STAT-S 670: Exploratory Data Analysis

Final Project: Global Suicide Rate Analysis

Uma Maheswari Gollapudi, Shreyas Bhujbal, Astha Hurkat

Project Description

The World Health Organization reports that close to 800,000 people die by committing suicide every year, which is almost 1 person every 40 seconds. This is very disturbing as it is a preventable health problem, but still a leading cause of death worldwide.

Reports from the American Psychological Association show us that the USA alone has recorded a substantial increase in the rate of suicide, increasing by 33 percent from 1999 through 2017 or from 10.5 to 14 suicides per 100,000 people. It ranks as the second most common cause for deaths among people in the 10 –34 year age range, and is the fourth leading cause of death for people ages 35 to 54 in the US. Sadly, 90% of those who died by suicide had a diagnosable mental health condition at the time of their death.

Globally, suicide ranks as the 10th leading cause of death overall. Our dataset includes the global numbers and rate of suicides per country, and includes characteristics such as age, gender, generation, Gross Domestic Product and location. Here, we explore the relationship between per capita GDP, age, gender, location with respect to suicide rates between 1985 and 2016 using the suicide rates overview data set. We found that GDP per capita, gender and age have effects on the suicide rate, though the causal relationship is not very strong, and varies across these factors.

Dataset Description

The dataset we are using is the Suicide Rates Overview-1985 to 2016 dataset from Kaggle. [Link to Dataset : https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016]

We will use it to derive correlations between different attributes to understand the variation in suicide rates among different countries globally. Our data contains a total of 12 columns and 27,821 observations.

- Country - Each country to which the observations belong to
- Year - When these observations were recorded
- Sex – The gender of the subjects
- Age – Age range groupings of the subjects
- Suicides no – The total number of cases of suicides recorder per country per age group
- Population - Total population of the country per age group
- Suicides/100k pop – Ratio of number of suicides per 100,000 people
- Country-year - Contains the country name and the year
- HDI for a year - Human development index for that year

- GDP for year ($) - GDP for a particular year
- GDP per capita ($) - GDP over the population for that year
- Generation - splits the population based on the generation they were born in - G.I Generation, Silent Generation, Boomers, Generation X, Millennials, Generation Z

Out of these variables Year, Suicides no, Population, Suicides/100k pop, HDI for a year, GDP for year and GDP per capita are quantitative variables, while Sex, Age and Generation are categorical variables.

Data Preprocessing

1. We have eliminated the year 2016 from the data, since it had a lot of missing data for most countries.
2. We also eliminated the HDI and Generation variables, since the former was missing about 70% of the data, while the latter included the same data as the Age variable.
3. The African continent has very few countries within this dataset - Cabo Verde, Mauritius, Seychelles and South Africa, so it is not really a true representation of the country's suicide statistics.
4. We created a new variable called Suicides_100K, which is the ratio of suicides by population among 100,000 people. We use this and the Suicides_no variable, which just gives the number of suicides, for the rest of the analysis.
5. We created numerical categories for the Age and Gender categorical values, for better correlation analyses later on.

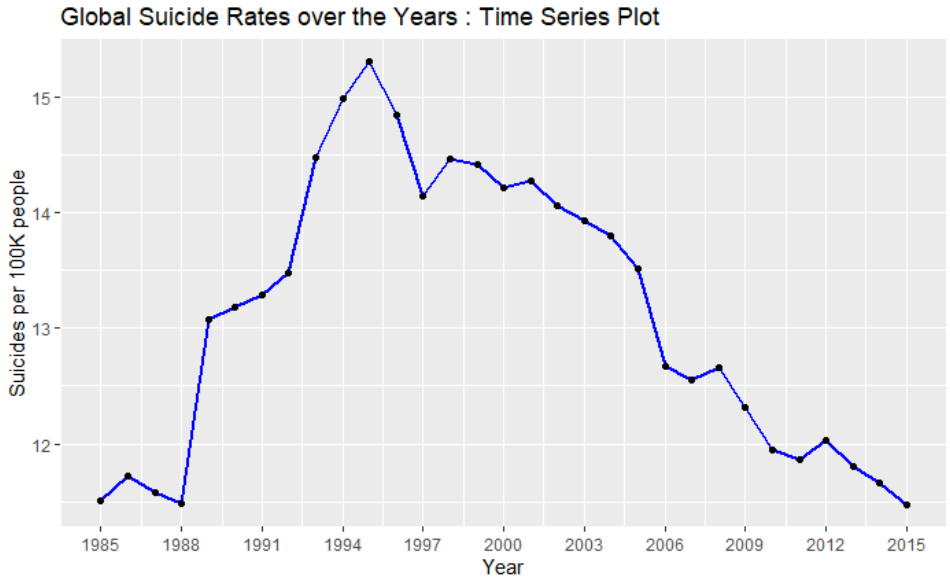
Exploratory Analysis

Considering the suicides recorded per 100K people as the response variable, we perform an exploratory analysis of the relationship it shares with age and across generations, gender, the per capita GDP, across countries.

For ease of analysis, we split this question into three research questions:

1. How does the suicide rate vary across age and gender, over time?
2. How does the suicide rate change across continents and countries, over time?
3. What is the relationship between GDP and suicide rate overall, and in particular, during recession periods?

