Suicides_Analytics

Ng Wei Xiang

5/25/2021

Suicides Analytics

This is an R Markdown document of the personal work/project I am for practice purposes. This dataset was obtained from kaggle and is cleaned and altered by myself for my own analytics needs.

Loading of libraries

These libraries are used and required as tools for analytics to be performed further down.

```
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 4.0.3
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('ra
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

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

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

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

```r
library(psych) # for describeBy
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(tidyverse) # for ancova
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.3     v purrr   0.3.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'stringr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x psych::%+%()    masks ggplot2::%+%()
## x psych::alpha()  masks ggplot2::alpha()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tibble::view()  masks summarytools::view()
```

```r
library(ggpubr) # for ancova
```

```
## Warning: package 'ggpubr' was built under R version 4.0.3
```

```r
library(rstatix) # for ancova
```

```
## Warning: package 'rstatix' was built under R version 4.0.3
```

```
##
## Attaching package: 'rstatix'
```

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

```r
library(broom) # for ancova
```

```
## Warning: package 'broom' was built under R version 4.0.3
```

Dataset loading, cleaning and modifying

Firstly, the csv file is read into a dataframe called "master". Then, a copy of "master" was created with the name "dataset1" to create a new dataset and prevent overriding of the master and original version. Then, a new dataset was created to get the total suicides number in the country. The new dataset named "sum" was then merged with "dataset1" to add the new variable into the main dataset and the variables were then renamed for more readibility. Then, the variable "gdp_for_year" was converted from character to numeric and removed the ","(comma). It is then renamed properly for better usability.

```r
# reading the CSV file
master <- read.csv('master.csv')

# to remove the column country.year because is unused
dataset1 = master
dataset1$country.year <- NULL

# add all suicides_no by country as a new variable
dataset2 = master
sum2 <- aggregate(x=dataset2$suicides_no, by=list(dataset2$ï..country), FUN=sum)
names(sum2)[1] <- "ï..country"

# merging the new dataset with the main and renaming some variables
dataset1 <- merge(dataset1, sum2, by="ï..country")
names(dataset1)[12] <- "total_suicides_country"
names(dataset1)[1] <- "country"

# covert gdp_for_year to numeric
dataset1$gdp_for_year.... <- as.numeric(gsub(",", "", dataset1$gdp_for_year....))
names(dataset1)[9] <- "gdp_for_year"
```

Descriptive Analysis

Variables

Variable 1 - country This dataset consists a total of 101 countries recorded. It is a nominal categorical variable

```r
dfSummary(dataset1$country)
```

```
## dataset1$country was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27719
##
```

```
## ----------------------------------------------------------------------------------------------------
## No   Variable      Stats / Values     Freqs (% of Valid)   Graph               Valid      Missing
## ---- ------------- ------------------ -------------------- ------------------- ---------- ---------
## 1    country       1. Austria           382 ( 1.4%)                           27820      0
##      [character]   2. Iceland           382 ( 1.4%)                           (100.0%)   (0.0%)
##                    3. Mauritius         382 ( 1.4%)
##                    4. Netherlands       382 ( 1.4%)
##                    5. Argentina         372 ( 1.3%)
##                    6. Belgium           372 ( 1.3%)
##                    7. Brazil            372 ( 1.3%)
##                    8. Chile             372 ( 1.3%)
##                    9. Colombia          372 ( 1.3%)
##                    10. Ecuador          372 ( 1.3%)
##                    [ 91 others ]      24060 (86.5%)        IIIIIIIIIIIIIIIII
## ----------------------------------------------------------------------------------------------------
```

Variable 2 - year This dataset collected records since year 1985 up till 2016. It is a continuous variable.

```
dfSummary(dataset1$year)
```

```
## dataset1$year was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27788
##
## ----------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values              Freqs (% of Valid)   Graph                       Valid
## ---- ----------- --------------------------- -------------------- --------------------------- ------
## 1    year        Mean (sd) : 2001.3 (8.5)    32 distinct values         . : : : : : .          27820
##      [integer]   min < med < max:                                  : . : : : : : : : :          (100.
##                  1985 < 2002 < 2016                               : : : : : : : : : :
##                  IQR (CV) : 13 (0)                                : : : : : : : : : :
##                                                                   : : : : : : : : : :
## ----------------------------------------------------------------------------------------------------
```

Variable 3 - sex This dataset collected records of 2 sex, female and male. It is a categorical binary variable.

```
dfSummary(dataset1$sex)
```

```
## dataset1$sex was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27818
##
## ----------------------------------------------------------------------------------------------
## No   Variable      Stats / Values   Freqs (% of Valid)   Graph           Valid      Missing
```

```
## ---- ------------- ---------------- -------------------- --------------- ---------- ---------
## 1    sex           1. female        13910 (50.0%)        IIIIIIIIII      27820      0
##      [character]   2. male          13910 (50.0%)        IIIIIIIIII      (100.0%)   (0.0%)
## ------------------------------------------------------------------------------------------------
```

Variable 4 - age This dataset collected 6 age groups. It is a nominal categoricalvariable. Ascendingly, 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years,

```
dfSummary(dataset1$age)
```

```
## dataset1$age was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27814
##
## ------------------------------------------------------------------------------------------------
## No   Variable      Stats / Values     Freqs (% of Valid)   Graph    Valid      Missing
## ---- ------------- ------------------ -------------------- -------- ---------- ---------
## 1    age           1. 15-24 years     4642 (16.7%)         III      27820      0
##      [character]   2. 25-34 years     4642 (16.7%)         III      (100.0%)   (0.0%)
##                    3. 35-54 years     4642 (16.7%)         III
##                    4. 5-14 years      4610 (16.6%)         III
##                    5. 55-74 years     4642 (16.7%)         III
##                    6. 75+ years       4642 (16.7%)         III
## ------------------------------------------------------------------------------------------------
```

Variable 5 - suicides_no This dataset collected suicide numbers as continuous variable from the above country and each row of record is represented by each sex, generation, year, and country. It recorded with a minimum of 0 suicides, median of 25 suicides, and maximum of 22,338 suicides.

```
dfSummary(dataset1$suicides_no)
```

```
## dataset1$suicides_no was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 25736
##
## ------------------------------------------------------------------------------------------------------
## No   Variable       Stats / Values             Freqs (% of Valid)     Graph   Valid      Missing
## ---- -------------- -------------------------- ---------------------- ------- ---------- ---------
## 1    suicides_no    Mean (sd) : 242.6 (902)    2084 distinct values   :       27820      0
##      [integer]      min < med < max:                                  :       (100.0%)   (0.0%)
##                     0 < 25 < 22338                                    :
##                     IQR (CV) : 128 (3.7)                              :
##                                                                       :
## ------------------------------------------------------------------------------------------------------
```

Variable 6 - population This dataset recorded population of the countries as a single column with the largest country having a population of 43,805,214 and a median of 430,150. It is a discrete variable.

```
dfSummary(dataset1$population)
```

```
## dataset1$population was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 2256
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable      Stats / Values                    Freqs (% of Valid)      Graph   Valid      Missi
## ---- ------------- --------------------------------- ----------------------- ------- ---------- ------
## 1    population    Mean (sd) : 1844794 (3911779)     25564 distinct values   :       27820      0
##      [integer]     min < med < max:                                          :       (100.0%)   (0.0%
##                    278 < 430150 < 43805214                                   :
##                    IQR (CV) : 1388645 (2.1)                                  :
##                                                                              : .
## ------------------------------------------------------------------------------------------------------------
```

Variable 7 - suicides.100k.pop This dataset recorded the number of suicides in every 100 thousand of the population in every countries. It is a continous variable with a median of 6 cases and a maximum of 225 cases.

