```
---
title: "Final Project - Suicide rate Analysis."
author: "Bhushan Suryawanshi"
date: '2020-08-05'

---
```

```
library("ggplot2")
library("dplyr")
library("doBy")
library("countrycode")
```

```
suicide_ny_df <- read.csv("suicide-data-ny.csv", header= TRUE, sep= ",", stringsAsFactors = FALSE)

head(suicide_ny_df)
```

```
##   Year Region  Race.or.Ethnicity Sex Age.Group Firearm.Deaths
## 1 2015    NYC Black Non Hispanic   F       1-9              1
## 2 2015    NYC Black Non Hispanic   M       1-9              0
## 3 2015    NYC Black Non Hispanic   F     10-19              2
## 4 2015    NYC Black Non Hispanic   M     10-19             18
## 5 2015    NYC Black Non Hispanic   F     20-24              0
## 6 2015    NYC Black Non Hispanic   M     20-24             34
##   Alcohol.Related.Deaths Suicide.Deaths
## 1                      0              0
## 2                      0              0
## 3                      0              1
## 4                      0              4
## 5                      0              3
## 6                      0              9
```

```
summary(suicide_ny_df)
```

```
##       Year         Region          Race.or.Ethnicity      Sex
##  Min.   :2003   Length:2076        Length:2076        Length:2076
##  1st Qu.:2007   Class :character   Class :character   Class :character
##  Median :2012   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
##  3rd Qu.:2015
##  Max.   :2017
##    Age.Group         Firearm.Deaths   Alcohol.Related.Deaths Suicide.Deaths
##  Length:2076        Min.   :  0.00   Min.   :   0.00        Min.   :  0.00
##  Class :character   1st Qu.:  0.00   1st Qu.:   0.00        1st Qu.:  0.00
##  Mode  :character   Median :  0.00   Median :   1.00        Median :  0.00
##                     Mean   : 10.81   Mean   :  59.75        Mean   : 21.56
##                     3rd Qu.:  2.00   3rd Qu.:  26.00        3rd Qu.: 11.00
##                     Max.   :597.00   Max.   :1832.00        Max.   :913.00
```

```
str(suicide_ny_df)
```

```
## 'data.frame':    2076 obs. of  8 variables:
```

```
##  $ Year                  : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ Region                : chr  "NYC" "NYC" "NYC" "NYC" ...
##  $ Race.or.Ethnicity     : chr  "Black Non Hispanic" "Black Non Hispanic" "Black Non Hispanic" "Blac
##  $ Sex                   : chr  "F" "M" "F" "M" ...
##  $ Age.Group             : chr  "1-9" "1-9" "10-19" "10-19" ...
##  $ Firearm.Deaths        : int  1 0 2 18 0 34 4 65 2 29 ...
##  $ Alcohol.Related.Deaths: int  0 0 0 0 0 0 10 4 9 20 ...
##  $ Suicide.Deaths        : int  0 0 1 4 3 9 5 20 4 6 ...
```

```
suicide_who_df <- read.csv("who_suicide_statistics.csv", header= TRUE, sep= ",", stringsAsFactors = FALS

head(suicide_who_df)
```

```
##   country year    sex         age suicides_no population
## 1 Albania 1985 female 15-24 years          NA     277900
## 2 Albania 1985 female 25-34 years          NA     246800
## 3 Albania 1985 female 35-54 years          NA     267500
## 4 Albania 1985 female  5-14 years          NA     298300
## 5 Albania 1985 female 55-74 years          NA     138700
## 6 Albania 1985 female   75+ years          NA      34200
```

```
summary(suicide_who_df)
```

```
##    country               year          sex                age
##  Length:43776       Min.   :1979   Length:43776       Length:43776
##  Class :character   1st Qu.:1990   Class :character   Class :character
##  Mode  :character   Median :1999   Mode  :character   Mode  :character
##                     Mean   :1999
##                     3rd Qu.:2007
##                     Max.   :2016
##
##   suicides_no        population
##  Min.   :    0.0   Min.   :     259
##  1st Qu.:    1.0   1st Qu.:   85113
##  Median :   14.0   Median :  380655
##  Mean   :  193.3   Mean   : 1664091
##  3rd Qu.:   91.0   3rd Qu.: 1305698
##  Max.   :22338.0   Max.   :43805214
##  NA's   :2256      NA's   :5460
```

```
str(suicide_who_df)
```

```
