ISSS616 Applied Statistics with R – G3

Global Suicide Analysis with R

Prepared by G3-Group 1:

Xia Qingquan

Sun Yuancheng

Suo Tianwei

Zhao Chenyang

Xiao Chuan

Liu Cuiyi

Li Yanning

21 Nov 2019

Section 1: Introduction

Suicide has long been regarded as a serious public safety problem. According to the report of World Health Organization (WHO), about 800,000 people die of suicide every year, and the number of attempted suicides is much higher. The report also points out that the suicide phenomenon has a youth-oriented tendency that it has become the second leading cause of death after traffic accidents among teenagers. Faced with this trend, WHO has appealed to countries to take notice and carry out measures to prevent suicides.

According to the article reported by the Economist in 2018, global suicide rates have fallen by 29% since 2000, thanks to urbanization and government policies. The United States, however, is the exception with the figure rising by 18%. The differences in suicide rates among countries may owe to diverse demographic characteristics and economic factors. One of the tasks in the project is to visualize the differences.

Although the suicide rate shows different tendencies among countries, as we concerned, it always shows similar patterns in all the countries. We believe that in most countries of the world, older people are more likely to commit suicides than youngsters, and men are more likely to commit suicide than women. What we need to do is to verify and confirm the pattern by doing statistical analysis.

In a word, it is helpful to understand the changes in suicide mortality rates among countries and generations, so effective suicide prevention strategies can be developed for vulnerable populations.

In this project, we use the suicides dataset provided by Kaggle to analyse the pattern and trends of suicides in different countries. The dataset contains global, regional and 195-country suicide mortality patterns from 1985 to 2016 in the form of age, gender, demographic and economic indicators, etc. We will analyse the global suicide rates throughout the period and discover the differences among countries and continents. The project aims to explore the relationship between suicide rates and age groups, gender, countries, continents, and GDP.

We mainly apply R programming to derive statistics and plot the relationships between factors. Then we use Shiny to interactively show the results. We want to draw inspiration from the analysis of suicide rates to make the public pay attention to personal mental health and take action to reduce the occurrence of suicides.

Section 2: Overall Concept

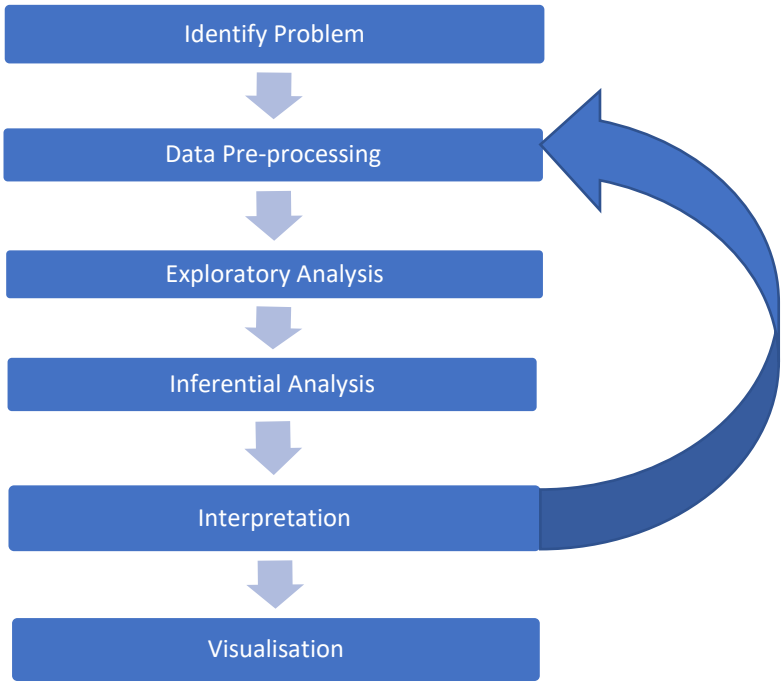
Our study aims to explore the suicides pattern from the year 1985 to the year 2016 through global, continent and country scales, as well as gender, age, GDP and generation factors. The descriptive exploration obtained will then be verified through statistical inferential analysis. We set the target audience as World Health Organization officer who are concerned to track the trend and factors associated with suicides, as we wish the analysis will provide insights and inspire necessary actions taken on suicides prevention.

Problems we want to identify may include and not limited to:

- What is the geographical distribution for suicides if we search by continents? What if by countries?

- What is the suicides trend according to year?
- Are men far more prone to committing suicide than women?
- Are young people or old people more likely to commit suicide?

These problems are brought up and we will proceed with the analysis following the model below.



Section 3: Data Sources

The dataset we use is Suicide Rates Overview 1985 to 2016 dataset from Kaggle[1].

The dataset has 12 attributes, being:

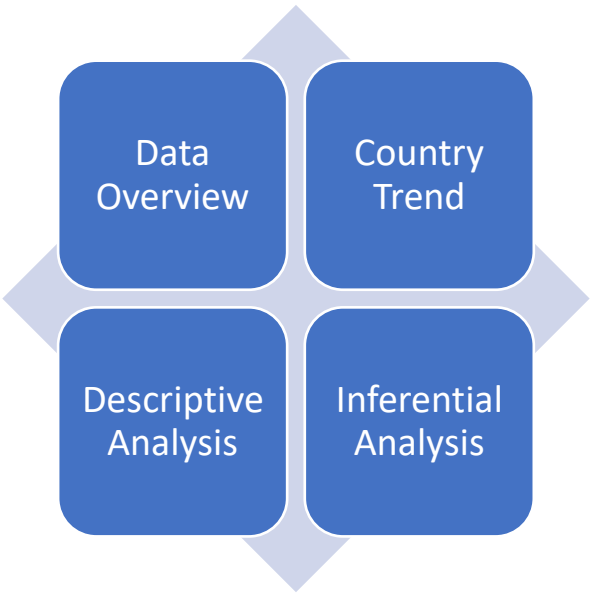
Variables	Descriptions
Country	The personal information when the individual committed suicide.
Year	Year of suicides.
Sex	Female or Male indicator.
Age	All the ages are divided into six groups.

