EDA IEOR 142, Final Project

3033342158

December 16, 2019

```
#install.packages("Rcpp")
#install.packages("purrr")
#install.packages("dplyr")

library(softImpute)
```

```
## Warning: package 'softImpute' was built under R version 3.5.3
```

```
## Loading required package: Matrix
```

```
## Loaded softImpute 1.4
```

```
library(gridExtra, verbose=FALSE, warn.conflicts=FALSE, quietly=TRUE)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.5.3
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
library(ranger)
```

```
## Warning: package 'ranger' was built under R version 3.5.3
```

```
##
## Attaching package: 'ranger'
```

```
## The following object is masked from 'package:randomForest':
##
##     importance
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

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

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

```
library(reshape2)
library("caTools")
```

```
## Warning: package 'caTools' was built under R version 3.5.3
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.5.3
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.5.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```r
restore = list(repr.plot.width=8, repr.plot.height=3)
PALETTE = c("#00A9FF", "#F8766D", "#7CAE00", "#C77CFF", "#CD9600", "#00BE67", "#FF61CC", "#00BFC4")
theme.x_axis_only = theme(axis.title.y=element_blank(), axis.text.y=element_blank(), axis.ticks.y=element_blank(), panel.grid.major.y=element_blank(), panel.grid.minor.y=element_blank())
theme.no_legend = theme(legend.position="none")
theme.legend_title = theme(legend.title=element_text(size=7))
data <- read.csv("us_suicides_merged_no_na.csv")
```

EDA

Our cleaned and merged data consists of 372 observations and 14 variables.

```r
nrow(data)
```

```
## [1] 372
```

```
ncol(data)
```

```
## [1] 14
```

```
#First look at the first 6 and last 6 observations of our data
head(data)
```

```
##         country year    sex         age suicides_no population
## 1 United States 1985 female 15-24 years         854   19589000
## 2 United States 1985   male 15-24 years        4267   19962000
## 3 United States 1985 female 25-34 years        1242   21041000
## 4 United States 1985   male 25-34 years        5134   20986000
## 5 United States 1985 female 35-54 years        2105   27763000
## 6 United States 1985   male 35-54 years        6053   26589000
##   suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 1              4.36 United States1985        0.841     4.346734e+12
## 2             21.38 United States1985        0.841     4.346734e+12
## 3              5.90 United States1985        0.841     4.346734e+12
## 4             24.46 United States1985        0.841     4.346734e+12
## 5              7.58 United States1985        0.841     4.346734e+12
## 6             22.77 United States1985        0.841     4.346734e+12
##   gdp_per_capita....   generation depression_percentage drug_death_rate
## 1              19693 Generation X              6.519361         0.00000
## 2              19693 Generation X              3.520442         0.00000
## 3              19693      Boomers              6.519361         0.00000
## 4              19693      Boomers              3.520442         0.00000
## 5              19693       Silent              6.519361         0.00000
## 6              19693       Silent              3.520442        10.69853
```

```
tail(data)
```

```
##           country year    sex         age suicides_no population
## 367 United States 2015 female  5-14 years         158   20342901
## 368 United States 2015   male  5-14 years         255   21273987
## 369 United States 2015 female 55-74 years        2872   35115610
## 370 United States 2015   male 55-74 years        9068   32264697
## 371 United States 2015 female   75+ years         540   11778666
## 372 United States 2015   male   75+ years        3171    8171136
##     suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 367              0.78 United States2015         0.92     1.812071e+13
## 368              1.20 United States2015         0.92     1.812071e+13
## 369              8.18 United States2015         0.92     1.812071e+13
## 370             28.11 United States2015         0.92     1.812071e+13
## 371              4.58 United States2015         0.92     1.812071e+13
## 372             38.81 United States2015         0.92     1.812071e+13
##     gdp_per_capita....   generation depression_percentage drug_death_rate
## 367              60387 Generation Z                  6.03             0.3
## 368              60387 Generation Z                  3.51             0.2
## 369              60387      Boomers                  6.03            23.7
## 370              60387      Boomers                  3.51            34.7
## 371              60387       Silent                  6.03             7.4
## 372              60387       Silent                  3.51             8.9
```

The dataset contains 31 unique years from 1985 to 2015, the suicide rate per 100k has a variance of 175.0296.

```
unique(sort(data$year))
```

```
##  [1] 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
## [15] 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
## [29] 2013 2014 2015
```

```
length(unique(sort(data$year)))
```

```
## [1] 31
```

```
var(data$suicides.100k.pop)
```

```
## [1] 175.0296
```

```
#table(data$suicides.100k.pop) / length(data$suicides.100k.pop)) # relative frequencies
as.numeric(names(table(data$suicides.100k.pop))[table(data$suicides.100k.pop) == max(table(data
$suicides.100k.pop))]) # mode for suicide rate
```

```
## [1] 0.34
```

```
as.numeric(names(table(data$gdp_per_capita....))[table(data$gdp_per_capita....) == max(table(dat
a$gdp_per_capita....))]) # mode for gdp per capita
```

```
##  [1] 19693 20588 21631 23103 24654 26004 26503 27760 28891 30375 31518
## [12] 32928 34644 36164 38072 39218 40018 40845 42468 44867 47423 49666
## [23] 50563 51585 51989 52128 53452 55170 56520 58531 60387
```

```
range(data$suicides.100k.pop)
```

```
## [1]  0.26 58.95
```

```
data[data$suicides.100k.pop == min(data$suicides.100k.pop), ]
```

```
##           country year    sex        age suicides_no population
## 175 United States 1999 female 5-14 years          50   19275566
##     suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 175              0.26 United States1999        0.885     9.660624e+12
##     gdp_per_capita.... generation depression_percentage drug_death_rate
## 175              38072 Millenials                  5.92             0.1
```

```
data[data$suicides.100k.pop == max(data$suicides.100k.pop), ]
```

```
##          country year  sex       age suicides_no population
## 36 United States 1987 male 75+ years        2532    4295000
##    suicides.100k.pop      country.year HDI.for.year gdp_for_year....
## 36             58.95 United States1987         0.85     4.870217e+12
##    gdp_per_capita....      generation depression_percentage
## 36              21631 G.I. Generation               3.51864
##    drug_death_rate
## 36        7.466624
```

Investigating Suicide rate and Sex

There are 186 males and 186 females. There is also 62 records for every age range provided in the data. The data seems to be split evenly thus far except for the generation variable.Generation X has the highest amount of records and Generation Z has the least. The Suicide rate had a decline from about the late 1990’s to the mid 2000’s but has been steadily increasing since around the year 2008.