## 'data.frame':    43776 obs. of  6 variables:
##  $ country    : chr  "Albania" "Albania" "Albania" "Albania" ...
##  $ year       : int  1985 1985 1985 1985 1985 1985 1985 1985 1985 1985 ...
##  $ sex        : chr  "female" "female" "female" "female" ...
##  $ age        : chr  "15-24 years" "25-34 years" "35-54 years" "5-14 years" ...
##  $ suicides_no: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ population : int  277900 246800 267500 298300 138700 34200 301400 264200 296700 325800 ...
```

```r
suicide_df <- read.csv("final_suicide_data.csv", header= TRUE, sep= ",", stringsAsFactors = FALSE)

head(suicide_df)
```

```
##   ï..country year    sex         age suicides_no population suicides.100k.pop
## 1    Albania 1987   male 15-24 years          21     312900              6.71
## 2    Albania 1987   male 35-54 years          16     308000              5.19
## 3    Albania 1987 female 15-24 years          14     289700              4.83
## 4    Albania 1987   male   75+ years           1      21800              4.59
## 5    Albania 1987   male 25-34 years           9     274300              3.28
## 6    Albania 1987 female   75+ years           1      35600              2.81
##   country.year HDI.for.year gdp_for_year.... gdp_per_capita....      generation
## 1  Albania1987           NA    2,156,624,900                796    Generation X
## 2  Albania1987           NA    2,156,624,900                796          Silent
## 3  Albania1987           NA    2,156,624,900                796    Generation X
## 4  Albania1987           NA    2,156,624,900                796 G.I. Generation
## 5  Albania1987           NA    2,156,624,900                796         Boomers
## 6  Albania1987           NA    2,156,624,900                796 G.I. Generation
```

```r
summary(suicide_df)
```

```
##   ï..country             year          sex                age
##  Length:27820       Min.   :1985   Length:27820       Length:27820
##  Class :character   1st Qu.:1995   Class :character   Class :character
##  Mode  :character   Median :2002   Mode  :character   Mode  :character
##                     Mean   :2001
##                     3rd Qu.:2008
##                     Max.   :2016
##
##   suicides_no        population       suicides.100k.pop country.year
##  Min.   :    0.0   Min.   :     278   Min.   :  0.00    Length:27820
##  1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92    Class :character
##  Median :   25.0   Median :  430150   Median :  5.99    Mode  :character
##  Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
##  3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
##  Max.   :22338.0   Max.   :43805214   Max.   :224.97
##
##   HDI.for.year   gdp_for_year....   gdp_per_capita....  generation
##  Min.   :0.483   Length:27820       Min.   :   251     Length:27820
##  1st Qu.:0.713   Class :character   1st Qu.:  3447     Class :character
##  Median :0.779   Mode  :character   Median :  9372     Mode  :character
##  Mean   :0.777                      Mean   : 16866
##  3rd Qu.:0.855                      3rd Qu.: 24874
##  Max.   :0.944                      Max.   :126352
##  NA's   :19456
```

```r
str(suicide_df)
```

```
## 'data.frame':    27820 obs. of  12 variables:
##  $ ï..country        : chr  "Albania" "Albania" "Albania" "Albania" ...
##  $ year              : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987 ...
##  $ sex               : chr  "male" "male" "female" "male" ...
```

```
##  $ age               : chr  "15-24 years" "35-54 years" "15-24 years" "75+ years" ...
##  $ suicides_no       : int  21 16 14 1 9 1 6 4 1 0 ...
##  $ population        : int  312900 308000 289700 21800 274300 35600 278800 257200 137500 311000 ...
##  $ suicides.100k.pop : num  6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0 ...
##  $ country.year      : chr  "Albania1987" "Albania1987" "Albania1987" "Albania1987" ...
##  $ HDI.for.year      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ gdp_for_year....  : chr  "2,156,624,900" "2,156,624,900" "2,156,624,900" "2,156,624,900" ...
##  $ gdp_per_capita....: int  796 796 796 796 796 796 796 796 796 796 ...
##  $ generation        : chr  "Generation X" "Silent" "Generation X" "G.I. Generation" ...
```