[1] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Suicides_no	The sum of people who committed suicide in a specific age group.
Population	The population of the country.
Suicides/100k pop	The suicide numbers per 100k population.
Country-year	The join of Column “Country” and “Year”, no other information.
HDI for year	Nearly 20k rows are missing out of 28k rows.
GDP for year	The Gross Domestic Product of a country.
GDP per capita	The Gross Domestic Product per capita of a country.
Generation	People born in different eras are categorized into six groups, namely, "Generation X", "Silent", "G.I.Generation", "Boomers","Millennials","Generation Z”.

The variables we use are: Country, Year, Sex, Age, Suicides no, population, GDP Per capita, and generation. We recode country as some of the country names are wrong or not standardized and use it to add a new column named Continent based on country. We reclassify countries that have been coded as 'Americas' into 'North America' and 'South America'.

We developed an interactive dashboard for demo purpose, providing a user-friendly window that the user can filter data as he/she wants. Our shiny app consists of four modules:



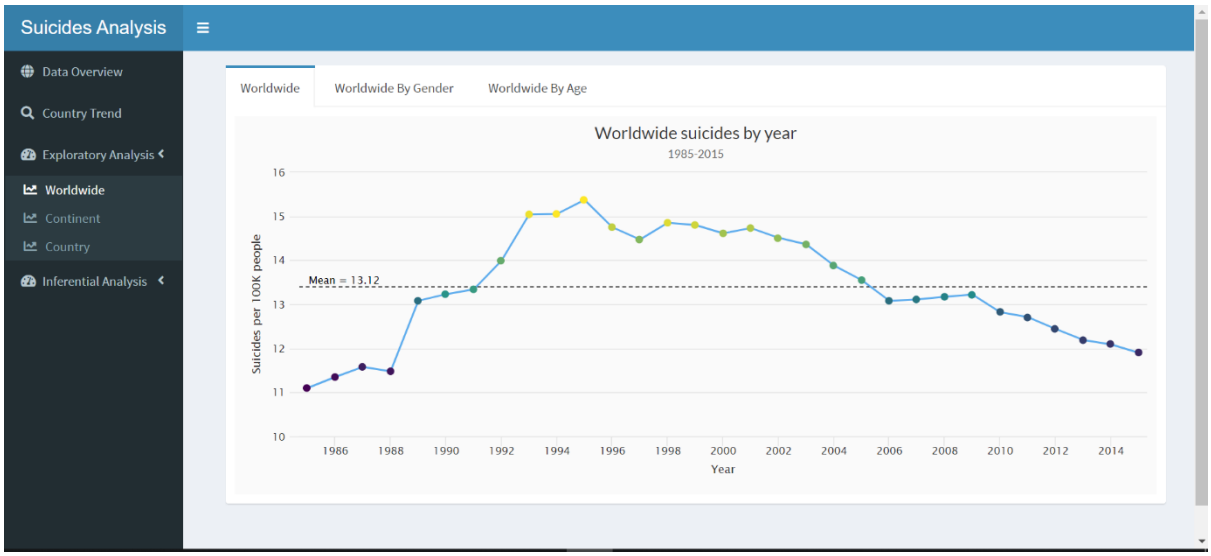
Data Overview – Allow the user to filter out suicides socio-economic data table based on country, year and age. The data extracted can be examined by search function and can be sorted by different variables.

Country Trend – Allow the user to visualize suicides trend by a specific country.

Descriptive Analysis – Interactive dashboard with bar chart, pie chart and line graph are included in the demo to how suicide rate is related to each factor.

Inferential Analysis – Confidence Interval and Linear Regression techniques are applied.

An example of the app may look like the picture below:



Section 4: Specific Methodology

Methodology

Employed R packages

dplyr: A grammar of data manipulation. It's useful for more efficient data cleansing, data analysis.

tidyverse: It's useful for data analysis processing and visualizing.

ggalt: A compendium of new geometries, coordinate systems, statistical transformations, scales and fonts for ggplot2.

countrycode: Standardize country names, convert them into one of eleven coding schemes, convert between coding schemes, and assign region descriptors.

rworldmap: Enables mapping of country level and gridded user datasets.

gridExtra: Provides a number of user-level functions to work with grid graphics, notably to arrange multiple grid-based plots on a page and draw tables.

broom: Summarizes key information about statistical objects in tidy tibbles.

readxl: Import excel files into R.

DT: Data objects in R can be rendered as HTML tables using the JavaScript library 'DataTables' (typically via R Shiny).

Highcharter: Shortcut functions to plot R objects and offer numerous interactive chart types with a simple configuration syntax.

Viridis: Set better color maps and browsers.

ggplot2: A great and popular graphic creating module.

Shiny & shinydashboard : Make users to build interactive web applications easily with R. Automatic "reactive" binding between inputs and outputs and extensive prebuilt widgets make it possible to build beautiful, responsive, and powerful applications with minimal effort.

Data Preparation

After importing the data to RStudio, firstly we amended names of columns and some countries to a standardized format.

```
# Standardize names of columns
colnames(data) = c("country","year","sex","age","suicides_no","population",
                   "country_year","gdp_for_year","gdp_per_capita","generation")
```

Besides, we removed rows for year 2016 because of the shortage of valid data. And countries “Dominica” and “Saint Kitts and Nevis” with too much missing values also have been removed.

```
# Filter out 2016 and countries with 0 data.
data <- data %>%
  filter(year != 2016,
         country != 'Dominica',
         country != 'Saint Kitts and Nevis')
```

Meanwhile, column “HDI.for.year” has been also excluded due to the extremely large amount of missing values existed.

```
# Remove unvalid variables
data <- data %>%
  select(-c(`HDI.for.year`, `suicides.100k.pop`)) %>%
  as.data.frame()
```

We employed “countrycode” package in RStudio to match various countries with the continent that they belong to. As a result, a new column named “continent” has been established.

```
# Fix the names of some of the countries in our data to match the country names
# used by our map later on so that they'll be interpreted and displayed.
data <- data %>%
  mutate(country = fct_recode(country, "The Bahamas" = "Bahamas"),
         country = fct_recode(country, "Cape Verde" = "Cabo Verde"),
         country = fct_recode(country, "South Korea" = "Republic of Korea"),
         country = fct_recode(country, "Russia" = "Russian Federation"),
         country = fct_recode(country, "Republic of Serbia" = "Serbia"),
         country = fct_recode(country, "United States of America" = "United States"))

# Create new column in our data for continent. Use countrycode() to extract continents from country names.
data$continent <- countrycode(sourcevar = data$country,
                              origin = "country.name",
                              destination = "continent")

# Reclassify countries that have been coded as 'Americas', by countrycode(), into 'North America' and 'South America'.
south_america <- c('Argentina', 'Brazil', 'Chile', 'Colombia', 'Ecuador', 'Guyana', 'Paraguay', 'Suriname', 'Uruguay')
data$continent[data$country %in% south_america] <- 'South America'
data$continent[data$continent=='Americas'] <- 'North America'
```