```
table(data$sex)
```

```
## 
## female   male 
##    186    186
```

```
table(data$age)
```

```
## 
## 15-24 years 25-34 years 35-54 years  5-14 years 55-74 years   75+ years 
##          62          62          62          62          62          62
```

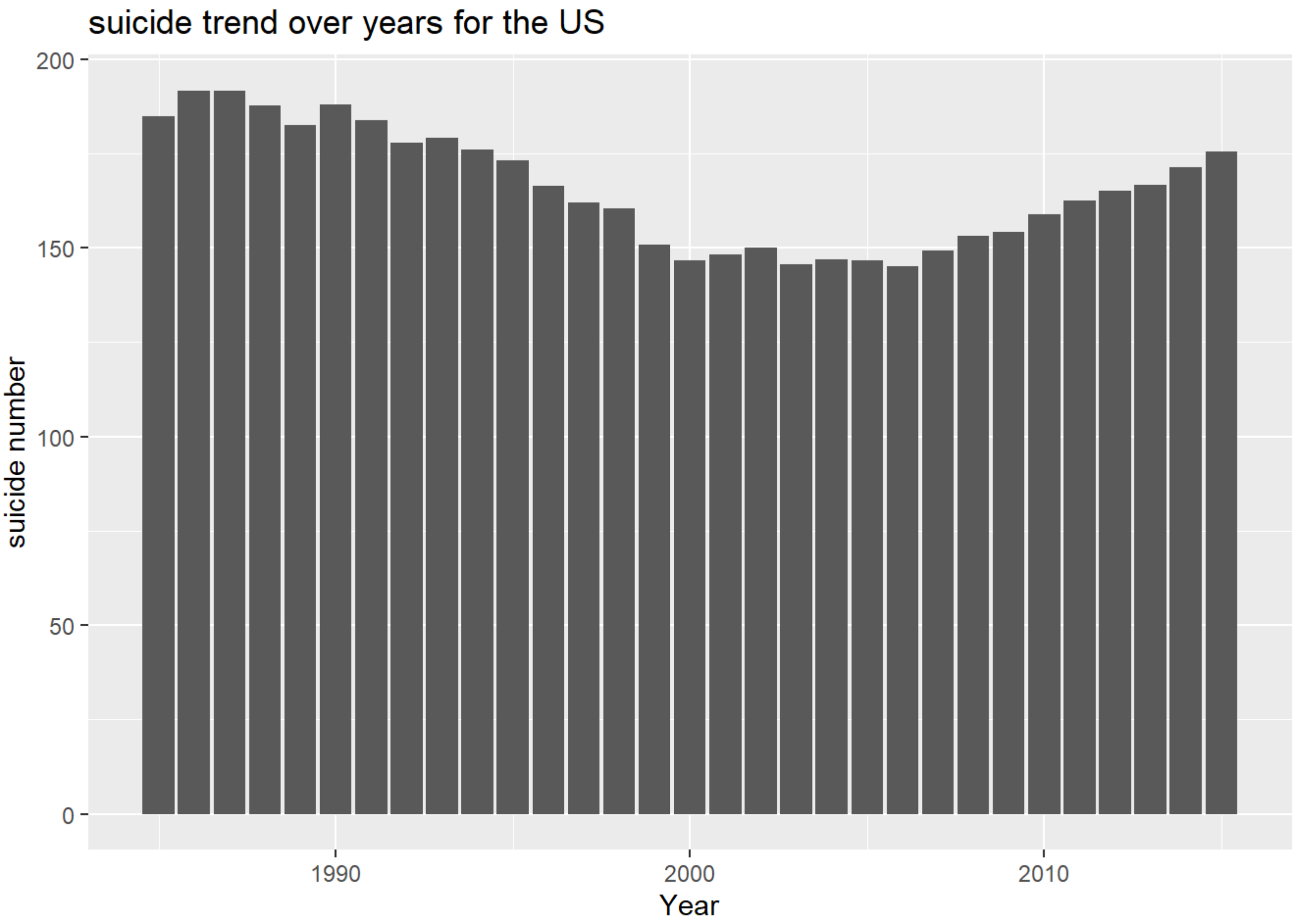
```
table(data$generation)
```

```
## 
##         Boomers G.I. Generation    Generation X    Generation Z 
##              68              44              88              18 
##      Millenials          Silent 
##              72              82
```

```
max(table(data$generation))
```

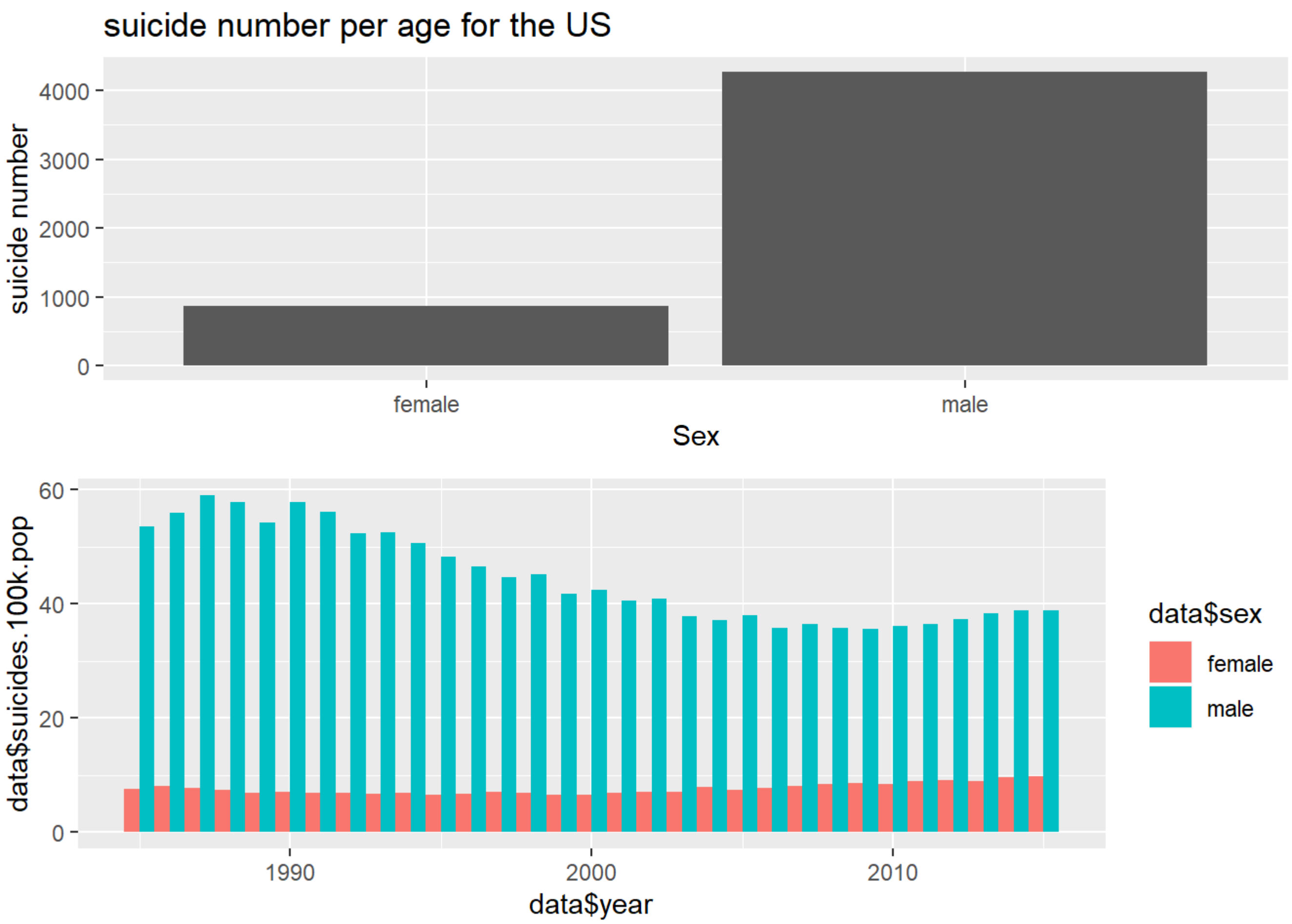
```
## [1] 88
```

```
ggplot(data)  + ggtitle("suicide trend over years for the US") +
geom_col(aes(x=data$year, y=data$suicides.100k.pop)) + xlab("Year") + ylab("suicide number")
```



```
p1 = ggplot(data)  + ggtitle("suicide number per age for the US") +
geom_col(aes(x=data$sex, y=data$suicides.100k.pop)) + xlab("Sex") + ylab("suicide number")