- I managed to search and get different datasets online for suicide data. Those data set include - The data set received from health data had data related to only New York state. Also this dataset does not have GDP data to know the economic impact. However, the questions I selected was more broader to address country level statistics and comparison between different geo locations across the globe. The other data set received from Kaggle (WHO statistic) was giving me required details of country level statistics for about 100 different countries. However that did not have the GDP data. The third dataset used for the suicide analysis was a large dataset same as WHO dataset with GDP data and hence used it as my final dataset for analysis.
- Checking and cleaning dataset was another step which was required as the data received has HDI column with missing data. As for the current analysis it was not required and hence dropped that column. There was additional column added as "suicide rate" which was not present earlier. I also required to split large dataset to create country specific datasets so that those can be used for comparison. Boxplot and histogram shows the data distribution.

```
# Checking duplicate - if one row is identical to another
distinctdata <- distinct(suicide_df)
nrow(suicide_df) == nrow(distinctdata)
```

Data Cleaning

```
## [1] TRUE
```

```
suicide_sorted_df <- suicide_df %>% select(-(HDI.for.year), -(country.year),
                                           -(gdp_for_year....)) %>%
  rename(country = ï..country, gdp_per_cap = gdp_per_capita....) %>%
  filter(year < c(2016))


suicide_sorted_df <- suicide_sorted_df %>%
mutate(suicideRate = (suicides_no/population)*100)


#gdp_per_capita - create a log transformed column
suicide_sorted_df<- suicide_sorted_df %>% mutate(gdp_per_cap_log = log10(suicide_sorted_df$gdp_per_cap))

suicide_sorted_df$age <- as.factor(suicide_sorted_df$age)

suicide_sorted_df$continent <- countrycode(sourcevar = suicide_sorted_df[, "country"],
                              origin = "country.name",
                              destination = "continent")
```

```
data_nominal <- c('country', 'sex', 'continent')
suicide_sorted_df[data_nominal] <- lapply(suicide_sorted_df[data_nominal],
                                          function(x){factor(x)})

suicide_sorted_df$age <- gsub(" years", "", suicide_sorted_df$age)


suicide_sorted_df$age <- factor(suicide_sorted_df$age,
                  ordered = T, levels = c("5-14", "15-24", "25-34", "35-54",
                                          "55-74", "75+"))

suicide_sorted_df$generation <- factor(suicide_sorted_df$generation,
                  ordered = T,
                  levels = c("G.I. Generation", "Silent", "Boomers",
                             "Generation X", "Millenials", "Generation Z"))

global_average <- (sum(as.numeric(suicide_sorted_df$suicides_no)) /
                     sum(as.numeric(suicide_sorted_df$population))) * 100000

summary(suicide_sorted_df)
```

```
##       country           year          sex           age        suicides_no
##  Argentina:  372   Min.   :1985   female:13830   5-14 :4610   Min.   :    0.0
##  Austria  :  372   1st Qu.:1994   male  :13830   15-24:4610   1st Qu.:    3.0
##  Belgium  :  372   Median :2002                  25-34:4610   Median :   25.0
##  Brazil   :  372   Mean   :2001                  35-54:4610   Mean   :  243.4
##  Chile    :  372   3rd Qu.:2008                  55-74:4610   3rd Qu.:  132.0
##  Colombia :  372   Max.   :2015                  75+  :4610   Max.   :22338.0
##  (Other)  :25428
##    population       suicides.100k.pop  gdp_per_cap              generation
##  Min.   :     278   Min.   :  0.00    Min.   :   251   G.I. Generation:2744
##  1st Qu.:   97535   1st Qu.:  0.91    1st Qu.:  3436   Silent         :6332
##  Median :  430725   Median :  5.98    Median :  9283   Boomers        :4958
##  Mean   : 1850689   Mean   : 12.81    Mean   : 16816   Generation X   :6376
##  3rd Qu.: 1491041   3rd Qu.: 16.60    3rd Qu.: 24796   Millenials     :5780
##  Max.   :43805214   Max.   :224.97    Max.   :126352   Generation Z   :1470
##
##   suicideRate       gdp_per_cap_log    continent
##  Min.   :0.000000   Min.   :2.400   Africa  :  840
##  1st Qu.:0.000911   1st Qu.:3.536   Americas: 9204
##  Median :0.005977   Median :3.968   Asia    : 5316
##  Mean   :0.012813   Mean   :3.936   Europe  :11328
##  3rd Qu.:0.016605   3rd Qu.:4.394   Oceania :  972
##  Max.   :0.224972   Max.   :5.102
##
```

```
str(suicide_sorted_df)
```

