FIT5147 Data Exploration Assignment - 1

Jaimon Thyparambil Thomas
Student ID : 29566428
Email : jthy0001@monash.student.edu
Monash University
April 29, 2019

Contents

Introduction

Every year approximately 800000 people die due to suicide which is approximately 1 person per second. There is also sources which say that for every person who has committed suicide there is approximately 20 persons who attempted and failed. Suicides happen globally across all age groups, gender etc. at different rates. Suicide is a serious problem worldwide but it is one of the most preventable when comparing to other cause of death.

Here I have analysed the suicide rates based on various factors to answer the questions give below:

1. Find patterns in suicide rates based on country, age, sex etc.
2. Check the influence of economic factors such as GDP over suicide rates
3. Check the influence of happiness rating over suicide rates for different countries

Data Source used for analysing these question

```
Suicide related data source for years 1985 - 2016
Type - tabular data (CSV)
URL - https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
Spatial Info - 27.8k rows x 12 columns

Country wise happiness rating data source for years in 2015 - 2017
Type - Tabular data (CSV)
URL - https://www.kaggle.com/unsdsn/world-happiness#2017.csv
Spatial Info - 2015 - 158 rows x 12 columns
                    2016 - 157 rows x 13 columns
```

Data Wrangling and Checking

The suicide rates data from 1985 – 2016 looks like

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

```
Country                object
Year                    int64
sex                    object
age                    object
suicides_no             int64
population              int64
suicides/100k pop     float64
country-year           object
HDI for year          float64
 gdp_for_year ($)      object
gdp_per_capita ($)      int64
generation             object
dtype: object
```

Figure 1

On examination of data types of suicide rates table (figure 1), GDP for a year is understood to correspond of data type object

Further on examination it was understood that the values of GDP have ‘,’ in between the numbers (as in Figure 2). So we will remove the ‘,’ and convert the character into respective numeric value.

```
0    2,156,624,900
1    2,156,624,900
2    2,156,624,900
3    2,156,624,900
4    2,156,624,900
Name:  gdp_for_year ($) , dtype: object
```

Figure 2

```
print("Unique gdp for Year count:",len(main[' gdp_for_year ($) '].unique()))
print("Unique gdp per capita count:",len(main['gdp_per_capita ($)'].unique()))
```

```
Unique gdp for Year count: 2321
Unique gdp per capita count: 2233
```

	Year	suicides_no	population	suicides/100k pop	HDI for year	gdp_for_year ($)	gdp_per_capita ($)
0.1	1989.0	0.0	17303.3	0.000	0.648	1.658688e+09	1524.0
0.2	1993.0	1.0	56100.0	0.410	0.692	5.726898e+09	2733.0
0.3	1996.0	4.0	170110.5	1.600	0.724	1.280345e+10	4117.0
0.4	1999.0	11.0	293370.4	3.540	0.753	2.443288e+10	6166.0
0.5	2002.0	25.0	430150.0	5.990	0.779	4.811469e+10	9372.0
0.6	2004.0	49.0	646413.6	9.090	0.814	9.962714e+10	14193.0
0.7	2007.0	94.3	1155718.6	13.563	0.837	1.938704e+11	21219.0
0.8	2010.0	190.0	2367638.4	20.530	0.872	3.647565e+11	28733.0
0.9	2013.0	496.0	4960713.5	33.291	0.897	1.002219e+12	43487.0

Figure 1

Further on examining it was understood that the no of unique values or GDP for year and for per capita is in range of 2000 which will make it difficult to analyse. So in order to make analysis process easier some new columns with values based on its quantile values as in Figure 3 is been added

	Country	Region	Happiness Rank	Happiness Score	Lower Confidence Interval	Upper Confidence Interval	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Denmark	Western Europe	1	7.526	7.460	7.592	1.44178	1.16374	0.79504	0.57941	0.44453	0.36171	2.73939
1	Switzerland	Western Europe	2	7.509	7.428	7.590	1.52733	1.14524	0.86303	0.58557	0.41203	0.28083	2.69463
2	Iceland	Western Europe	3	7.501	7.333	7.669	1.42666	1.18326	0.86733	0.56624	0.14975	0.47678	2.83137
3	Norway	Western Europe	4	7.498	7.421	7.575	1.57744	1.12690	0.79579	0.59609	0.35776	0.37895	2.66465
4	Finland	Western Europe	5	7.413	7.351	7.475	1.40598	1.13464	0.81091	0.57104	0.41004	0.25492	2.82596

The Happiness data for 2016 looks like figure in left

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201
2	Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.49204
3	Norway	Western Europe	4	7.522	0.03880	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	2.46531
4	Canada	North America	5	7.427	0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.45176

The Happiness data for 2015 looks like figure in left

Further on examination it is understood that in 2016 table standard error is given as a range of confidence interval i.e. Upper confidence interval and the lower confidence interval. So in order to maintain consistency a new column in 2016 with name Standard Error which is actually the difference between upper and lower confidence interval has been added. After that those two columns have been removed.

