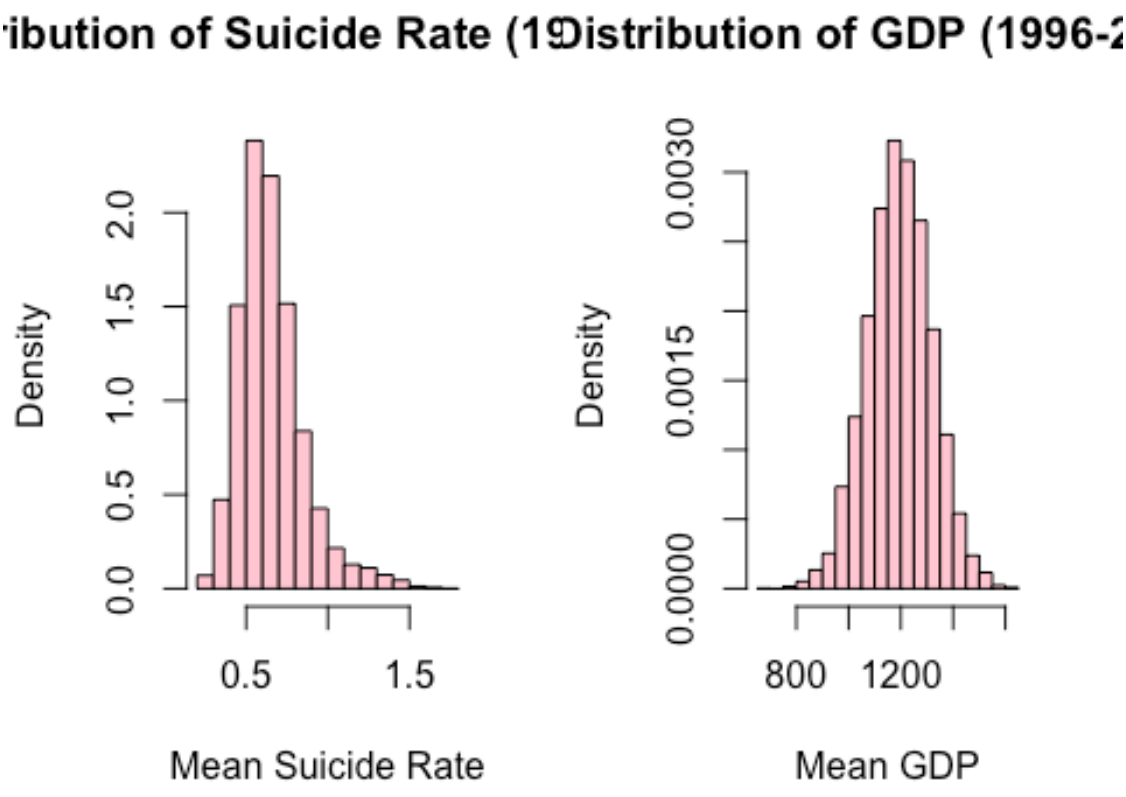
6. Central Limit Theorem

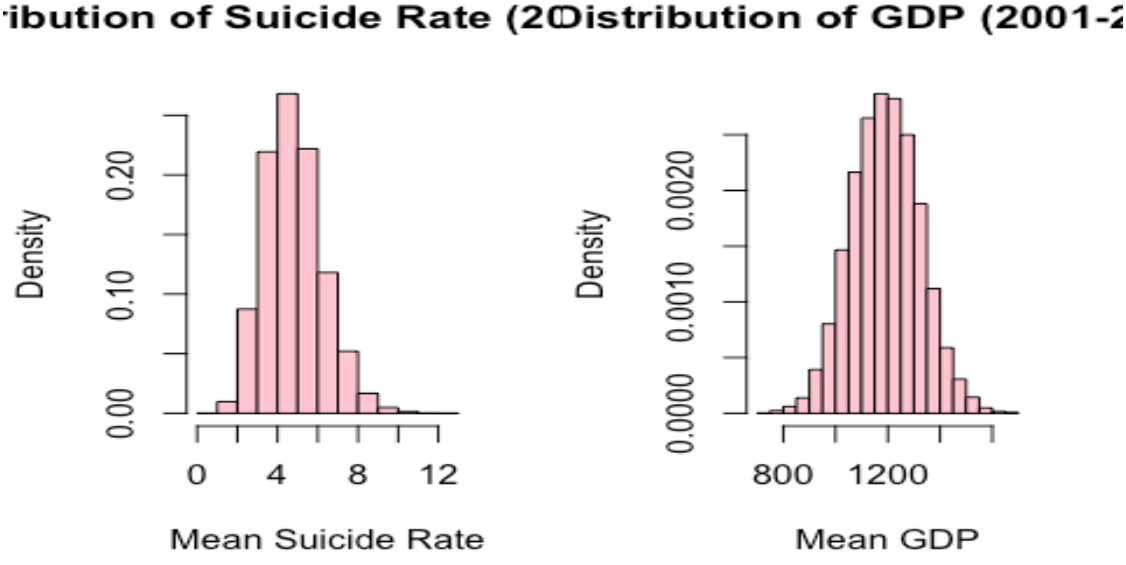
```
# Central Limit Theorem
samples = 1e4
sample.size = 30
data_Mil_1 = subset(data, (generation == "Millenials")&(year < 1996))
data_Mil_2 = subset(data, (generation == "Millenials")&(year >= 1996)&
(year < 2001))
data_Mil_3 = subset(data, (generation == "Millenials")&(year >= 2001)&
(year < 2006))
data_Mil_4 = subset(data, (generation == "Millenials")&(year >= 2006)&
(year < 2011))
data_Mil_5 = subset(data, (generation == "Millenials")&(year >= 2011))
data_Mil = list(data_Mil_1, data_Mil_2, data_Mil_3, data_Mil_4, data_Mi
```

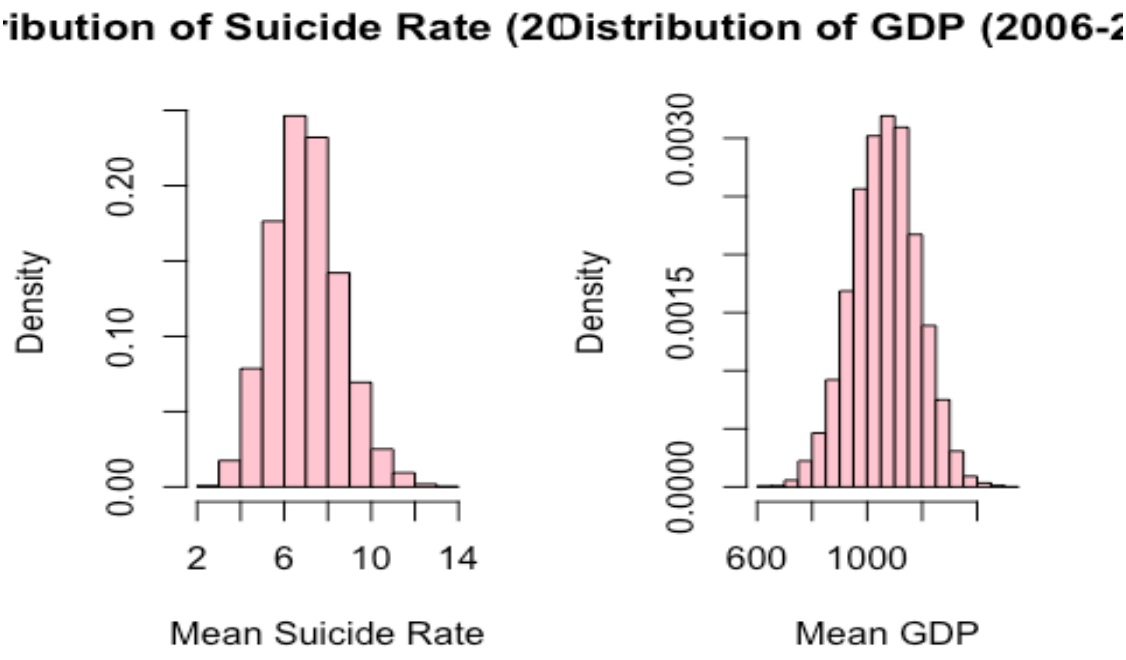
```
l_5)

year_interval = c("(1991-1995)", "(1996-2000)", "(2001-2005)", "(2006-2
010)","(2011-2015)")
average_rate_gdp = matrix(nrow = 5, ncol = 2)
sd_rate_gdp = matrix(nrow = 5, ncol = 2)
for (j in 1:5){
  rate_gdp = cbind(data_Mil[[j]]$rate, data_Mil[[j]]$gdp_for_year....)
  N = dim(rate_gdp)[1]
  mean_rate_gdp = matrix(nrow = samples, ncol = 2)
  for (i in 1:samples){
    ord = sample(1:N, size = sample.size, replace = FALSE)
    mean_rate_gdp[i,] = apply(rate_gdp[ord,], 2, FUN = mean)
  }
  par(mfrow = c(1,2))
  hist(mean_rate_gdp[,1], col = hcl(0), xlab = "Mean Suicide Rate", pro
b = TRUE
       , main = paste("Distribution of Suicide Rate", year_interval
[j]))
  hist(mean_rate_gdp[,2], col = hcl(0), xlab = "Mean GDP", prob = TRUE,
 main =
         paste("Distribution of GDP", year_interval[j]))
  average_rate_gdp[j, ] = apply(mean_rate_gdp, 2, FUN = mean)
  sd_rate_gdp[j, ] = apply(mean_rate_gdp, 2, FUN = sd)
```