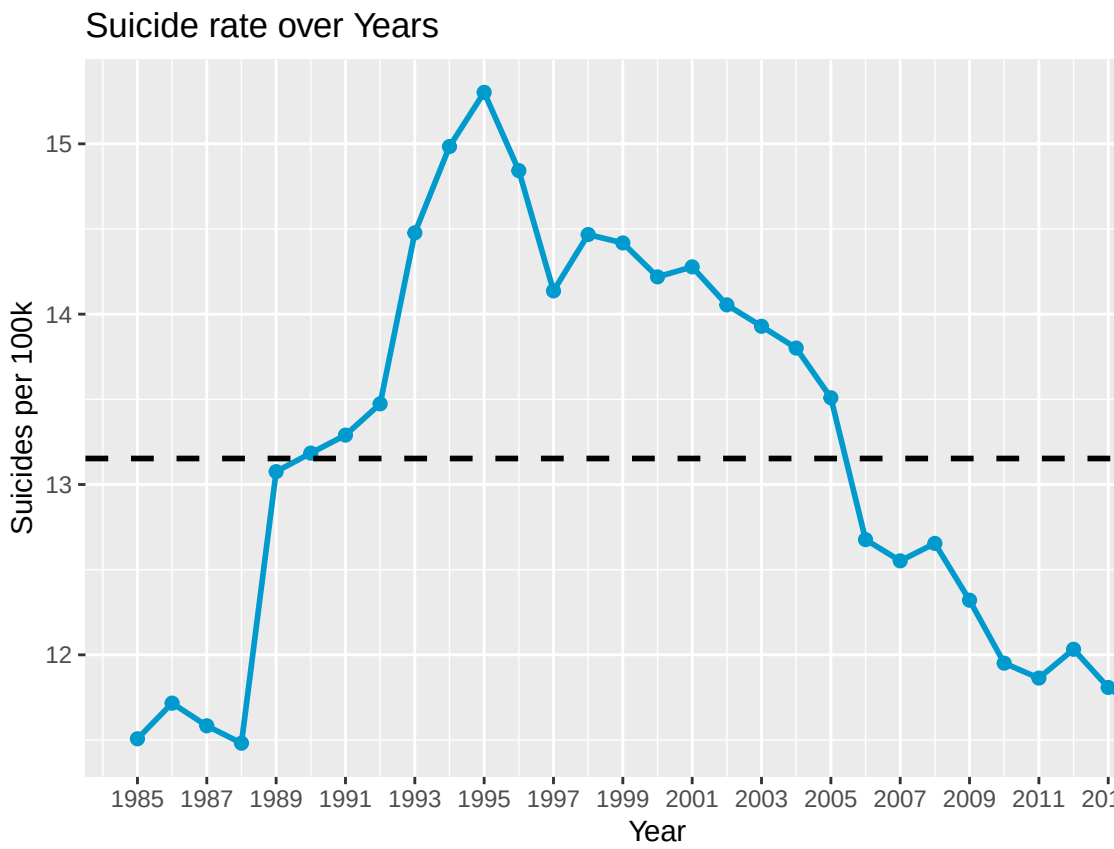
```
## 'data.frame':    27660 obs. of  12 variables:
##  $ country          : Factor w/ 100 levels "Albania","Antigua and Barbuda",..: 1 1 1 1 1 1 1 1 1 1 .
##  $ year             : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987 ...
```

```
##  $ sex              : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 1 2 1 ...
##  $ age              : Ord.factor w/ 6 levels "5-14"<"15-24"<..: 2 4 2 6 3 6 4 3 5 1 ...
##  $ suicides_no      : int  21 16 14 1 9 1 6 4 1 0 ...
##  $ population       : int  312900 308000 289700 21800 274300 35600 278800 257200 137500 311000 ...
##  $ suicides.100k.pop: num  6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0 ...
##  $ gdp_per_cap      : int  796 796 796 796 796 796 796 796 796 796 ...
##  $ generation       : Ord.factor w/ 6 levels "G.I. Generation"<..: 4 2 4 1 3 1 2 3 1 4 ...
##  $ suicideRate      : num  0.00671 0.00519 0.00483 0.00459 0.00328 ...
##  $ gdp_per_cap_log  : num  2.9 2.9 2.9 2.9 2.9 ...
##  $ continent        : Factor w/ 5 levels "Africa","Americas",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
head(suicide_sorted_df)
```

```
##   country year    sex   age suicides_no population suicides.100k.pop
## 1 Albania 1987   male 15-24          21     312900              6.71
## 2 Albania 1987   male 35-54          16     308000              5.19
## 3 Albania 1987 female 15-24          14     289700              4.83
## 4 Albania 1987   male   75+           1      21800              4.59
## 5 Albania 1987   male 25-34           9     274300              3.28
## 6 Albania 1987 female   75+           1      35600              2.81
##   gdp_per_cap      generation suicideRate gdp_per_cap_log continent
## 1         796    Generation X 0.006711409        2.900913    Europe
## 2         796          Silent 0.005194805        2.900913    Europe
## 3         796    Generation X 0.004832585        2.900913    Europe
## 4         796 G.I. Generation 0.004587156        2.900913    Europe
## 5         796         Boomers 0.003281079        2.900913    Europe
## 6         796 G.I. Generation 0.002808989        2.900913    Europe
```

```
suicide_sorted_df %>%
  group_by(year) %>%
  summarize(population = sum(population),
            suicides = sum(suicides_no),
            suicides_per_100k = (suicides / population) * 100000) %>%
  ggplot(aes(x = year, y = suicides_per_100k)) +
  geom_line(col = "deepskyblue3", size = 1) +
  geom_point(col = "deepskyblue3", size = 2) +
  geom_hline(yintercept = global_average, linetype = 2, size = 1) +
  labs(title = "Suicide rate over Years",
       x = "Year",
       y = "Suicides per 100k") +
  scale_x_continuous(breaks = seq(1985, 2015, 2)) +
  scale_y_continuous(breaks = seq(10, 20))
```

