**UNIVERSITY OF COLORADO DENVER**

**Data Visualization Final Project Report**

BANA - 6800

Avi Manawat (Student Id: 109638446)

**Abstract:**

Suicide is the 10th leading cause of death in the United States, taking the lives of approximately 47,000 Americans each year. Thus, it becomes a matter of concern that needs to be investigated

As the' American Association of Suicidology ' tagline states, I strongly believe that suicide prevention is the business of everyone. Keeping in mind that anything can be resolved with the help of adequate support and lifestyle, the act of ending one's own life specifying the reasons for being depression, alcoholism, social reasons or any other mental illness in that matter is not a good idea.

We can choose to stand together in the face of a world that may often feel like a lonely and disconnected place, and we can choose to make a difference by making lives more livable for those who struggle to cope. Through this project, I look forward to identifying the trends in suicidal rates by region, gender, age and ethnicity, and to relate these trends to the possible reasons that lead to the drastic decision that could enable us to curb the thinking at the very start.

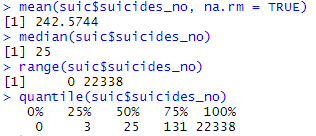
**Exploratory Data Analysis in R**



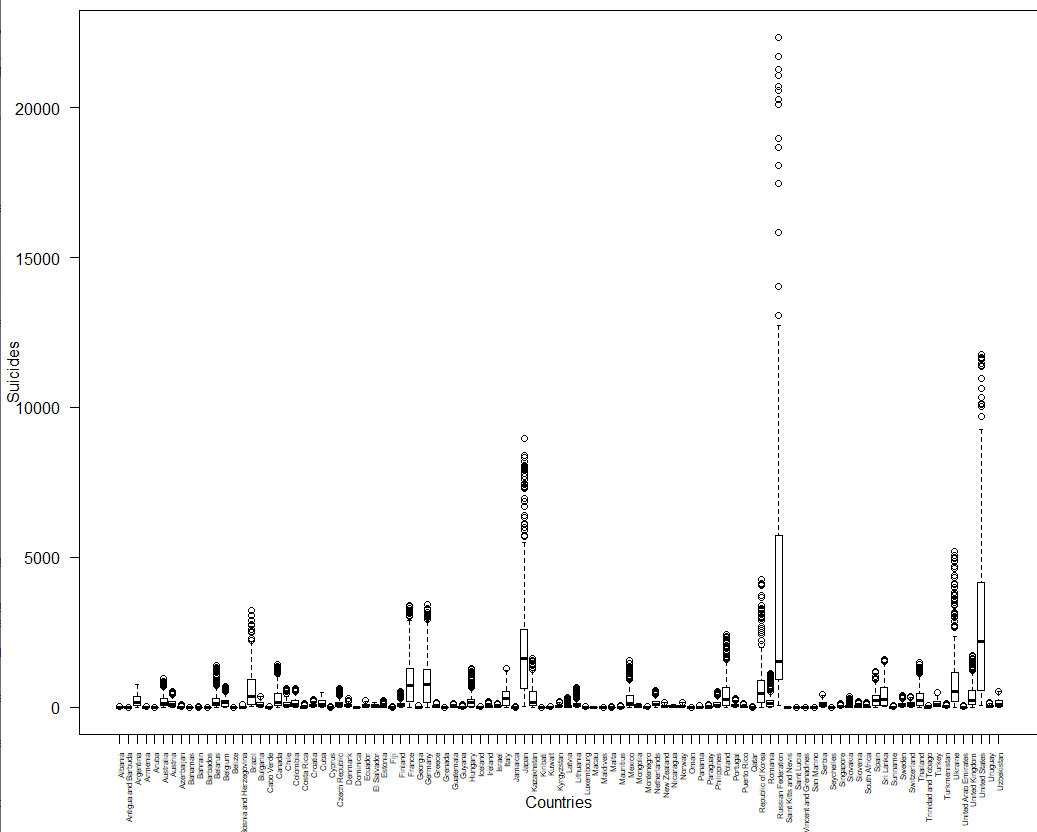
* First step was to identify the number of Nan values present in the data in order to deal with them.



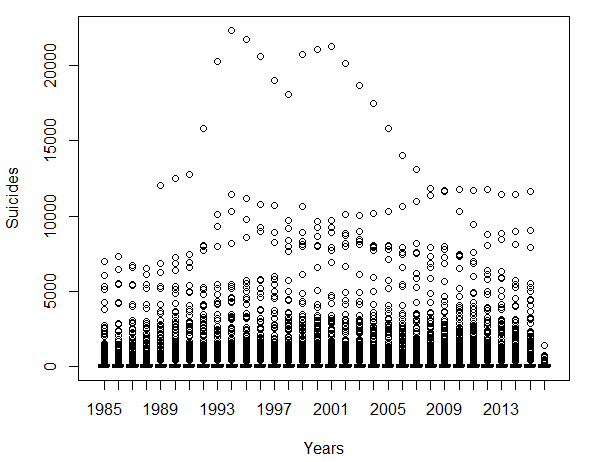
* On analyzing, it was found that all these Nan values were present in a single column. The whole column was not taken into consideration in order to get a clean data for analysis



* The mean, median, range and quantile of the suicidal data were calculated in order to get a hint on presence of any outliers. With median = 25, 3rd quartile = 132 and maximum value being 22338 definitely indicates presence of outlier in the data.