GDP per capital has been graded and classified into five groups, generating a new column called “gdp_per_capita_grade”.

```
# Group data by graded GDP
y <- quantile(data$gdp_per_capita, c(0.8, 0.6, 0.4, 0.2))
data$gdp_per_capita_grade[data$gdp_per_capita >= y[1]] <- "very high"
data$gdp_per_capita_grade[data$gdp_per_capita < y[1]] <- "high"
data$gdp_per_capita_grade[data$gdp_per_capita < y[2]] <- "moderate"
data$gdp_per_capita_grade[data$gdp_per_capita < y[3]] <- "low"
data$gdp_per_capita_grade[data$gdp_per_capita < y[4]] <- "very low"
```

Lastly, we designed and customized a theme including color, font size and layout for the data visualization.

```
# Create a custom theme for the plots.
custom_theme <- hc_theme(
  colors = c('#5CACEE', 'green', 'red'),
  chart = list(
    backgroundColor = '#FAFAFA',
    plotBorderColor = "black"),
  xAxis = list(
    gridLineColor = "E5E5E5",
    labels = list(style = list(color = "#333333")),
    lineColor = "#E5E5E5",
    minorGridLineColor = "#E5E5E5",
    tickColor = "#E5E5E5",
    title = list(style = list(color = "#333333"))),
  yAxis = list(
    gridLineColor = "#E5E5E5",
    labels = list(style = list(color = "#333333")),
    lineColor = "#E5E5E5",
    minorGridLineColor = "#E5E5E5",
    tickColor = "#E5E5E5",
    tickWidth = 1,
    title = list(style = list(color = "#333333"))),
  title = list(style = list(color = '#333333', fontFamily = "Lato")),
  subtitle = list(style = list(color = '#666666', fontFamily = "Lato")),
  legend = list(
    itemStyle = list(color = "#333333"),
    itemHoverStyle = list(color = "#FFF"),
    itemHiddenStyle = list(color = "#606063")),
  credits = list(style = list(color = "#666")),
  itemHoverStyle = list(color = 'gray'))
```

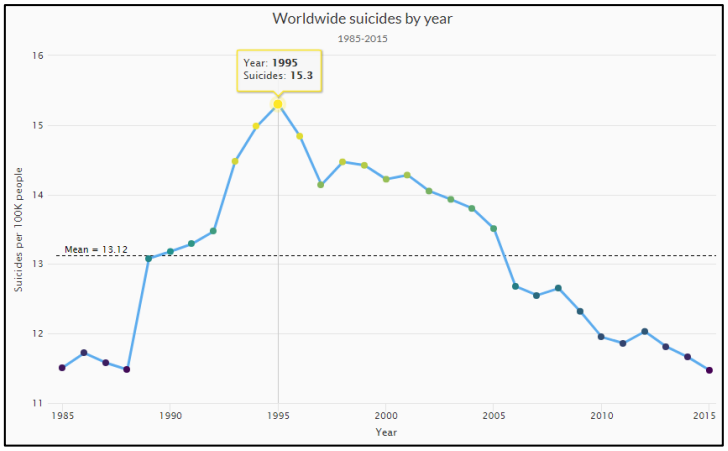
Descriptive Analysis

In general, we intend to analyze data from three main scales: worldwide, continent and country.

Worldwide

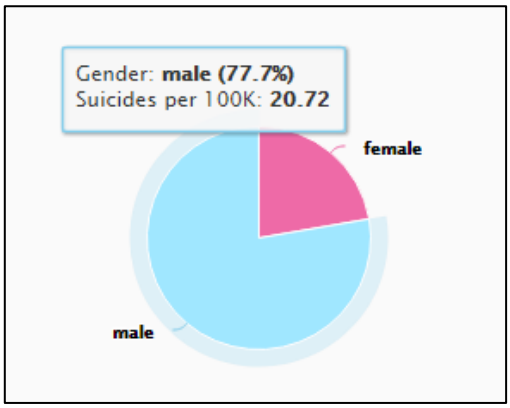
Firstly, we grouped data by years, gender and age and drawn interactive graphs by using “Highcharter” package to display the overall trend that how the number of suicides would change with time. As figure shows below, it reached a peak in 1995 that there were about 15 people out of every 100k to choose ending their lives.

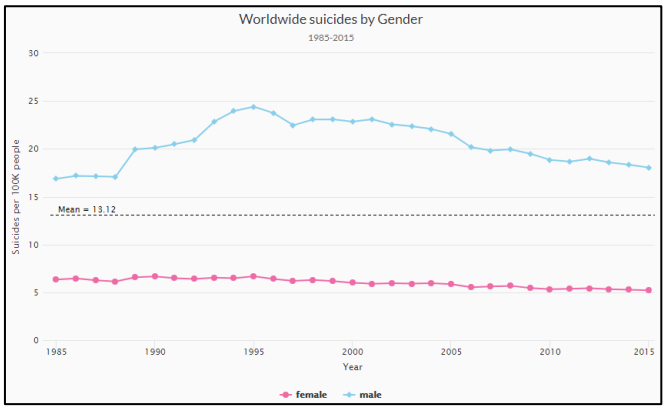
```
# Create tibble for our line plot.
overall_tibble <- data %>%
  select(year, suicides_no, population) %>%
  group_by(year) %>%
  summarise(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2))
```



Besides, we can notice that the suicides number of males had always outnumbered that of females. It accounted for about 77.7% of the total suicides number from 1985 to 2015 and increased sharply since 1988.