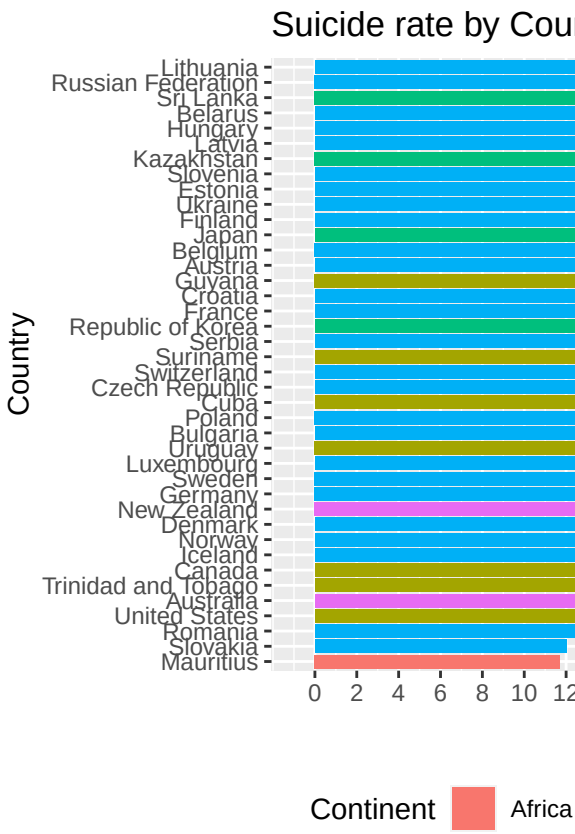


Suicide Rate over time

- Average suicide rate for available data from 1985 to 2015 shows 13.5 per 100k of population.
- Highest rate of suicide was observed in 1995.
- This answers our question " Is Suicide rate affected recently (decreasing/increasing)?". Yes suicide rate decreased since 1995. If we compare suicide rate in 1995 and 2015 there is approximately 25% decrease.

```
country <- suicide_sorted_df %>%
  group_by(country, continent) %>%
  summarize(n = n(),
            suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %
  arrange(desc(suicide_per_100k))

country$country <- factor(country$country,
                          ordered = T,
                          levels = rev(country$country))

country1 <- country[1:40,]

ggplot(country1, aes(x = country, y = suicide_per_100k, fill = continent)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Suicide rate by Country",
       x = "Country",
       y = "Suicides per 100k",
```

```
      fill = "Continent") +
  coord_flip() +
  scale_y_continuous(breaks = seq(0, 45, 2)) +
  theme(legend.position = "bottom")
```



Lets look at sucide rate per 100k population for top 40 countries

```
country_year_gdp <- suicide_sorted_df %>%
  group_by(country, year) %>%
  summarize(gdp_per_cap = mean(gdp_per_cap))

country_year_gdp_corr <- country_year_gdp %>%
  ungroup() %>%
  group_by(country) %>%
  summarize(year_gdp_correlation = cor(year, gdp_per_cap))


summary(country_year_gdp_corr)
```