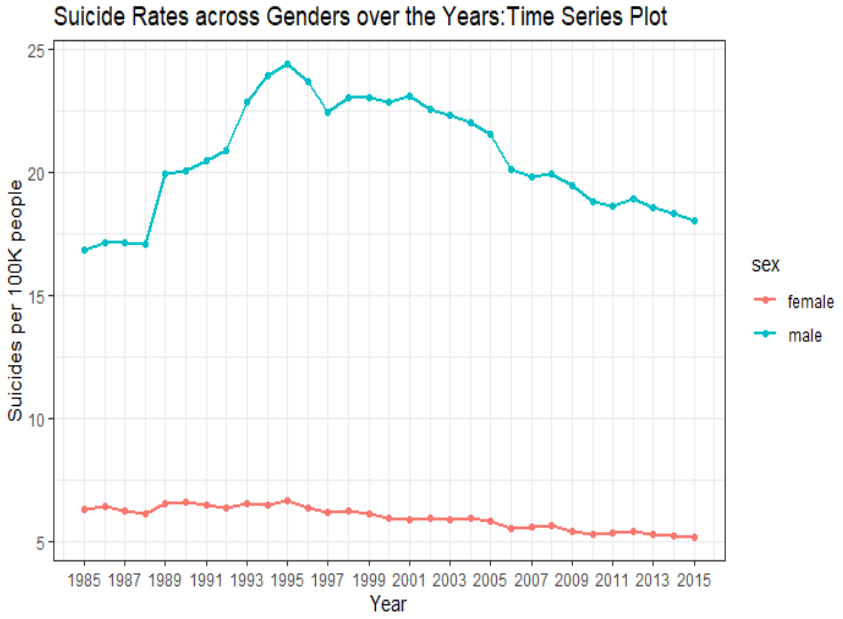
Global Suicide rate over time:



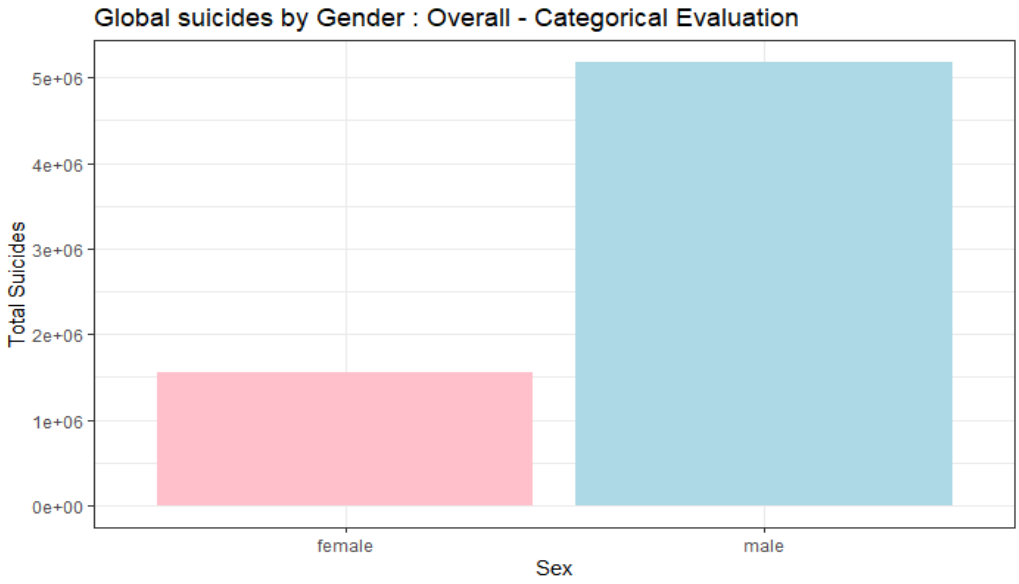
This time series plot depicts the overall trend of suicides per 100K globally. We can see that over the years the constant trend has been one of decline - global suicide rates are reducing. Suicides are generally underreported due to the stigma surrounding them, so we can say that while there have definitely been more deaths due to suicide than visible on this plot, it is also possible that there was very less data collected during the earlier years (1980's) which is why there is such a steep incline by the 1990s. According to this data, the deadliest year with respect to suicides was 1995.

Section 1:

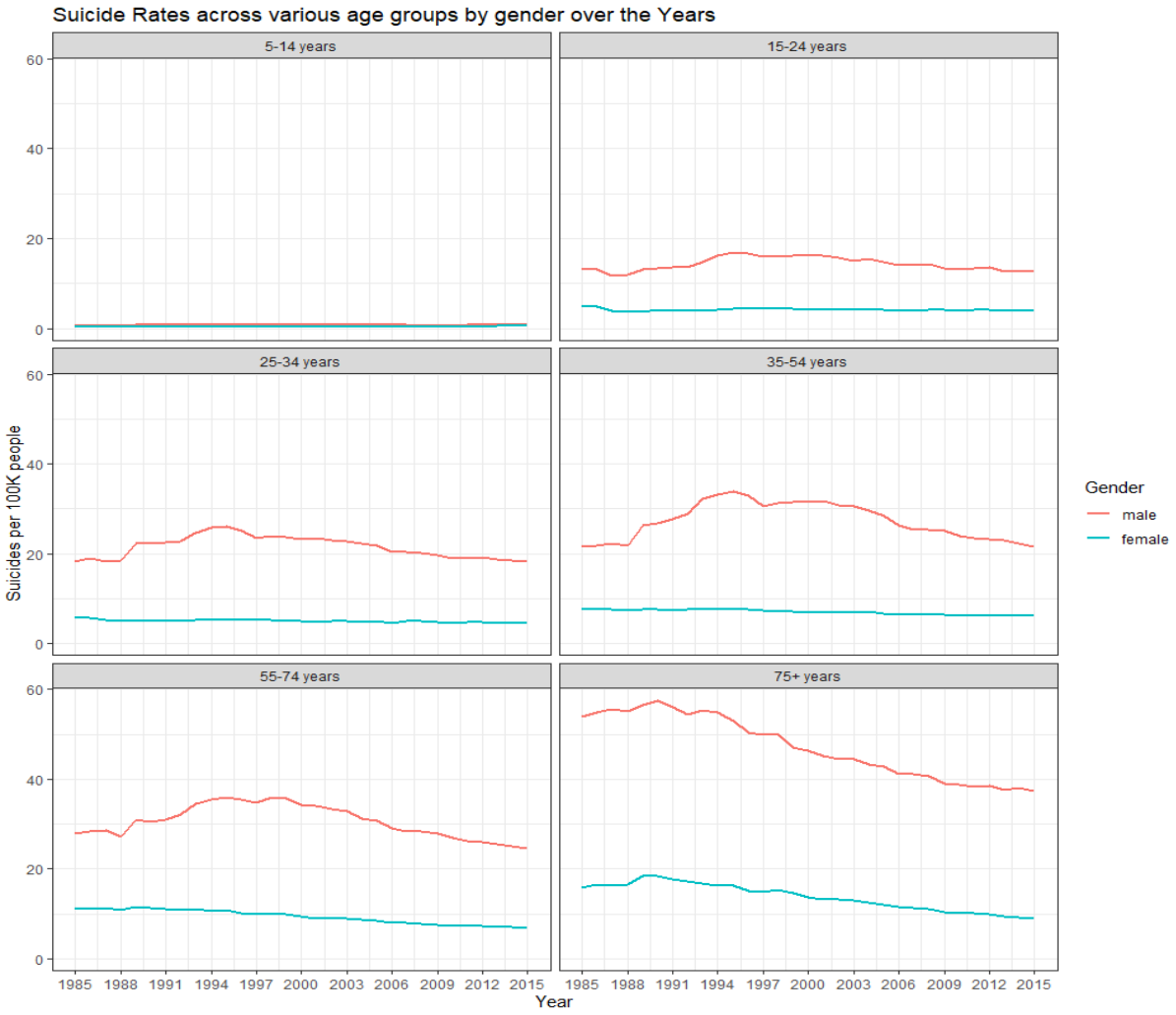
Suicide Rate versus Gender and Age



From this plot, we can see that men have a staggeringly high number of suicides as opposed to women, throughout the three decades. 1995 had the highest spate of suicides across both the genders, although the rate of suicides in men was almost 4 times as high as that in women. In general, the male suicide rate is at least about 2.83 or more times higher as compared to women. This could be due to a larger underrepresentation of women suicides – or other factors such as general mental health and well-being etc.



As we can see from the bar graphs, approximately 1.5 million women have perished to suicide from 1985 to 2015, as compared to almost over 5.2 million deaths due to suicide for men.