Further on examining we can see that some countries present in 2015 is not present in 2016 table and vice versa.

```
{'Central African Republic',
 'Djibouti',
 'Lesotho',
 'Mozambique',
 'Oman',
 'Somaliland region',
 'Swaziland'}
```

Countries Present in 2015 but not in2016

```
{'Belize',
 'Namibia',
 'Puerto Rico',
 'Somalia',
 'Somaliland Region',
 'South Sudan'}
```

Countries Present in 2016 but not in2015

Now we will remove all countries present in 2015 and not present in 2016 and vice versa.

After that a new column Year is been added into both these tables with the respective year. After this the two tables have been merged into a single table

	Dystopia Residual	Economy (GDP per Capita)	Family	Freedom	Generosity	Happiness Rank	Happiness Score	Health (Life Expectancy)	Standard Error	Trust (Government Corruption)	Year
0.1	1.450509	0.345612	0.600457	0.189262	0.088571	15.7	3.8723	0.233309	0.037800	0.029332	2015.0
0.2	1.649935	0.557800	0.756392	0.257657	0.126916	30.4	4.2940	0.357696	0.045718	0.047980	2015.0
0.3	1.813638	0.740523	0.867657	0.325728	0.169016	47.0	4.6774	0.514723	0.066145	0.063995	2015.0
0.4	1.942728	0.889179	0.952936	0.379380	0.195906	62.0	5.0932	0.592166	0.134000	0.078564	2016.0
0.5	2.083533	1.008205	1.034090	0.415895	0.220263	78.0	5.2980	0.637660	0.154000	0.093665	2016.0
0.6	2.207530	1.107824	1.107134	0.452073	0.248132	93.0	5.7684	0.686386	0.168513	0.118048	2016.0
0.7	2.326653	1.198268	1.202717	0.490468	0.287110	109.9	5.9870	0.731061	0.183648	0.146578	2017.0
0.8	2.518948	1.313313	1.274386	0.543748	0.333098	125.0	6.4798	0.804462	0.199800	0.183430	2017.0
0.9	2.795270	1.440177	1.396831	0.586455	0.430668	141.0	6.9860	0.851783	0.229720	0.306678	2017.0

Figure 3

After that here also new the columns have been added with values as a range based on their respective quantile values as in Figure 3

On Further examining it was understood that some of the country name in happiness data is different from the value present in the suicide data like those in figure 4. Those values have been fixed in suicide data with the country name from happiness data table.

```
Country Name in happiness Data --  Suicide Data
Dominican Republic -- Dominica
South Korea -- Republic of Korea
Russia -- Russian Federation
```

Figure 4

After that both the tables i.e. suicide data and the happiness ranking table have been merged on inner join based on columns like country and year

On Further examining if there are any null values in any of the columns it is understood that the columns HDI has null values (From Figure 5).

```
Country                          False
Year                             False
sex                              False
age                              False
suicides_no                      False
population                       False
suicides/100k pop                False
country-year                     False
HDI for year                      True
 gdp_for_year ($)                False
gdp_per_capita ($)               False
generation                       False
gdp_block($1e9)                  False
gdp_per_capita_block($1e9)       False
Dystopia Residual                False
Economy (GDP per Capita)         False
Family                           False
Freedom                          False
Generosity                       False
Happiness Rank                   False
Happiness Score                  False
Health (Life Expectancy)         False
Region                           False
Standard Error                   False
Trust (Government Corruption)    False
dystopiaResidualBlock            False
familyBlock                      False
freedomBlock                     False
generosityBlock                  False
happinessRankBlock               False
happinessScoreBlock              False
trustBlock                       False
healthBlock                      False
dtype: bool
```

Figure 5

On further analysis it is understood that now HDI only has null values. So that column has been removed from the final table