Suicide rate by GDP

```
##                 country   year_gdp_correlation
##  Albania            : 1   Min.   :-0.5458
##  Antigua and Barbuda: 1   1st Qu.: 0.8665
```

```
##  Argentina          : 1   Median : 0.9104
##  Armenia            : 1   Mean   : 0.8796
##  Aruba              : 1   3rd Qu.: 0.9552
##  Australia          : 1   Max.   : 1.0000
##  (Other)            :94   NA's   :3
```

```
country_mean_gdp <- suicide_sorted_df %>%
  group_by(country, continent) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000,
            gdp_per_capita = mean(gdp_per_cap))

summary( lm(suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp))
```

```
##
## Call:
## lm(formula = suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.755  -6.606  -2.556   4.748  30.815
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.028e+01  1.229e+00   8.370  4.1e-13 ***
## gdp_per_capita 4.036e-05  5.374e-05   0.751    0.454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.769 on 98 degrees of freedom
## Multiple R-squared:  0.005723,   Adjusted R-squared:  -0.004423
## F-statistic: 0.5641 on 1 and 98 DF,  p-value: 0.4544
```

```
summary(suicide_sorted_df$suicideRate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000911 0.005977 0.012813 0.016605 0.224972
```

```
summary(suicide_sorted_df$gdp_per_cap_log)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.400   3.536   3.968   3.936   4.394   5.102
```

```
summary(lm(suicide_sorted_df$suicideRate~suicide_sorted_df$gdp_per_cap_log,data=suicide_sorted_df))
```

```
##
## Call:
## lm(formula = suicide_sorted_df$suicideRate ~ suicide_sorted_df$gdp_per_cap_log,
##     data = suicide_sorted_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

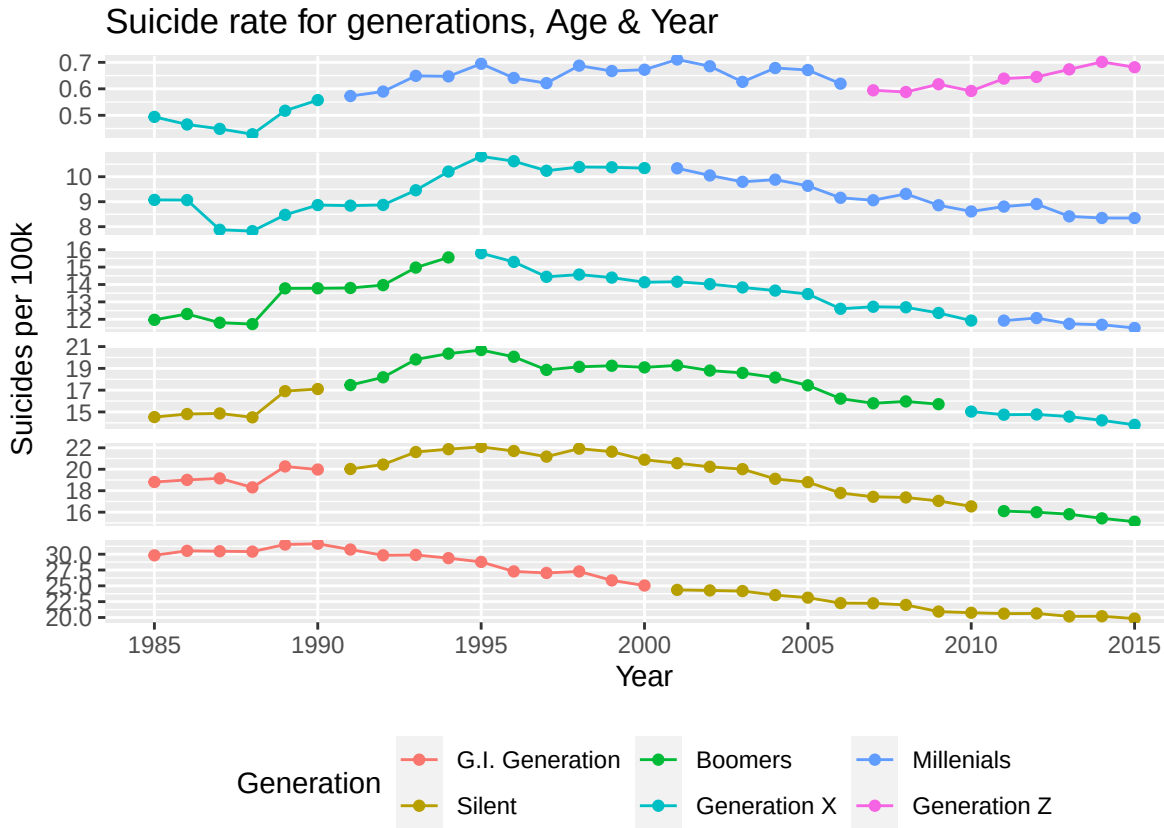
```
## -0.013313 -0.011844 -0.006840  0.003758  0.212023
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.0111241  0.0008228  13.519   <2e-16 ***
## suicide_sorted_df$gdp_per_cap_log 0.0004290  0.0002070   2.072   0.0383 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01897 on 27658 degrees of freedom
## Multiple R-squared:  0.0001552,  Adjusted R-squared:  0.0001191
## F-statistic: 4.294 on 1 and 27658 DF,  p-value: 0.03827
```

