Northeastern University
College of Engineering

Analysis of Suicide Rate in United States

IE6200 Engineering Probability & Statistics
Spring 2020

Project Report
By

Arati R Patil (001404989)
Ethan Neilan (001203623)
Jinay Neetin Shah (001085725)
Prem Mulchandani (001029556)
Vishal Baliga (001027278)

Under the Guidance of,
Prof. Rajesh Jugulum, Ph.D.

Objective:

The project's objective is to find out if there is any difference in number of suicides among sex and age groups. The project would be helpful giving a broad picture of suicide's predictors and hence it would help governments and relevant institutions controlling the suicide rates globally. Due to the limitation of data set regarding limited variables, we will only discuss few variables.

Gather and Compile the Data:

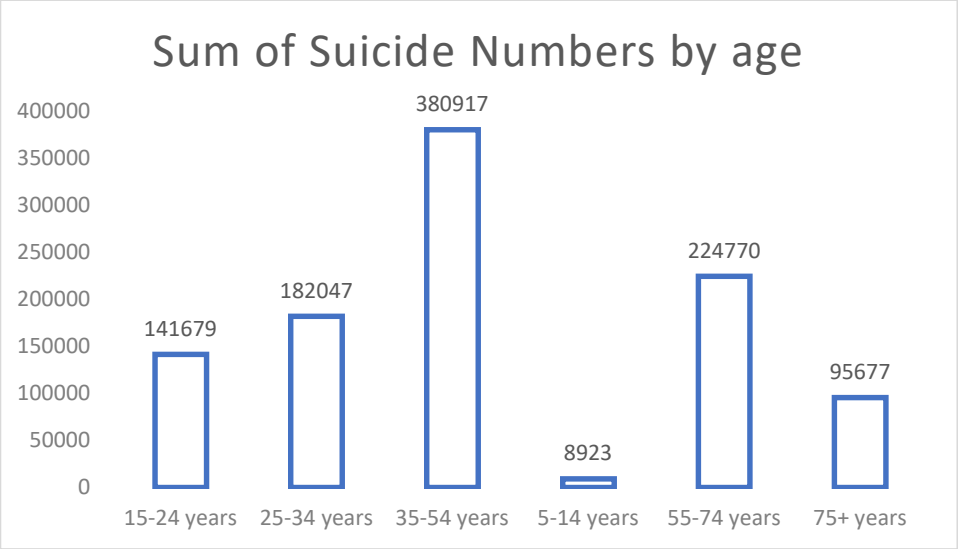
Data has been collected from Kaggle's dataset which is online and free dataset. It can be downloaded free from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016. Dataset was compiled from four different datasets linked through time and place for the better understanding of suicides globally. The source of those datasets is WHO, World Bank, UNDP and a dataset published in Kaggle.

Prepare for Analysis:

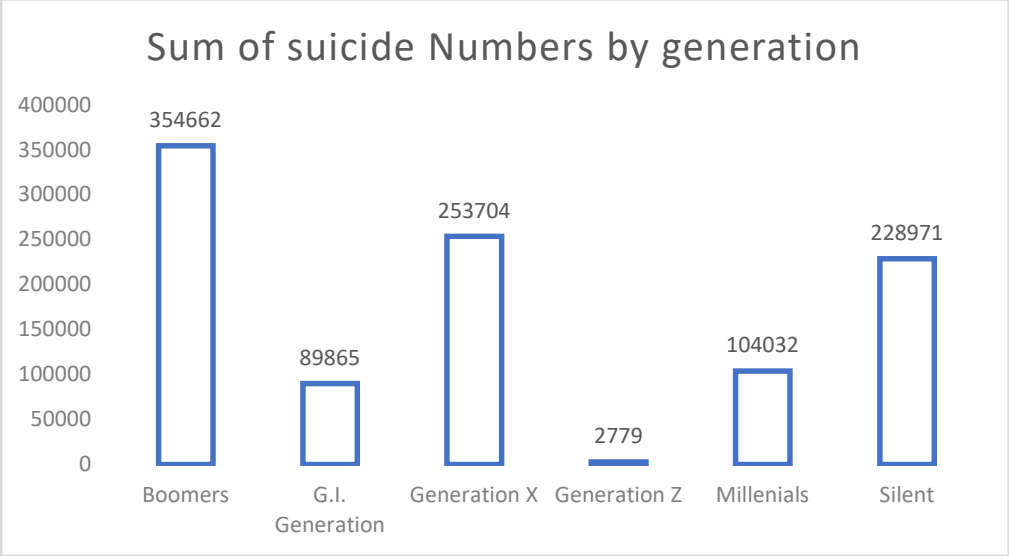
The dataset contains the number of suicides from 1985 to 2016 from different countries around the world. It has 27,820 observations and 12 observations including both predictors and dependent variable. Number of suicides and suicides in 100K (Suicide Rate) are the outcomes and year, sex, age, population, GDP, GDP per capita and Generation are the predictors. This study is an observational study as the data set already exists and we will conduct few statistical tests to conclude the study. The main objective of this study is to find out if there is any difference in suicide rates among the sex and age groups in the United States over the years from 1985 to 2015. Although there are many other predictors of suicides as well but due to the limitation of data set, we would only consider the above-mentioned variables for this project. This data is cross-sectional study as it has many representatives of the population i.e. different countries and many variables but for the sake of simplicity, we will conduct the analysis only for the United States only.