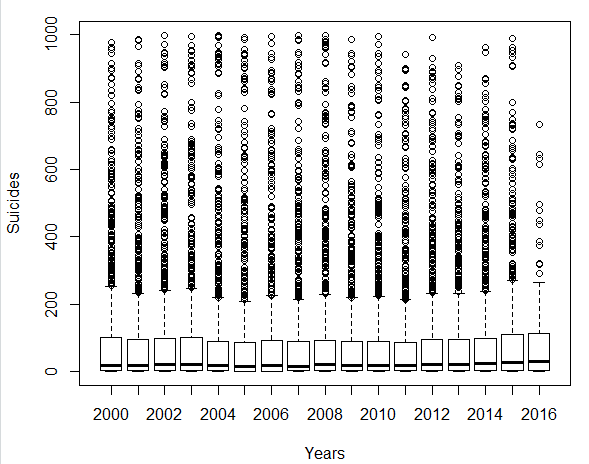
* Plotting a country wise box plot of the suicidal data clearly showed the presence of outliers in the data (seen in the above figure) in the form of countries like United States and Russian Federation that hindered the process of data visualization.



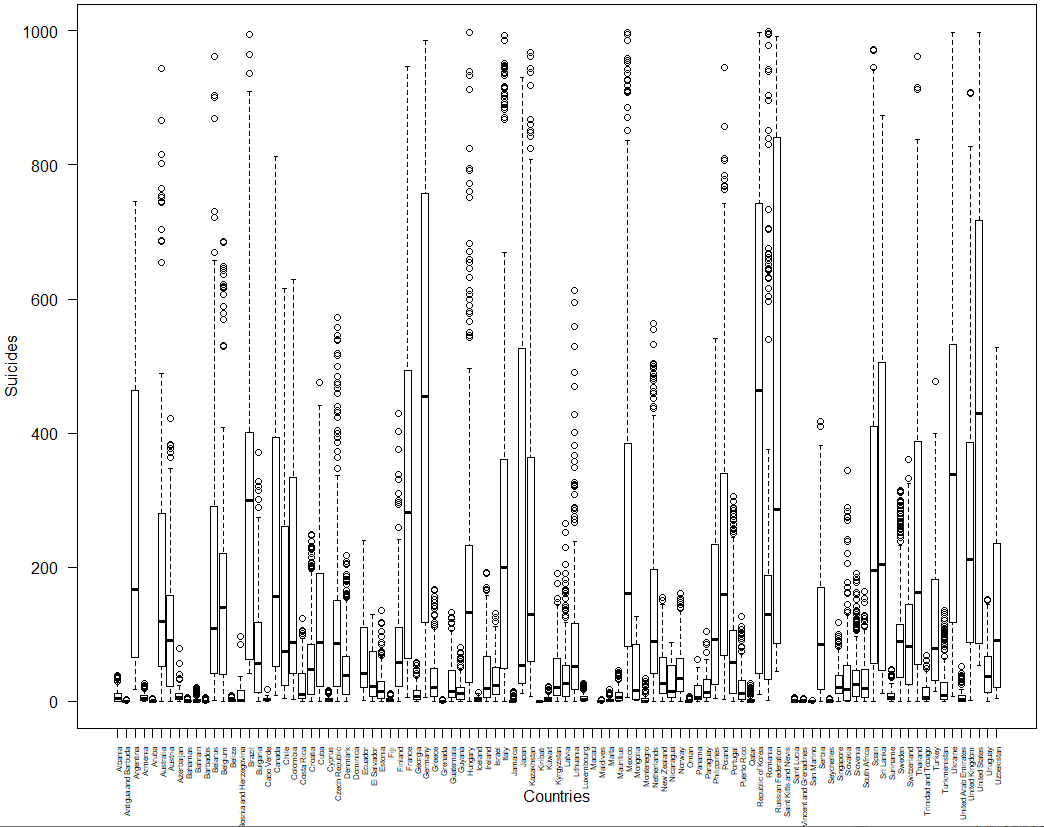
* These are the year wise boxplots of the suicidal data per age group. This graph of the original data is depiction of how the outliers affected the analysis.



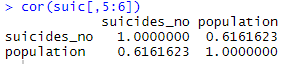
* A subset of the data was created by using the above code and it was decided to treat the above mentioned countries separately for the purpose of initial analysis for better visualization.



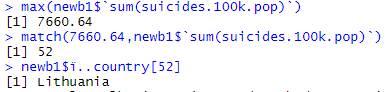
* After dealing with the outliers, the boxplots were clearly visible and interesting information can be drawn from the graph. Year 2007 had the least suicidal death mean per age group while the mean of suicidal deaths per age group in all the countries kept increasing since the year 2009.



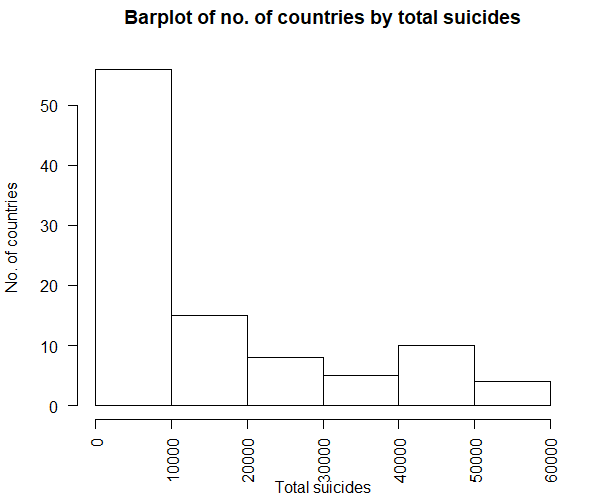
* The above is a country-wise depiction of suicidal deaths per age group. The number of deaths vary largely depending on the country. For ex., country like Albania has really low suicidal death stats as compared to the other countries.



* The correlation between population of a country and number of suicides is 0.61 which states that there is some positive relationship between the two. But since the correlation is not too strong, there are other factors as well affecting the number of suicides in the country.