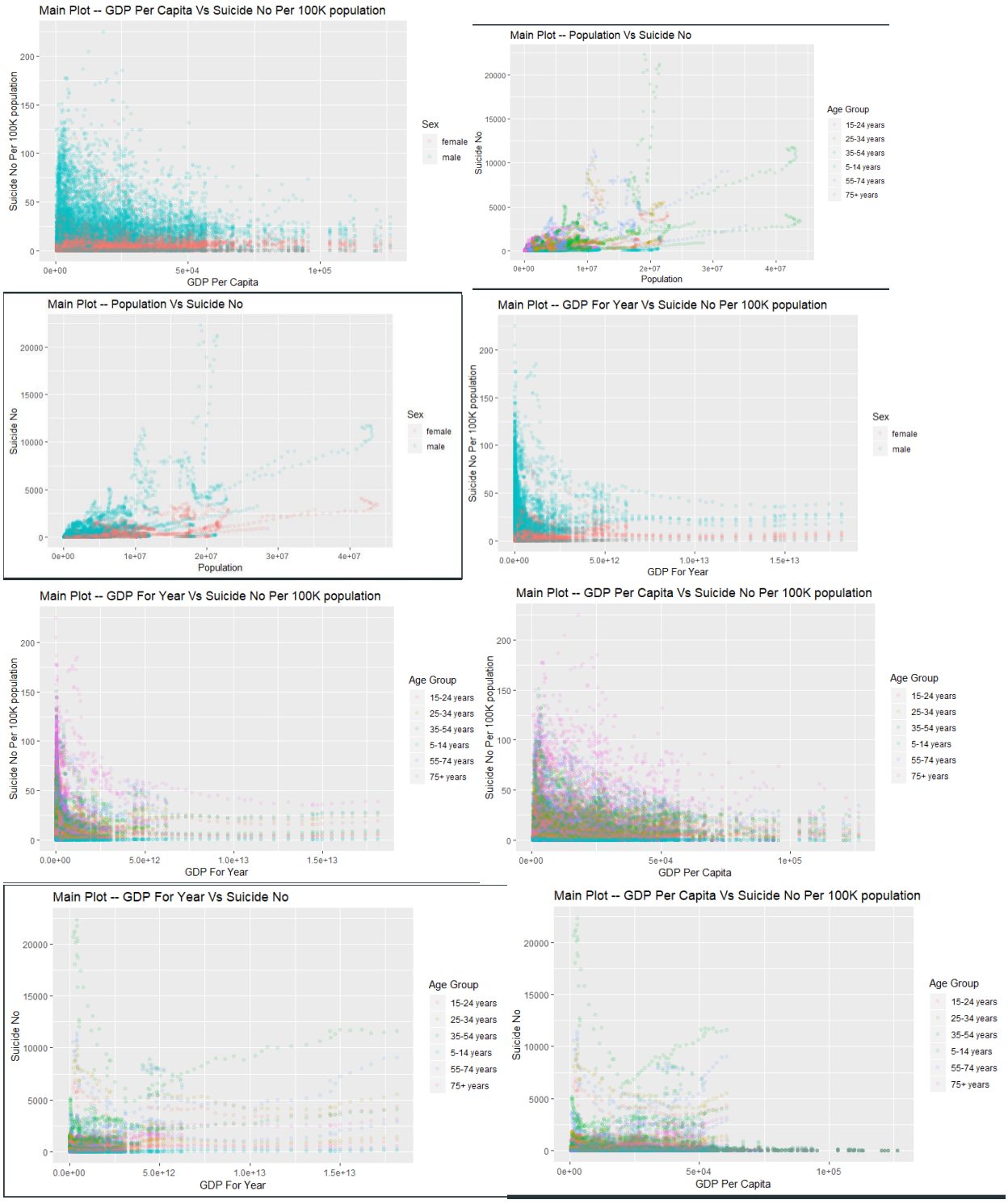
```
overalInfo['HDI for year'].unique()
```

```
array([nan])
```

Figure 6

Data Exploration

Initially to get a general idea regarding the data few of the generic plots is been plot



Since not much insight is being got about the trends related to suicide rates from these figures other options are been used like

Plotting a country wise heat map for suicide no as in Figure 7. From this figure it pretty much evident that countries like Russia, USA has more suicides rates. But what if the no of suicide rates is more since the population is more in these countries.

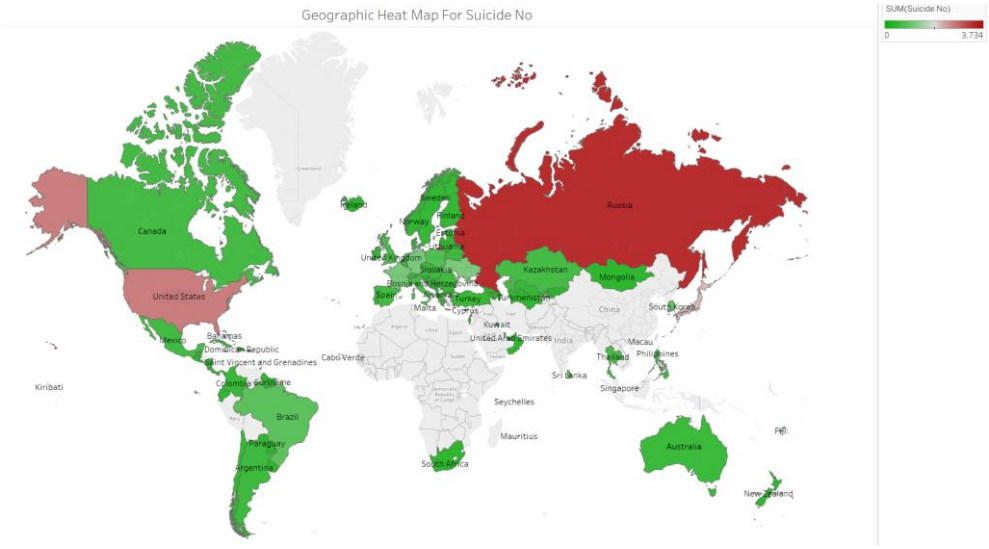
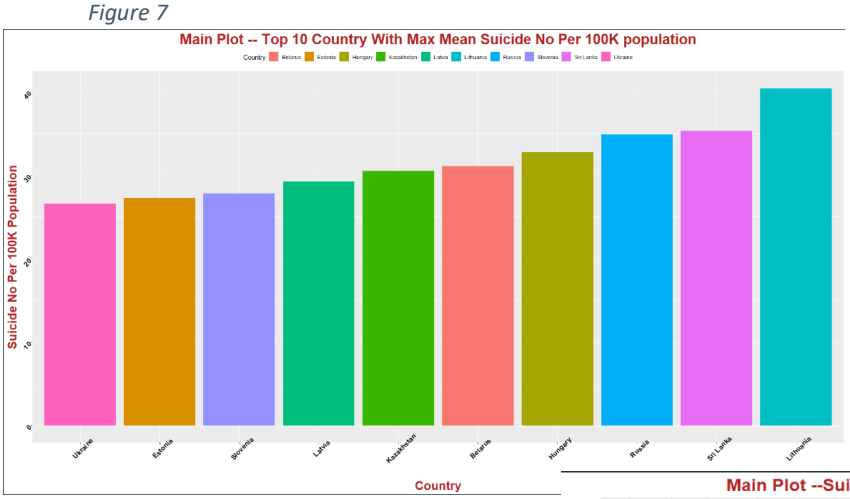
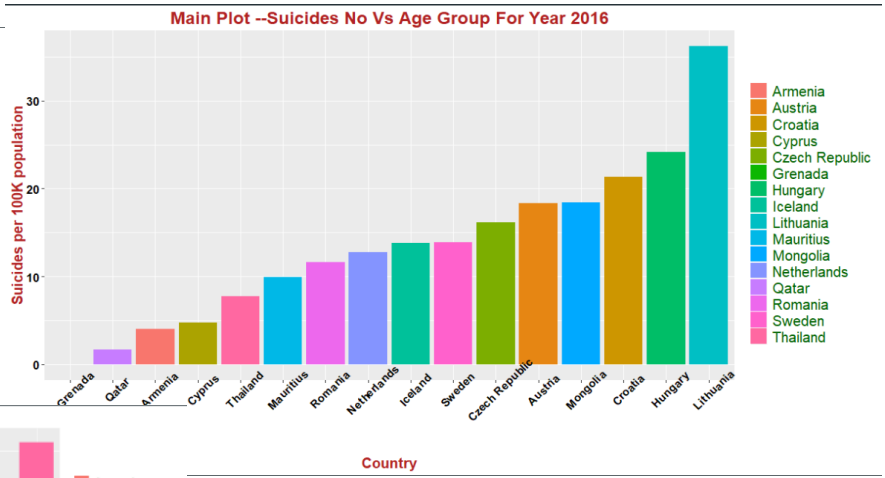