Descriptive Analysis:

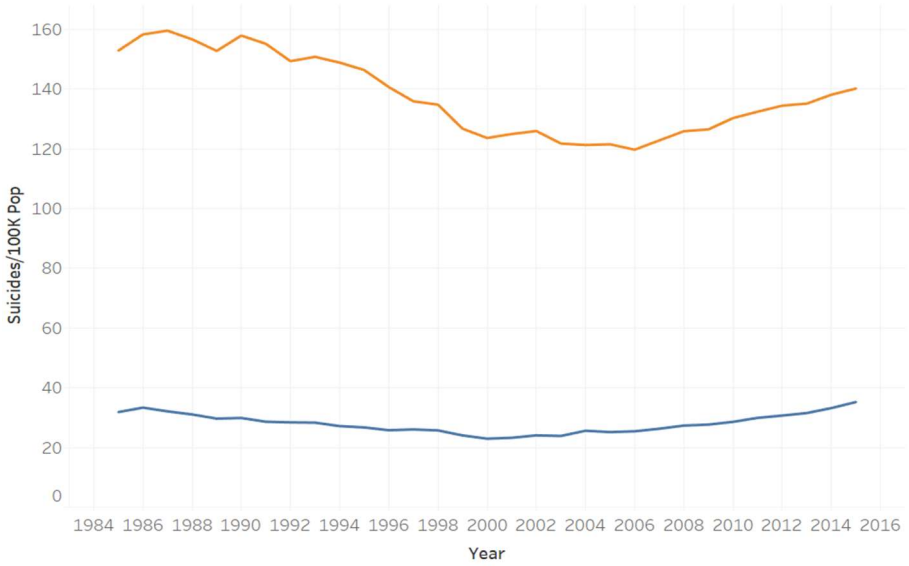
Now, let's explore the data to see the pattern of suicide number and suicide rate with reference to age group and sex in the United States from the year 1985 to 2015.



The above graph represents the total number of suicides from 1985 to 2015 in each age group. The number of suicides is maximum in the 35-54 age group.



The above graph represents the total number of suicides from 1985 to 2015 in each generation. The number of suicides is maximum in the Boomers generation. Generation are defined as per the groups such as Gen Z, iGen, or Centennials: Born 1996 – TBD, Millennials or Gen Y: Born 1977 – 1995, Generation X: Born 1965 – 1976, Baby Boomers: Born 1946 – 1964, Traditionalists or Silent Generation: Born 1945 and before.