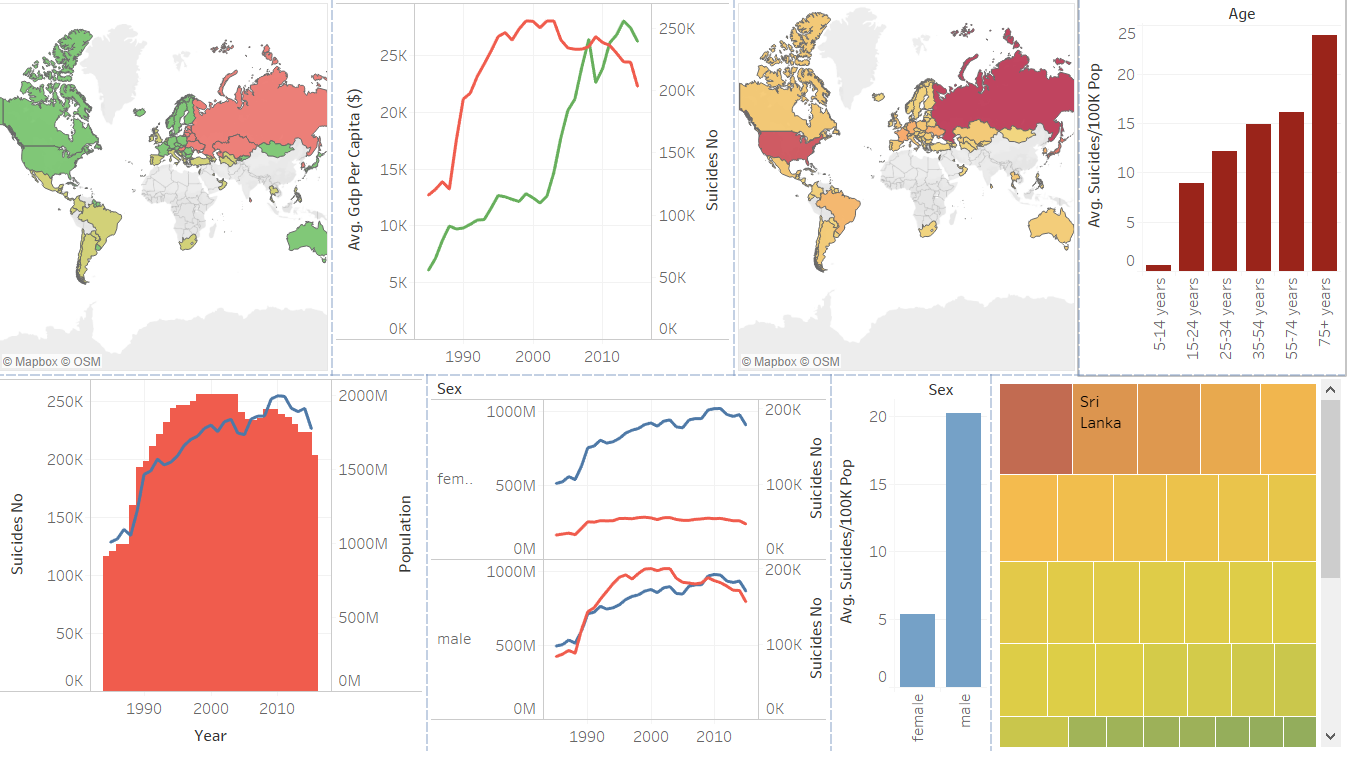
* The maximum cases of suicide per 100k population of the country was found to be in Lithuania. There were 7660.64 suicide cases per 100k population in this country, which was the highest.



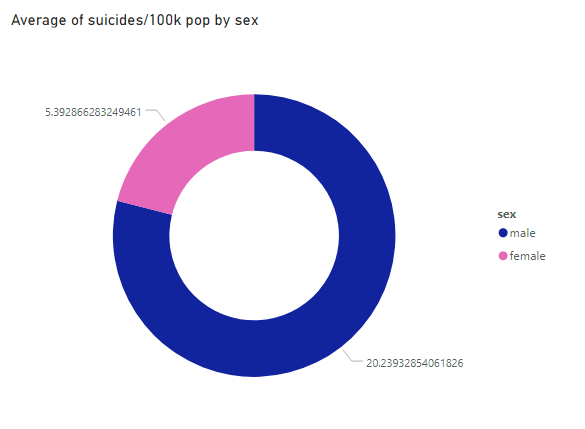
* This is a bar chart representing total no. of countries that fall in the given range of suicidal cases. As observed here, about 56 countries have witnessed suicidal cases between 0-10000, 15 countries have observed between 10000 to 20000 suicidal cases and only 5 countries have witnessed more than 50000 suicide cases.