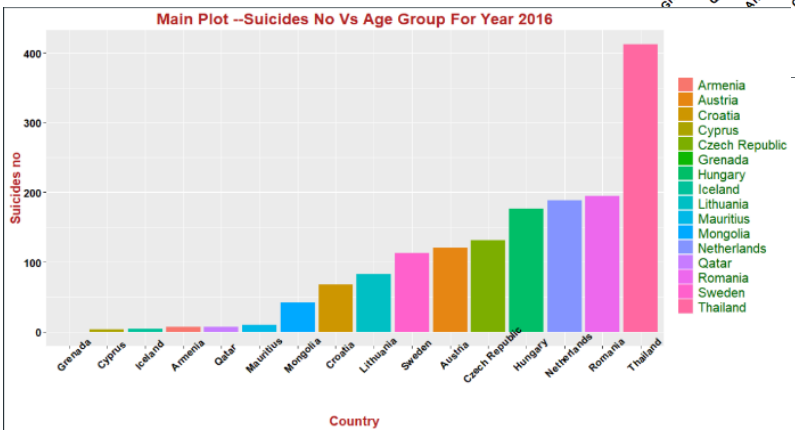
Figure 7



From this Figure It clear that the country with max mean suicide no per 100K population as per the data we have is Lithuania

From the figure in the Right side it is clear that the Country that has to be most concerned as per the year 2016 based on suicide per 100K population is Lithuania





From the figure in the Left side it is clear that the Country that has to be most concerned as per the year 2016 based on no of suicides is Thailand

For further examination relation between factors such as Age group and its respective mean suicide rates is being analysed.