```
dfSummary(dataset1$suicides.100k.pop)
```

```
## dataset1$suicides.100k.pop was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 22522
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable             Stats / Values           Freqs (% of Valid)     Graph   Valid      Missing
## ---- -------------------- ------------------------ ---------------------- ------- ---------- --------
## 1    suicides.100k.pop    Mean (sd) : 12.8 (19)    5298 distinct values   :       27820      0
##      [numeric]            min < med < max:                                :       (100.0%)   (0.0%)
##                           0 < 6 < 225                                     :
##                           IQR (CV) : 15.7 (1.5)                           :
##                                                                           : .
## ------------------------------------------------------------------------------------------------------------
```

Variable 8 - HDI.for.year This dataset recorded the HDI for year for the countries in the dataset.

```
dfSummary(dataset1$HDI.for.year)
```

```
## dataset1$HDI.for.year was converted to a data frame
```

```
## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27514
##
## ------------------------------------------------------------------------------------------------------
## No   Variable        Stats / Values           Freqs (% of Valid)    Graph                          Va
## ---- --------------- ------------------------ --------------------- ------------------------------ --
## 1    HDI.for.year    Mean (sd) : 0.8 (0.1)    305 distinct values            : . . . .              83
##      [numeric]       min < med < max:                                      . : : : : :              (3
##                      0.5 < 0.8 < 0.9                                     . : : : : : : :
##                      IQR (CV) : 0.1 (0.1)                                : : : : : : : :
##                                                                        . : : : : : : : : :
## ------------------------------------------------------------------------------------------------------
```

Variable 9 - gdp_for_year This dataset recorded the gdp of a country for the given year as a continuous variable.

```
dfSummary(dataset1$gdp_for_year)
```

```
## dataset1$gdp_for_year was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 25499
##
## ------------------------------------------------------------------------------------------------------
## No   Variable        Stats / Values                              Freqs (% of Valid)     Graph   Valid
## ---- --------------- ------------------------------------------- ---------------------- ------- --------
## 1    gdp_for_year    Mean (sd) : 4.45581e+11 (1.45361e+12)       2321 distinct values   :       27820
##      [numeric]       min < med < max:                                                   :       (100.0%)
##                      46919625 < 48114688201 < 1.812071e+13                              :
##                      IQR (CV) : 251217076318 (3.3)                                      :
##                                                                                         :
## ------------------------------------------------------------------------------------------------------
```

Variable 10 - gdp_per_capita This dataset recorded the gdp of a country for the given year as a continuous variable. It recorded a median of 9372 and a maximum of 126352.

```
dfSummary(dataset1$gdp_per_capita)
```

```
## dataset1$gdp_per_capita was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 25587
##
## ------------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values                  Freqs (% of Valid)     Graph       Valid      Missi
```

```
## ---- ----------- -------------------------------- ---------------------- ----------- ---------- -----
## 1    dataset1    Mean (sd) : 16866.5 (18887.6)    2233 distinct values   :           27820      0
##      [integer]   min < med < max:                                          :           (100.0%)   (0.0%
##                  251 < 9372 < 126352                                      :
##                  IQR (CV) : 21427 (1.1)                                   : .
##                                                                           : : : . .
## ---------------------------------------------------------------------------------------------------
```

Variable 11 - generation This dataset recorded 6 generations which corresponds to a certain age group. Namingly, age 5 to 14 years is **Gen Z**, age 15 to 24 is **Millenial**, age 25 to 34 is **Gen X**, age 35 to 54 is **Boomers**, age 55 to 74 is **Silent**, and age 75 and above is **G.I. Gen**.

```
dfSummary(dataset1$generation)
```

```
## dataset1$generation was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 27814
##
## ---------------------------------------------------------------------------------------------
## No   Variable      Stats / Values         Freqs (% of Valid)   Graph     Valid      Missing
## ---- ------------- ---------------------- -------------------- --------- ---------- ---------
## 1    generation    1. Boomers             4990 (17.9%)         III       27820      0
##      [character]   2. G.I. Generation     2744 ( 9.9%)         I         (100.0%)   (0.0%)
##                    3. Generation X        6408 (23.0%)         IIII
##                    4. Generation Z        1470 ( 5.3%)         I
##                    5. Millenials          5844 (21.0%)         IIII
##                    6. Silent              6364 (22.9%)         IIII
## ---------------------------------------------------------------------------------------------
```

Variable 12 - total_suicides_country Lastly, this variable is a self created variable to record the total number of suicides happened in a country in total as a continuous variable. It recorded a median of 9372 cases and a maximum of 126,352 cases.

```
dfSummary(dataset1$gdp_per_capita)
```

```
## dataset1$gdp_per_capita was converted to a data frame

## Data Frame Summary
## dataset1
## Dimensions: 27820 x 1
## Duplicates: 25587
##
## ---------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values                   Freqs (% of Valid)     Graph       Valid      Miss
## ---- ----------- -------------------------------- ---------------------- ----------- ---------- -----
## 1    dataset1    Mean (sd) : 16866.5 (18887.6)    2233 distinct values   :           27820      0
##      [integer]   min < med < max:                                        :           (100.0%)   (0.0%
##                  251 < 9372 < 126352                                    :
```

```
##                   IQR (CV) : 21427 (1.1)                                        : .
##                                                                                 : : : . .
## ---------------------------------------------------------------------------------------------
```

Diagnostic Analysis

Largest amount of suicides recorded by country in the dataset

The first result showed the largest amount of suicdes recorded was at 1,209,742 cases which happened in Russia. The next result showed the largest number of suicides happened in 1994 and consist of male boomers. This single data was orginated back in russia and the suspected responsible cause is the Russian constitutional crisis about the political stand-off thats was resolved by military force according to wikipedia.

```
# to return the country that has the largest number of suicides in total across all years
dataset1[which.max(dataset1$total_suicides_country),]
```

```
##                   country year  sex       age suicides_no population
## 20937 Russian Federation 1989 male 75+ years        1393    1349100
##       suicides.100k.pop HDI.for.year gdp_for_year gdp_per_capita....
## 20937            103.25           NA 506500173960               3740
##            generation total_suicides_country
## 20937 G.I. Generation                1209742
```