p2= ggplot(data, aes(x=data$year, y=data$suicides.100k.pop, fill=data$sex), xlab("Year"), ylab(
"Suicide Rate")) +
   geom_bar(stat="identity", width=1, position = "dodge")
grid.arrange(p1, p2, nrow=2, ncol = 1.2)
```

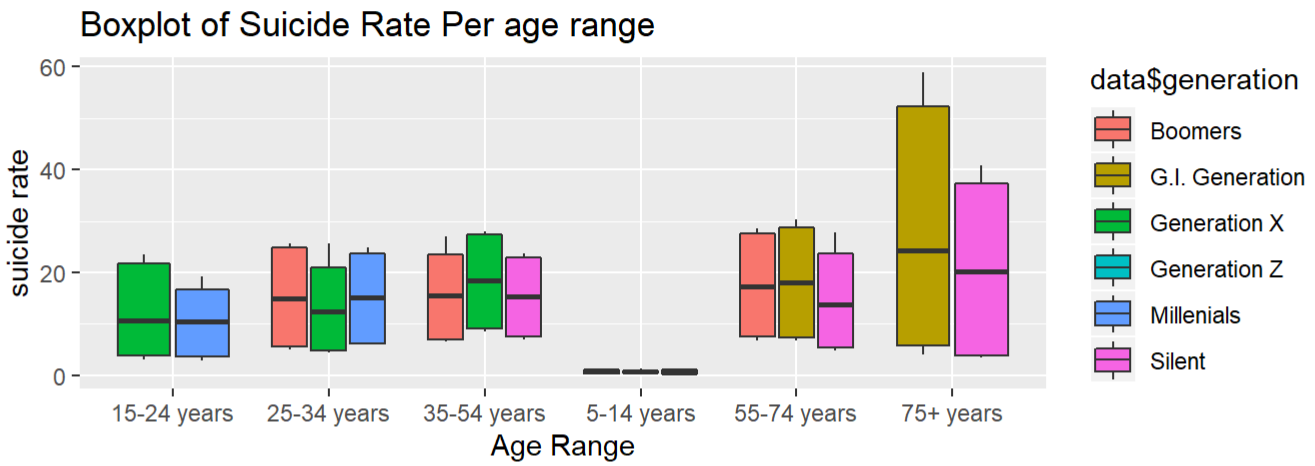


Investigating Suicide rate and Age

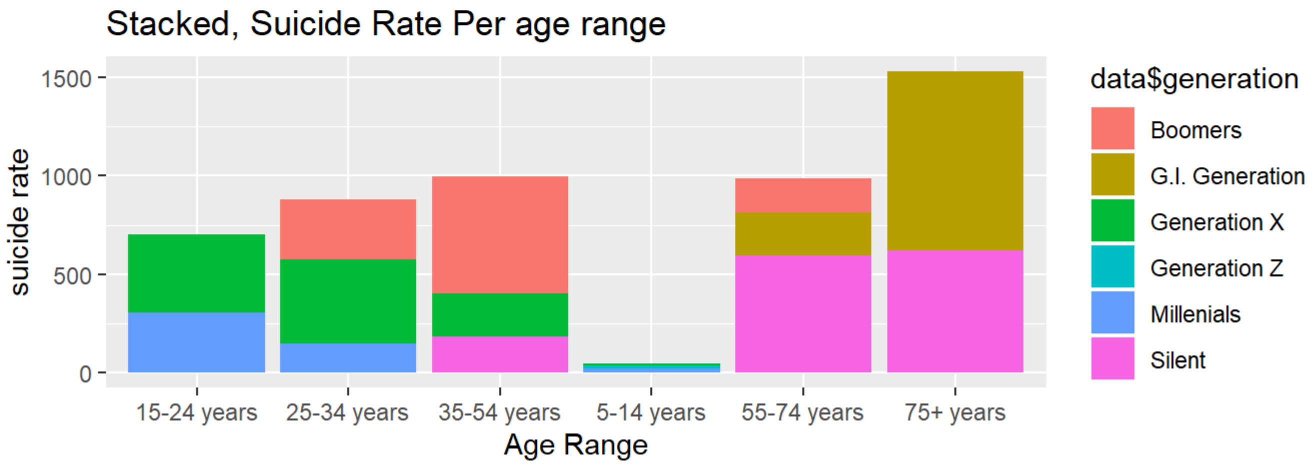
Suicide rates are highest among indivduals in the age group 75+(This is mostly people considered to be from the G.I. generation(1901-1924) and Silent generation(1925-1945)) and the lowest rates occur in the age group 5-14(generation X and generation Z).

```
p3 = ggplot(data) + ggtitle("Boxplot of Suicide Rate Per age range") + geom_boxplot(aes(x= data
$age, y=data$suicides.100k.pop, fill = data$generation)) +
xlab("Age Range") + ylab("suicide rate")

p4 = ggplot(data) + ggtitle("Stacked, Suicide Rate Per age range") + geom_col(aes(x= data$age, y
=data$suicides.100k.pop, fill = data$generation)) +
xlab("Age Range") + ylab("suicide rate")
grid.arrange(p3, p4, nrow=2, ncol = 1.1)
```

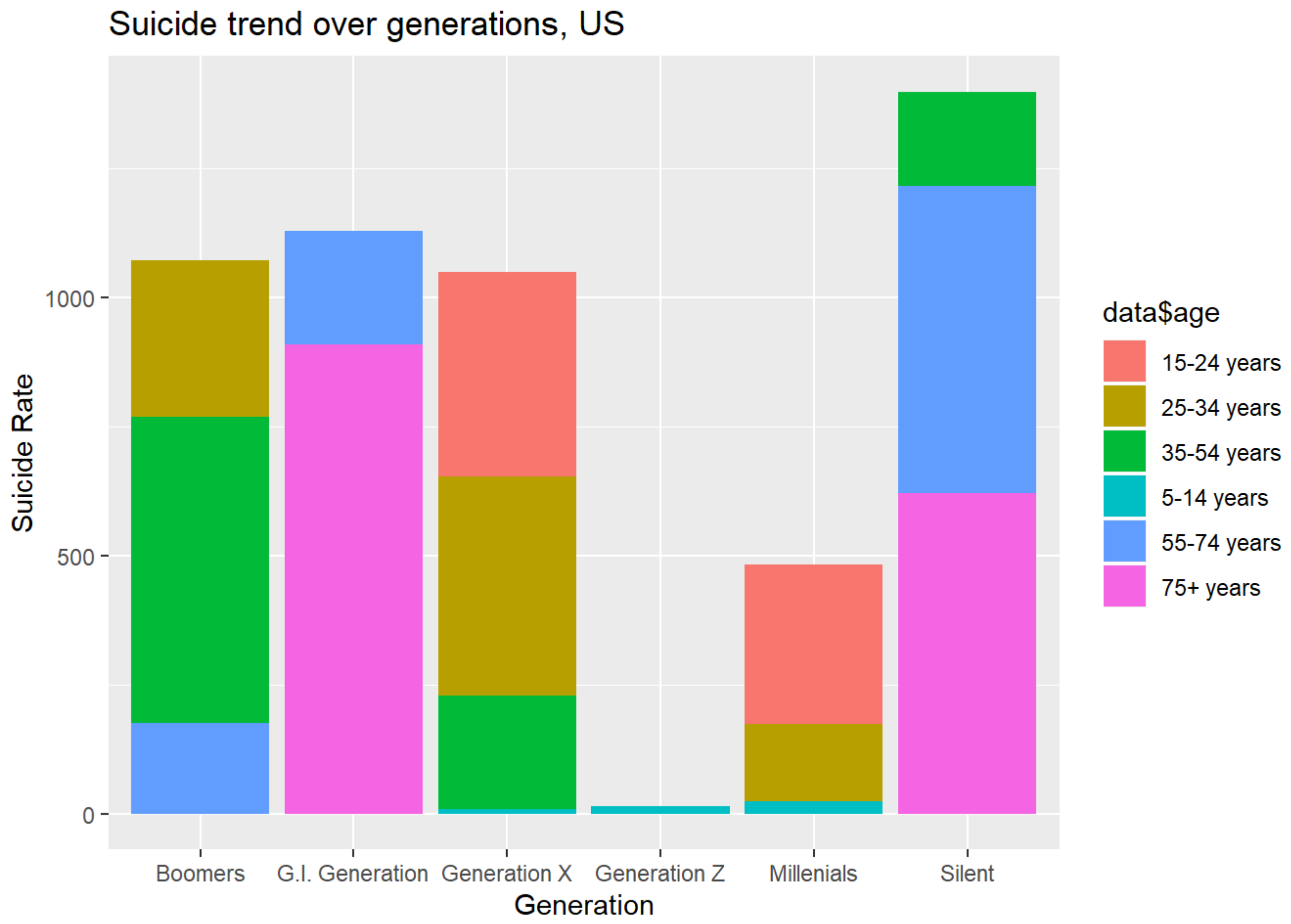


Boxplot of Suicide Rate Per age range



Stacked, Suicide Rate Per age range