```
cor.test(suicide_sorted_df$suicideRate, suicide_sorted_df$gdp_per_cap_log, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  suicide_sorted_df$suicideRate and suicide_sorted_df$gdp_per_cap_log
## t = 2.0721, df = 27658, p-value = 0.03827
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0006736194 0.0242397371
## sample estimates:
##        cor
## 0.01245841
```

- The mean correlation was 0.878, indicates that there is strong positive linear relationship between GDP and number of suicides.
- This shows increase in wealth per person also affect suicide rate positively.
- The regression model shows p-value of 0.0057 which means there is little variance in suicide rate because of GDP.
- This means there is positive relation but is a weak relationship.
- However earlier graph of suicide rate over time shows that the suicide rates are decreasing over time.
- Hence more answers can be found by analyzing each country separately.

```
suicide_sorted_df %>%
  group_by(generation, age, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %
  ggplot(aes(x = year, y = suicide_per_100k, col = factor(generation, ordered = F))) +
  geom_point() +
  geom_line() +
  facet_grid(age ~ ., scales = "free_y") +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = NULL) +
  labs(title = "Suicide rate for generations, Age & Year",
       x = "Year",
       y = "Suicides per 100k",
       col = "Generation") +
  theme(legend.position = "bottom")
```



Suicide rate by age

```
by(suicide_sorted_df$suicides_no, suicide_sorted_df$age, mean)
```

```
## suicide_sorted_df$age: 5-14
## [1] 11.33709
## ------------------------------------------------------------
## suicide_sorted_df$age: 15-24
## [1] 175.1295
## ------------------------------------------------------------
## suicide_sorted_df$age: 25-34
## [1] 243.3497
## ------------------------------------------------------------
## suicide_sorted_df$age: 35-54
## [1] 530.667
## ------------------------------------------------------------
## suicide_sorted_df$age: 55-74
## [1] 358.7262
## ------------------------------------------------------------
## suicide_sorted_df$age: 75+
## [1] 141.2714
```

- The above number shows that age group 5-14 has lowest number of suicides.
- Age group 35-54 has highest number of suicides globally.

```r
by(suicide_sorted_df$suicides_no, suicide_sorted_df$age, length)
```

```
## suicide_sorted_df$age: 5-14
## [1] 4610
## ------------------------------------------------------------
## suicide_sorted_df$age: 15-24
## [1] 4610
## ------------------------------------------------------------
## suicide_sorted_df$age: 25-34
## [1] 4610
## ------------------------------------------------------------
## suicide_sorted_df$age: 35-54
## [1] 4610
## ------------------------------------------------------------
## suicide_sorted_df$age: 55-74
## [1] 4610
## ------------------------------------------------------------
## suicide_sorted_df$age: 75+
## [1] 4610
```

```r
anova <- aov(suicides_no ~ age, data= suicide_sorted_df)
summary(anova)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## age             5 7.596e+08 151914991   192.1 <2e-16 ***
## Residuals   27654 2.187e+10    790783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value shown above is < 0.05. Hence we can say that there is significant difference in suicide rate across age group.

