**STAT 515**

**Midterm Project**

**Suicide Rates Overview 1985 to 2016**

**INSTRUCTOR: STERLING THOMAS**

**BY**

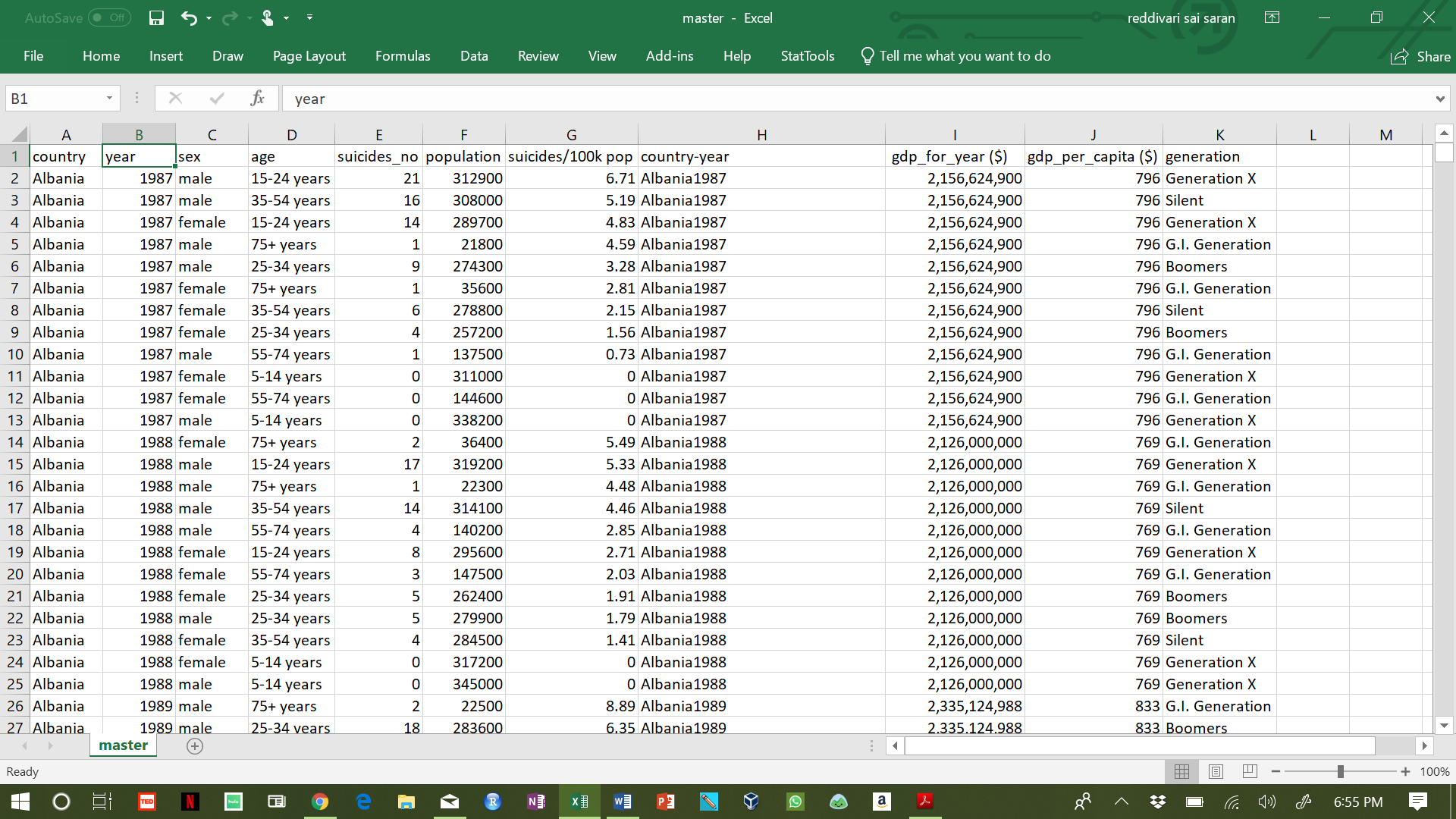
**REDDIVARI SAI SARAN**

**G01142501**

**Introduction**

One of the global public problems is suicide. Annually, 800,000 deaths are suicides. Among those people, nearly one-third are youngsters. According to the World Health Organization (WHO) “Overall global age-standardized suicide rate of 10.5 per 100,000 population in 2016, 13.7 per 100,000 for males and 7.5 per 100,000 for females”. Suicide has become the second leading cause of death in the world. Suicides are mostly occurred between 15-29 years old and 15-19 years aged female. The primary reasons for these suicides are a loss of job, depression, stress, schizophrenia, bipolar disorder and the loss of a relationship [1].