```
ggplot(data)  + ggtitle("Suicide trend over generations, US") +
geom_col(aes(x= data$generation, y= data$suicides.100k.pop, fill=data$age), position="stack") +
 xlab("Generation") + ylab("Suicide Rate")
```

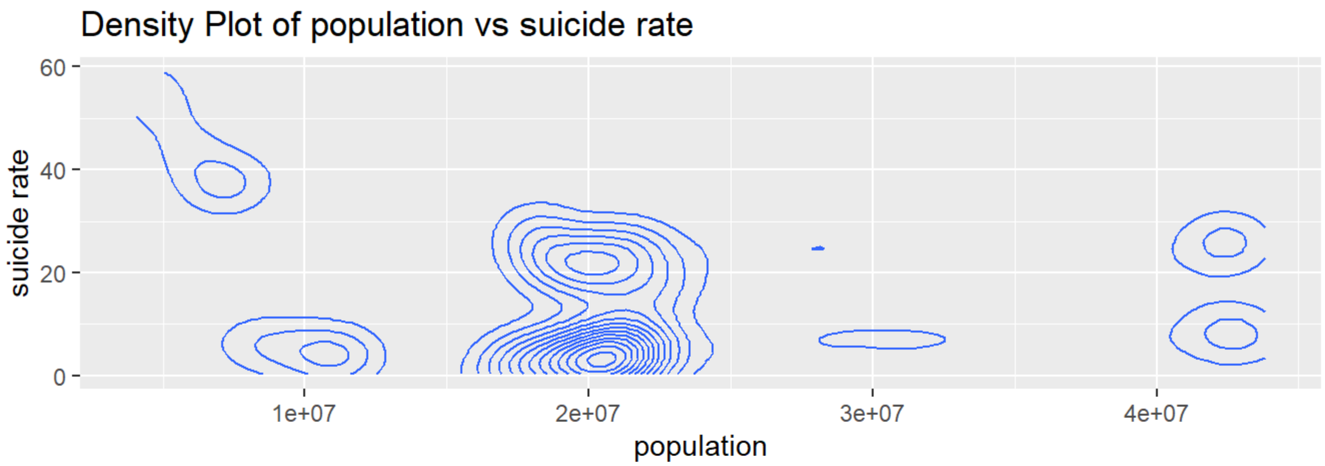


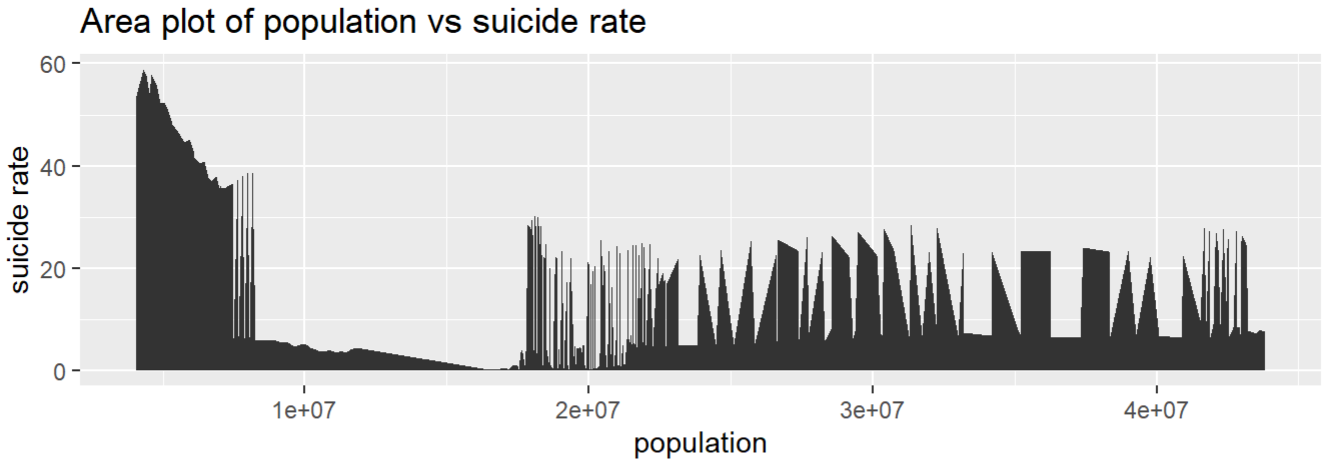
Investigating Suicide rate and Population

```
p5 = ggplot(data)  + ggtitle("Density Plot of population vs suicide rate") +
geom_density_2d(aes(x=data$population, y=data$suicides.100k.pop)) + xlab("population") + ylab("suicide rate")

p6 = ggplot(data)  + ggtitle("Area plot of population vs suicide rate") +
geom_area(aes(x=data$population, y=data$suicides.100k.pop)) + xlab("population") + ylab("suicide rate")

grid.arrange(p5, p6, nrow=2, ncol = 1.1)
```


Density Plot of population vs suicide rate


Area plot of population vs suicide rate

```
var(data$population)
```

```
## [1] 8.92766e+13
```

```
cor(data$suicides.100k.pop, data$population)
```

```
## [1] -0.1703968
```

```
summary(data$population)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##  4064000 18185450 20375469 21650611 22616944 43805214
```

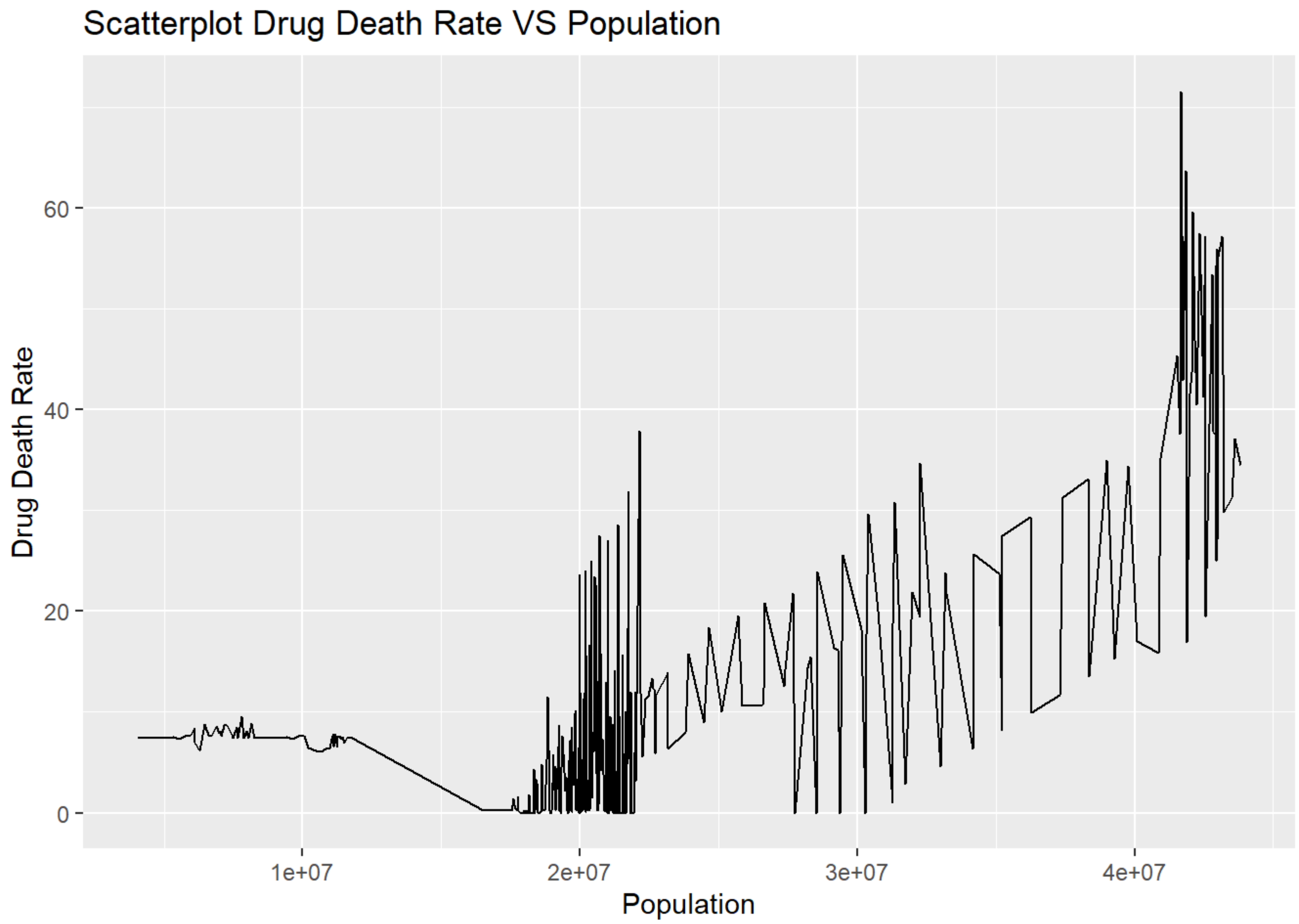
```
cor(data$population, data$depression_percentage)
```

