**UNIVERSITY OF COLORADO DENVER**

**Exploratory Data Analysis with Project Data in R**

ASSIGNMENT – 10

BANA - 6800

Avi Manawat (student id: 109638446)

**Abstract:**

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. 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.

Initially, data cleaning and preprocessing is done in order to deal with the missing and anomalies in the data. I have used Excel for data preparation. Next step was looking out for outliers in the data which may impact visualization.

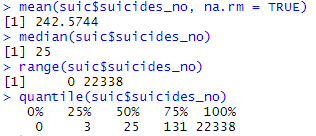
**Details**



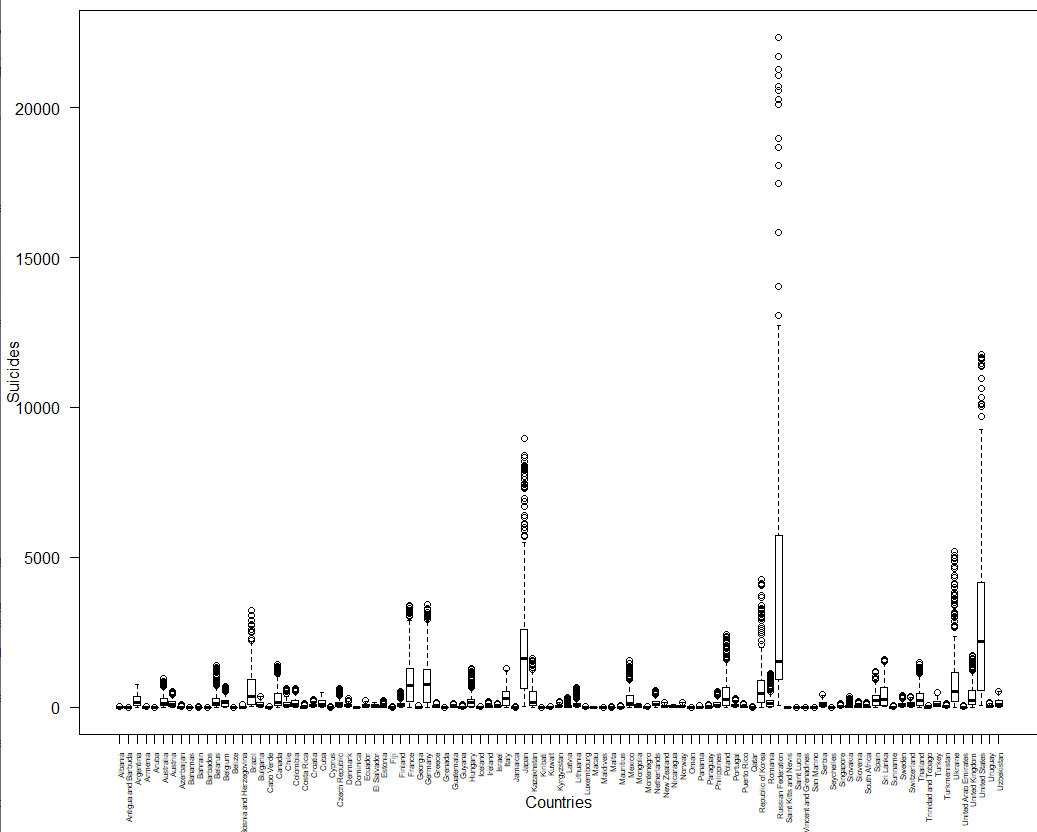
* First step was to identify the number of Nan values present in the data in order to deal with them.



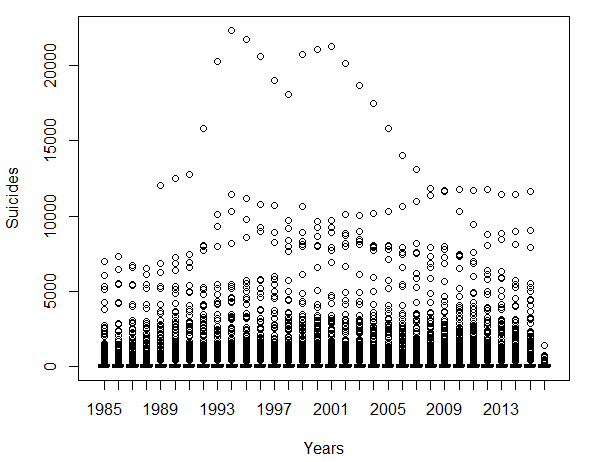
* On analyzing, it was found that all these Nan values were present in a single column. The whole column was not taken into consideration in order to get a clean data for analysis



* 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.