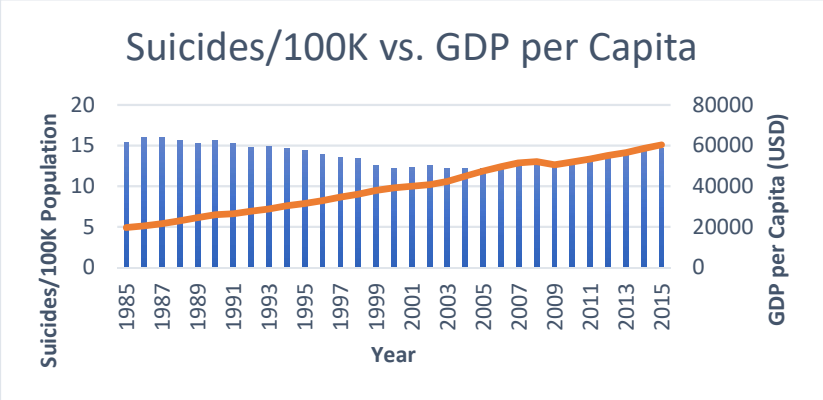
Suicide rate in Male and Female over years

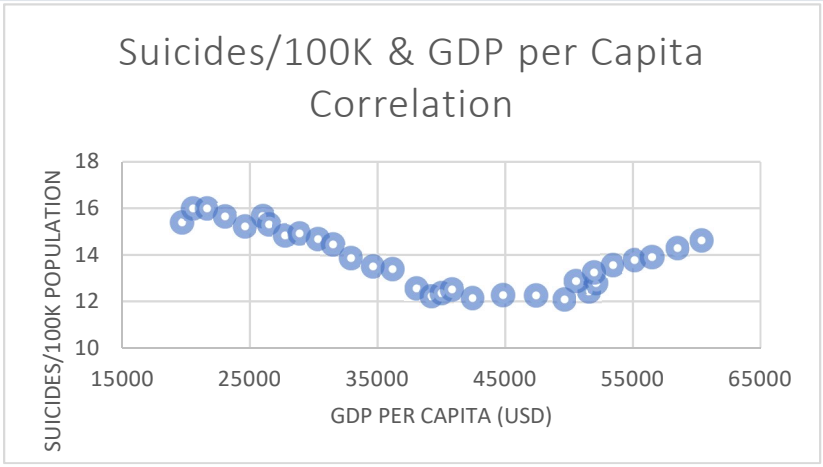
The above graph shows the trend of suicide rate in male (Orange graph) and female (Blue graph) in the United states from 1985 to 2015 (plotted in Tableau). As we can see, the suicide rate in female is much less than that od male. To support our understanding, we will conduct the ANOVA and Post-hoc analysis for better understanding of insights.

Correlation Analysis

This test was performed to determine the degree of any potential linear relationship between two variables. Specifically, the relationship between suicides per 100,000 population and the GDP per capita. The correlation value for the United States between the number of suicides per 100,000 population and the gross domestic product (GDP) per capita was -0.63474, or -63.474%, and was found using Excel. The correlation value can be categorized as a strong negative correlation between the two variables. Correlation is dimensionless, and the following table was used to describe the results of the correlation tests.

Correlation	Interpretation of Correlation Value
-1.0 to -.8	There is a very strong negative correlation
-.6 to -.79	There is a strong negative correlation
-.4 to -.59	There is a moderate negative correlation
-.2 to -.39	There is a weak negative correlation
-.01 to -.19	There is a very weak negative correlation
0 to .19	There is a very weak positive correlation
.2 to .39	There is a weak positive correlation
.4 to .59	There is a moderate positive correlation
.6 to .79	There is a strong positive correlation
.8 to 1.0	There is a very strong positive correlation





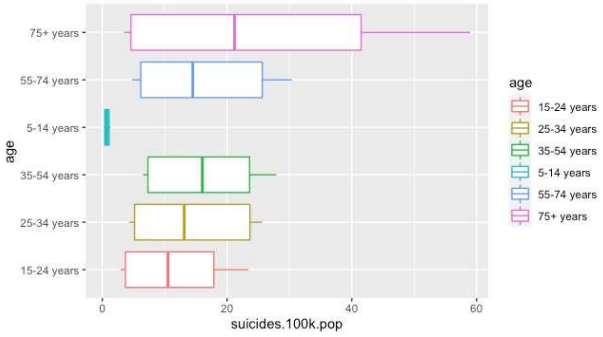
This is a reasonable result as GDP per capita is often directly related to overall quality of life. Therefore, it would be reasonable to assume that as people become wealthier, they are less likely to commit suicide.