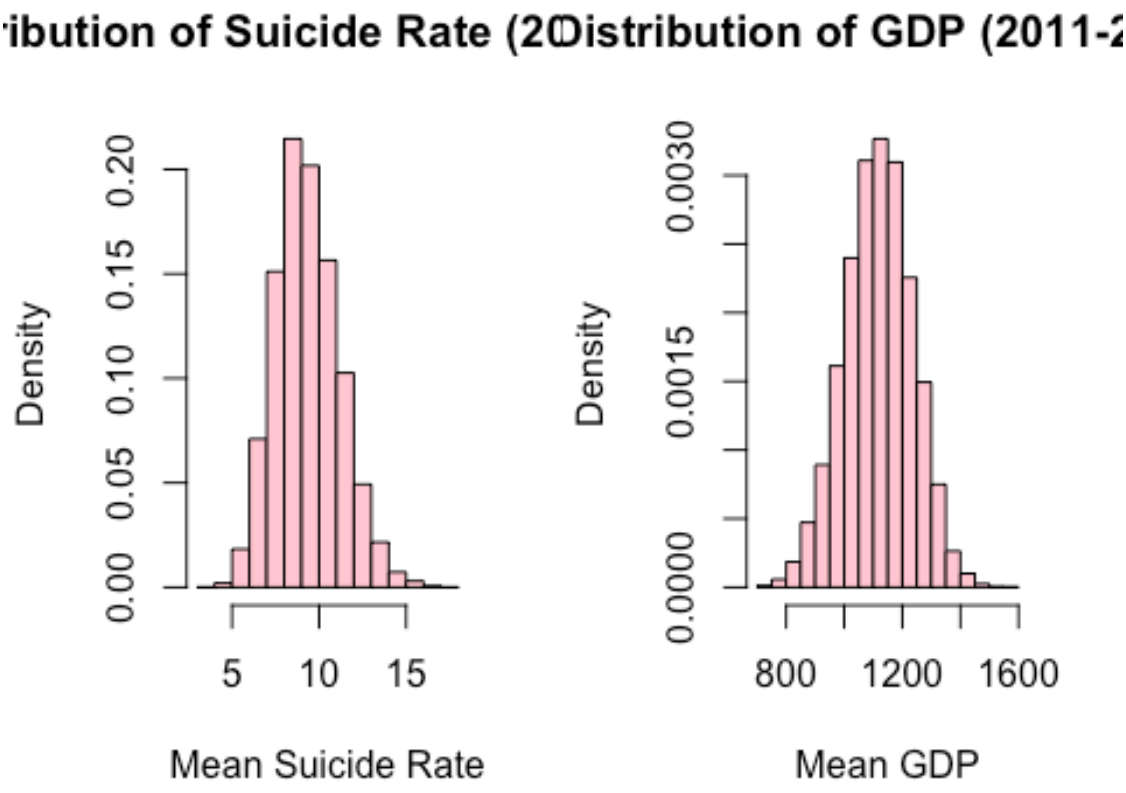



```
}
```


ibution of Suicide Rate (19
Distribution of GDP (1996-2
Density
0.0
0.5
1.0
1.5
2.0
0.5
1.5
Mean Suicide Rate
Density
0.0000
0.0015
0.0030
800
1200
Mean GDP