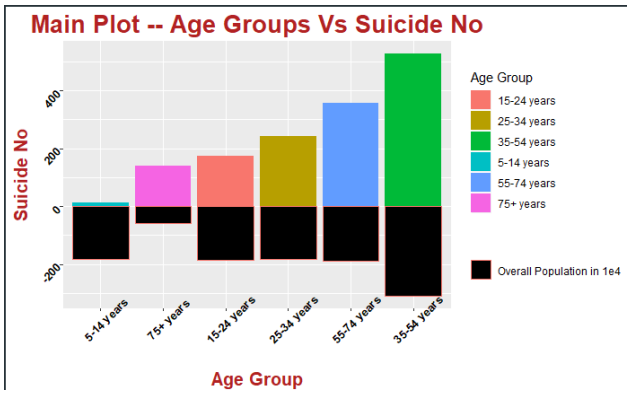


Figure 8

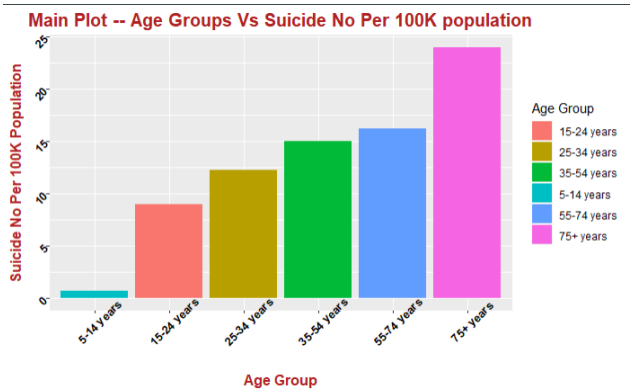


Figure 10

From Figure 9 it is understood that age group 35-54 has the highest no of suicides. It is also the age group with max population. From Figure 10 it is understood that while considering the total no of population for a particular age group and its respective no of suicides. The maximum number of suicides are for the age group 75 years plus. In both these cases it is also evident that the least no of suicides are for kids that is between age 5 – 14. From Figure 10 it is also clear that the as the age group keeps on increasing the no of suicide rates also keeps on increasing which shows people are less willing to live as they keep on growing which is a serious issue.
For further analysis, relation between factors such as Sex and year and its respective mean suicide value is being examined

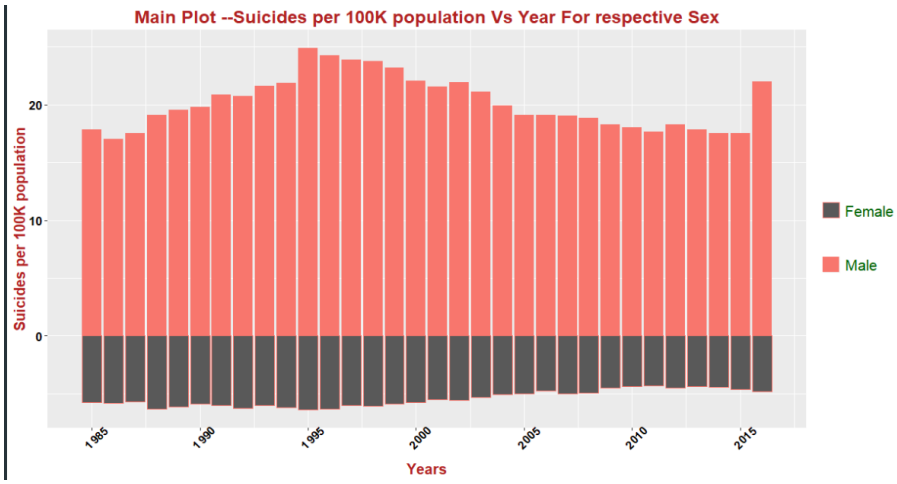


Figure 11

From Figure 11 it is pretty much clear that the no of suicide done by male is far more than that done by women. From this figure it is also understood that the year which had maximum number of suicide is 1995. It also evident that there is a sudden spike in the number of suicide rate for the year 2016 which is a serious threat that has to be handled immediately.

Figure 12

From figure 12 it is pretty much clear that the no of suicide rates is decreasing over the for both males and females which is a good thing.

Now Let's Analyse the influence of GDP over the suicide rates

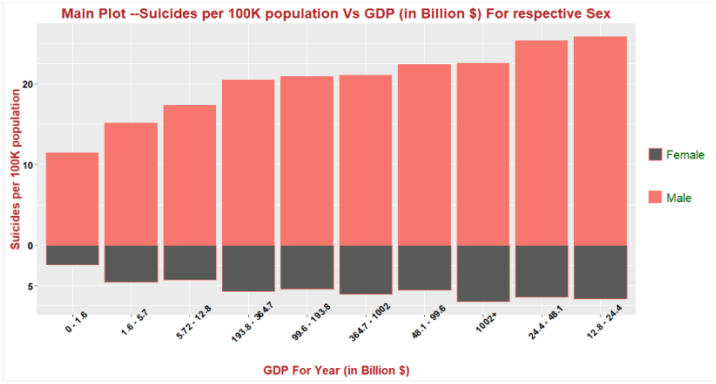


Figure 13

From figure 13 nothing is been understood as it is a mix of values. Maybe it's because the countries with higher GDP has more population which might have influenced those values. So in order to verify that we will examine GDP with respect to suicide no.

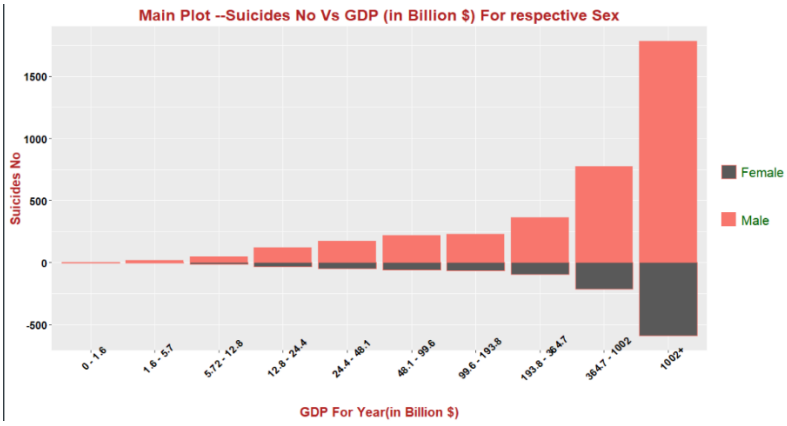


Figure 14

From the Figure 14 it is pretty much clear that as the GDP of the country is being increasing the no of suicide rate is also being increased.