The increase in the death rate by suicide in the world makes me more concern and interesting to choose this topic. This dataset contains different attributes like country, year, sex, age group, count of suicides, population, suicide rate, country year composite key, gdp\_for\_year, gdp\_per\_captia, and generation.



**File Size:**

* The file size is 396 KB.
* The data set contain 11 rows and 27,821 columns.

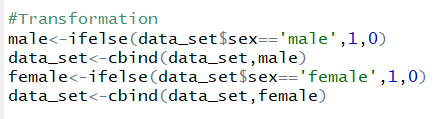
**Types of Data and Measurement Scales:**

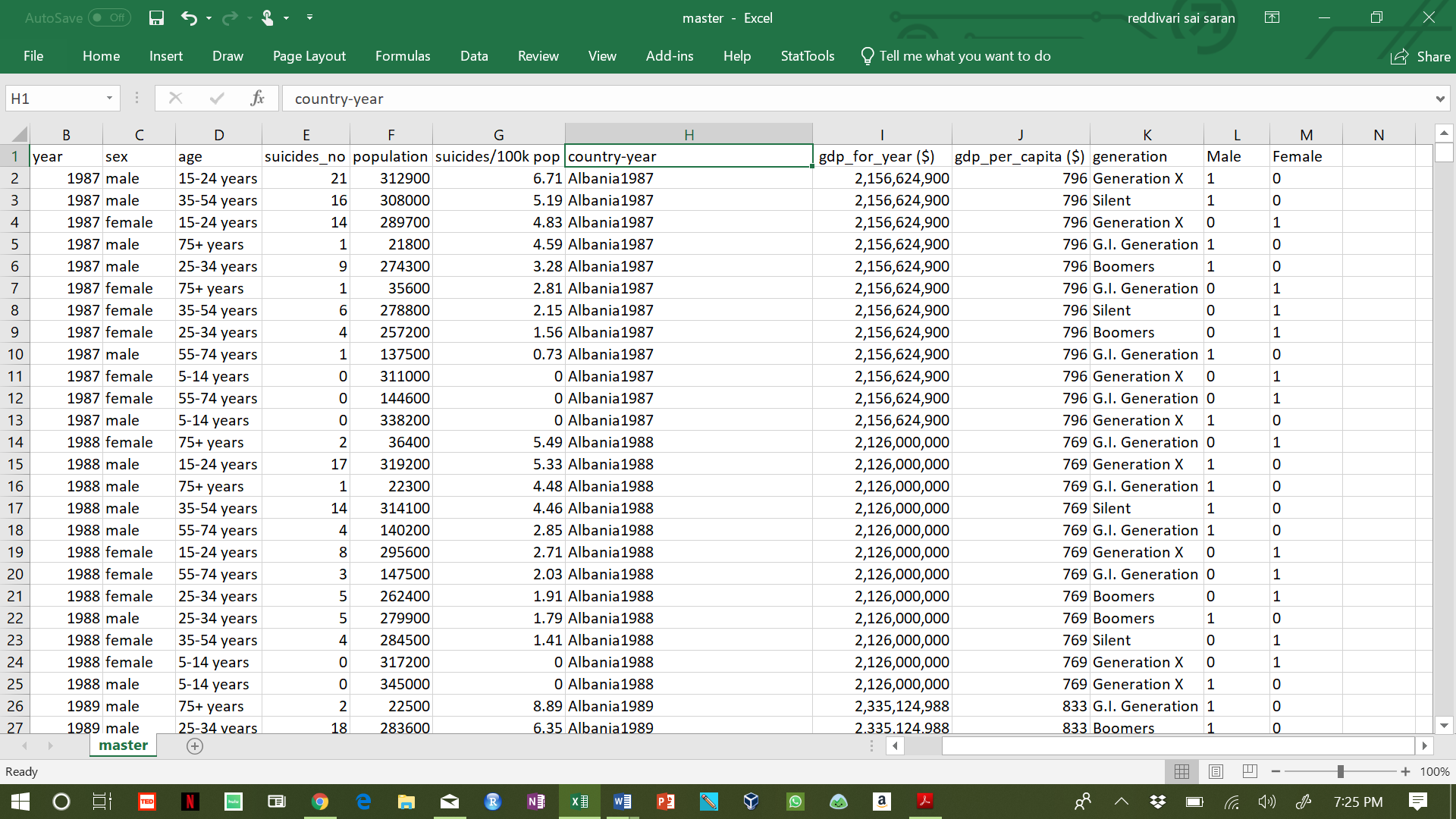
* Country- Nominal
* Year- Ordinal
* Sex-Nominal
* Population-Ratio
* Suicide-Ratio
* Age-Interval
* Suicide\_no-Ratio
* Suicide/100k pop- Ratio
* Gdp\_for\_year- Ratio
* Gdp\_per\_captial- Ratio
* Generation- Nominal