```
## [1] 0.05065976
```

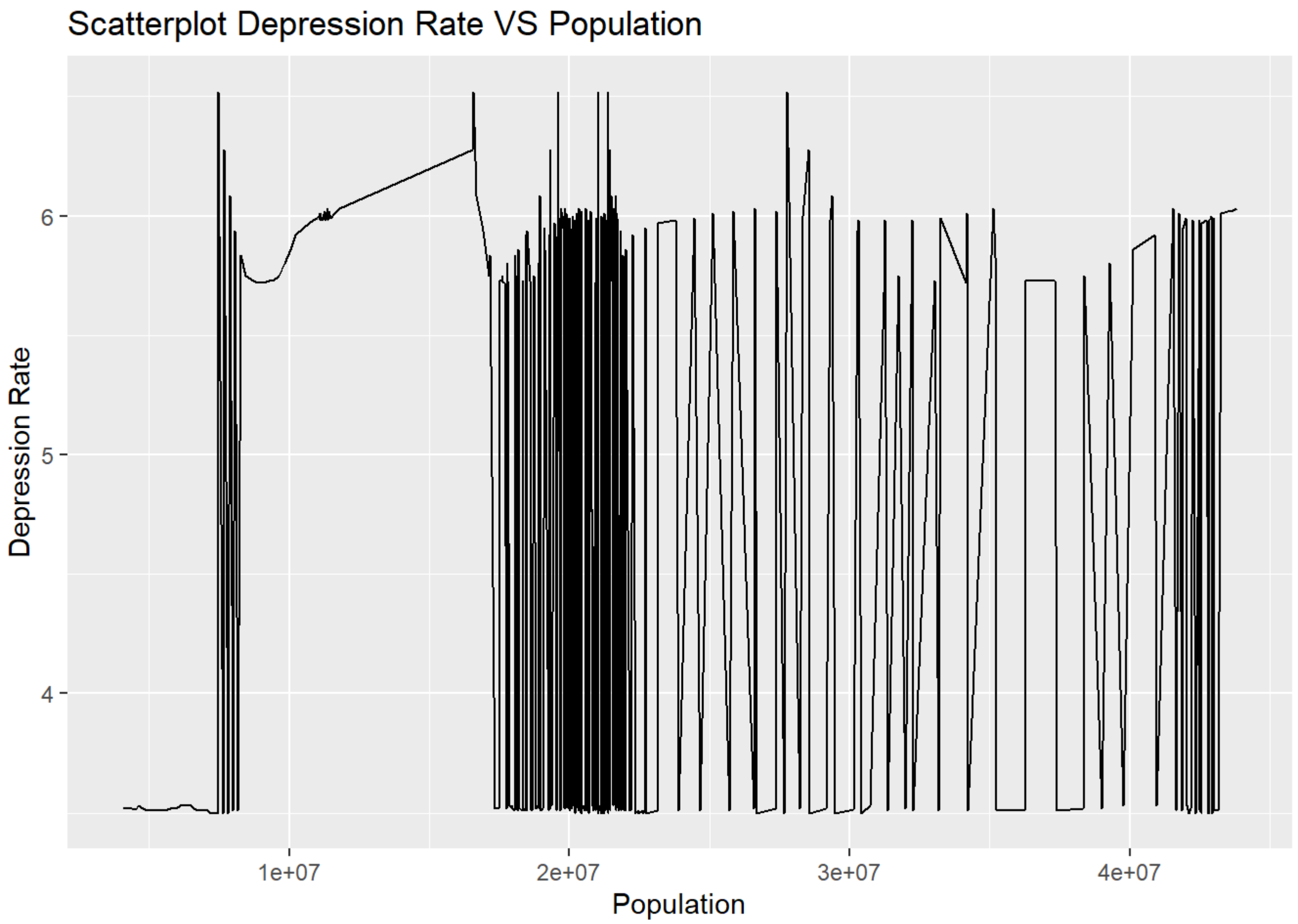
```
cor(data$population, data$drug_death_rate)
```

```
## [1] 0.6774055
```

```
ggplot(data) + geom_line(aes(x = data$population, y = data$drug_death_rate)) + xlab("Population"
) +ggtitle("Scatterplot Drug Death Rate VS Population") + ylab("Drug Death Rate")
```



```
ggplot(data) + geom_line(aes(x = data$population, y = data$depression_percentage)) + xlab("Popul
ation") +ggtitle("Scatterplot Depression Rate VS Population") + ylab("Depression Rate")
```



###Investigating Suicide rate and HDI for year

```
var(data$HDI.for.year) #Very low variance for HDI year to year
```

```
## [1] 0.0005165123
```

```
cor(data$suicides.100k.pop, data$HDI.for.year) #Barley negatively correlated
```

```
## [1] -0.06456609
```

```
cor(data$population, data$HDI.for.year)
```

```
## [1] 0.2177246
```

```
cor(data$gdp_per_capita...., data$HDI.for.year)#Sanity check: has a positive correlation
```

```
## [1] 0.9853092
```

```
cor(data$depression_percentage, data$HDI.for.year)
```

```
## [1] -0.0009472623
```

```
cor(data$drug_death_rate, data$HDI.for.year) # correlation: 0.4429688 somewhat positivley correl
ated
```

```
## [1] 0.4429688
```

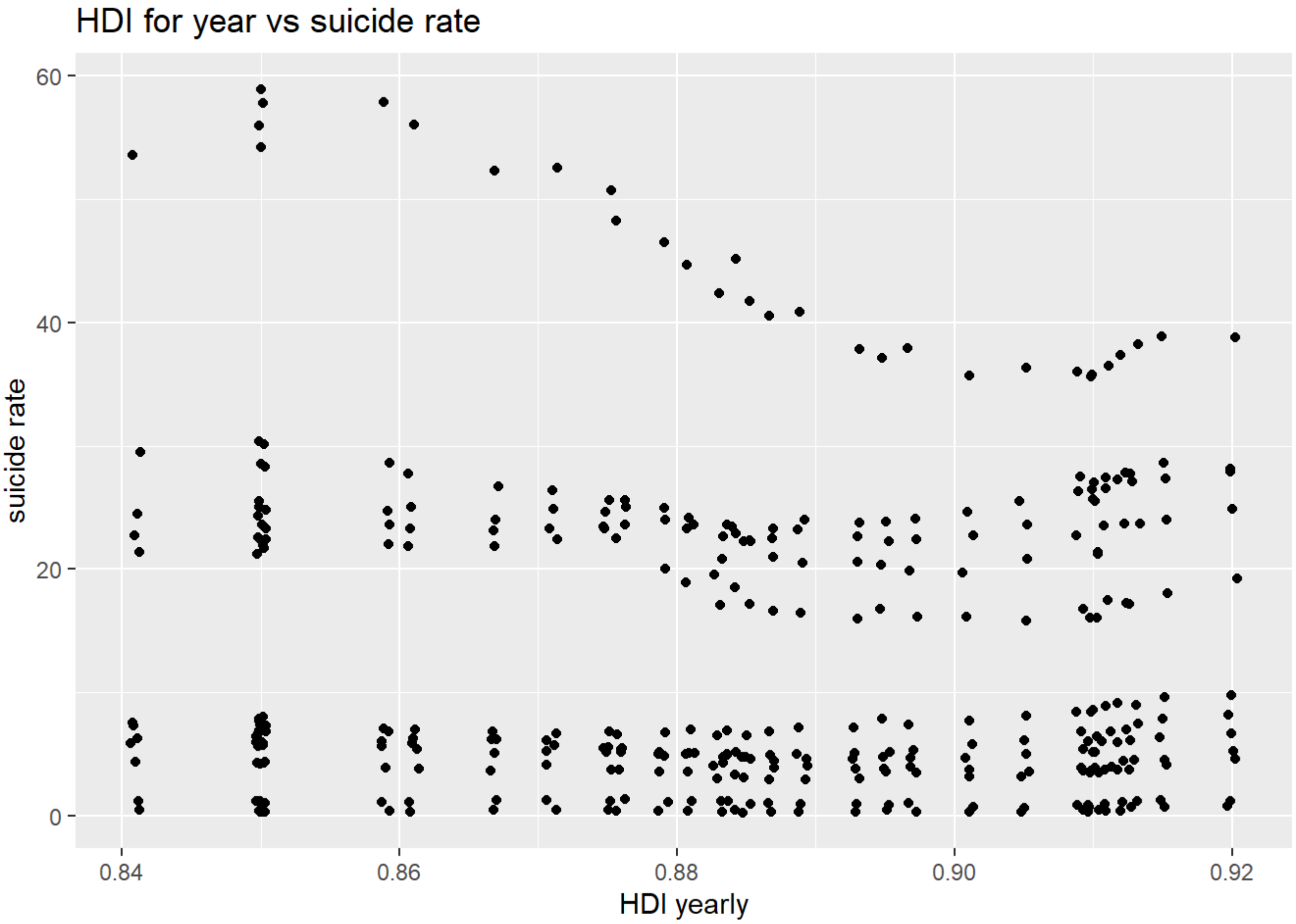
```
summary(data$HDI.for.year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.8410  0.8670  0.8850  0.8848  0.9090  0.9200
```