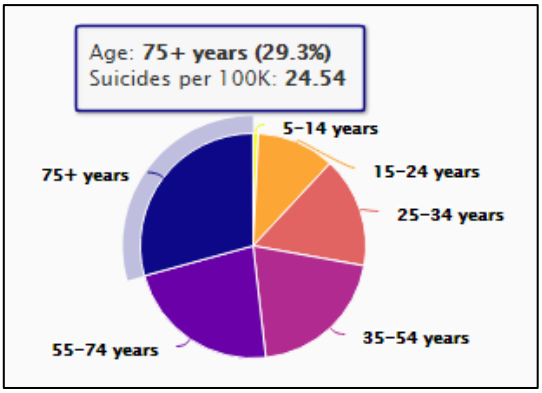
```
# Create tibble for sex and year.
sex_tibble <- data %>%
  select(year, sex, suicides_no, population) %>%
  group_by(year, sex) %>%
  summarise(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2))
```

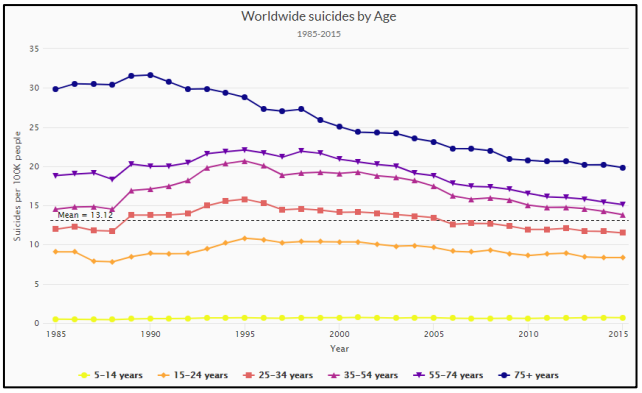




It is extremely necessary to mention that people above age 75 had stronger intention to commit suicide than others. They nearly made up ¼ of the total suicides. But fortunately, it seems like the suicides number of them had experienced a smooth decrease from 1900 to 2015.

```
# Create tibble for age and year.
age_tibble <- data %>%
  select(year, age, suicides_no, population) %>%
  group_by(year, age) %>%
  summarise(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2))
```





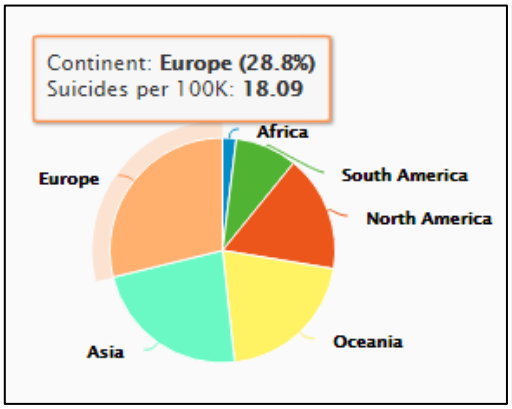
Continent

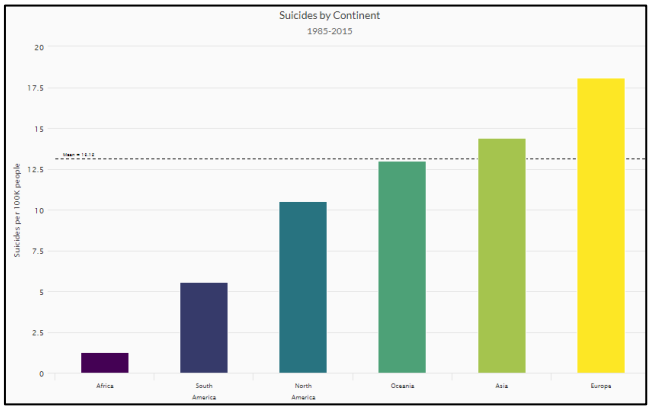
In this section, we are still trying to figure out that how the number of suicides would change with time for different gender and age but based on the continental classification.

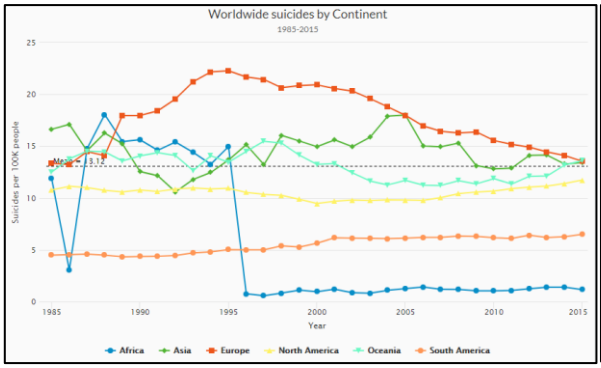
```
# Create a tibble for continent.
continent_tibble <- data %>%
  select(continent, sex, suicides_no, population) %>%
  group_by(continent) %>%
  summarize(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2)) %>%
  arrange(suicide_capita)
```

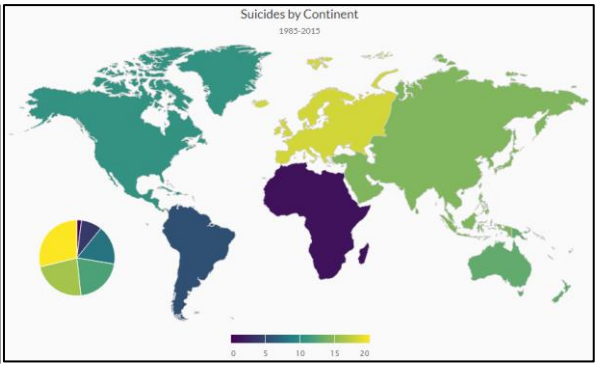
Generally, people in Europe are more prone to commit suicide. As graphs indicate below, the average number of suicides in Europe from 1985 to 2015, about 18.09 per 100k population which accounted for 28.8%, ranked number one around the world. On the other hand, we can observe that the suicides number in Africa experienced a sharp rise from 1986 and significantly drop from 1995. It will be very beneficial for the prevention actions' formulation and application if factors that caused such shifts can be identified.

```
# Create a tibble for continent and year.
continent_year_tibble <- data %>%
  select(continent, year, suicides_no, population) %>%
  group_by(continent, year) %>%
  summarize(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2))
```