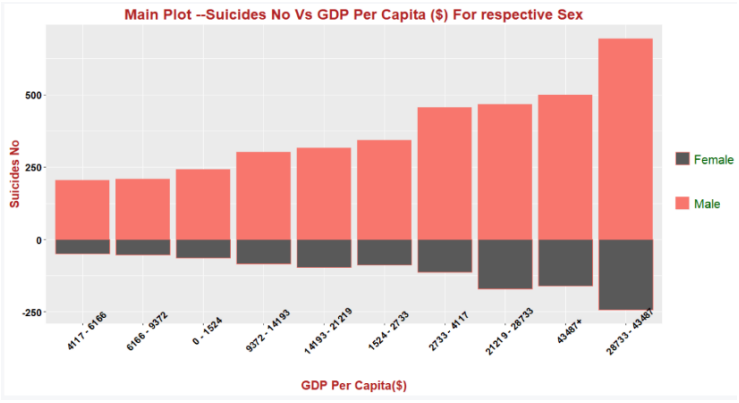


Figure 15

From the figure 14 we can also notice that the most no of suicides are for countries with higher GDP per capita.

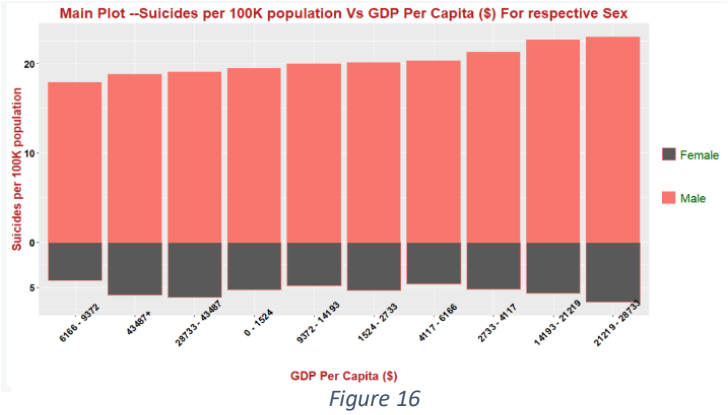


Figure 16

From Figure 16 we can notice that the no of suicides is higher for counties with higher GDP per capita.

Now Let's analyse the influence of Happiness ranking factors along with the no of suicides

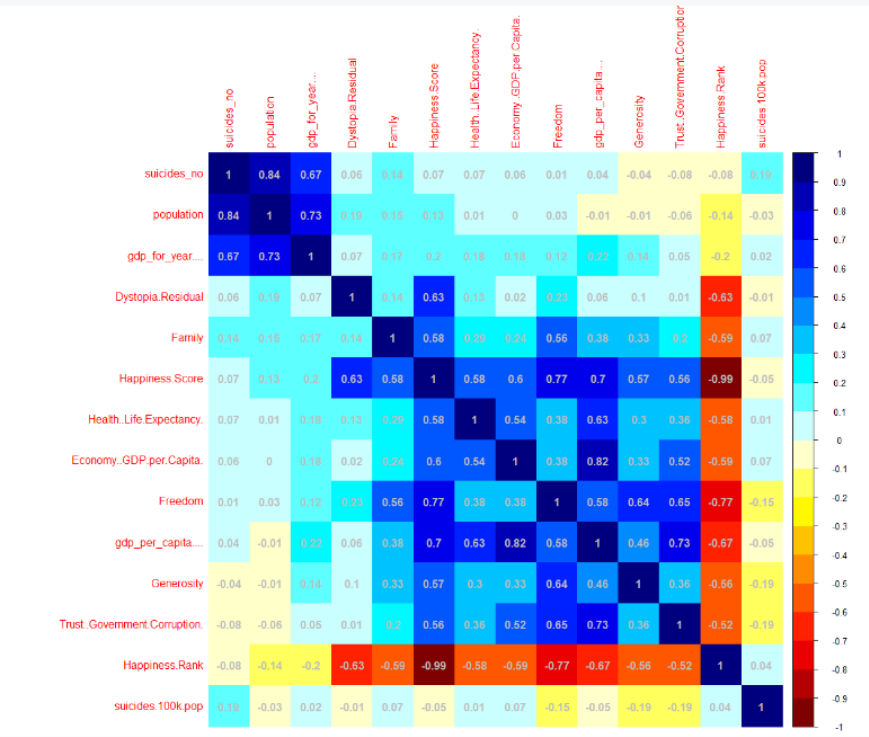


Figure 17

From figure 17 we can see that in general there is not a significant amount of correlation between the happiness ranking or its factor along with the suicide no or suicides per 100k population as all the values are within the range of -0.2 to 0.2. Out of these values which have comparatively more significance is generosity and trust so let's examine that