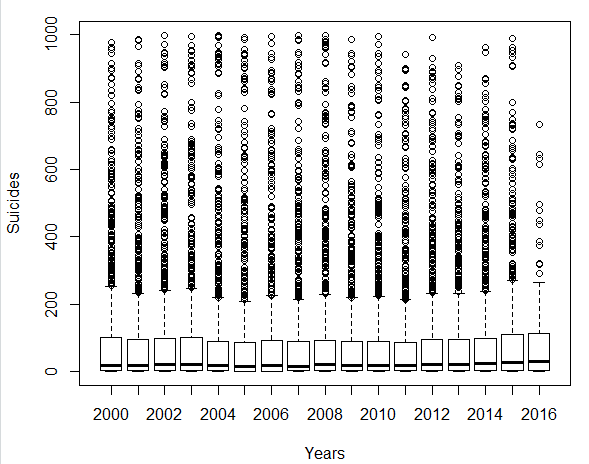
* Plotting a country wise box plot of the suicidal data clearly showed the presence of outliers in the data (seen in the above figure) in the form of countries like United States and Russian Federation that hindered the process of data visualization.



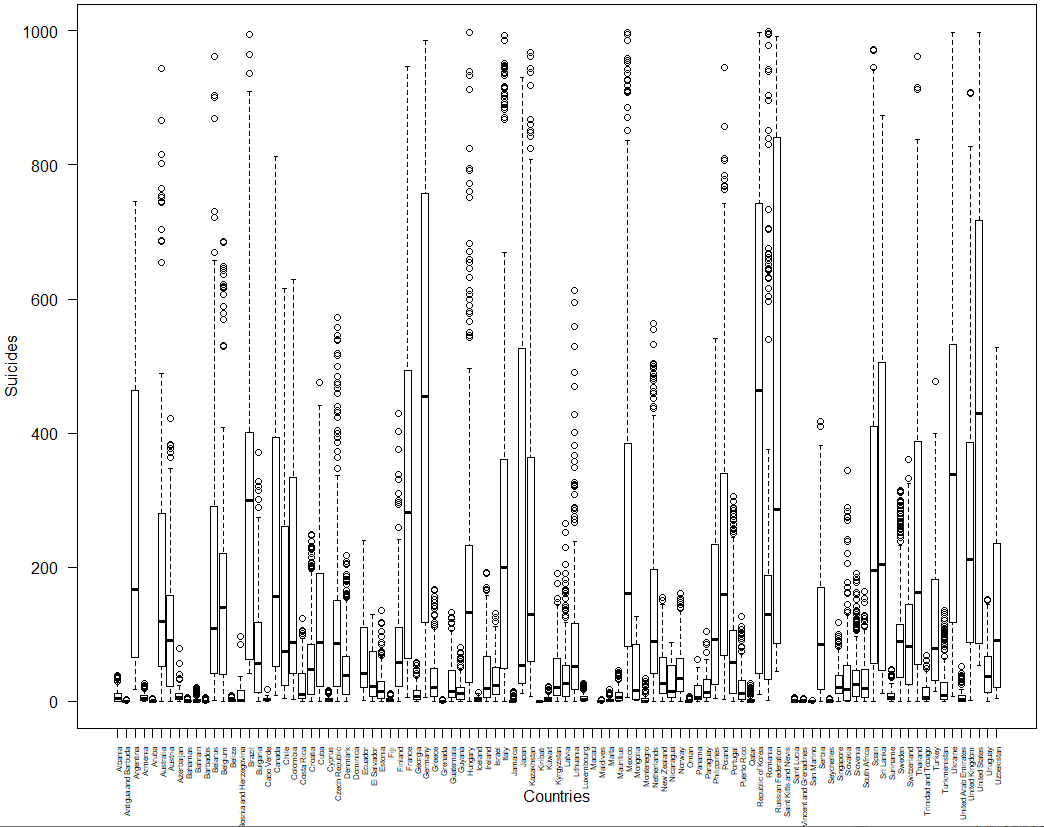
* These are the year wise boxplots of the suicidal data per age group. This graph of the original data is depiction of how the outliers affected the analysis.



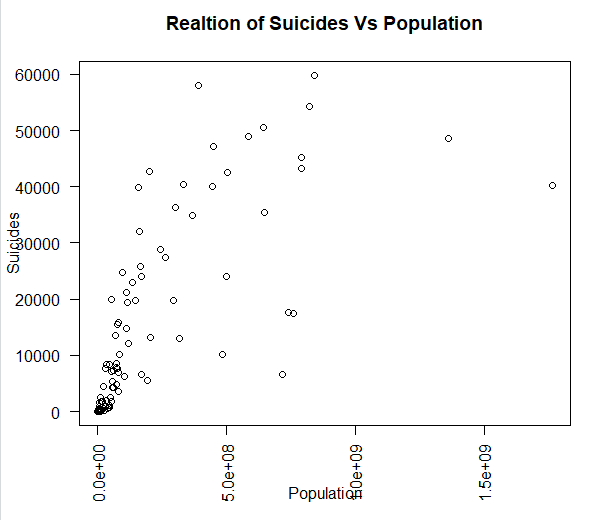
* A subset of the data was created by using the above code and it was decided to treat the above mentioned countries separately for the purpose of analysis for better visualization.