```
# to see the details of the largest number of suicides in the country
dataset1[which.max(dataset1$suicides_no),]
```

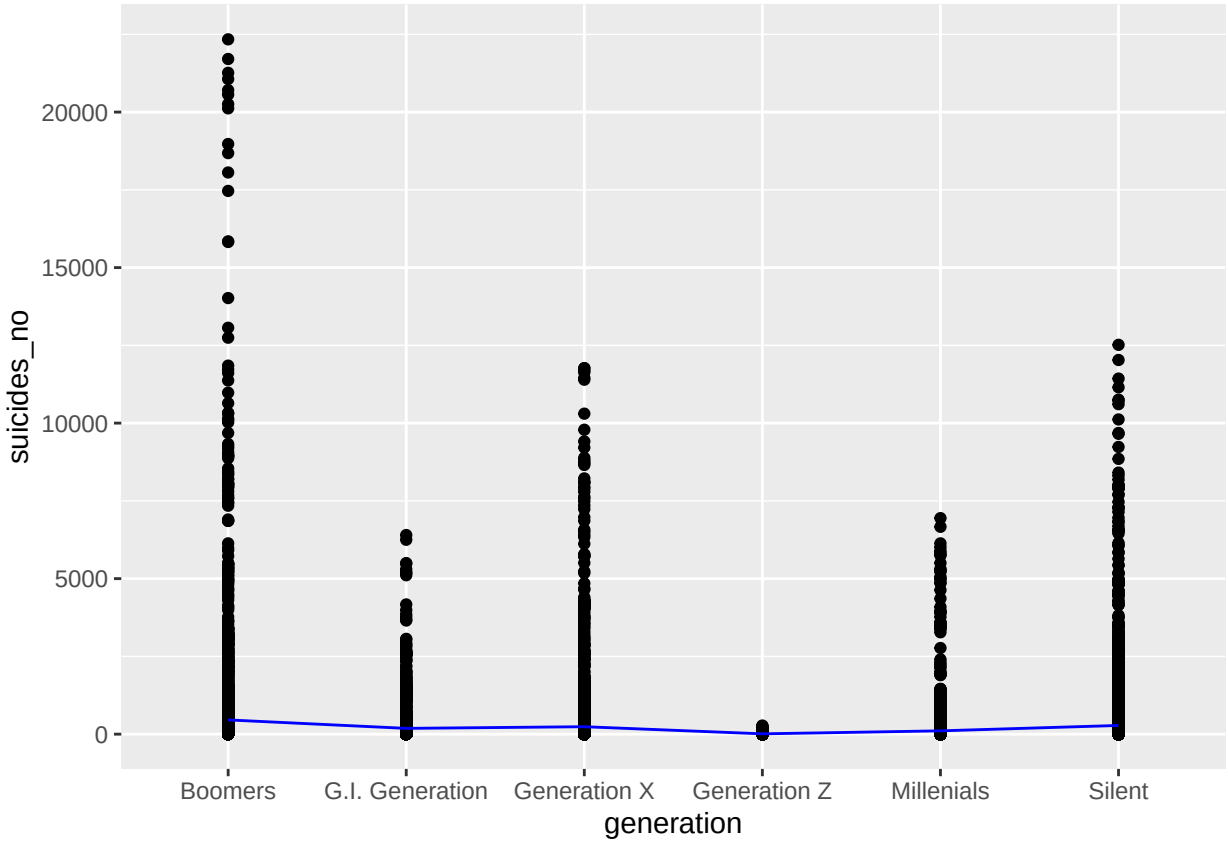
```
##                   country year  sex         age suicides_no population
## 20997 Russian Federation 1994 male 35-54 years       22338   19044200
##       suicides.100k.pop HDI.for.year gdp_for_year gdp_per_capita.... generation
## 20997             117.3           NA 395077301248               2853    Boomers
##       total_suicides_country
## 20997                1209742
```

Suicide numbers in different generations/age groups

The result of the scatter plot shows that "Boomers"(Age 35 to 54) had the highest suicide numbers and rates which at most reaching more than 20000. It is followed by the "Silent" generation(Age 55 to 74) and then Gen X(Age 25 to 34). G.I generation(Age 75 and above) was the second last and the least was Gen Z(Age 5 to 14). The blue line was the mean line indicating the mean number of suicides in generations. Based on the result shown, it is obvious to see the boomers having a very high numbers of suicides at an average of 457 cases. It could be due to the stress of work, life, family and the combination of it. With price inflation getting serious each generation, the responsibility to take good care of the family is constant but the pay from work might not provide sufficient support. The "Silent" generation followed next at a mean of 279 cases which could be due to the same reason as the boomers. Interestingly, both of this age group was in the right age to undergo midlife crisis which could be the reason.

```
# scatter plot to see suicides numbers in different generations
ggplot(dataset1) + aes(x=generation, y=suicides_no) + geom_point() + stat_summary(fun.y=mean, aes(group=
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



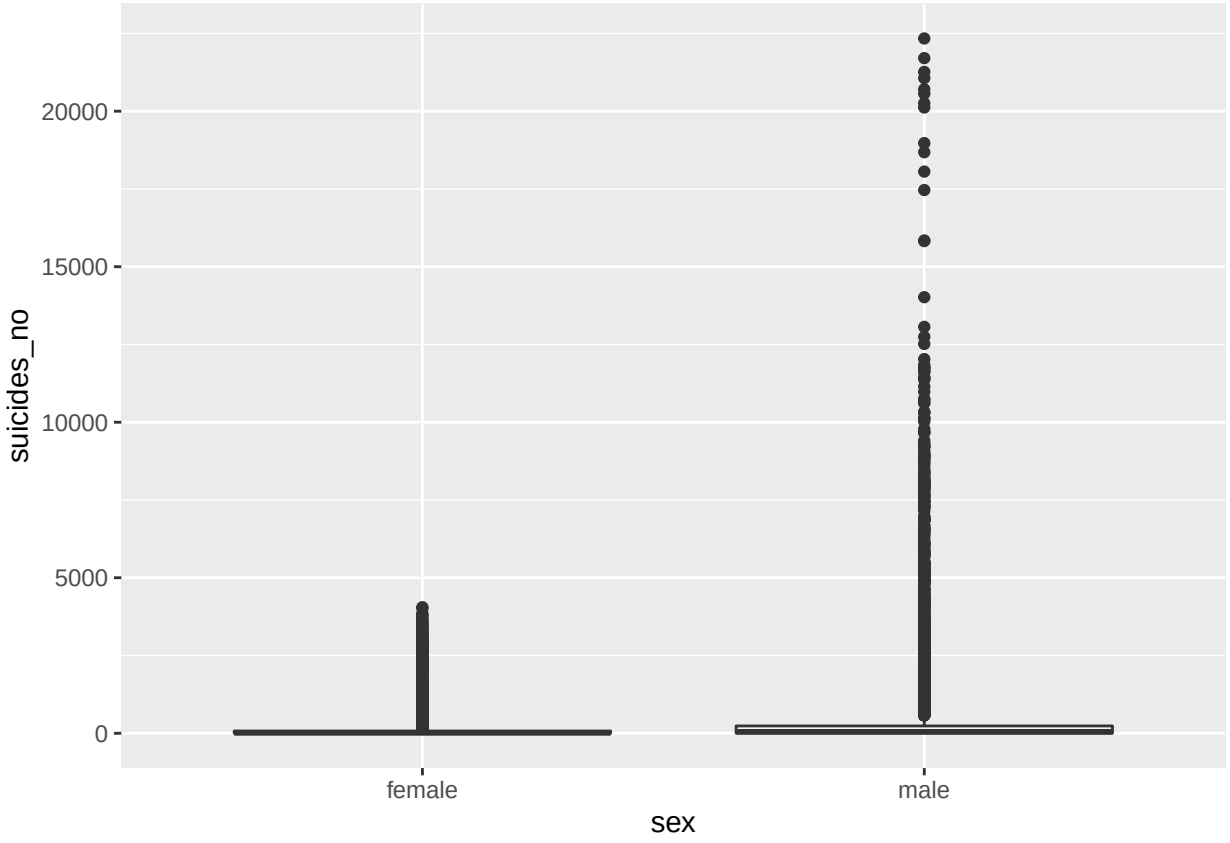

```
# mean suicide numbers in different generations
aggregate(x = dataset1$suicides_no,
          by = list(dataset1$generation),
          FUN = mean)
```

```
##           Group.1         x
## 1         Boomers 457.81523
## 2 G.I. Generation 185.86334
## 3    Generation X 239.20162
## 4    Generation Z  10.82041
## 5      Millenials 106.68361
## 6          Silent 279.97234
```

Suicide numbers in different sex

The result of the scatter plot shows that male has far higher amount of suicide numbers as compared to female. The highest amount of female suicides was shy of 5000 while the highest amount of suicides in male was at 25000. This amount is higly imbalance and showed that male appeared to be much weaker than what we thought in mind. The amountof stress and pressure they had to go through is huge and unimaginable. The result of descriptive statistics by sex showed the in average, females had a mean of 112 suicides and maximum of 4053. AT the same time, males had a mean of 373 and maximum of 22338. Mental health in male should be a focus and issue to be solved or not the problem will continue to worsen in the future whjich no one would like to see it happen.

```r
# box plot to see suicides numbers in different sex
ggplot(dataset1) + aes(x=sex, y=suicides_no) + geom_boxplot()
```