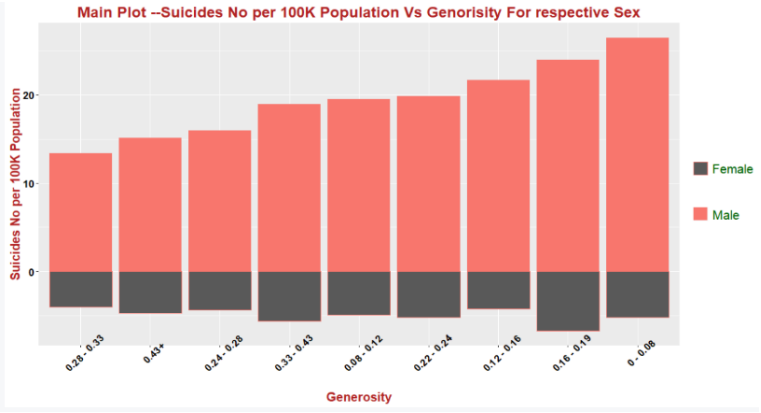


Figure 18

From Figure 18 it is looks like most of suicides occur when the value of generosity is low

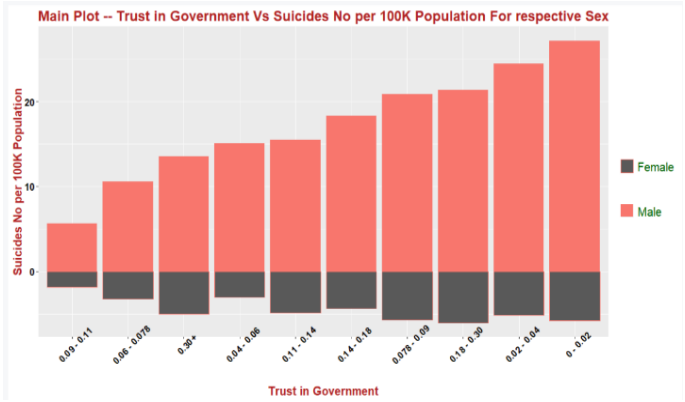
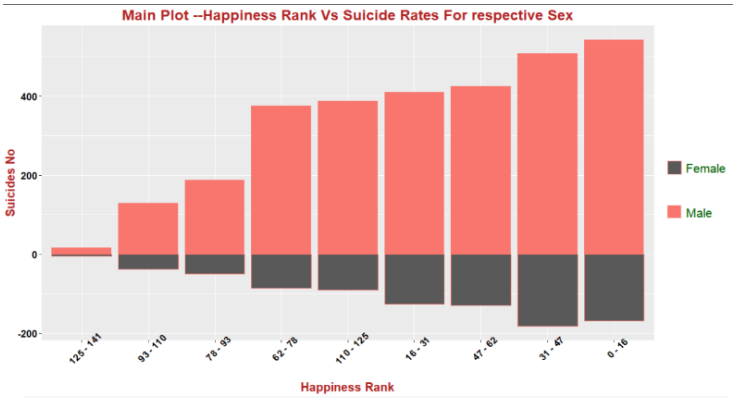


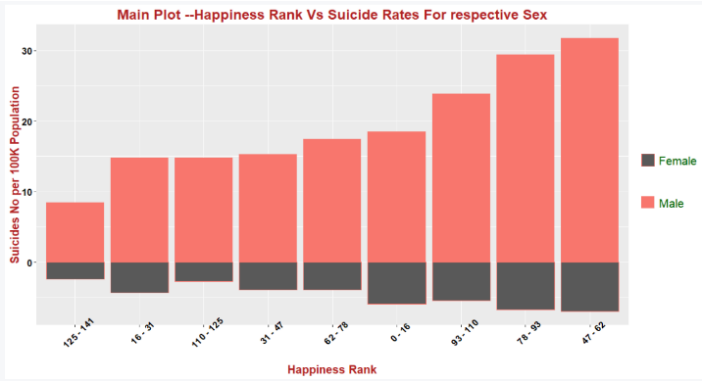
Figure 19

From figure 19 it looks like most of the suicides occur when the value of Trust in government is low

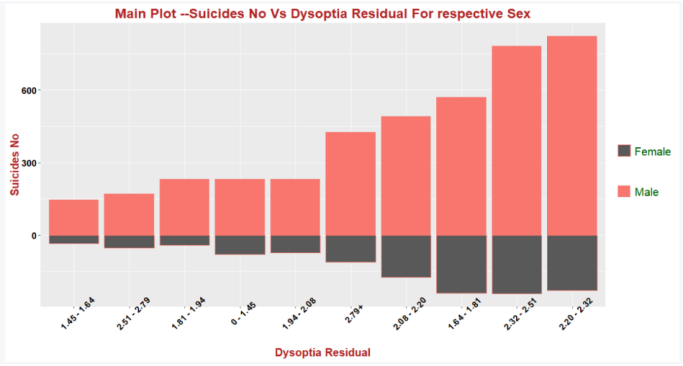
From figure 19 it looks like most of the suicides occur when the value of Trust in government is low