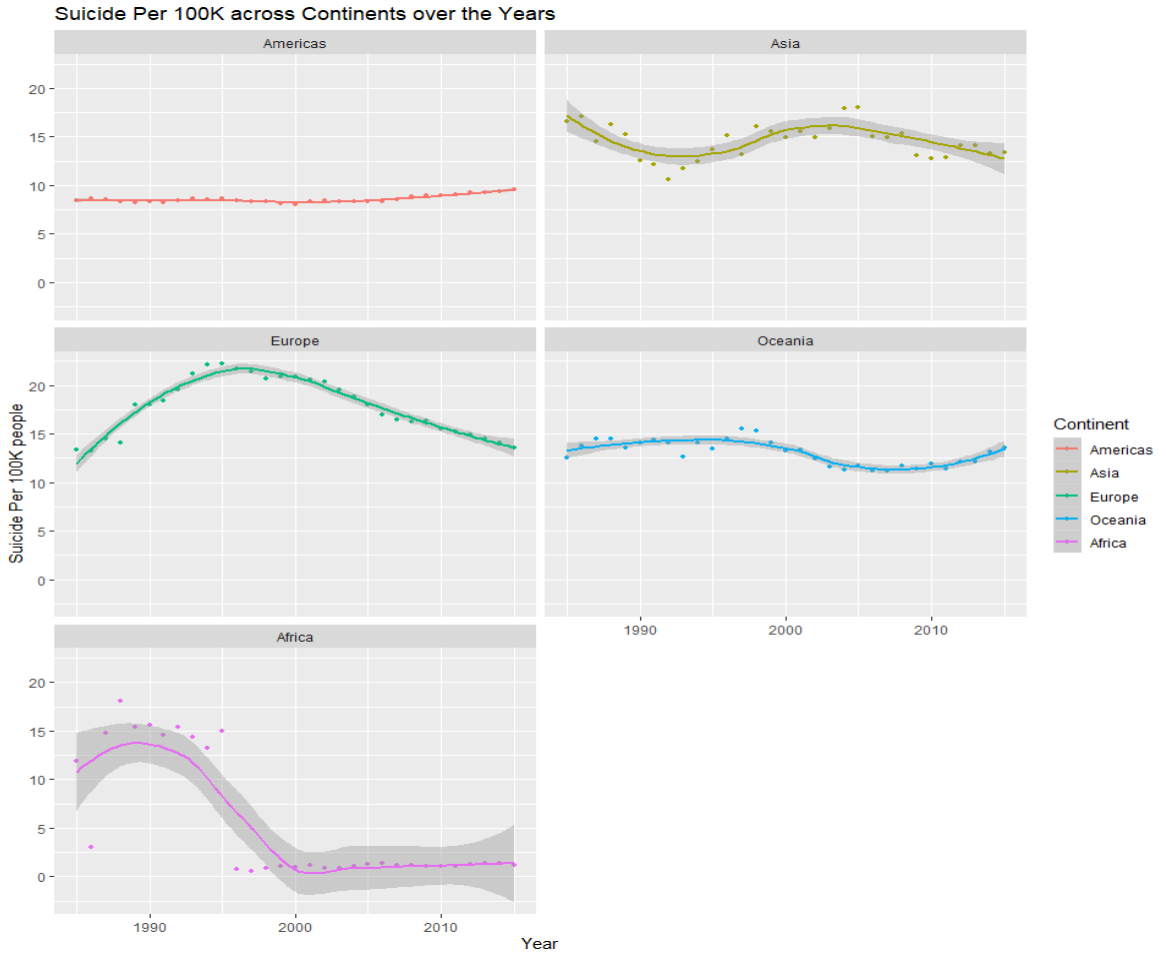


The above plot compares the trends of suicides per 100K people across different age groups. The lowest number of suicides are observed amongst children, in the 5-12 years group. It is interesting to observe that there is a gradual increase in suicide rates across the groupings, eventually culminating with people aged 75 years or more, particularly men, recording the highest suicides numbers. We can say that age has a positive relationship with suicide rates, although, there has been an overall decrease in the number of suicides of most of these groups (excepting 5-24 year-olds, which has remained low and constant throughout) in the recent years.

Section 2:

How has the suicide rate varied across continents over the years?

Our dataset has suicide numbers for different countries but not for different continents. So, we used the inbuilt “countrycode” library to group the data by continent. Here, Oceania is a geographic region that includes Australasia, Melanesia, Micronesia, and Polynesia.