```
as.numeric(names(table(data$HDI.for.year))[table(data$HDI.for.year) == max(table(data$HDI.for.ye
ar))]) # mode for HDI
```

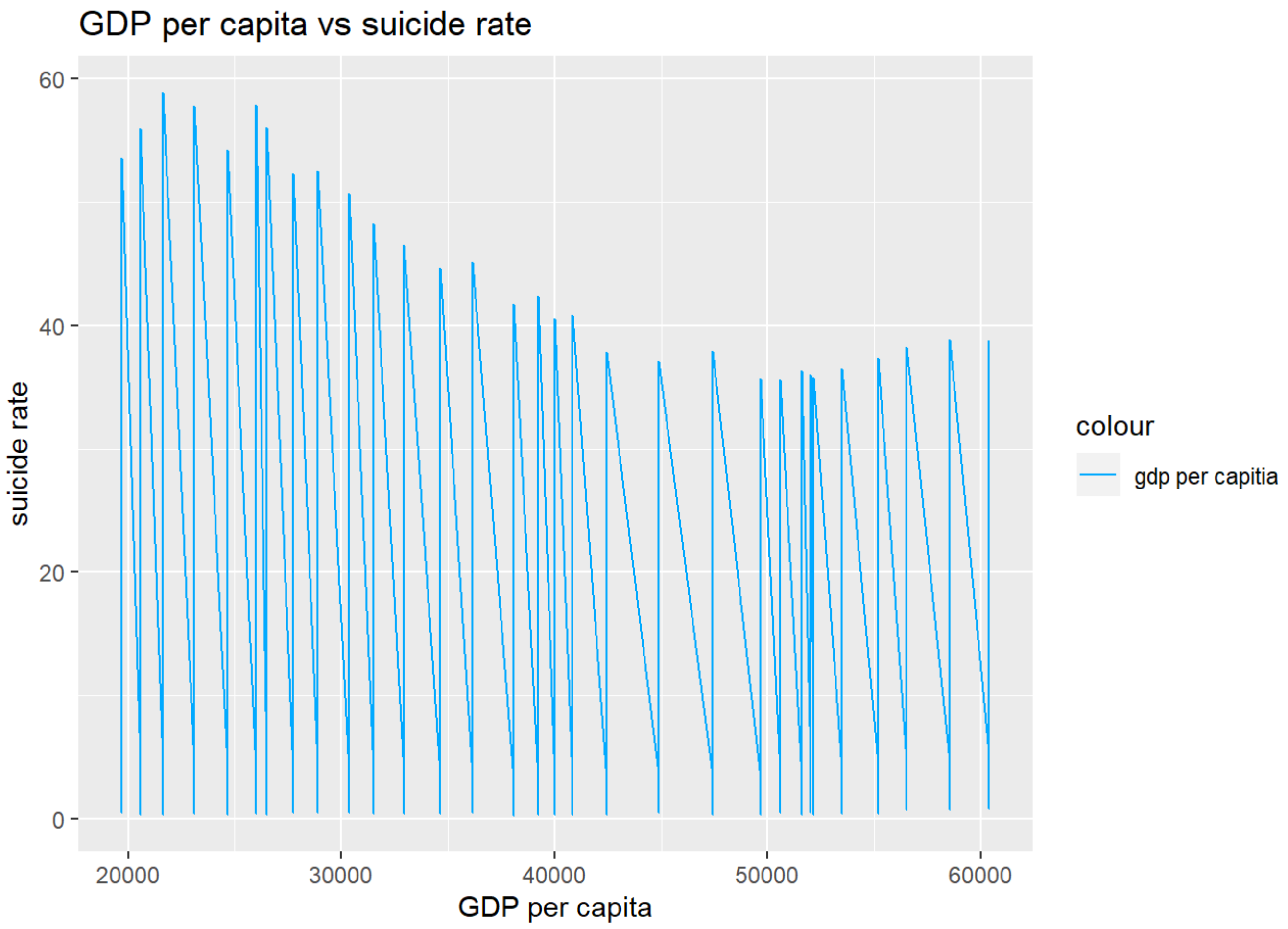
```
## [1] 0.85
```

```
ggplot(data)  + ggtitle("HDI for year vs suicide rate") +
geom_jitter(aes(x=data$HDI.for.year, y=data$suicides.100k.pop)) + xlab("HDI yearly") + ylab("sui
cide rate")
```



###Investigating Suicide rate and GDP per capita

```
ggplot(data)  + ggtitle("GDP per capita vs suicide rate") +
geom_line(aes(x=data$gdp_per_capita...., y=data$suicides.100k.pop, color = "gdp per capitia")) +
xlab("GDP per capita") + ylab("suicide rate") + scale_color_manual(values=PALETTE[1:3])
```



```
var(data$suicides.100k.pop, data$gdp_per_capita....)
```

```
## [1] -9979.495
```

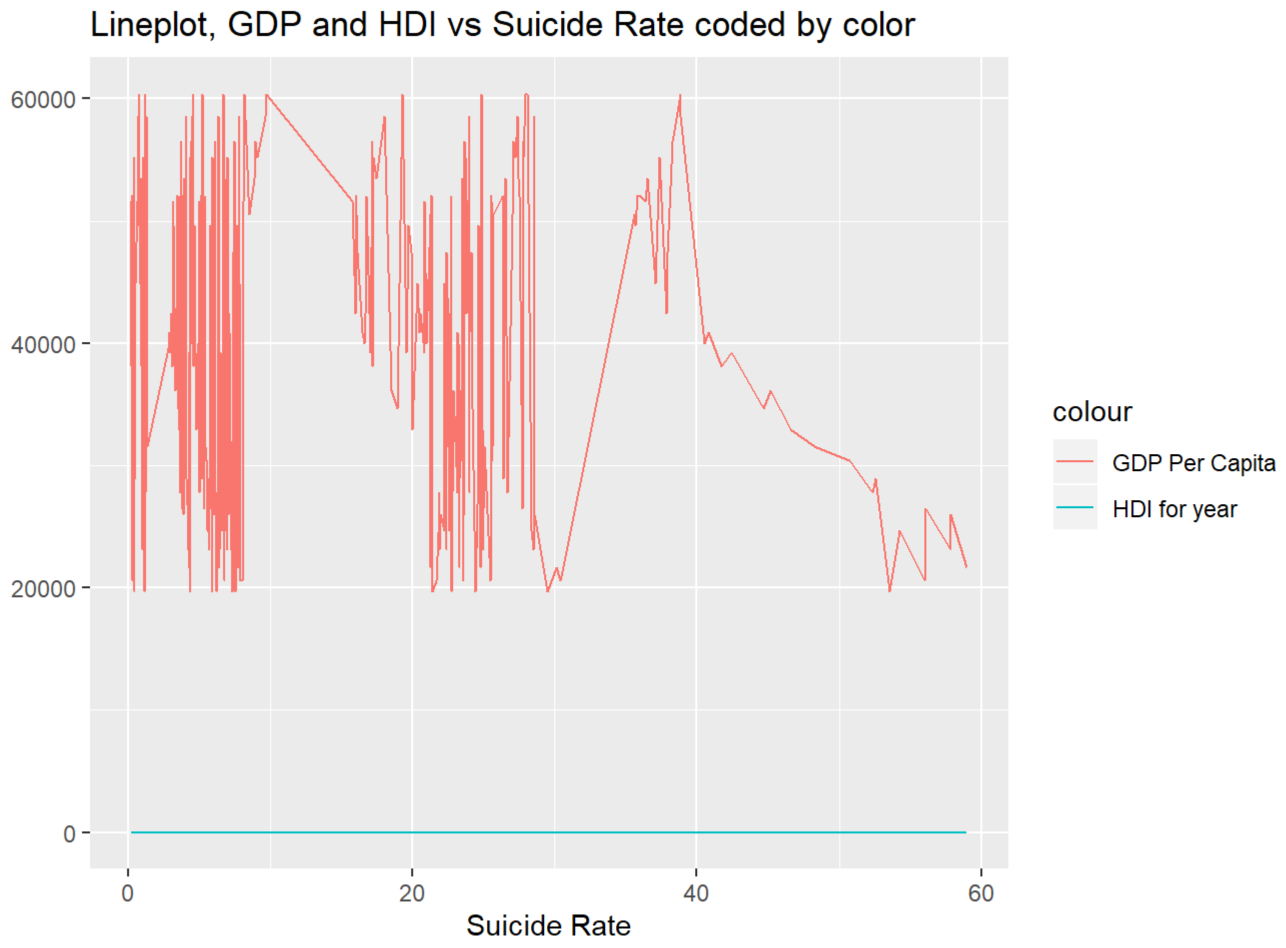
```
cor(data$suicides.100k.pop, data$gdp_per_capita....)
```