**Transformation**

The categorical fields ‘sex’ is transformed into two fields ‘male’ and ‘female’ containing only binary values 0 and 1. This transformation is done using R-studio. First, two vectors male and female are created with values 0, 1 and then combined with the original data frame using the ‘cbind’ function.

The code for the transformation is:



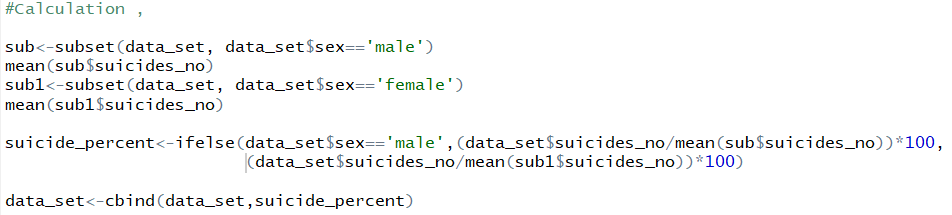


The above image is a snapshot of the dataset after transformation.

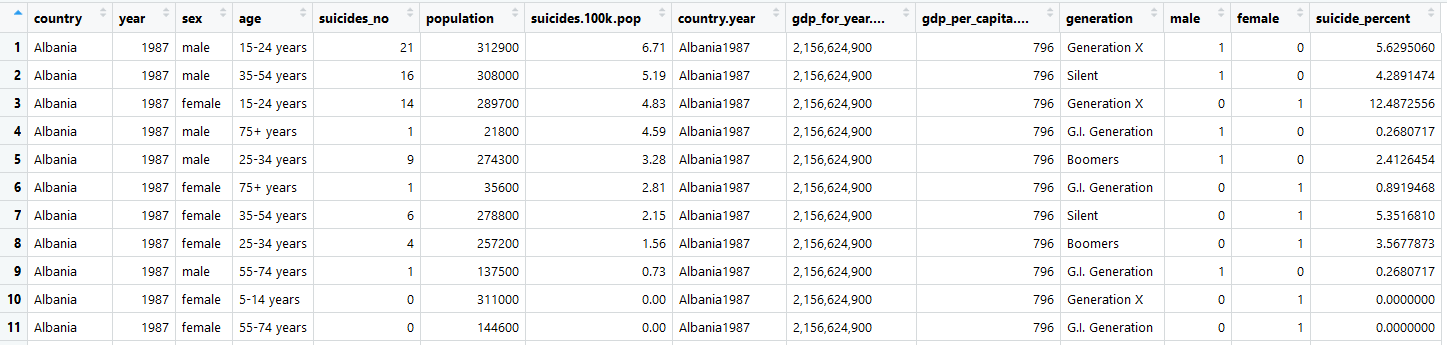
**Calculation**

A new variable ‘suicide\_percent’ is derived from ‘suicides\_no’ as a percentage of mean value of suicides grouped by gender is created. First, the mean values for number of suicides for male and female are calculated. Then the percentage of suicides with respect to mean value is computed and the new field is added to the existing data frame. This calculation is made using R studio.