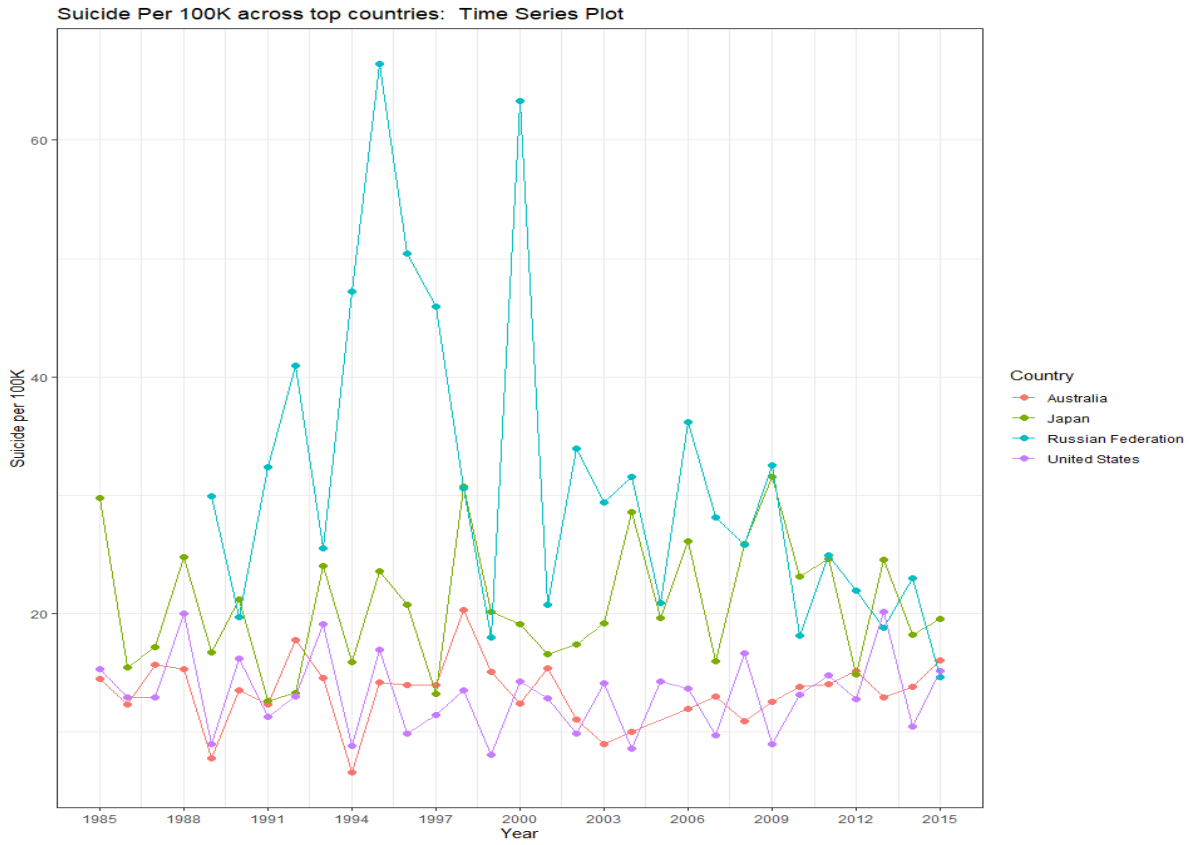


The above plots show us the continent wise split of the suicides per 100k people. Here, we can see that the there is a decreasing trend in Europe and Asia, while suicides in Oceania are once again rising. The Americas, on the other hand, have a nearly consistent number of suicides over the years, if not slightly increasing. Out of all the countries, Europe showcases the highest number of suicides overall, and particularly in the mid-90s, followed by Asia and then Oceania.

Due to a lack of data, Africa’s trend is not very reliable to interpret, since it only contains the suicide count of four countries. The data was best fit by a Loess plot, as the number of suicides across continents do not share a linear relationship.

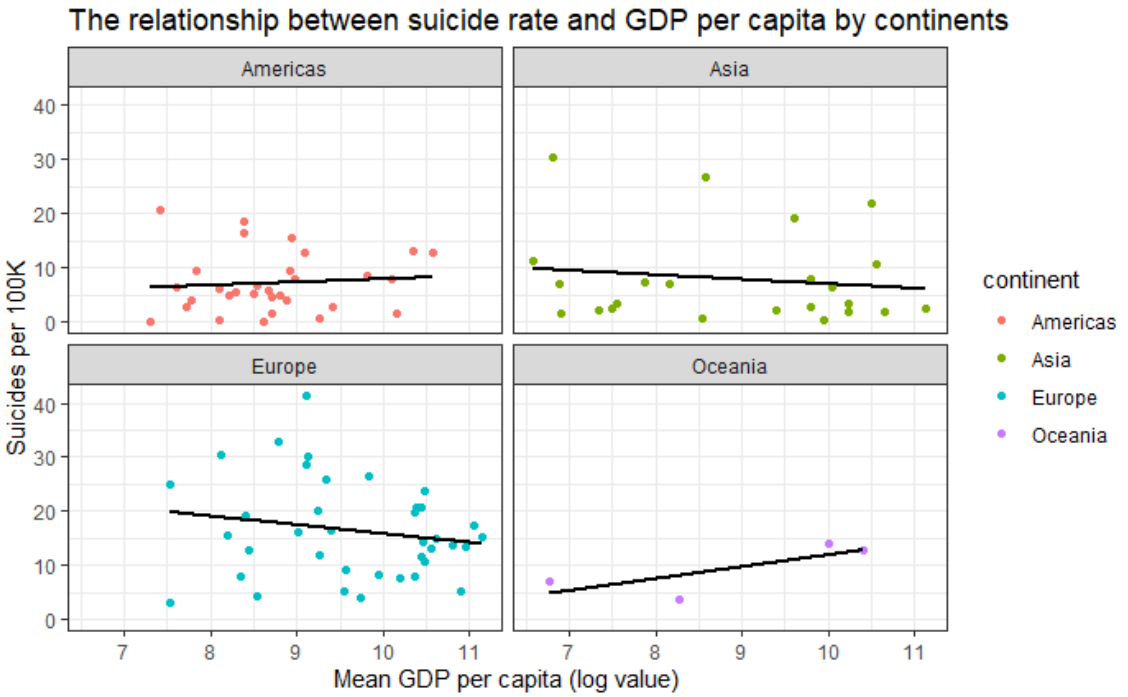
Continent <chr>	Total_Suicides <int>	Country <fctr>
Africa	7321	South Africa
Americas	1034013	United States
Asia	806902	Japan
Europe	1209742	Russian Federation
Oceania	70111	Australia

The most dangerous countries for deaths by suicide are Russia in Europe, the US in the Americas, Australia in Oceania and Japan in Asia, according to this data. The graph below depicts the various trends in suicide rate for each of these countries. We have omitted Africa from this analysis. We can see that Russia has the highest number of suicides per 100k people among all the countries, and peaked in 1995, with a second, similarly large peak in 2000. Japan too, had a peak in 2009 (possible due the economic recession). We can see that in the years following the recessions in 1991 and 2008, Japan, Australia and the US states record increases in the suicides rates. This is possibly because they were hit pretty hard by the effects of the recession.



Section 3:

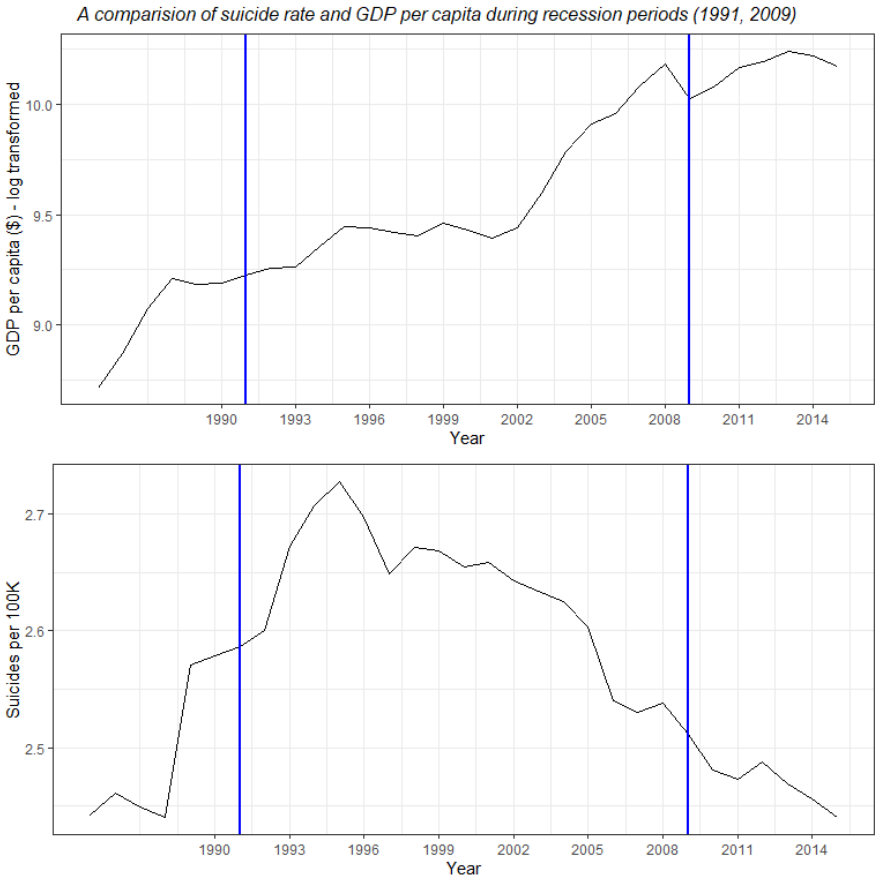
How has the suicide rate varied with respect to the GDP per capita?



Since the GDP per capita was heavily right skewed, we performed log transformations on it, and used that data for the rest of this analysis.

The above image captures the relationship between the mean GDP per capita and number of suicides per 100k people over various continents. Different continents practice different cultures and have various habits, social interactions and ideologies, all of which affect the well-being of an individual. Hence, we thought it would be more meaningful to capture this relationship by countries. We have fit the plot us the 'lm' function, since the linear model is providing a general sense of the trend between GDP and suicides per 100k per continent.

We can see that for Europe, very low GDP per capita records a higher number of suicides. As the GDP increases, we note that there is a declining trend for the suicides per 100k people. This shows that as people earn more money, they tend to have lower rates of suicide. The same is true for Asia, which shows an inverse relationship between the two. The Americas on the other hand, establish a slightly positive relationship between GDP and suicide rate, which is surprising because the notion would be that as the population becomes richer and has a better lifestyle, the rate of suicides would decrease. Oceania also shows a positive trend, but this can be attribute to the fact that it contains only four countries (fewer overall data).



The side-by-side comparison of the suicide per 100k and GDP across time is to try and analyze their relationships during recession periods. The first recession period, in 1991, shows us that there is an increase in both the GDP per capita, and the number of suicides per 100K people, while the second recession period, 2007-2009, shows us that there was a drop in the GDP trend, along with a drop in the suicide rate.

In order to understand the effect of each factor on the suicide rate, we fitted a linear model to the data, and checked the R-squared and p values. The correlation values for the factors with the number of suicides committed differed, as suicide rate had a correlation value of –0.226 with year, 0.080 with age and 0.1447295 with gender (where men were assigned higher value). The correlation value for GDP and suicide number is negative, but this only when grouped by year. They have different relationships for different countries and continents, as seen above (negative for Europe and Asia, positive for Americas and Oceania)

As we fit the linear model with each of these variables – GDP_per_capita, age, gender, year and Country, the p-values remained very small throughout (< 2.2e-16), and along with the correlation values, we can determine that that each of these have a weak relationship with the number of suicides committed per 100K people.

Limitations

Primarily, we were missing a majority of the data for the African continent, and data for other countries such as India, China etc. Having the data for all these countries would have increased our ability to understand the data better. Next, we the HDI data was missing, but since this is shares a mainly positive relationship with GDP, we can say that it possibly shares a similar relationship as GDP with suicide rate.

Future Scope

For a more in-depth analysis of suicide rate, we should have a better data set that contains variables such as marital status, unemployment, physical and mental health of people who committed suicide, drug and alcohol abuse tendencies, and so on. If we have this data, a more comprehensive analysis can be made about suicidal tendencies and rates across the world, and we can gather better insights on the factors that create conducive circumstances, or even cause suicides.

Conclusion:

From our analysis, we can conclude that there is a general trend of decline in the number of suicides committed over time. The average number of suicides per 100k people in 2015 was only around 11.3, as compared to the average suicides per 100k people of 13.12 from 1985 to 2015.

The suicide variable is highly correlated with gender. We observed that men have very high numbers of suicide over time, consistently surpassing women by at least 2.83 times. This could either be due to a large underreporting of suicide rates among women, less collection of their data for this particular dataset, or other social factors that negatively affect men. As people get older, we observed that the number of suicides is increasing, i.e., suicide increases with age. This is true for both the genders, and people aged more than 75 years are particularly vulnerable.