```
## [1] -0.0611568
```

```
summary(data$suicides.100k.pop, data$gdp_per_capita....)
```

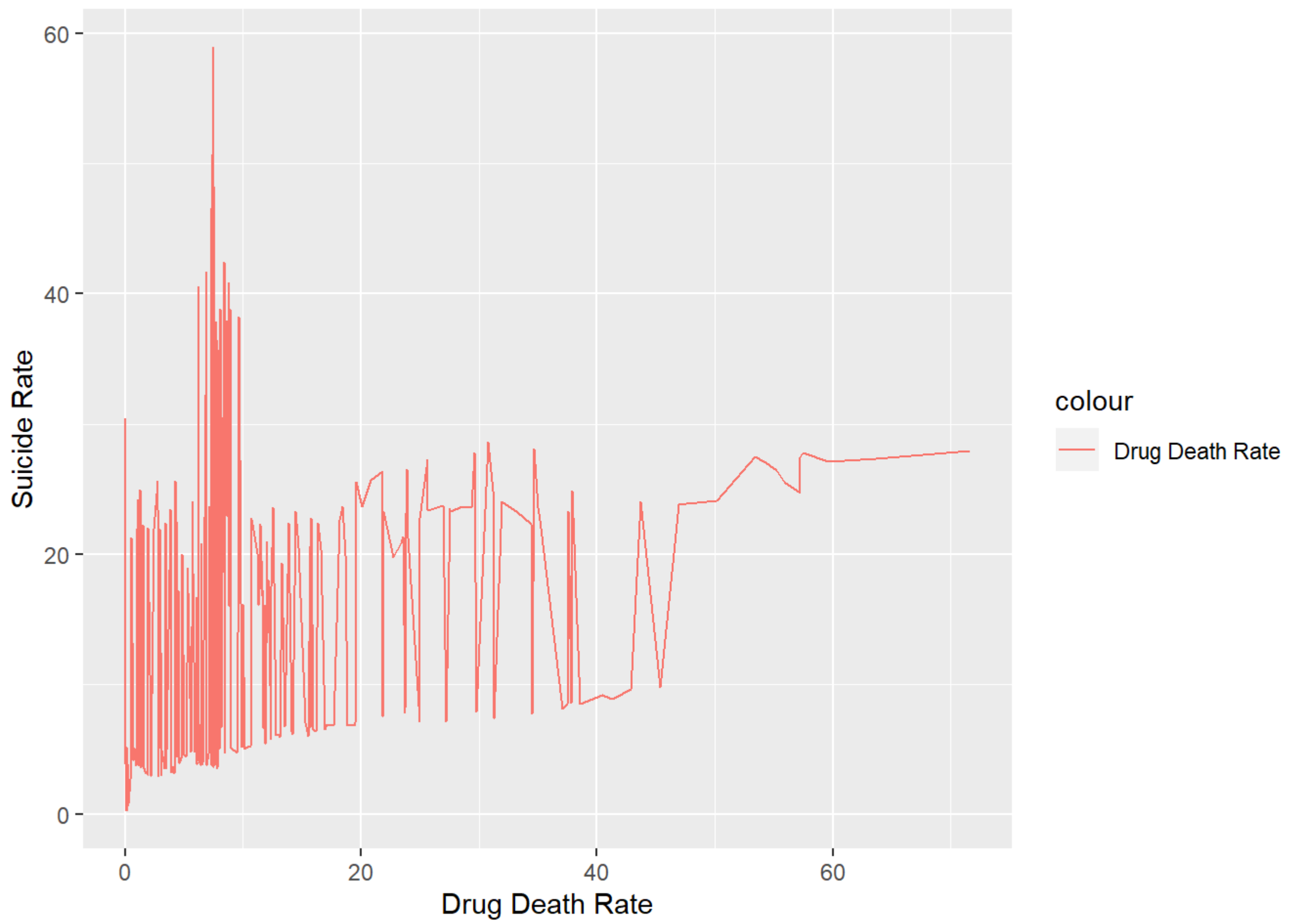
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.260   3.973   6.890  13.820  23.305  58.950
```

```
ggplot(data) + ylab("") + xlab("Suicide Rate")+ ggtitle("Lineplot, GDP and HDI vs Suicide Rate c
oded by color") +
      geom_line(aes(x=data$suicides.100k.pop, y= data$HDI.for.year, color = "HDI for year")) +
     geom_line(aes(x=data$suicides.100k.pop, y= data$gdp_per_capita...., color = "GDP Per Capit
a"))
```



Investigating Suicide rate and the drug death rate

```
ggplot(data)+ geom_line(aes(x= data$drug_death_rate, y= data$suicides.100k.pop, color = "Drug De
ath Rate")) +
ylab("Suicide Rate") + xlab("Drug Death Rate")
```



```
var(data$suicides.100k.pop, data$drug_death_rate)
```

```
## [1] 51.17861
```

```
cor(data$suicides.100k.pop, data$drug_death_rate)
```

```
## [1] 0.2891455
```

```
cor(data$population, data$drug_death_rate)
```

```
## [1] 0.6774055
```

```
cov(data$suicides.100k.pop, data$drug_death_rate)
```