**Data Analysis**

**Final Dashboard in Tableau**

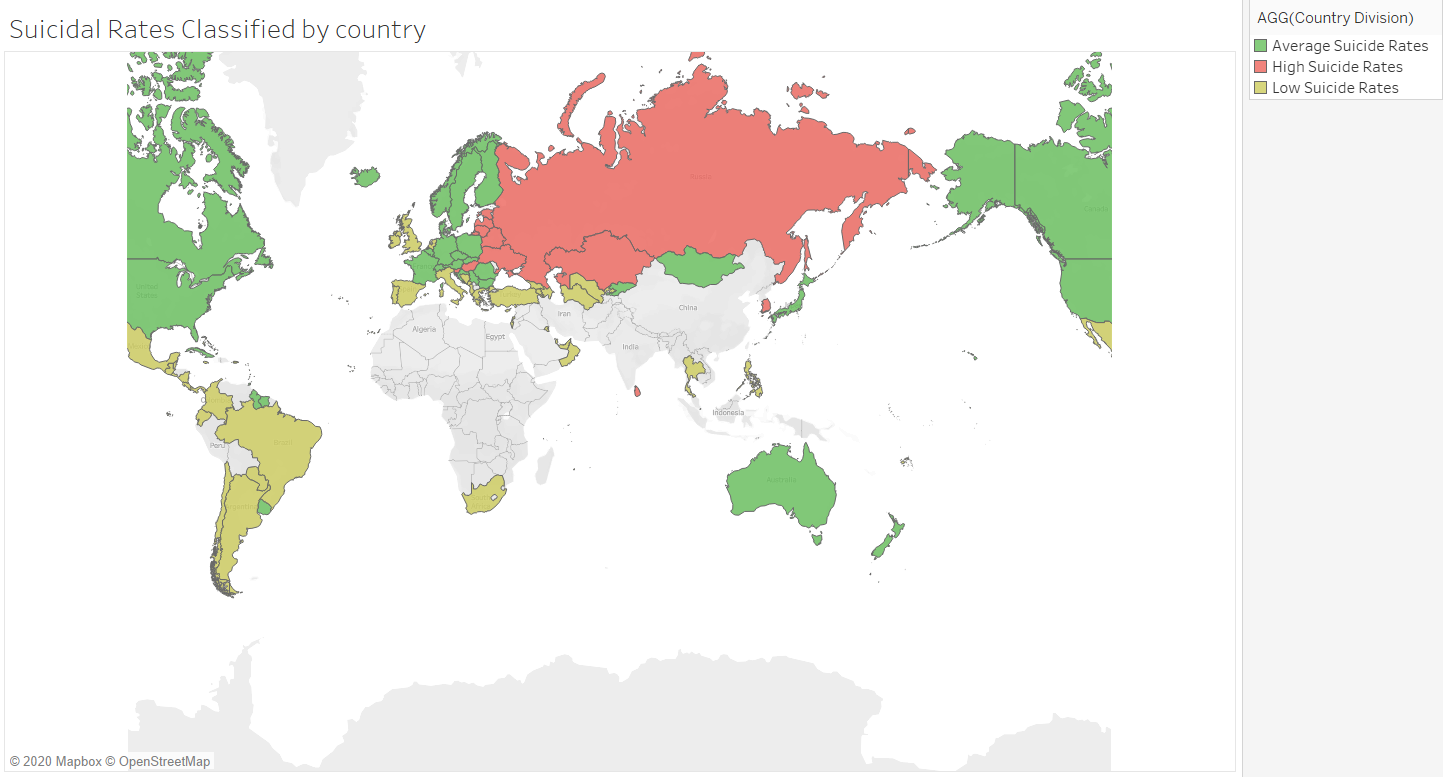


**Average of suicides per 100k population by gender**

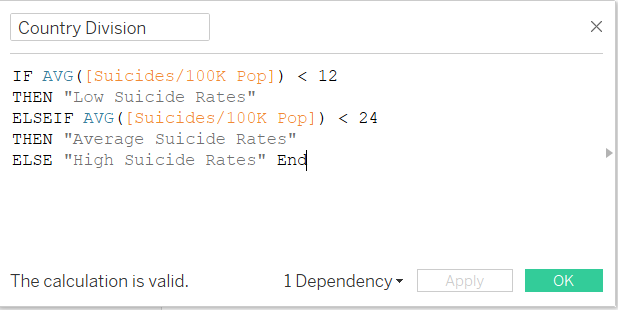


* The above donut chart represents suicides per 100k population by gender. It is seen that around 20 out of every 100k males commit suicide as compared to only 5 in case of females. We can conclude that male have a higher tendency of committing suicide than female in general.

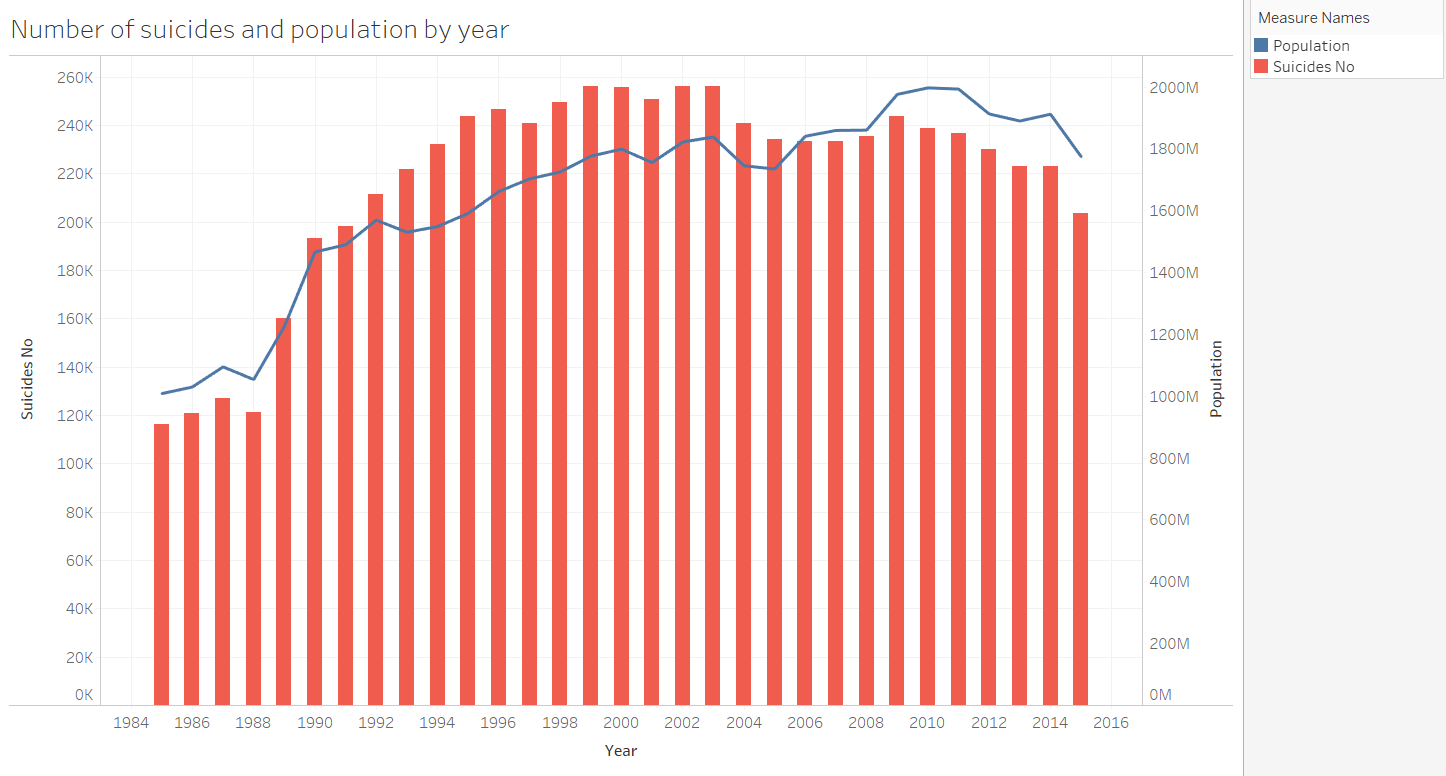
**Suicidal rates classified by country**