The most dangerous countries for deaths by suicide are Russia in Europe, the US in the Americas, Australia in Oceania and Japan in Asia, according to this dataset, with Europe ranking as the deadliest continent for the same. We can see that Japan, Australia and the US record increases in the number of suicides per 100K people after both the recession periods in 1991 and 2008.

The relationship between GDP per capita for each continent and the suicides per 100k differs across continents. While Europe and Asia show a decline in suicide rates with increasing GDP, both the Americas and Oceania show a positive relationship. Overall, there is a weak relationship between a country's GDP per capita, age, gender and the number of suicides. However, there is not enough information to establish a causal relationship between these variables and the suicide rate.

Appendix:

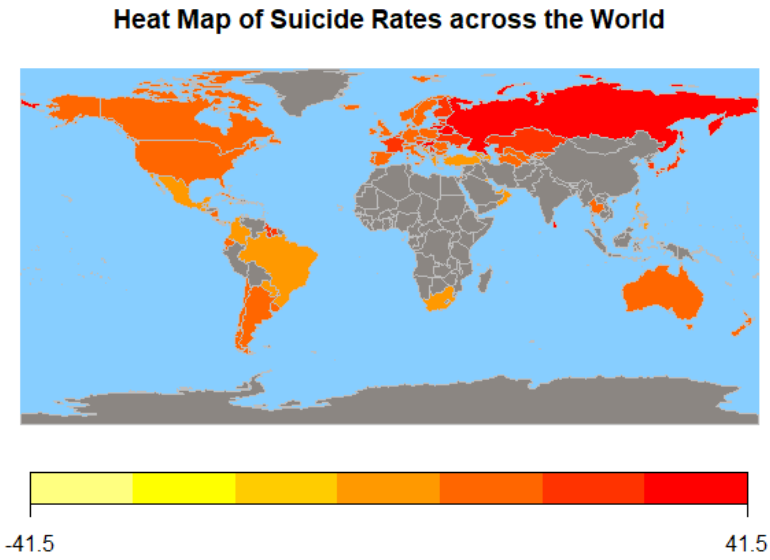


Fig 1: Heat map of suicide rates in the world.

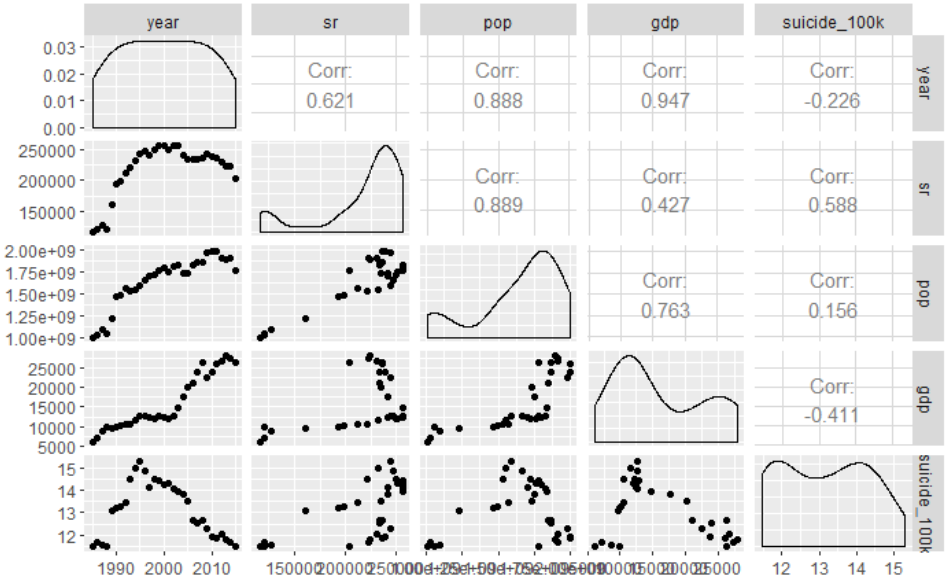


Fig 2: Correlation between suicides per 100k and year, gdp and population size.

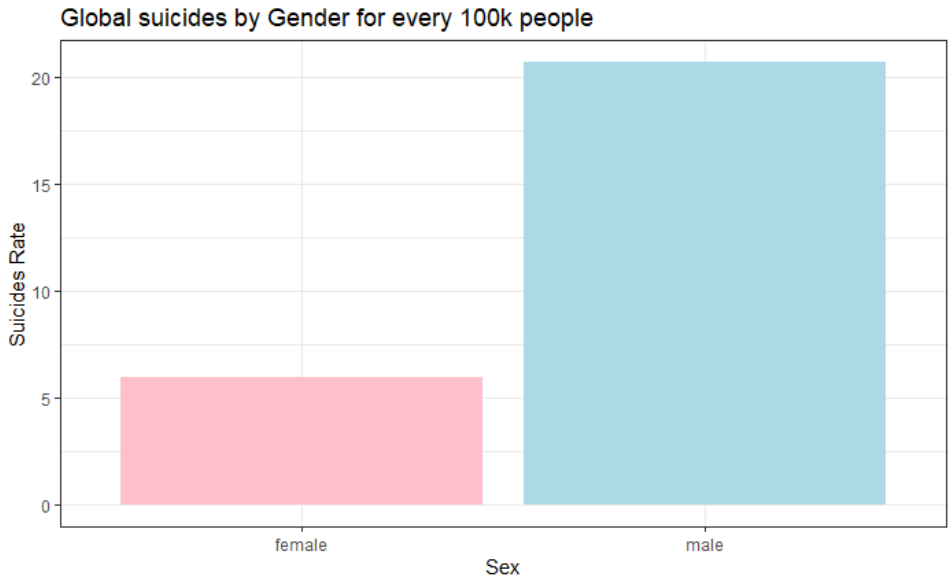


Fig 3: Global suicides by Gender for every 100k people.