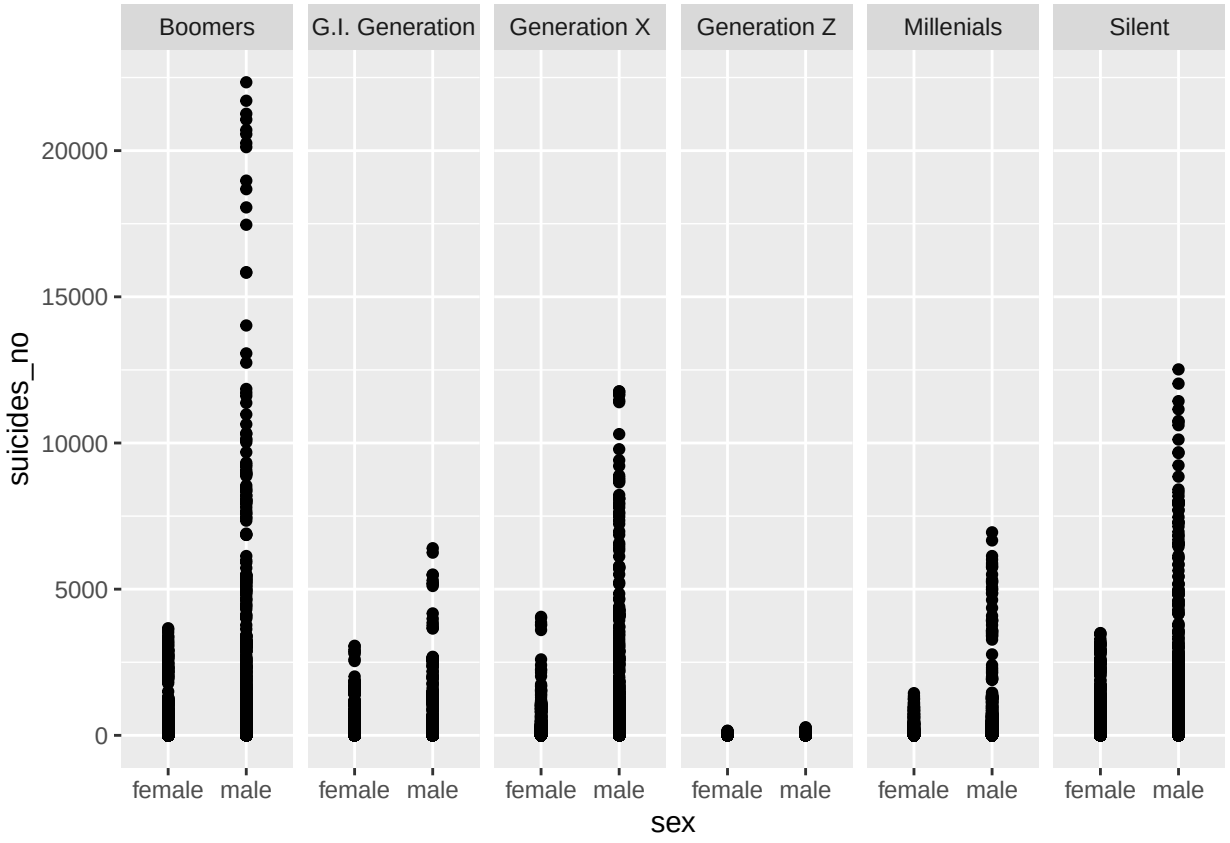



```r
# desxcriptive statistics of suicide numbers by sex
describeBy(dataset1$suicides_no, dataset1$sex)
```

```
##
##  Descriptive statistics by group
## group: female
##    vars     n   mean     sd median trimmed   mad min  max range skew kurtosis
## X1    1 13910 112.11 333.49     14   37.63 20.76   0 4053  4053 5.95    42.92
##      se
## X1 2.83
## ------------------------------------------------------------
## group: male
##    vars     n   mean      sd median trimmed   mad min   max range skew kurtosis
## X1    1 13910 373.03 1217.45     48  124.15 71.16   0 22338 22338    8    89.81
##       se
## X1 10.32
```

The next scatter plot is the combination of both above, which is the suicide numbers in sex grouped by generation. By combining the insights from above, the plot shows same result whereby male generally have a higher suicide numbers and boomers are the generation had more serious numbers. Thus, it is seen obviously that boomer males take up a very big part and numbers in suicide numbers worldwide.

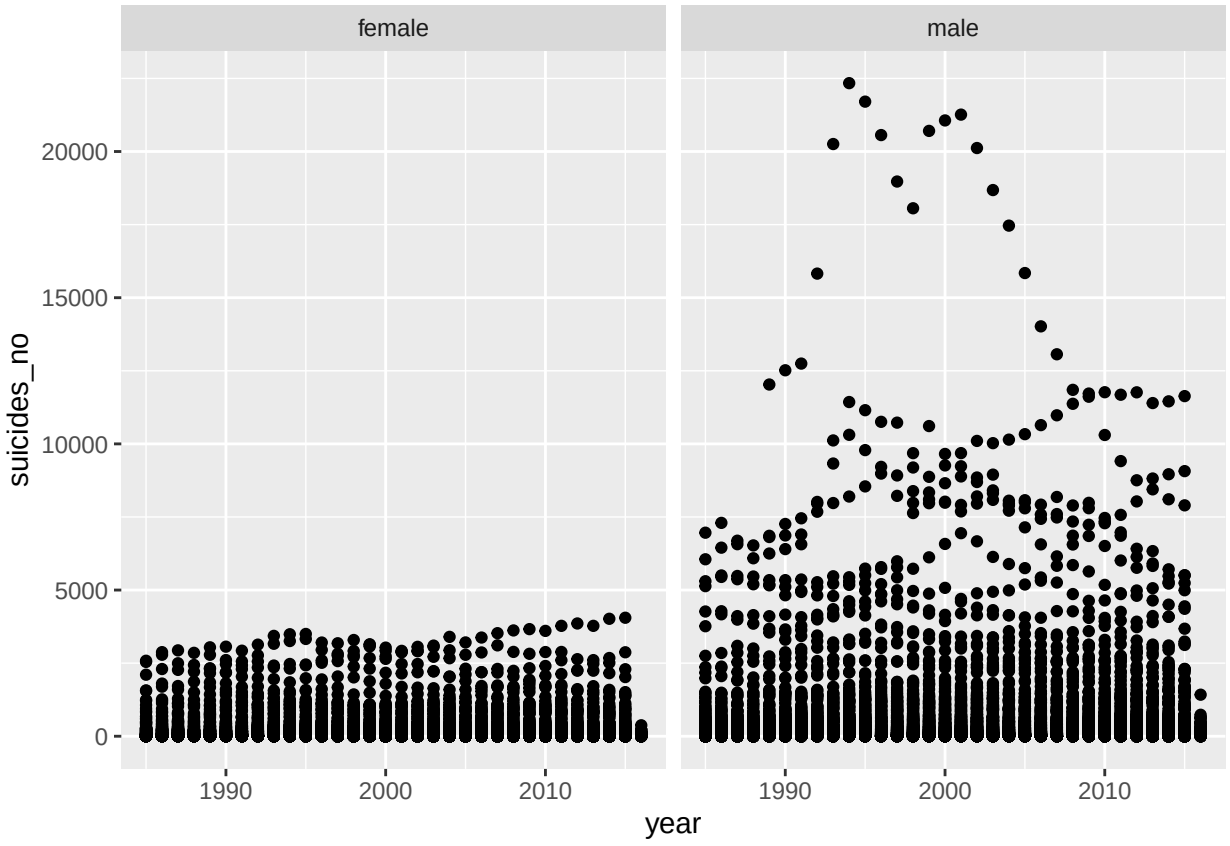
```
# scatter plot to see suicide numbers in sex grouped by generation
ggplot(dataset1) + aes(x=sex, y=suicides_no) + geom_point() + facet_grid(~ generation)
```