* This is a map that depicts average suicidal rate per 100k population classified by country. High suicidal rates are identified as more than 24 suicide cases per 100k population in that country and is observed in the north-west Asian countries. Medium and Low suicidal rates are depicted by green and yellow colors respectively. The classification was done as shown below:

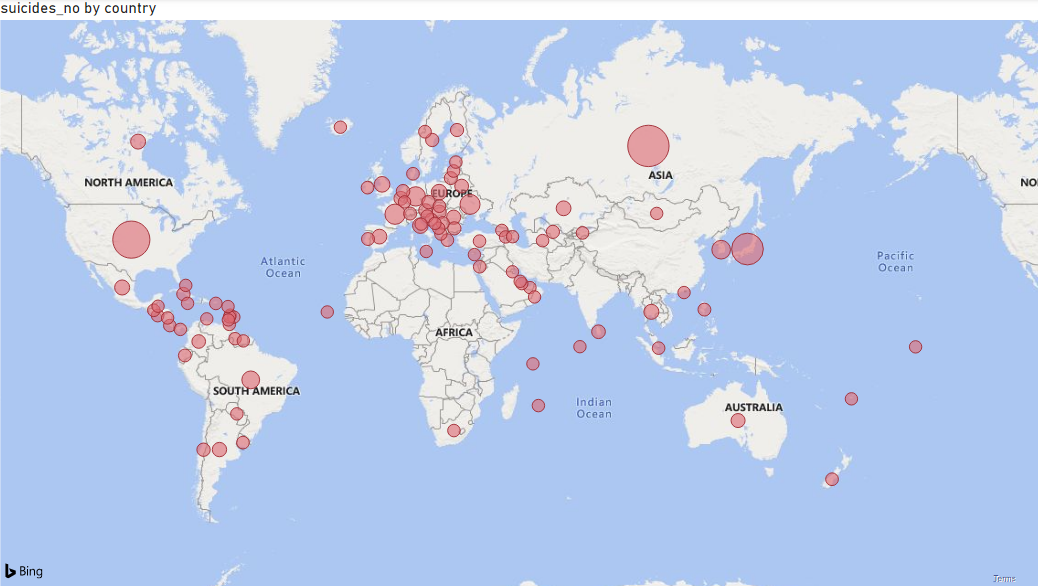


**Suicide Number and population by year**



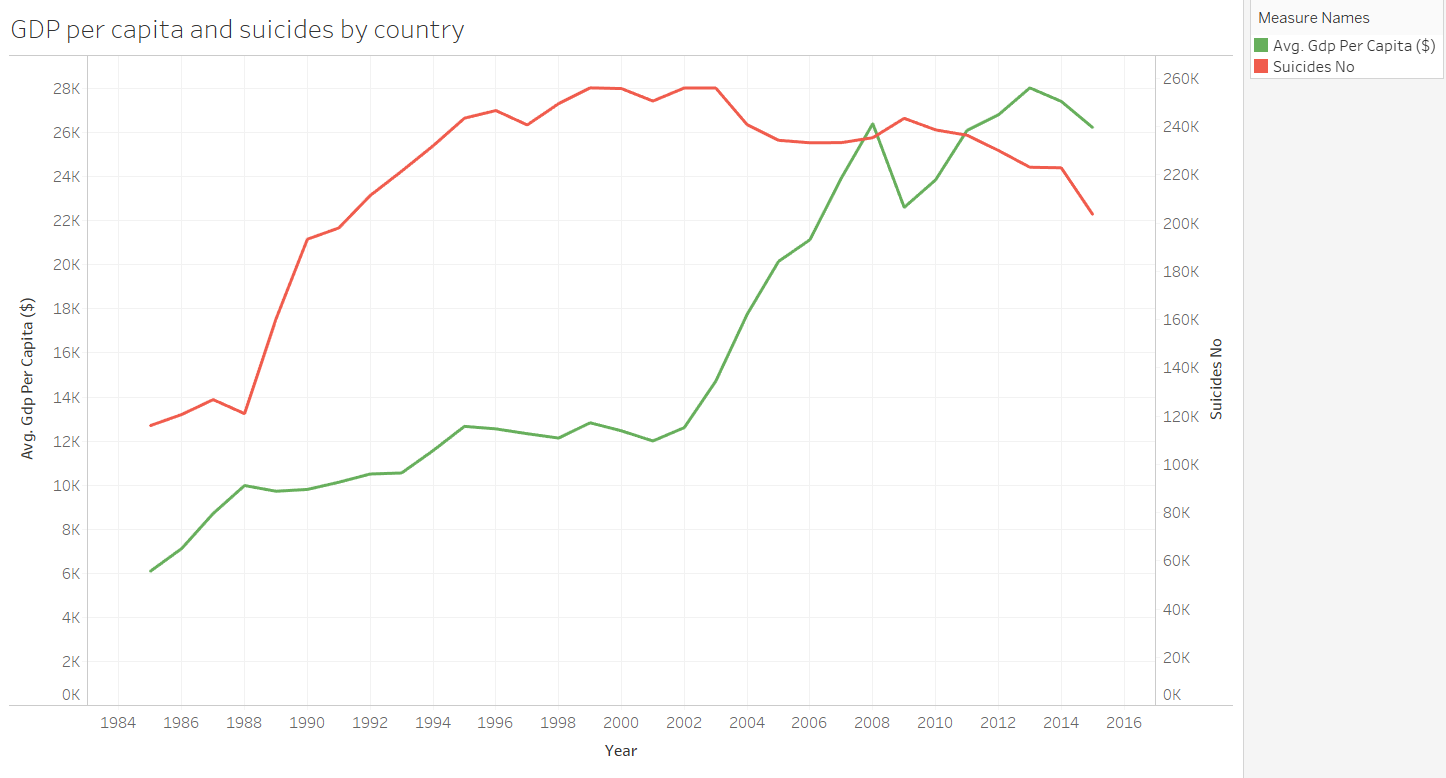
* This is a dual combination chart that shows the total number of suicides by year and the line represents the population over that period. The aim is to determine if population influences rising number of suicidal cases. As observed here, there is a significant increase in the number of suicide cases with the increase in population during the span of 30 years from 1985 to 2015.
* A general trend of more suicide cases in the country are registered if the population of the country itself is more. This might be reason for greater number of cases in countries like United States and Russia.