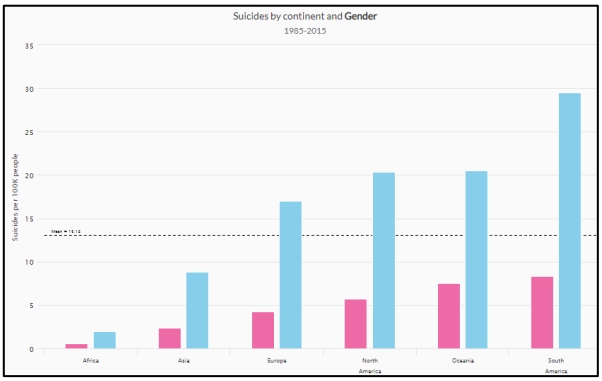


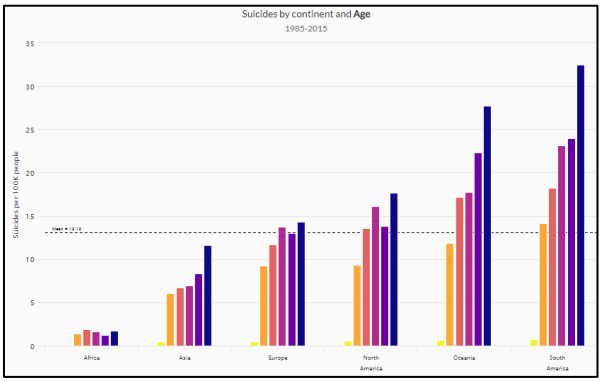




As we expected, the proportion of suicide number for gender and age holds still in different continent - males and elders were much more inclined to commit suicide.

```
# Create a tibble for continent and sex.
continent_sex_tibble <- data %>%
  select(continent, sex, suicides_no, population) %>%
  group_by(continent, sex) %>%
  summarize(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2)) %>%
  arrange(suicide_capita)

# Create a tibble for continent and sex.
continent_age_tibble <- data %>%
  select(continent, age, suicides_no, population) %>%
  group_by(continent, age) %>%
  summarize(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2)) %>%
  arrange(suicide_capita)
```

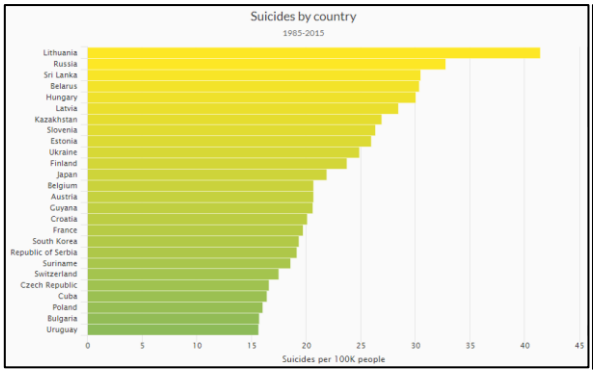


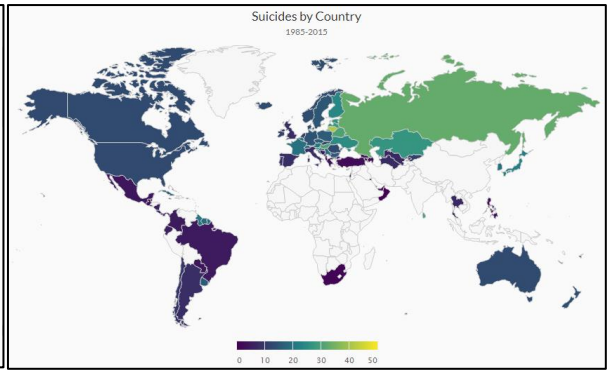


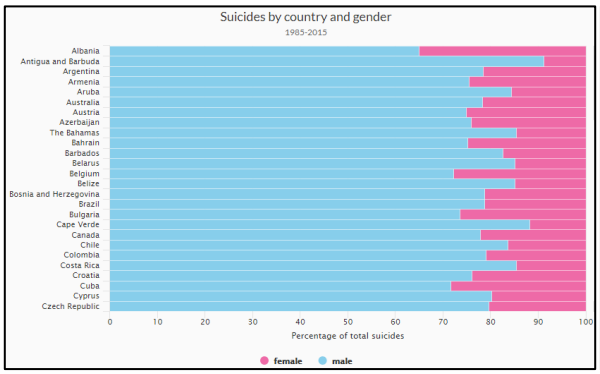
Country

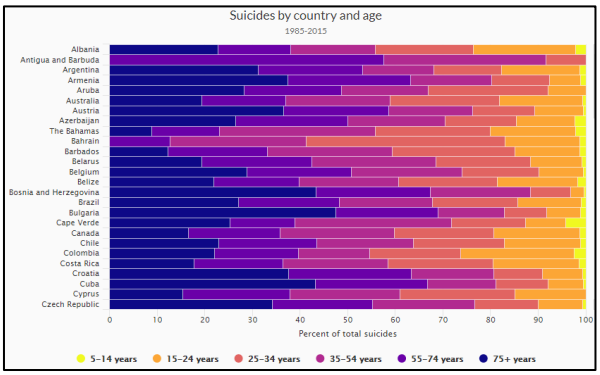
Lastly, we are going to discover the number of suicides in each country. Based on the figures, we can notice that Lithuania, Russia and Sri Lanka ranked top three for having the highest average number of suicides. And it can also be observed directly through the map below. Blank regions are caused by the shortage of data for certain countries.

```
# Create tibble for overall suicides by country.
country_bar <- data %>%
  select(country, suicides_no, population) %>%
  group_by(country) %>%
  summarise(suicide_capita = round((sum(suicides_no)/sum(population))*100000, 2)) %>%
  arrange(desc(suicide_capita))
```