ANOVA: Analysis of Variance

Analysis of Variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Null hypothesis:
Ho: There is no difference in suicide Rate between age groups for USA (1985-2015)
Ha: There is a difference in suicide Rate between age groups for USA (1985-2015)

Conditions:
1- Normality / Sample size - Seems like data is not normal but due to huge sample size, we can say that this condition is met.
2- Independence between the groups - Respondents could be in one of the age groups and hence this condition is met too.
3 - Independence within group – The dataset is well organised to maintain the integrity of the dataset. This condition is met too as we can see from the below box plot that suicide rate is different in each age group



Box plot as per Age Group and suicide Rate

Results of ANOVA:
ANOVA was performed in Excel and input variables to perform the test were Suicide Rate and Age group.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	18907.2918	5	3781.45836	30.0685055	1.3449E-25	2.23864886
Within Groups	46028.6847	366	125.761434			
Total	64935.9765	371				

From the results above we get,
F statistic – 30.0685
F critical – 2.2386
We can see that the F-value is greater than the F-critical value for the alpha level selected (0.05). Therefore, we have evidence to reject the null hypothesis and conclude that there is a difference in suicide Rate between age groups for USA (1985-2015).

Post-hoc Analysis:

Tukey Test is a single-step multiple comparison procedure and statistical test. It is a post-hoc analysis, what means that it is used in conjunction with ANOVA. It allows to find means of a factor that are significantly different from each other, comparing all possible pairs of means with a t-test like method. Post-hoc Analysis is performed in R and the results are as show below:

```
> tukeystest
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = suicides.100k.pop ~ age, data = suicide2)

$age
                               diff       lwr       upr     p adj
15-24 years-5-14 years   10.61258065  4.842261 16.382900 0.0000035
25-34 years-5-14 years   13.43903226  7.668713 19.209352 0.0000000
55-74 years-5-14 years   15.20967742  9.439358 20.979997 0.0000000
35-54 years-5-14 years   15.30725806  9.536939 21.077578 0.0000000
75+ years-5-14 years     23.95000000 18.179681 29.720319 0.0000000
25-34 years-15-24 years   2.82645161 -2.943868  8.596771 0.7250685
55-74 years-15-24 years   4.59709677 -1.173223 10.367416 0.2039053
35-54 years-15-24 years   4.69467742 -1.075642 10.464997 0.1843982
75+ years-15-24 years    13.33741935  7.567100 19.107739 0.0000000
55-74 years-25-34 years   1.77064516 -3.999674  7.540965 0.9513392
35-54 years-25-34 years   1.86822581 -3.902094  7.638545 0.9392508
75+ years-25-34 years    10.51096774  4.740648 16.281287 0.0000045
35-54 years-55-74 years   0.09758065 -5.672739  5.867900 1.0000000
75+ years-55-74 years     8.74032258  2.970003 14.510642 0.0002660
75+ years-35-54 years     8.64274194  2.872422 14.413061 0.0003268

>
```

The above diff values show the difference in the suicides among the different age groups. Used for multiple comparisons in ANOVA, the adjusted p-value indicates which factor level comparisons within a family of comparisons (hypothesis tests) are significantly different. If the adjusted p-value is less than alpha, then we reject the null hypothesis. We can notice that the adjusted p value for most of the pairs (except 3 pairs) is less than the considered alpha (0.05) value. Thus, we reject the null hypothesis.

Conclusion

The objective of project was to see if there is any significant difference in the suicide rates among male & female and different age groups. We had to use t-test to check the difference in suicides among male and female while we used ANOVA to see the difference among age groups. Result showed that male have more suicide rates than the female. On the other side, there is significant difference in the suicide rates in different age groups. People with age group of 5 - 14 years old have the least number of suicides while people with age group of 35 - 54 years old have the highest number of suicides in the United States.

This project has its own limitation and future data scientists may elaborate the other aspects of suicides. Other variables such as unemployment, economy, stress level, etc. should also be taken in consideration behind the reasons of suicides around the world.

References:

1. https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/
2. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.
3. https://www.schoolnet.org.za/twt/06/M6_Understanding_Correlation.pdf