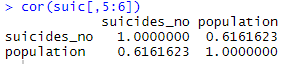
* After dealing with the outliers, the boxplots were clearly visible and interesting information can be drawn from the graph. Year 2007 had the least suicidal death mean per age group while the mean of suicidal deaths per age group in all the countries kept increasing since the year 2009.



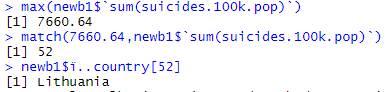
* The above is a country-wise depiction of suicidal deaths per age group. The number of deaths vary largely depending on the country. For ex., country like Albania has really low suicidal death stats as compared to the other countries. One possible reason for this could be low population of the country.



* It was assumed that population would play as a big factor contributing in the number of suicidal deaths of a country.
* For this reason, the above scatterplot was prepared to look at the relationship between population of the country and number of suicidal deaths in the country. The plot indicates some positive correlation between the two factors



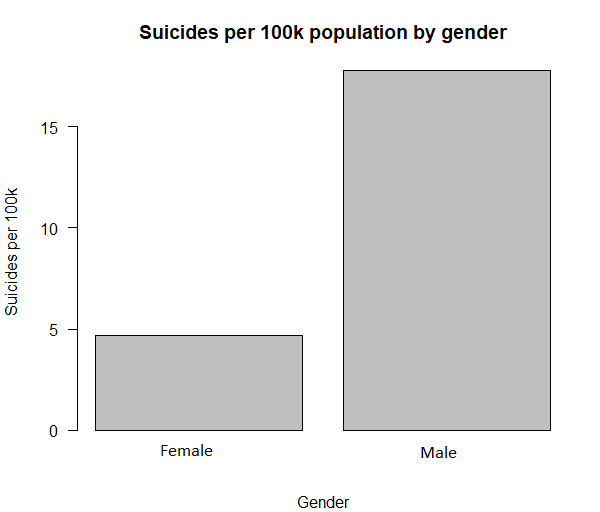
* The correlation between the two is 0.61 which further validated the above statement. But since the correlation is not too strong, there are other factors as well affecting the number of suicides in the country.

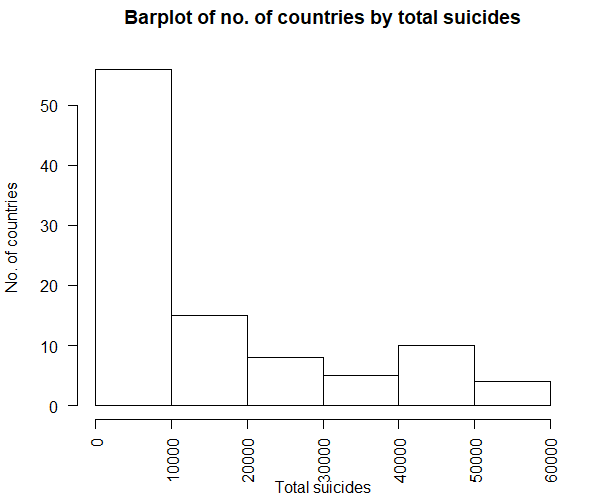


* The maximum cases of suicide per 100k population of the country was found to be in Lithuania. There were 7660.64 suicide cases per 100k population in this country, which was the highest.

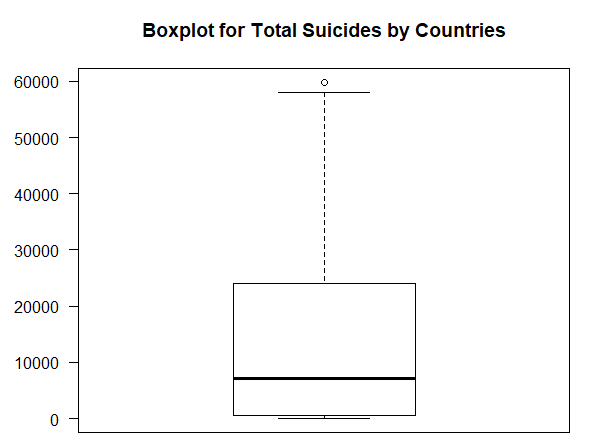


* Of all the suicide cases throughout the world, the cases of males committing suicide were four times more than female. Below is a representation of this fact





* 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.



The mean of suicide cases among the given countries is around 8200 with the highest being 60000 as depicted by the above boxplot.

**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")