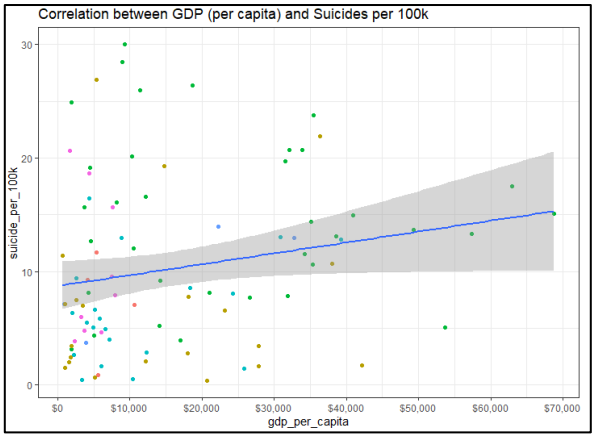


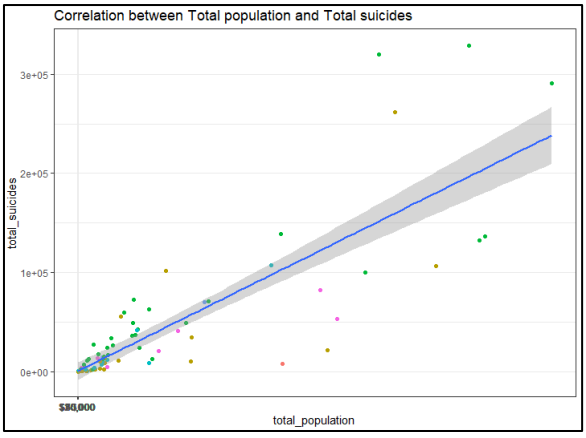
Inferential Analysis

Simple Linear Regression Model

In this part, two linear regression models have been created to discover the relationship between the number of suicides, GDP per capital and population by implementing the “lm” function in R. Extreme values has been excluded for more accurate results.

```
# Build the linear regression model.
# Remove outliers.
model1 <- lm(suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp)
gdp_suicide_no_outliers <- model1 %>%
  augment() %>%
  arrange(desc(.cooksd)) %>%
  filter(.cooksd < 4/nrow(.)) %>%
  inner_join(country_mean_gdp, by = c('suicide_per_100k', 'gdp_per_capita')) %>%
  select(country, continent, gdp_per_capita, suicide_per_100k)
model2 <- lm(suicide_per_100k ~ gdp_per_capita, data = gdp_suicide_no_outliers)
summary(model2)
```





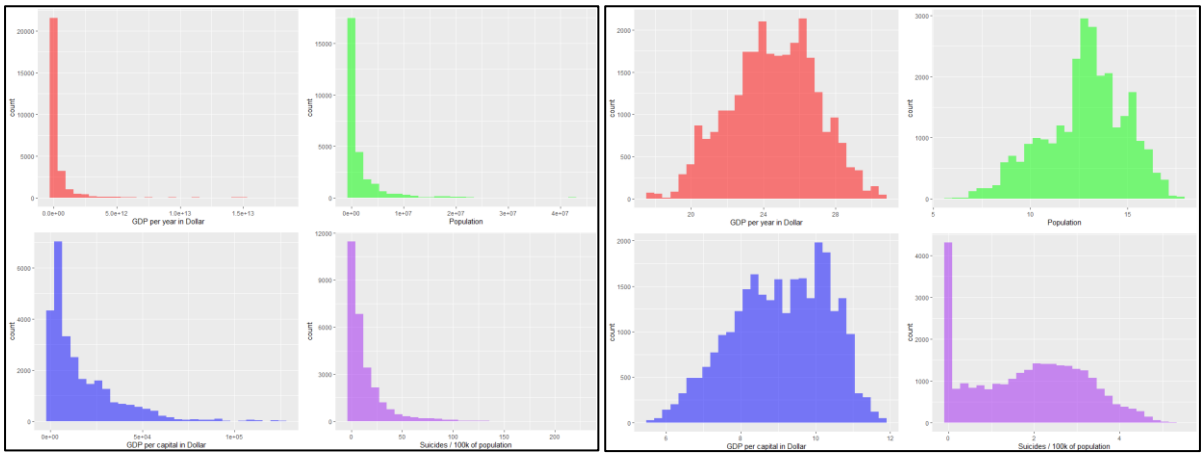
According to the summary report, we can conclude that the number of suicides had a strong positive correlation with the GDP per capital and population, due to the extremely low p-values (both <0.05). Consequently, we can predict the future suicides number based on the coefficients that these two models generated.

```
> summary(model2)

Call:
lm(formula = suicide_per_100k ~ gdp_per_capita, data = gdp_suicide_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-11.061  -5.108  -1.656   3.094  20.413

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.722e+00  1.086e+00   8.030 3.33e-12 ***
gdp_per_capita 9.581e-05  4.807e-05   1.993   0.0492 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.301 on 91 degrees of freedom
Multiple R-squared:  0.04184,   Adjusted R-squared:  0.03131
F-statistic: 3.974 on 1 and 91 DF,  p-value: 0.04922
```

```
> summary(model4)

Call:
lm(formula = total_suicides ~ total_population, data = population_suicide_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-104334   -5601    -602    5505  168241

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.731e+02  4.504e+03   0.083    0.934
total_population 1.176e-04  8.114e-06  14.497   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36420 on 91 degrees of freedom
Multiple R-squared:  0.6979,    Adjusted R-squared:  0.6945
F-statistic: 210.2 on 1 and 91 DF,  p-value: < 2.2e-16
```

Multivariate Linear Regression Model

We do multivariate regression analysis to find out which variables affect the number of suicides. Before doing regression analysis, it is essential to transform skewed variables to be normally distributed and select useful variables to build model.