Suicide numbers trend in years

The following scatter plot shows the trend of suicide numbers by year grouped by gender. We can see an obvious upgrowing trend whereby the maximum numbers in each year is growing. Obiously without a doubt that the plot shows male has higher suicide numbers as compared to female adding the observations we got from above already. The number of suicide numbers saw a peak at 1994.

```
ggplot(dataset1) + aes(x=year, y=suicides_no) + geom_point() + facet_grid(~ sex)
```



Predictive Analysis

Multiple linear regression model A multiple linear regression model was built with the formula of $Y = B_0 + b_1 Population + b_2 GdpForYear + b_3 GdpPerCapita$ which I took suicides_no as response variable, population, gdp_per_year and gdp_per_capita as the predictor variable. The result showed that all of the predictor variables are significant in predicting the suicide numbers. The model also recorded an R^2 value of 0.3914 which means the model fits in 39% well. It suggests that to increase the suicide numbers, the more the population is and the the lower the gdp is the country in that year. That sounds similar to those country where their GDP growth is low but population is constantly high.

```
set.seed(100) #generates random numbers to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(dataset1), 0.8*nrow(dataset1))  # row indices for training data
trainingData <- dataset1[trainingRowIndex, ]  # model training data
testData  <- dataset1[-trainingRowIndex, ]   # test data

lmMod <- lm(suicides_no ~ population + gdp_for_year + gdp_per_capita.... , data=trainingData)
distPred <- predict(lmMod, testData)

summary(lmMod)
```

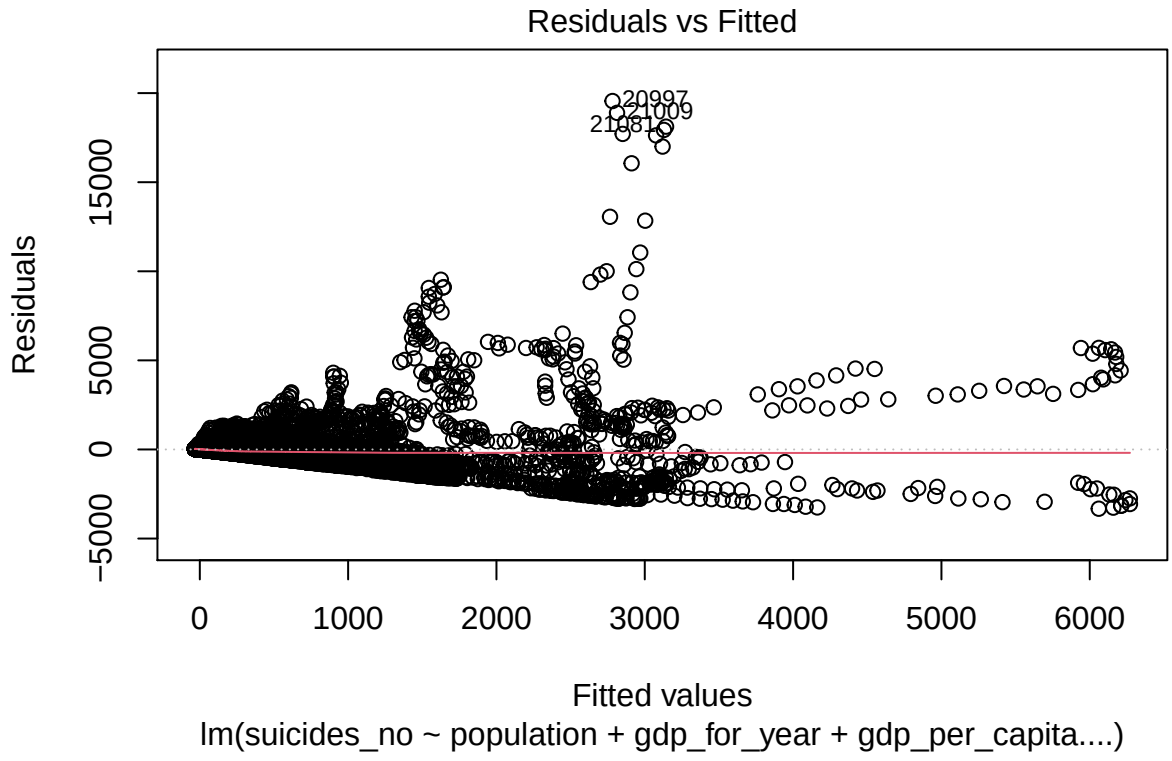
```
##
## Call:
## lm(formula = suicides_no ~ population + gdp_for_year + gdp_per_capita....,
##     data = trainingData)
```