**Map of number of suicides by country**



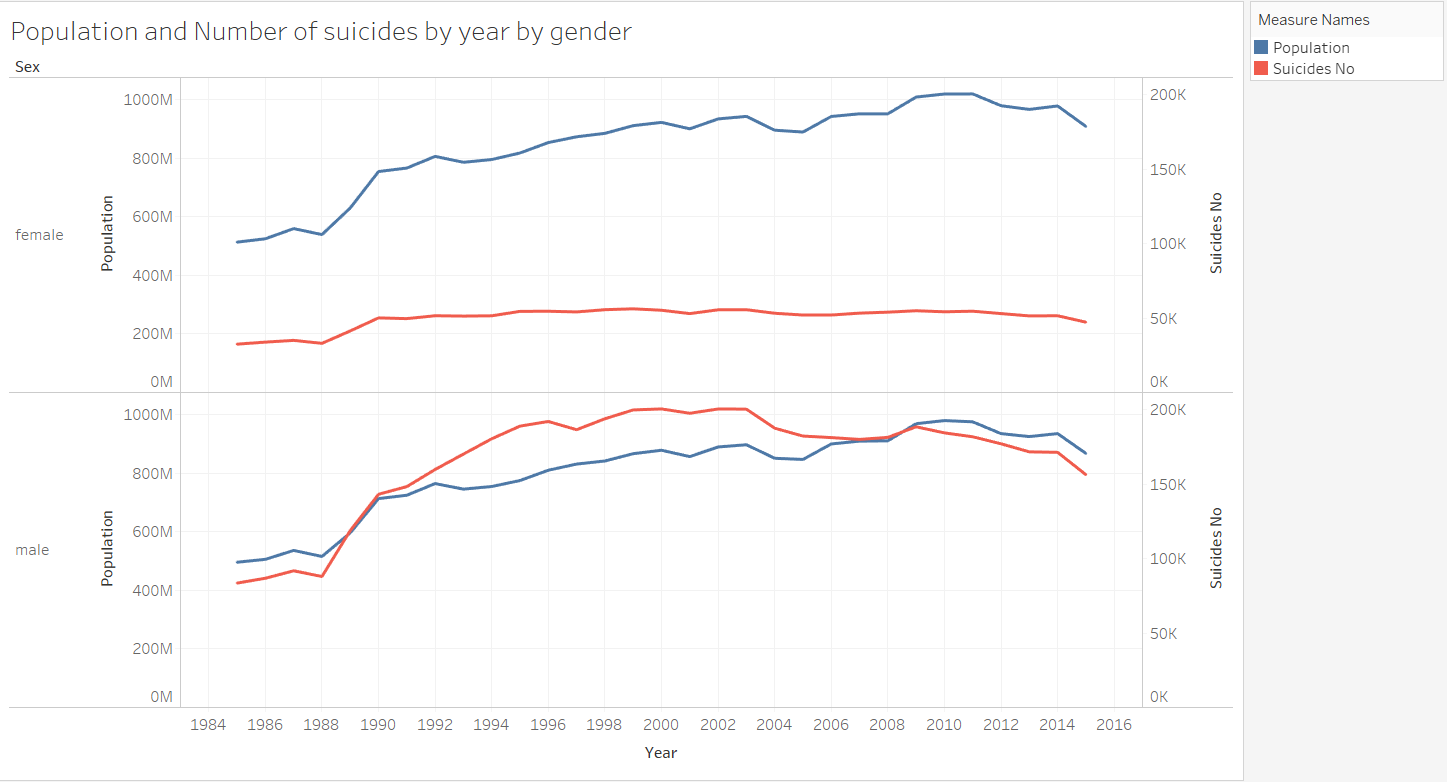
* This map shows the number of suicide cases registered by country. Here, the size of circles represents the number of cases in the country i.e., bigger circle means greater number of suicidal cases in that country. It can be clearly seen here that Russia, United States and Japan are the three countries with maximum number of cases.