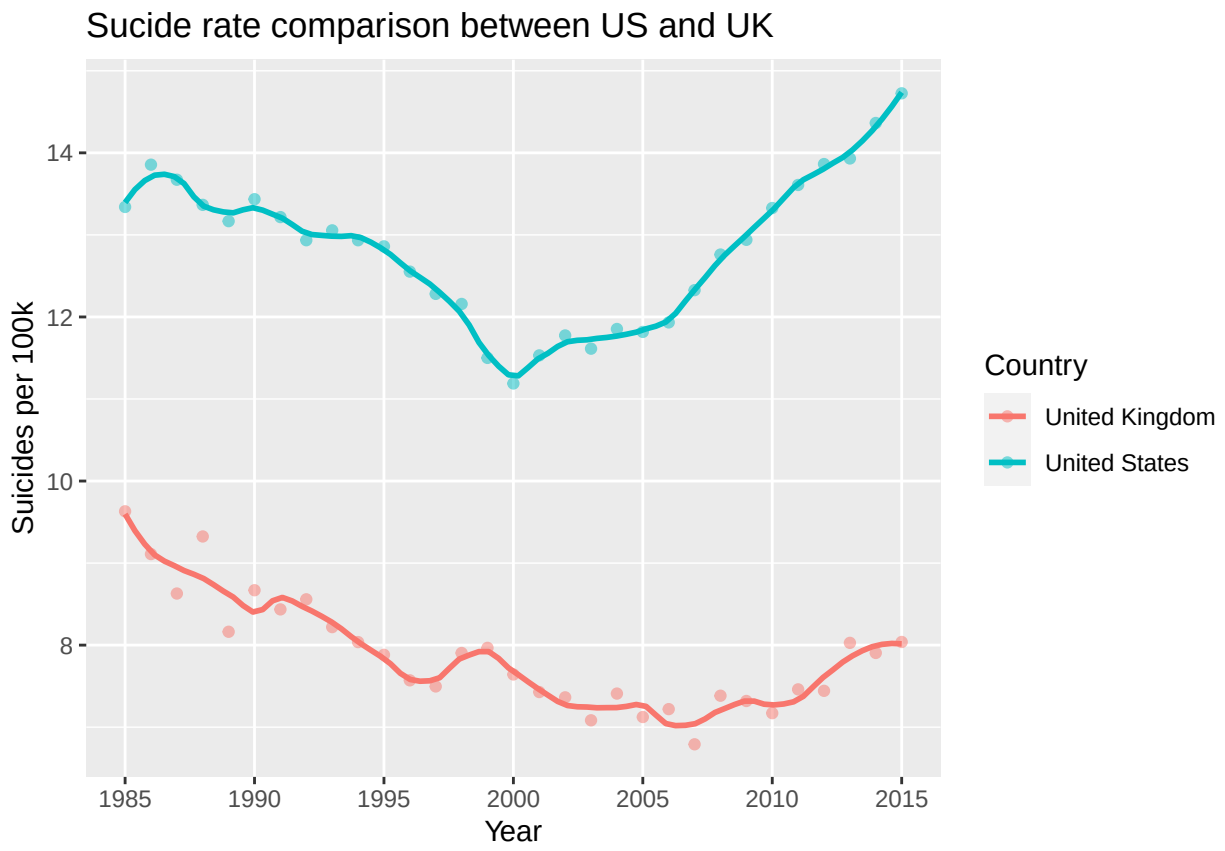
```r
suicide_us_uk_df <- suicide_sorted_df %>%
  filter(country %in% c("United Kingdom",
                        "United States"))

summary(suicide_us_uk_df)
```

Comparing suicide rates between USA and UK as they are similar economies

```
##                 country         year          sex         age
##  United Kingdom     :372   Min.   :1985   female:372   5-14 :124
##  United States      :372   1st Qu.:1992   male  :372   15-24:124
##  Albania            :  0   Median :2000                25-34:124
##  Antigua and Barbuda:  0   Mean   :2000                35-54:124
##  Argentina          :  0   3rd Qu.:2008                55-74:124
##  Armenia            :  0   Max.   :2015                75+  :124
##  (Other)            :  0
##   suicides_no      population       suicides.100k.pop  gdp_per_cap
```

```
##  Min.   :    0   Min.   : 1202838   Min.   : 0.000    Min.   : 9231
##  1st Qu.:  181   1st Qu.: 4112590   1st Qu.: 3.490    1st Qu.:24654
##  Median :  548   Median : 8164568   Median : 6.225    Median :35404
##  Mean   : 1574   Mean   :13162359   Mean   :10.661    Mean   :35589
##  3rd Qu.: 2185   3rd Qu.:20359185   3rd Qu.:16.648    3rd Qu.:47163
##  Max.   :11767   Max.   :43805214   Max.   :58.950    Max.   :60387
##
##            generation   suicideRate       gdp_per_cap_log    continent
##  G.I. Generation: 88   Min.   :0.000000   Min.   :3.965   Africa  :  0
##  Silent         :164   1st Qu.:0.003494   1st Qu.:4.392   Americas:372
##  Boomers        :136   Median :0.006223   Median :4.549   Asia    :  0
##  Generation X   :176   Mean   :0.010661   Mean   :4.517   Europe  :372
##  Millenials     :144   3rd Qu.:0.016648   3rd Qu.:4.674   Oceania :  0
##  Generation Z   : 36   Max.   :0.058952   Max.   :4.781
##
```

```
suicide_us_uk_df %>%
  group_by(country, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  labs(title = "Sucide rate comparison between US and UK",
       x = "Year",
       y = "Suicides per 100k",
       col = "Country")
```



- UK rates shows lower and also decreased consistently over the years.
- US rate of suicide is quite volatile and shows decrease between 1995 to 2000. However there is equally sharp increase between 2000 to 2015.