After transformation, we include independent variables to regression model by using "lm" function one by one to see the variation between the number of suicides and dependent variables. The result shows that with more variables included in the model, the adjusted R Squared gets larger. The model includes all the independent variables is the best fit model with the largest adjusted R Squared of 0.9149, meaning these variables can explain 92.49% of the variation of the suicides number. The ANOVA also shows that these variables have significant predictive power to the suicides number. The stepwise regression selects the same variables as the former analysis.

```
## Build Linear Regression Model
### one by one analysis
mod1 <- lm(data = data2, suicides_no_log ~ `HDI for year`)
summary(mod1)
mod2 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log)
summary(mod2)
mod3 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log + sex)
summary(mod3)
mod4 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log + sex + age)
summary(mod4)
mod5 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log + sex + age + generation)
summary(mod5)
mod6 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log + sex + age +  gdp_per_capital_log)
summary(mod6)
mod7 <- lm(data = data2 , suicides_no_log ~ `HDI for year` + gdp_for_year_log + sex + age +  gdp_per_capital_log
           + suicides_per_100k_pop_log)
summary(mod7)
```

```
> summary(mod7)

Call:
lm(formula = suicides_no_log ~ `HDI for year` + gdp_for_year_log +
    sex + age + gdp_per_capital_log + suicides_per_100k_pop_log,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6737 -0.4005 -0.0302  0.3803  2.4016

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)              -11.289640   0.084639 -133.386  < 2e-16 ***
`HDI for year`             0.815492   0.193881    4.206 2.62e-05 ***
gdp_for_year_log           0.746509   0.003989  187.152  < 2e-16 ***
sexmale                    0.085137   0.016014    5.316 1.09e-07 ***
age15-24 years             0.729626   0.026791   27.234  < 2e-16 ***
age25-34 years             0.756875   0.027414   27.609  < 2e-16 ***
age35-54 years             1.235280   0.028129   43.915  < 2e-16 ***
age55-74 years             0.794213   0.028168   28.196  < 2e-16 ***
age75+ years              -0.150460   0.028478   -5.283 1.30e-07 ***
gdp_per_capital_log       -0.746292   0.014696  -50.782  < 2e-16 ***
suicides_per_100k_pop_log  0.903921   0.007981  113.261  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6346 on 8353 degrees of freedom
  (19248 observations deleted due to missingness)
Multiple R-squared:  0.915,     Adjusted R-squared:  0.9149
F-statistic:  8993 on 10 and 8353 DF,  p-value: < 2.2e-16
```

```
> anova(mod7)
Analysis of Variance Table

Response: suicides_no_log
                           Df  Sum Sq Mean Sq F value    Pr(>F)
`HDI for year`              1  1284.8  1284.8  3190.5 < 2.2e-16 ***
gdp_for_year_log            1 17460.0 17460.0 43358.8 < 2.2e-16 ***
sex                         1  2050.9  2050.9  5093.0 < 2.2e-16 ***
age                         5  7181.4  1436.3  3566.7 < 2.2e-16 ***
gdp_per_capital_log         1  3069.8  3069.8  7623.3 < 2.2e-16 ***
suicides_per_100k_pop_log   1  5165.7  5165.7 12828.1 < 2.2e-16 ***
Residuals                8353  3363.6     0.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Stepwise regression
step1 <- stepAIC(mod7, direction ="backward")
summary(step1)
anova(step1)

step2 <- stepAIC(mod7, direction ="both")
summary(step2)
anova(step2)
```

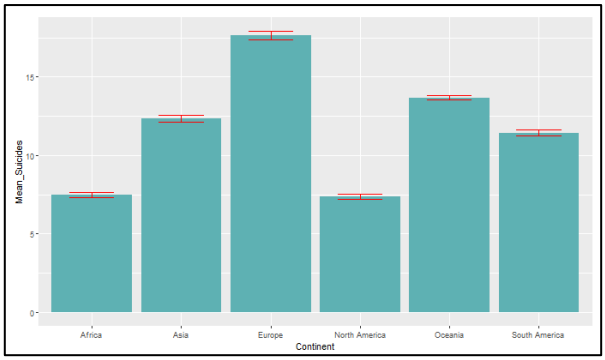
We use "summary(lm.beta())" function to see which variables have the greatest significance to the model. The result shows "gdp for year" has the greatest predictive power, followed by "suicides per 100k", "gdp per capital" and so on. From the coefficients, we can interpret that the number of suicides increases with GDP value while decreases with GDP per capita.

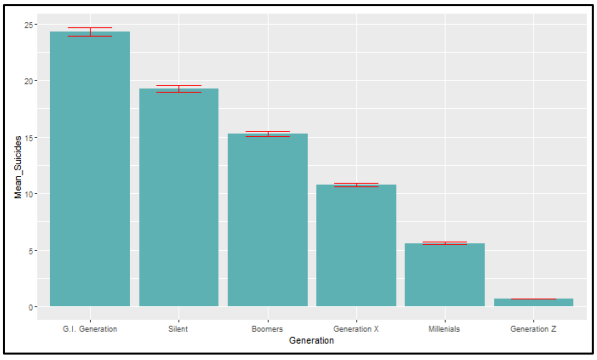
```
> summary(lm.beta(mod7))

Call:
lm(formula = suicides_no_log ~ `HDI for year` + gdp_for_year_log +
    sex + age + gdp_per_capital_log + suicides_per_100k_pop_log,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6737 -0.4005 -0.0302  0.3803  2.4016

Coefficients:
                           Estimate Standardized Std. Error  t value Pr(>|t|)
(Intercept)              -11.289640     0.000000   0.084639 -133.386  < 2e-16 ***
`HDI for year`             0.815492     0.035001   0.193881    4.206 2.62e-05 ***
gdp_for_year_log           0.746509     0.762857   0.003989  187.152  < 2e-16 ***
sexmale                    0.085137     0.019569   0.016014    5.316 1.09e-07 ***
age15-24 years             0.729626     0.125004   0.026791   27.234  < 2e-16 ***
age25-34 years             0.756875     0.129672   0.027414   27.609  < 2e-16 ***
age35-54 years             1.235280     0.211636   0.028129   43.915  < 2e-16 ***
age55-74 years             0.794213     0.136069   0.028168   28.196  < 2e-16 ***
age75+ years              -0.150460    -0.025778   0.028478   -5.283 1.30e-07 ***
gdp_per_capital_log       -0.746292    -0.416731   0.014696  -50.782  < 2e-16 ***
suicides_per_100k_pop_log  0.903921     0.521452   0.007981  113.261  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6346 on 8353 degrees of freedom
  (19248 observations deleted due to missingness)
Multiple R-squared:  0.915,     Adjusted R-squared:  0.9149
F-statistic:  8993 on 10 and 8353 DF,  p-value: < 2.2e-16
```