**GDP per capita and number of suicides by country**



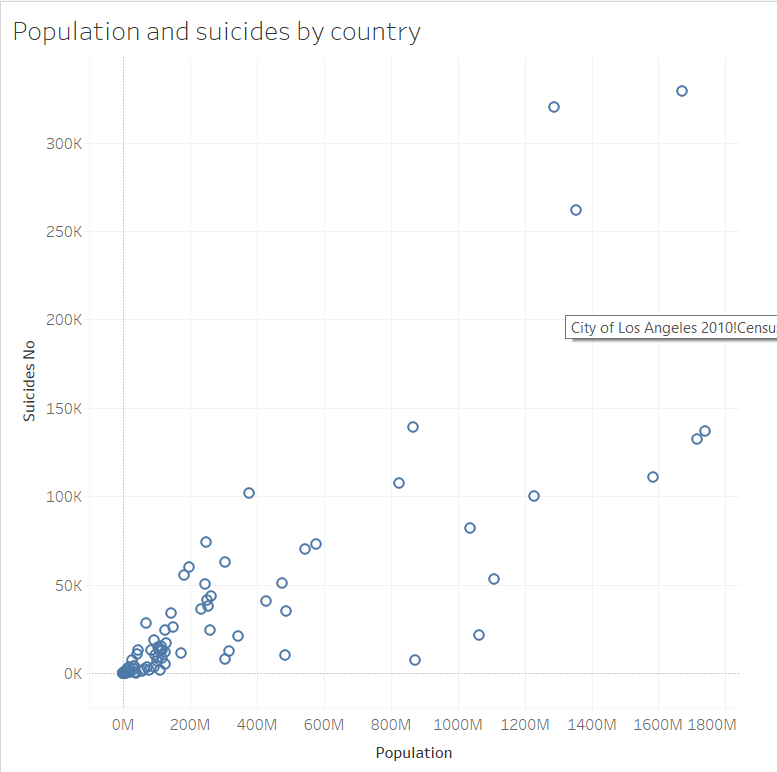
* The above is a Dual Lines chart. The green line shows the average GDP per capita by year while the red line represents the total number of suicide cases in that year. Observing the graph, we can say that there might me some positive relation of suicides and Gdp per capita that means that increasing number of suicide cases has a negative effect on GDP but no solid conclusion can be made.

**Populations and number of suicides by country by gender**



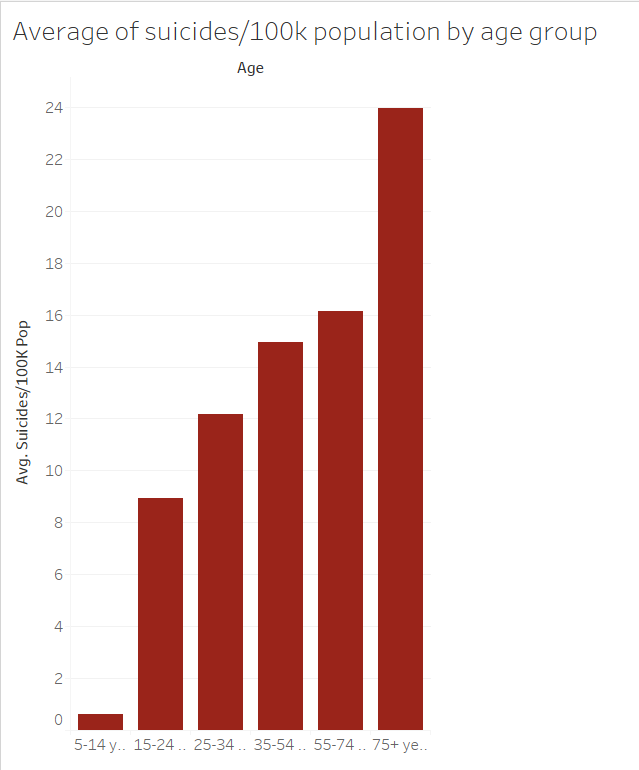
* This is a Line chart (discrete) with Dual Y axis representing population on the 1st axis (blue line) and number of suicide cases on the 2nd axis (red line) classified by gender. As seen here, population plays a major factor contributing to the number of suicidal cases.
* For males, increase in number of suicide cases seems to go hand in hand with population. With increase in population there is increase in the number of male suicidal rates while no such pattern is observed in case of females. The suicidal numbers for females look almost constant over time.

**Population vs Number of suicides by country**



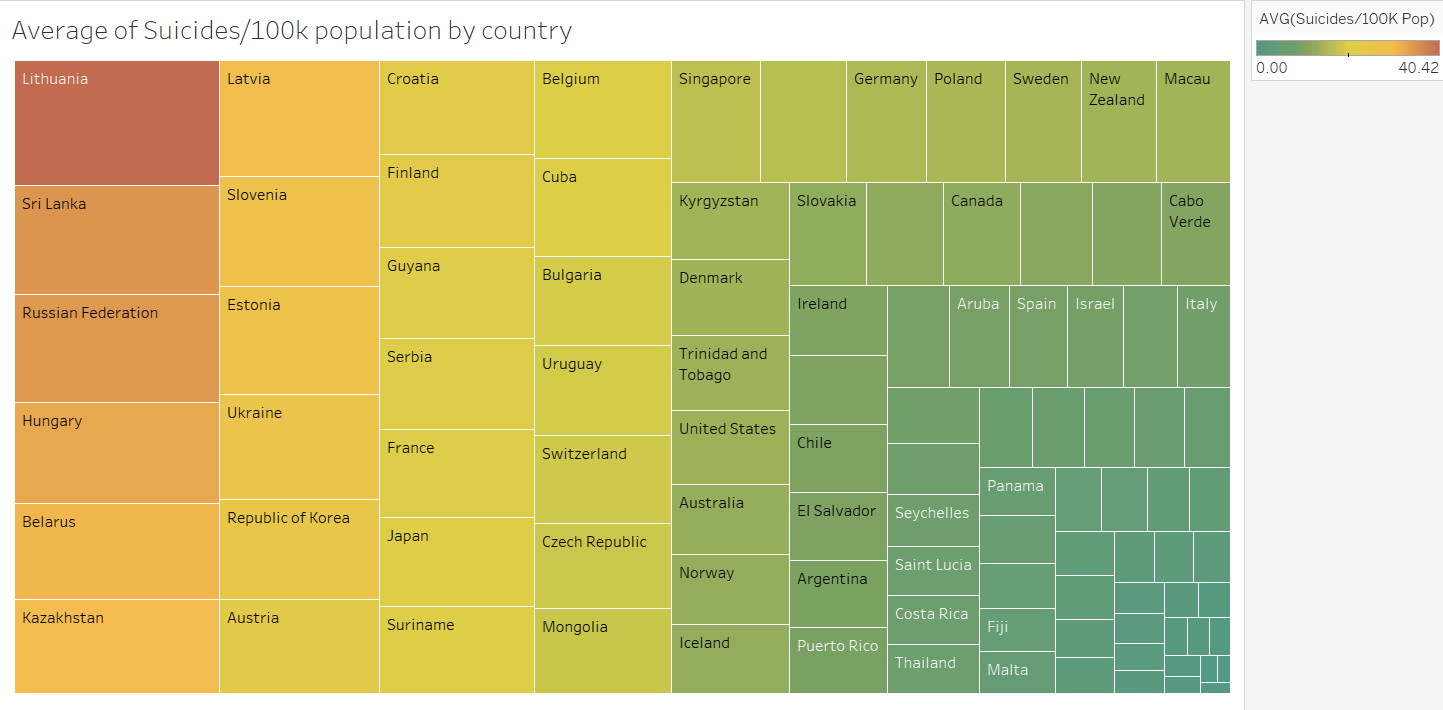
* This is a scatter chart between number of suicidal cases and population plotted to further justify the fact of relationship between population and number of suicides. This plot shows presence of positive correlation between the two variables that means that with increased population, there are more chances of having greater number of suicide cases.