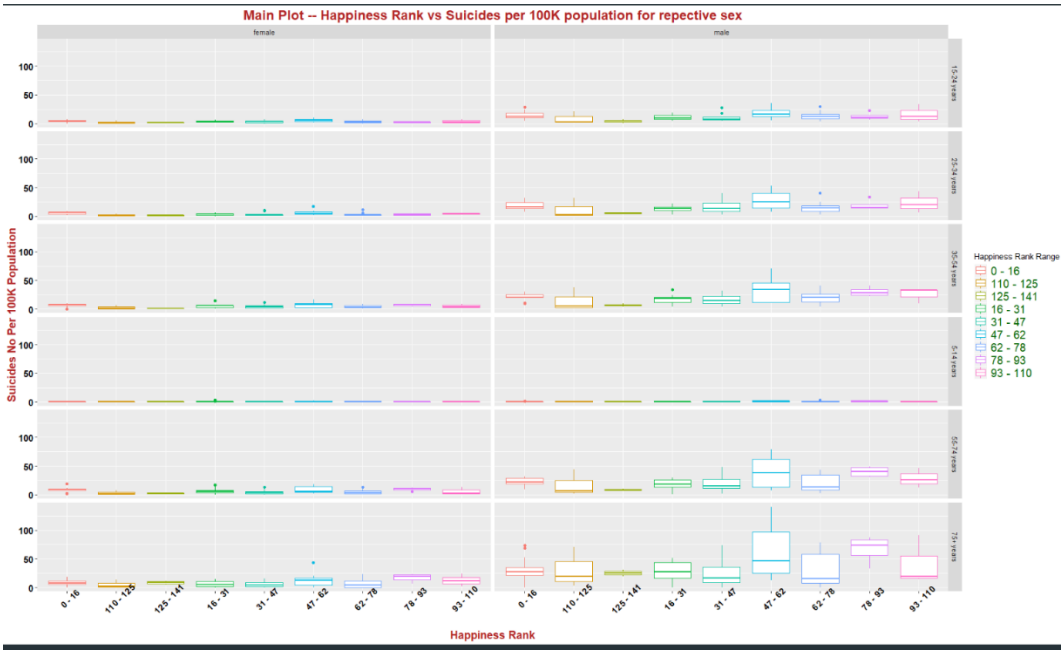
From figure in the left it looks like most of the suicides occur when the value of Happiness rank is low



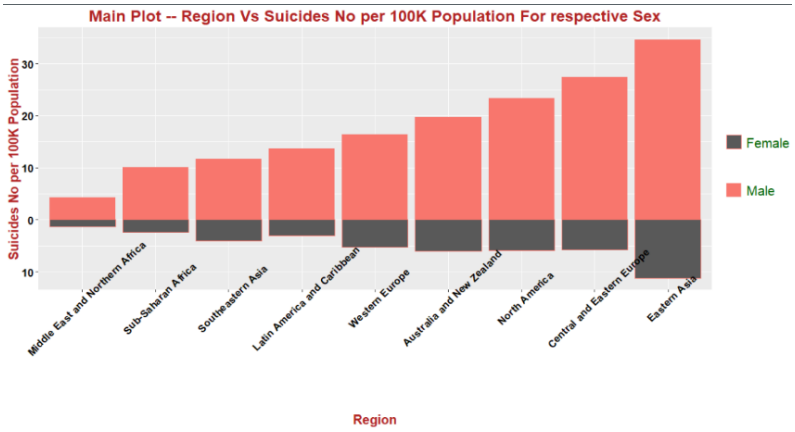
From figure in the left it looks like most of the suicides with respect to 100K population occurs when the value of Happiness rank is between 47 – 62



From the figure in the left it shows that the no of suicides is max for higher values of Dysoptia residual



From the figure in the left it shows that for almost all age groups for both male and female the max no of suicides per 100 K population is for the countries with happiness rank between 47 - 62



From the figure in the Left we can see that the region with max no suicides per 100K population for both males and females is in Eastern Asia

Conclusion

The conclusions made from the above analysis are

- Countries with max mean suicide no are Russia, US, Japan etc.
- Country that has max suicide no as of year 2016 is Thailand
- Country that has max suicide no per 100K population as of year 2016 is Lithuania
- Countries with max mean suicide no per 100K population are Lithuania, Russia, Srilanka etc
- 35 – 54 is the age group with max no of suicides for both male and female it also the age group with maximum population for both male and female.
- 75+ age group is the group with max no of suicides with respect to the population.
- In case of max no of suicides with respect to population as the age group increase the no of suicides also increases
- In general, the no of males that suicide is more compared to that of the female.
- 1995 was the year with most no of suicides with respect to population.
- The general trend of no of suicides is decreasing over the period of time.
- As the GDP increases the no of suicides also increases. Which implies the countries higher GDP has to take care of their citizens also.
- Similarly, more suicide is for countries with higher GDP per capita
- The region with max no of suicides per 100K population for both males and females is Eastern Asia
- Maximum number of suicides per 100 K population is for countries with happiness rank 47 – 62. Which implies even if a country is considered to be a happy country still has lots of suicides

Reflection

From doing this project we will get a clear idea about data exploration and analysis. Like

- from where we can get data.
- How to clean and check the data so that it can be used for analysis.
- What are the different methods that can be used for exploring the data?
- How to explore the data and make analysis?

In this project since happiness rank or its factors doesn't have much of the influence in the suicide rates. Other factors such as the unemployment rate should be analysed to find other sources which affects the no of suicides in a country.

Reference

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

https://www.kaggle.com/unsdsn/world-happiness#2017.csv

https://www.who.int/mental_health/suicide-prevention/en/