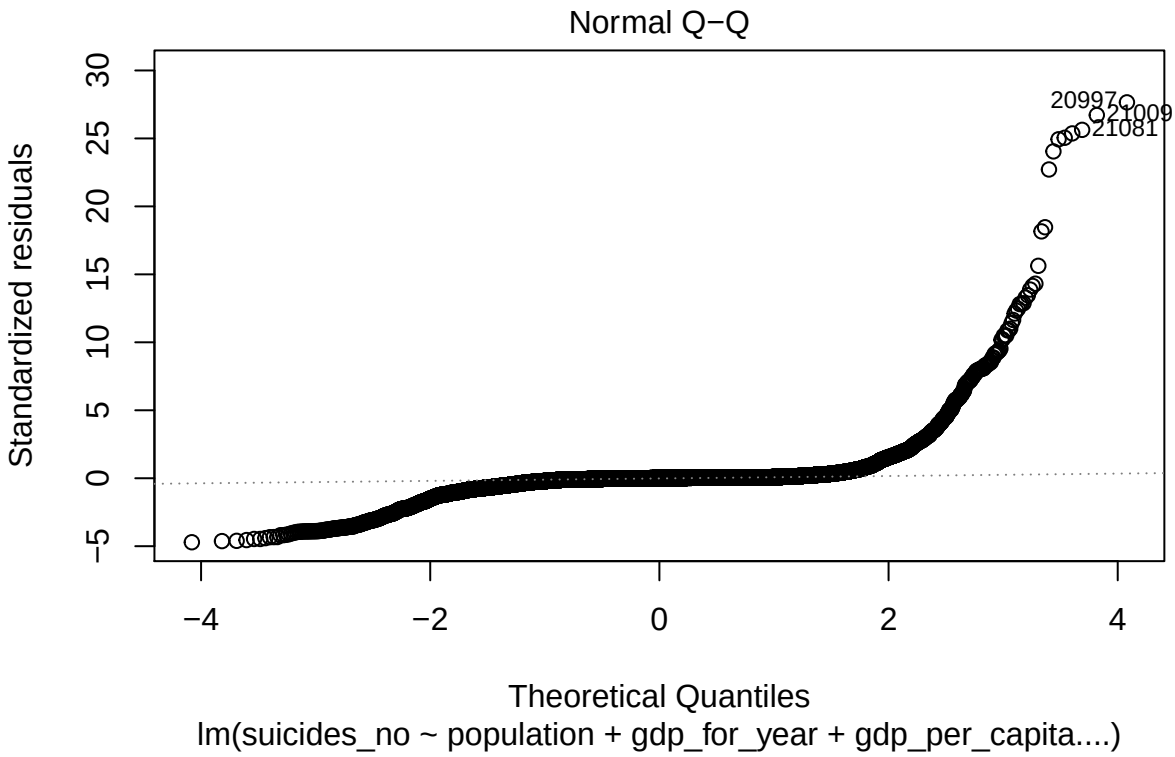
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3324.8   -53.7     8.2    31.5 19555.5
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.413e+01  6.851e+00  -4.982 6.35e-07 ***
## population          1.481e-04  1.779e-06  83.246  < 2e-16 ***
## gdp_for_year       -1.306e-11  5.046e-12  -2.589  0.00963 **
## gdp_per_capita....  7.181e-04  2.686e-04   2.674  0.00751 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 707.8 on 22252 degrees of freedom
## Multiple R-squared:  0.3914, Adjusted R-squared:  0.3913
## F-statistic:  4770 on 3 and 22252 DF,  p-value: < 2.2e-16
```

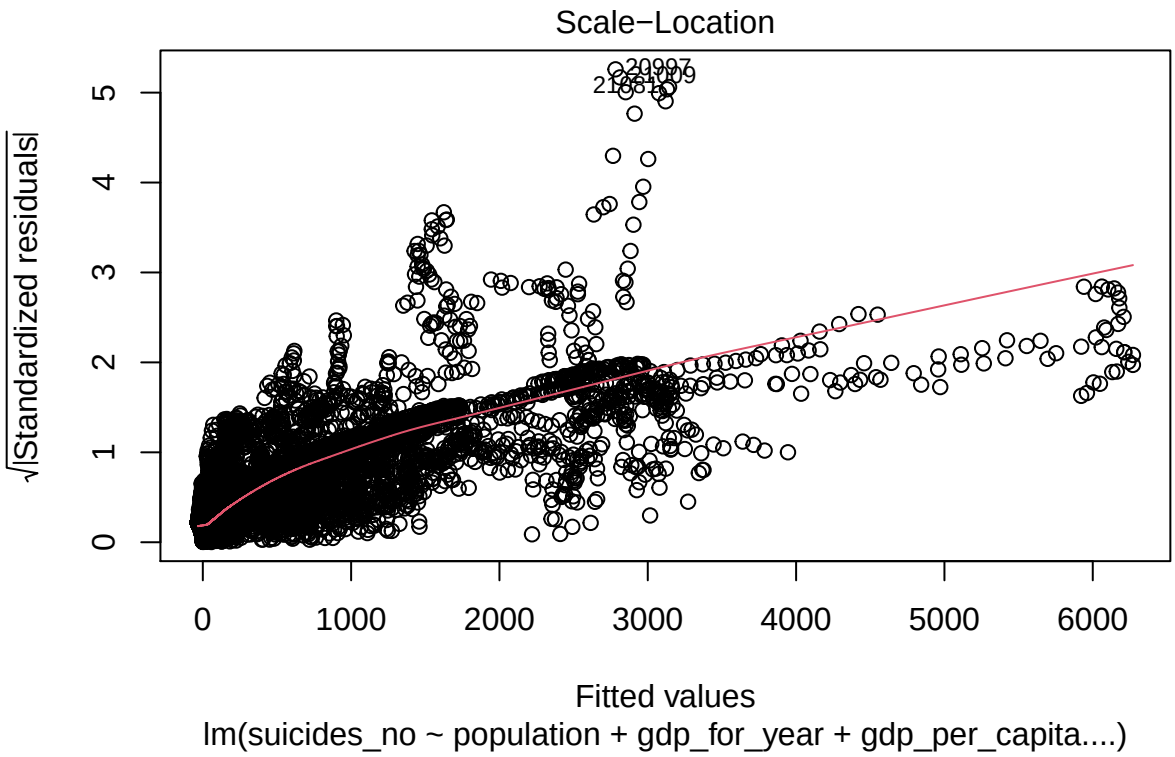
```
summary(distPred)
```

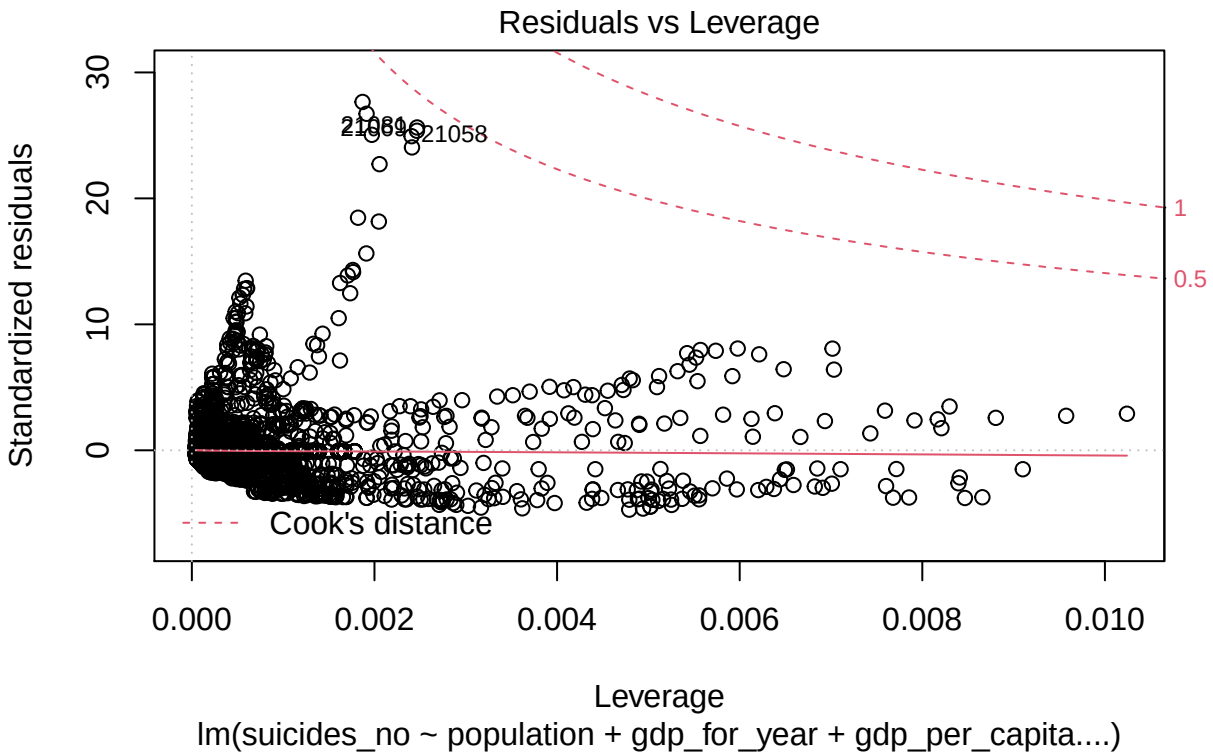
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  -33.537   -5.855   47.838  252.979  221.427 6306.528
```

```
plot(lmMod)
```




Normal Q−Q
Standardized residuals
−5
0
5
10
15
20
25
30
20997
21009
21081
−4
−2
0
2
4
Theoretical Quantiles
lm(suicides_no ~ population + gdp_for_year + gdp_per_capita....)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
5
0
1000
2000
3000
4000
5000
6000
Fitted values
lm(suicides_no ~ population + gdp_for_year + gdp_per_capita....)



1. The first plot of residuals vs fitted plot shows heteroscedasticity whereby the plots seems to be heavily skewed to the left. This means that the data we had in this dataset have a very large range between each other. According to the assumption of linear regression, heteroscedasticity cannot be accepted and thus we cannot put a high trust on the model and result.

2. The second plot shows a heavy-tailed Q-Q plot.This proves that we have extreme outlier values in our model. There is much more data located at the extremes of the distributions.

3. The third plot of scale-location also suggest similar results as the first plot which heteroscedasticity was found in the plot.

4. The fourth plot which is a residuals vs leverage detects heteroskedasticity similar plot 1 and plot 3. However, there are no influential points outside the dotted line(cooks'distance) and in this plot, there are none thus suggest that all the values here are respectable and usable.