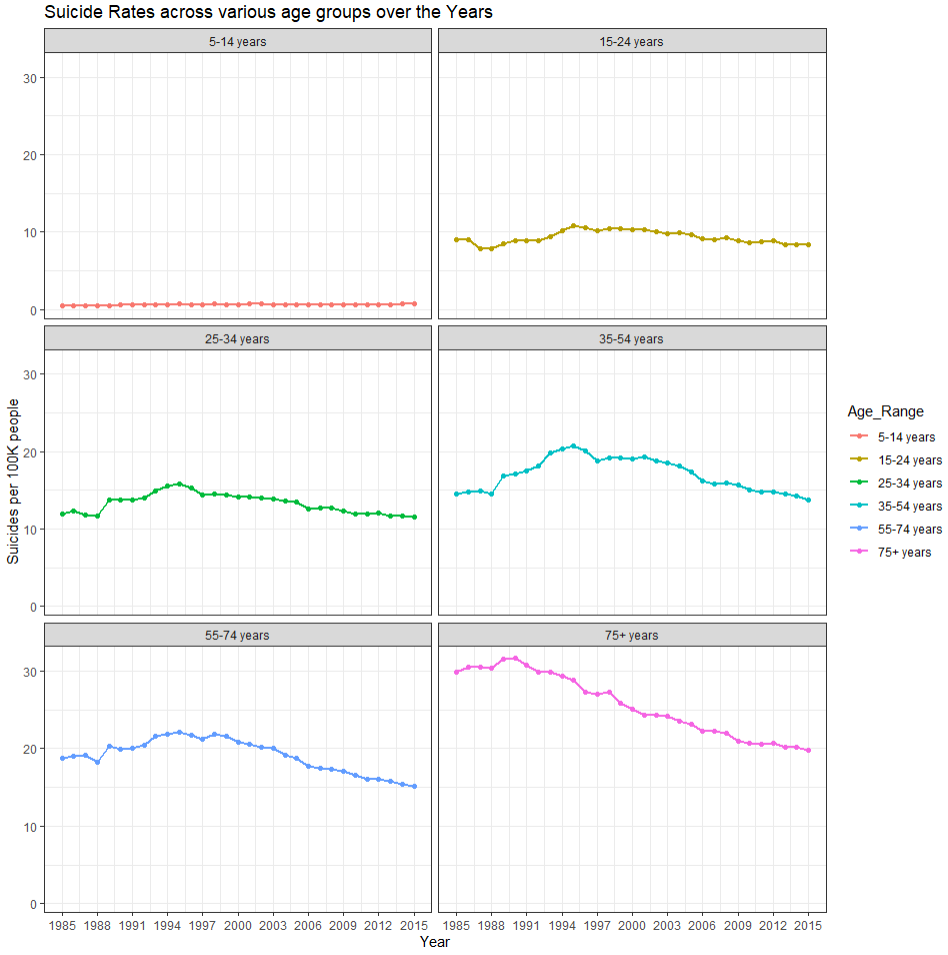


Fig 4: Global suicides by age groups for every 100k people.

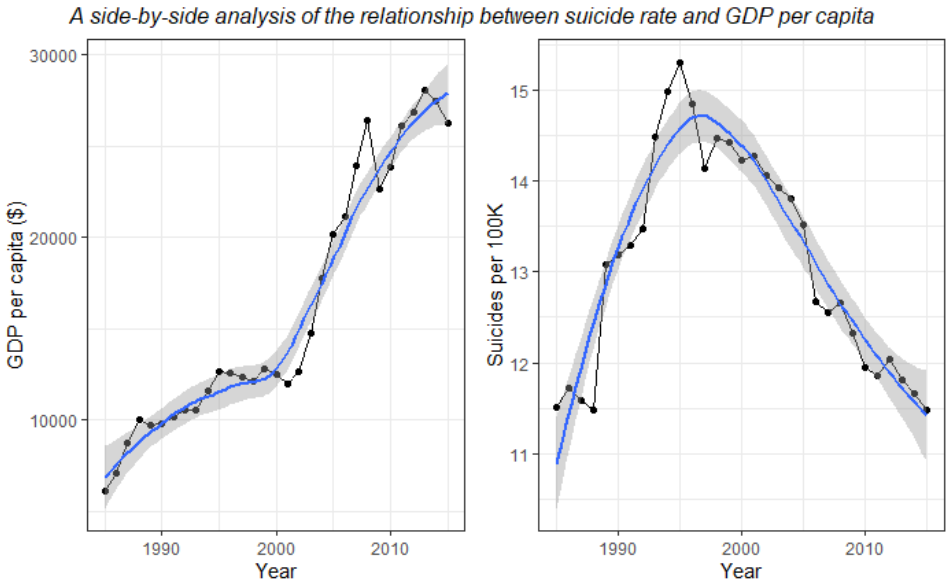


Fig 5: Comparison between GDP per capita and Suicide rates per 100K people