ibution of Suicide Rate (20
Distribution of GDP (2001-2
Density
0.00
0.10
0.20
0
4
8
12
Mean Suicide Rate
Density
0.0000
0.0010
0.0020
800
1200
Mean GDP


ibution of Suicide Rate (20
Distribution of GDP (2006-2
Density
0.00
0.10
0.20
2
6
10
14
Mean Suicide Rate
Density
0.0000
0.0015
0.0030
600
1000
Mean GDP


ibution of Suicide Rate (20
Distribution of GDP (2011-2
Density
0.00
0.05
0.10
0.15
0.20
5
10
15
Mean Suicide Rate
Density
0.0000
0.0015
0.0030
800
1200
1600
Mean GDP

```
par(mfrow = c(1,1))
row.names(average_rate_gdp) = year_interval
colnames(average_rate_gdp) = c("Mean Suicide Rate", "Mean GDP for
Year")
row.names(sd_rate_gdp) = year_interval
colnames(sd_rate_gdp) = c("SD of Suicide Rate", "SD of GDP for Year")
```

I filtered out the Millennials from the whole dataset, did the random sampling on the Millennials by a sample size of 30 and repeated for 10,000 times.

From the plot above, we can see that the distribution of the suicide rate changed from a right skewed distribution to a more normal distributed one, meanwhile, the distribution of GDP stayed stable overtime by a non-skewed distribution.

7. Sampling Methods

```
# Random Sampling
summary(data$rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.91    5.98   12.81   16.60  224.97

set.seed(233)
s1 = srswor(200, nrow(data))
s1[s1 != 0]

##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [71] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [106] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [141] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [176] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

rows = (1:nrow(data))[s1!=0]
rows = rep(rows, s1[s1 != 0])
sample1 = data[rows, ]
summary(sample1$rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.01    5.75   14.11   16.50  121.87

# Systematic Sampling
N = nrow(data)
n = 200
k = floor(N / n)
r = sample(k, 1)
```

```r
s2 = seq(r, by = k, length = n)
sample2 <- data[s2, ]
summary(sample2$rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.663   5.015  10.442  15.965  81.430
```

```r
# Stratified Sampling
freq = table(data$generation)
freq
```

```
##
##         Boomers G.I. Generation    Generation X    Generation Z
##            4958            2744            6376            1470
##      Millenials          Silent
##            5780            6332
```

```r
stsizes = 200 * freq / sum(freq)
st2 = strata(data, stratanames = c("generation"), size = stsizes, metho
d = "srswor", description = TRUE)
```

```
## Stratum 1
##
## Population total and number of selected units: 6376 35.8496
## Stratum 2
##
## Population total and number of selected units: 6332 19.84093
## Stratum 3
##
## Population total and number of selected units: 2744 46.10268
## Stratum 4
##
## Population total and number of selected units: 4958 10.62907
## Stratum 5
##
## Population total and number of selected units: 5780 41.7932
## Stratum 6
##
## Population total and number of selected units: 1470 45.78453
## Number of strata  6
## Total number of selected units 200
```

```r
sample3 = getdata(data, st2)
summary(sample3$rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.11    2.45   11.32   14.55  114.94
```

After applying three different methods of sampling (Random Sampling, Systematic Sampling and Stratified Sampling grouped by generation), I finally had a set of conclusion like this:

When applying Random Sampling

The mean of the sample is a bit larger than the mean of the whole dataset, and the median of the sample is very close to the median of the whole dataset

When applying Systematic Sampling

The mean of the sample is larger than the mean of the whole dataset, and the median of the sample is much larger to the median of the whole dataset

When applying Stratified Sampling

The mean of the sample is smaller than the mean of the whole dataset, and the median of the sample is smaller than the whole dataset.