ANCOVA test

An Ancova test was performed to compare all the variables and check that are ther any significant variables in affecting suicides number.

```
ancMod <- dataset1 %>% anova_test(suicides_no ~ country + sex + population + gdp_for_year + generation)
```

```
## Coefficient covariances computed by hccm()
```

```r
get_anova_table(ancMod)
```

```
## ANOVA Table (type II tests)
##
##         Effect DFn   DFd        F         p p<.05      ges
## 1      country 100 27711   82.214  0.00e+00     * 0.229000
## 2          sex   1 27711 1434.488 4.30e-306     * 0.049000
## 3   population   1 27711 6239.328  0.00e+00     * 0.184000
## 4 gdp_for_year   1 27711    3.383  6.60e-02       0.000122
## 5   generation   5 27711  172.892 9.22e-182     * 0.030000
```

Based on the result, I decided to remove gdp_for_year since it is close to the significant level of 0.05 even though it is just show of it. At the same time, country and population was also removed since the p-value does not seem to bring any meaning at 0. Then, I noticed that the remaining variables are sex and generation which both of them are categorical variable so another anova test will generate a more accurate result.

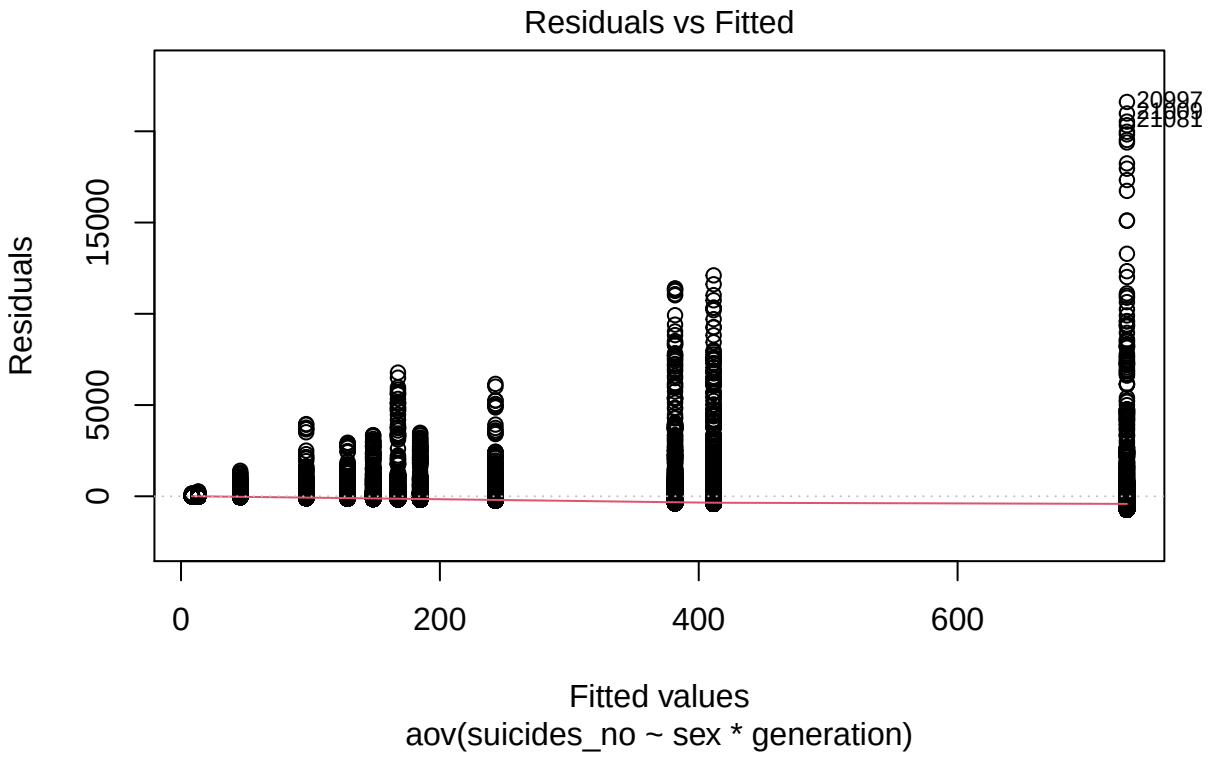
ANOVA test

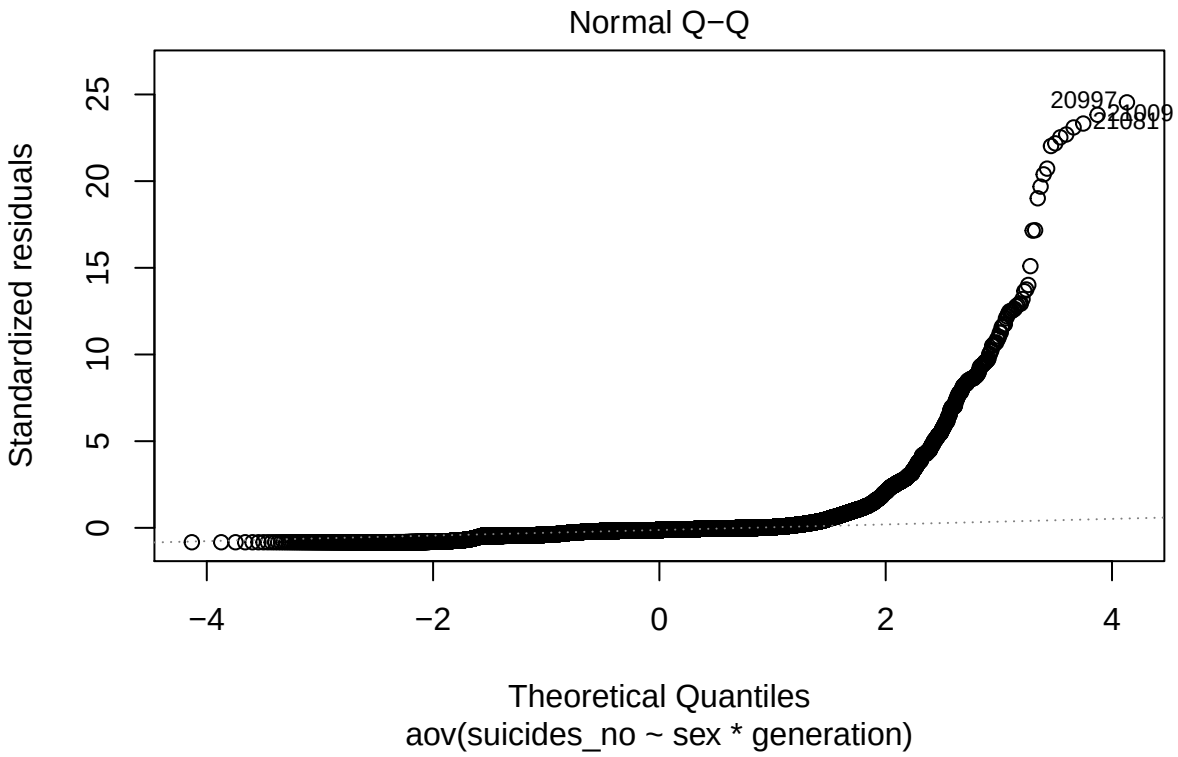
An Anova test was then performed to check for the variables against suicide numbers.