**Average of suicides per 100k population by age**



* This is a bar chart of average number of suicide cases per 100k population classified by age. It is very clear from this graph that the average number of suicide cases increase with increase in the age. People above the age of 75 years of age are most prone to commit suicide as compared to the other age groups.
* We might say that as a person grows older, the more he/she comes across the harsh realities of life and looses the will to live. Like a wise man said, “Every problem has a solution”, thus with proper guidance, we can see a significant decrease in these numbers and save lives of many innocent people.

**Treemap of average suicides per 100k population by country**



* This is a tree map of average number of suicides per 100k population by country. Lithuania tops the list with an average of around 40 suicide cases per 100 k population while countries in the bottom right corner of the tree map like Turkey, Oman and Jamaica has the lowest average number of suicide cases per 100k population.

**Takeaways**

* Country’s population plays a major factor contributing to the number of suicidal cases in that country
* Lithuania has the most number of suicide cases per 100k population
* Russia tops the list for most number of total suicide cases with more than 1,200,000 cases during the period
* People above the age of 75 years are most prone to commit suicide as compared to the other age groups
* Male have a higher tendency of committing suicide than female in general

**Warning signs that someone may attempt suicide**



**Recommended 3-Step Process To Be Followed**

1. Suicidal thoughts and behaviors should be considered a psychiatric emergency.
2. If you or someone you know is exhibiting either, you should seek immediate assistance from a healthcare provider.
3. If you’re concerned and don’t know what to do, you can get help from a crisis or suicide prevention hotline.

**Appendix**

Below is a summary of R code used for the purpose of exploratory data analysis:

setwd("C:/Users/avman/OneDrive/Desktop/excel files")

suic = read.csv("DV suic data.csv")

sum(is.na(suic[,9]))

suic = suic[,-9]

sum(is.na(suic[,9]))

mean(suic$suicides\_no, na.rm = TRUE)

median(suic$suicides\_no)

range(suic$suicides\_no)

quantile(suic$suicides\_no)

newdata = subset(suic & year >1999 & ï..country != c("United States", " Russian Federation"))

boxplot(suicides\_no ~ year , data=suic, xlab="Years", ylab="Suicides")

boxplot(suicides\_no ~ year , data=newdata, xlab="Years", ylab="Suicides")

par(las=2)

bp = boxplot(suicides\_no ~ ï..country , data=suic, xlab="Countries", ylab="Suicides", show.names = F)

axis(1, at=seq(length(bp$names)),

labels=bp$names,

cex.axis=0.50)

par(las=2)

bp1 = boxplot(suicides\_no ~ ï..country , data=newdata, xlab="Countries", ylab="Suicides", show.names = F)

axis(1, at=seq(length(bp$names)),

labels=bp$names,

cex.axis=0.5)

names(suic)

boxplot(suic$suicides\_no)

hist(suic$suicides\_no)

library(dplyr)

?group\_by

bycountry1 = group\_by(suic, ï..country)

newb2=summarize(bycountry1, sum(suicides\_no), sum(population))

newb3 = as.data.frame(newb2)

plot(newb3$`sum(population)`,newb3$`sum(suicides\_no)`)

bycountry = group\_by(newdata, ï..country)

newb=summarize(bycountry, sum(suicides\_no), sum(population), sum(suicides.100k.pop))

class(newb)

newb1 = as.data.frame(newb)

plot(newb1$`sum(population)`,newb1$`sum(suicides\_no)`, xlab = "Population", ylab = "Suicides")

cor(suic[,5:6])

max(newb1$`sum(suicides.100k.pop)`)

match(7660.64,newb1$`sum(suicides.100k.pop)`)

newb1$ï..country[52]

bysex = group\_by(newdata, sex)

newb5=summarize(bysex, mean(suicides\_no), mean(suicides.100k.pop))

newb6 = as.data.frame(newb5)

barplot(newb6$`mean(suicides.100k.pop)`, main = "Suicides per 100k population by gender", xlab = "Gender", ylab = "Suicides per 100k")

hist(newb1$`sum(suicides\_no)`, xlab = "Total suicides", ylab = "No. of countries")