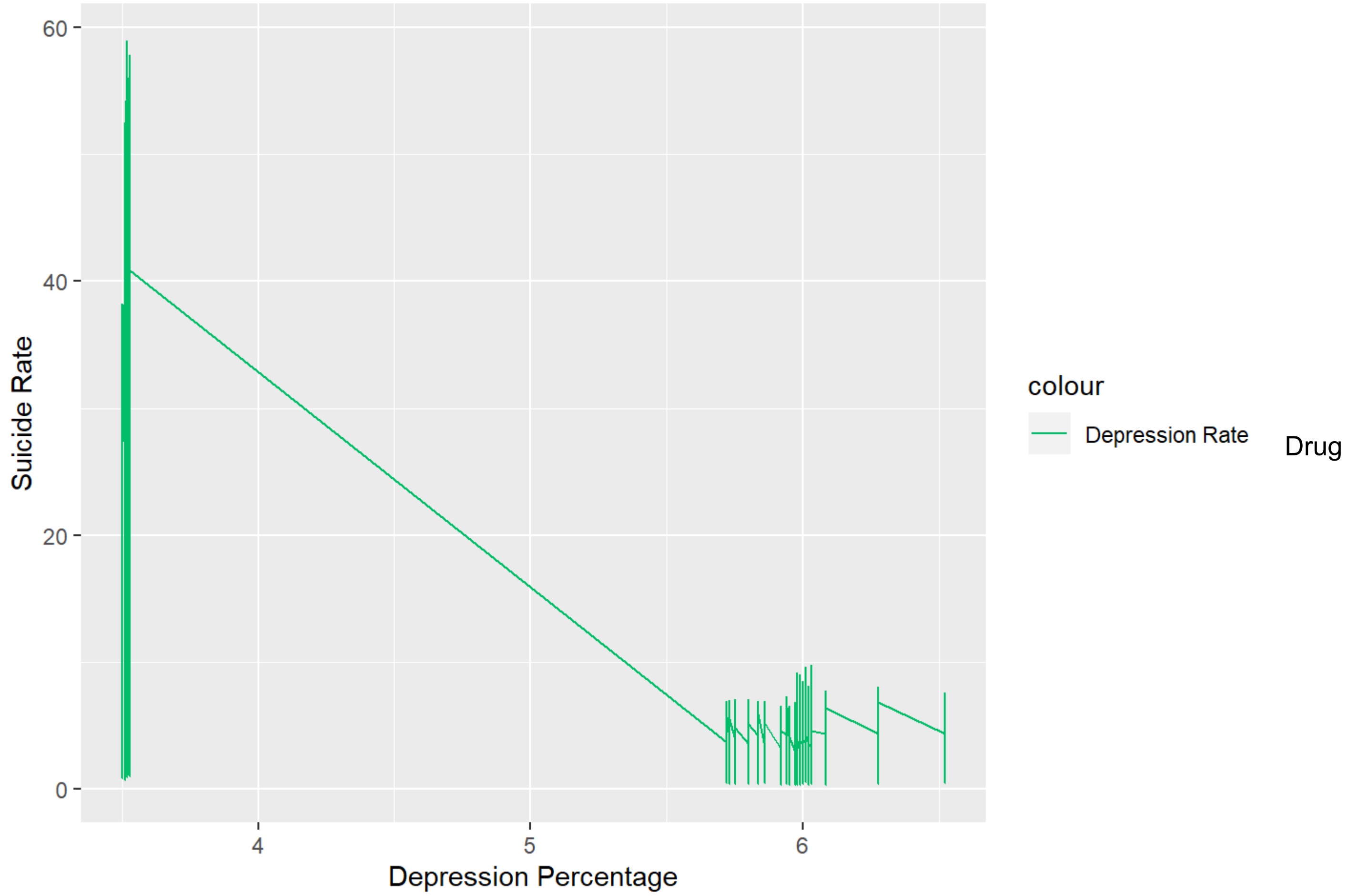
```
## [1] 51.17861
```

```
summary(data$drug_death_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.20    6.55   10.08   12.03   71.60
```

Investigating Suicide rate and the Depression Percentage

```
ggplot(data)+ geom_line(aes(x= data$depression_percentage, y= data$suicides.100k.pop, color = "D
epression Rate")) +
ylab("Suicide Rate") + xlab("Depression Percentage") + scale_color_manual(values=PALETTE[6])
```



death rate and depression rate are slightly negitively correlated as well as suicide rates and drepression percentages.

```
var(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -11.12504
```

```
cor(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -0.6878586
```

```
cor(data$depression_percentage, data$drug_death_rate)
```

```
## [1] -0.1956575
```

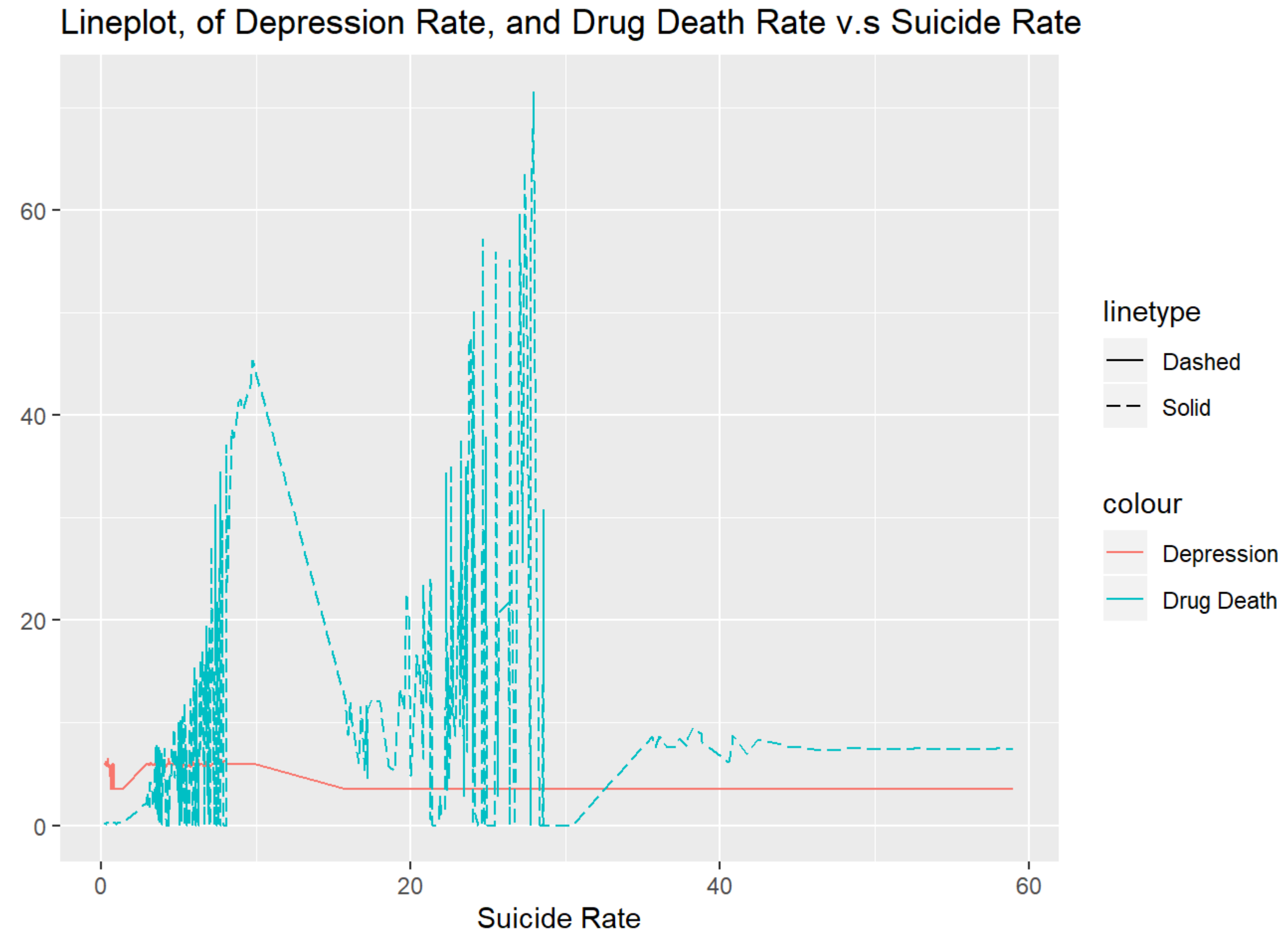
```
cov(data$suicides.100k.pop, data$depression_percentage)
```

```
## [1] -11.12504
```

```
summary(data$depression_percentage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.500   3.510   4.625   4.729   5.980   6.519 
```

```
ggplot(data) + ylab("") + ggtitle("Lineplot, of Depression Rate, and Drug Death Rate v.s Suicide
Rate") +
     geom_line(aes(x= data$suicides.100k.pop, y= data$depression_percentage, lty="Dashed", color
= "Depression")) +
     geom_line(aes(x=data$suicides.100k.pop, y= data$drug_death_rate, lty="Solid", color = "Drug
Death")) +
     #geom_line(aes(x=data$suicides.100k.pop, y= data$population, lty="x9", color = "Populatio
n")) +
     scale_linetype_manual(values=c("solid","longdash")) + xlab("Suicide Rate")
```


Lineplot, of Depression Rate, and Drug Death Rate v.s Suicide Rate
Suicide Rate
linetype
Dashed
Solid
colour
Depression
Drug Death