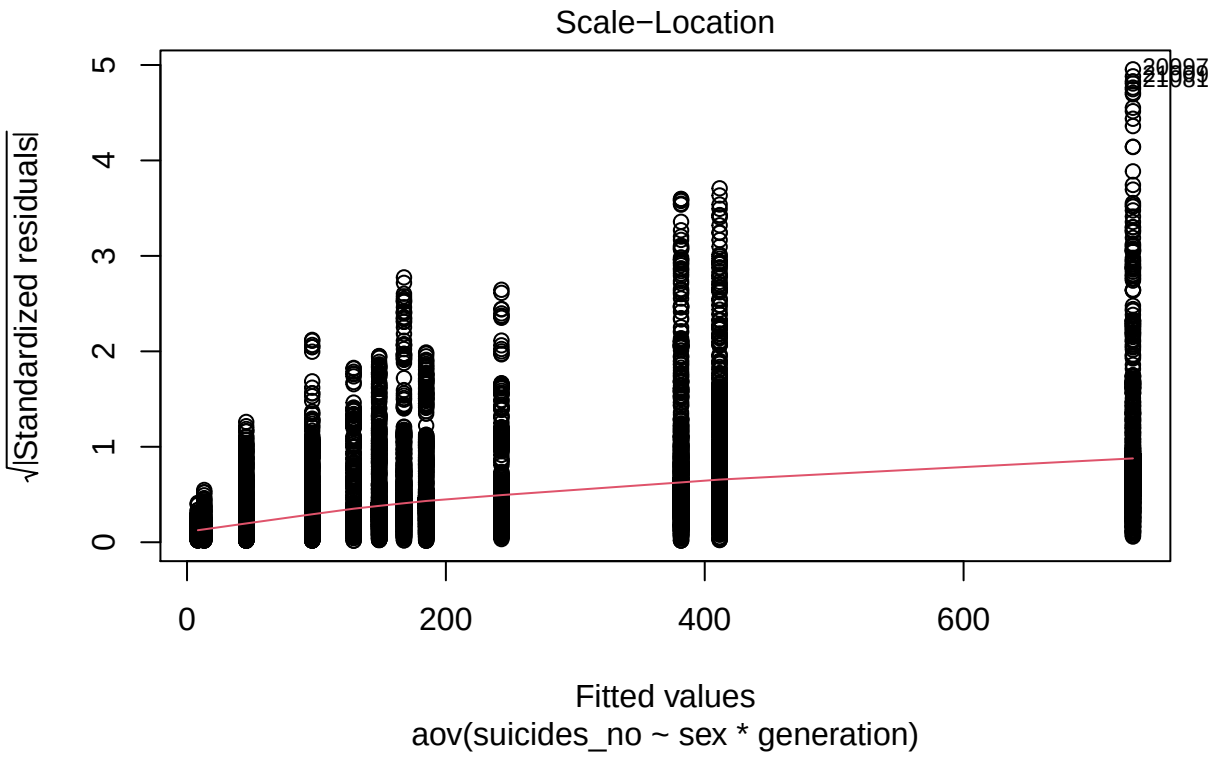
```r
anvMod <- aov(suicides_no ~ sex * generation, data=dataset1)
summary(anvMod)
```

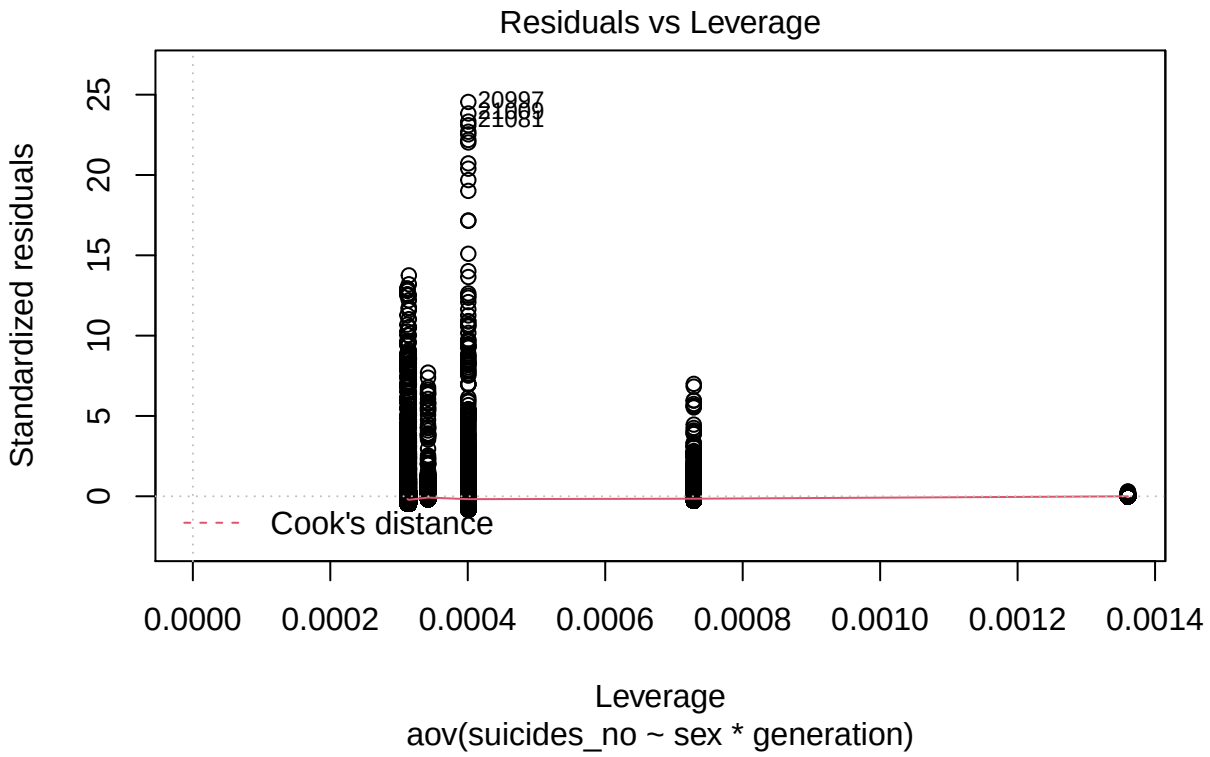
```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## sex                1 4.735e+08 473491889  610.79 <2e-16 ***
## generation         5 4.358e+08  87169838  112.45 <2e-16 ***
## sex:generation     5 1.695e+08  33896716   43.73 <2e-16 ***
## Residuals      27808 2.156e+10    775217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(anvMod)
```


Residuals vs Fitted
Residuals
0
5000
15000
0
200
400
600
Fitted values
aov(suicides_no ~ sex * generation)


Normal Q−Q
Standardized residuals
0
5
10
15
20
25
−4
−2
0
2
4
Theoretical Quantiles
aov(suicides_no ~ sex * generation)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
5
0
200
400
600
Fitted values
aov(suicides_no ~ sex * generation)




```
levene_test(suicides_no ~ sex * generation, data=dataset1)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic         p
##   <int> <int>     <dbl>     <dbl>
## 1    11 27808      117. 9.01e-263
```

The result showed that both variable "sex" and "generation" are significant based on the f-value which are less than 0.05. However, the levene's test showed a p-value that is lesser than 0.05 which we cannot assume homogeinity of residual variances for all groups.

1. The first plot showed a residuals vs fitted plot showed that point 20997, and 21081 are detected as outliers that can affect normality and homogeneity of varience.
2. The second plot of normal Q-Q plot showed a heavily skewed tail and does not follow along the reference line. This is a prove that normality is violated.

Based on all the explanations above, the result suggests that this is not a good model since it violates all the assumptions for a model. Thus, we can only take reference and not trsut fully and take action based on the results.

Conclusion

The dataset obtained from Kaggle was big and sufficient to discover many insights related about suicides. It is obvious that the data are highly skewed due to the appearance of extreme values in the dataset. However,

that was not a very big thing as there is no perfect data available. What we learnt was that most of the suicide cases happened among males in the Boomer generation which is in the age group of 35 to 54. And worryingly, the plots showed that the number of cases are growing year by year. Thus, in this era where we emphasize of gender equality and woman power, men do also need attention to help and ease their burden off. If not, the figures will just grow and getting more serious. The predictive model built showed that population, GDP of the country do affect the numbers. However, the model seemed to have heavily skewed data and we should not highly trust in the model and more research and data are required to obtain a more accurate and relevant prediction.