Confidence Interval

Last but not least, we intend to construct a confidence interval as an estimate of the mean suicides number for different continent and generation stated as a range with a lower and upper limit and a specific degree of certainty which is 95%. Although the true mean suicides numbers may or may not be in this interval, 95% of intervals formed in this manner will contain the true means.

```
analyseData <- data %>%
  select(generation, suicides_per_100k) %>%
  group_by(generation) %>%
  summarize(mean_suicides = mean(suicides_per_100k), sd_suicides = sd(suicides_per_100k),
            lower = mean_suicides-qnorm((1-confidence_level)/2+confidence_level)*sd_suicides/sqrt(nrow(data)),
            upper = mean_suicides+qnorm((1-confidence_level)/2+confidence_level)*sd_suicides/sqrt(nrow(data)))
ggplot(analyseData, aes(x = generation, y = mean_suicides)) +
  geom_bar(position=position_dodge(), stat="identity", fill = "#5EB1B3") +
  geom_errorbar(aes(ymin=lower, ymax=upper),
                width=.5,
                position=position_dodge(),
                color = "red") +
  xlab("Generation") +
  ylab("Mean_Suicides")
```





ANOVA

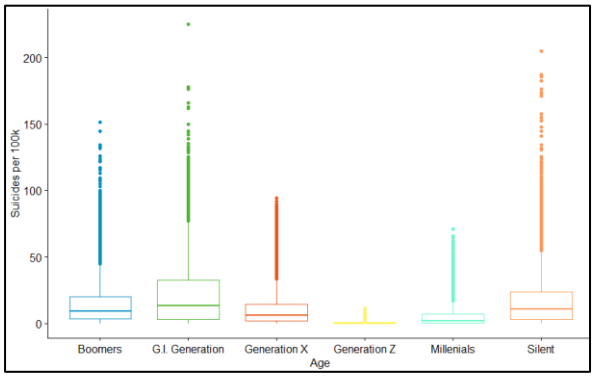
In this part, analysis of variance (ANOVA) has been performed to compare multiple means of suicides number for various generations and graded GDP per capital and evaluate whether the difference between them is significant or not.

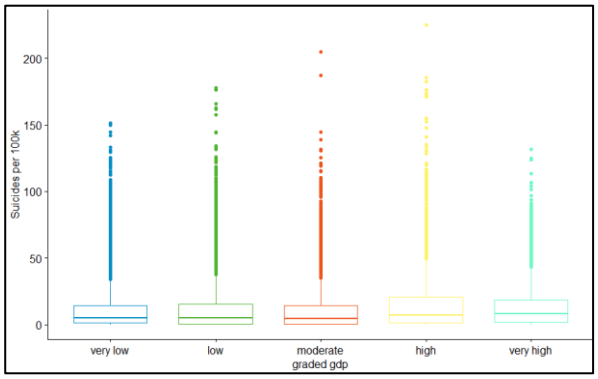
Null Hypothesis: All population means are equal

Alternate Hypothesis: At least one population mean is different

```
> # Implement ANOVA.
> res.aov <- aov(suicides_per_100k ~ generation, data = data)
> summary(res.aov)
              Df  Sum Sq Mean Sq F value Pr(>F)
generation     5 1135450  227090   711.4 <2e-16 ***
Residuals  27606 8812214     319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> # Implement ANOVA.
> res.aov <- aov(suicides_per_100k ~ gdp_per_capita_grade, data = data)
> summary(res.aov)
                     Df  Sum Sq Mean Sq F value   Pr(>F)
gdp_per_capita_grade     4   22486    5622   15.64 8.74e-13 ***
Residuals            27607 9925178     360
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





By implementing the “aov” function, we can conclude that the null hypotheses are rejected and we have sufficient evidence to support the claim that there are significant differences between the groups generation and between the groups graded GDP, due to the very low p-values (<0.005).

Shiny

In the Shiny dashboard, there are four different blocks. “Data Overview” allows the user to filter out the raw data by selecting country, year and age. This block also supports keyword search. In “Country Trend” block, the user can generate different trendlines of deaths numbers, by simply selecting different countries and time periods. A dropdown list and a sliding number axis are introduced in this sector. “Exploratory Analysis” allows the user to generally understand our analytic framework. We even dye a world map in different colours to show the

death number density of different countries. The last sector of “Inferential Analysis” shows the result of linear regression and confidence level.

Section 5: Summary

Suicide, as an act of intentionally causing one's own death, has become the 10th leading casue to death worldwide. In this project, we import a dataset of worldwide suicide rates from 1985 to 2016 from Kaggle and try to explore the patterns behind rows and columns. After analysis, we also build an R Shiny dashboard to better depict out result.

We first make a descriptive analysis using R. Through line chart, histogram, data table, pie chart and so forth, we illustrate the statistical parameters of suicide rates worldwide, grouped by years, genders, ages, countries and continents.

Globally, the suicide rate has been decreasing in the 30 years. The suicide rate trendline reaches a peak in 1995 and then shows a steady decreasing till now. In the histogram grouped by continents, Europe ranks the first and Asia as the second. However, the curve indicates that Europe, Asia and Africa are having fewer and fewer people who commit suicide in recent years, while the situation in Oceania and Americas is rather concerning.

As to sex and age, older groups have higher rates than younger ones, which is true throughout one’s life. The suicide rate of those aged 75+ has dropped by more than 50% since 1990 and the suicide rate in the ‘5-14’ category remains roughly static and small (< 1 per 100k per year). Men are about three times more likely to commit suicide than women. Although the absolute values are distinctive, the patterns of the both groups are similar. This means suicide rate is more likely to show a figure on a social level instead of within gender groups or age groups.

We also picture the bar chart of the suicide rate of each country. It shows Lithuania’s rate has been highest by a large margin. Also, there is a large overrepresentation of European countries with high rates, few with low rates, which is consistent with our former conclusion.

Besides descriptive exploration, we also conduct inferential analysis on the dataset. The regression model between rates and GDPs shows a weak positive relationship – richer the country is, more people are likely to end his or her life. In the assumption of normal distribution, we give out the confidence interval of the mean deaths of different generations and continents.

In conclusion, it is happy to see a decrement of the suicide rate in the last 30 years as well as the dropping trend still exists. Nevertheless, the result that older people tend to commit suicide than teenagers and middle-aged people do astonish us. We might think that the elderly has less stress and more leisure time, but the fact is that they are the ones that the society often ignore. With the aging problem becoming more and more serios, it is vital to build a more comprehensive system to take care of the elderly’s mental healthiness. What’s more, despite the dropping of total numbers, the ratio between men and women remains at a level of 3 throughout the 30 years. Under such circumstance, we strongly suggest boys go for a psychologist should he have depression, anxiety or other uncontrollable disorders.