The code for calculation:



Below is the snapshot of the data after including the calculated column.

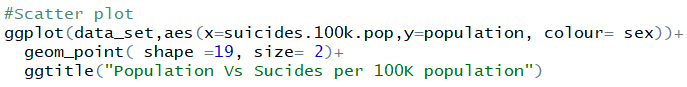


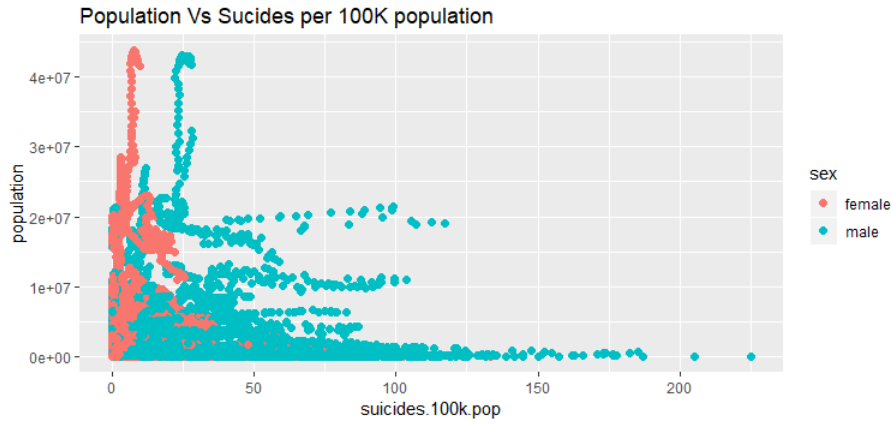
Here, the first value of the suicide\_percent represents that the number of suicides of 15-24 years males in Albania during 1987 is 5.629% of the total male suicides.

**Visualizations**

1. **Scatter plot**

A scatter plot for the number of suicides per 100k population against the population distinguished by gender is created using ggplot package of the R studio. The code for this plot is given below.

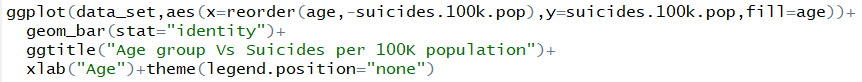


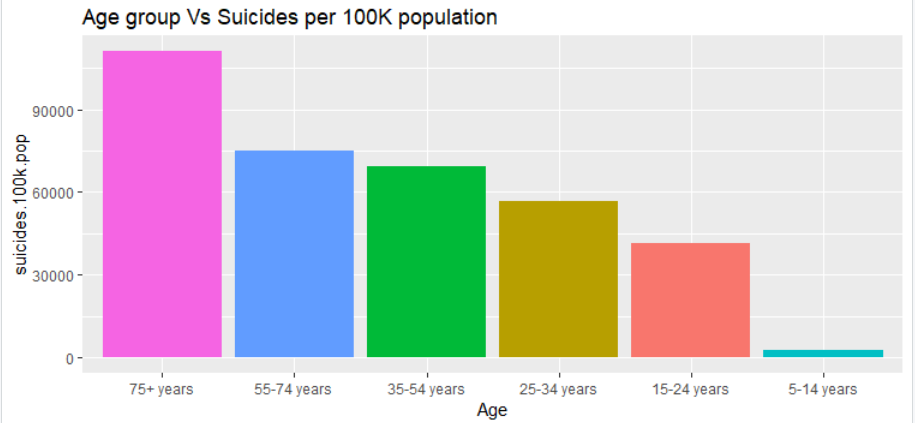


The above plot shows that the number of suicides per 100K population is greater for a population less than 10,000,000 and very low for the high value of population. Also, it is clear that the number of suicides is greater in males than in females.

1. **Bar plot**

A bar plot distinguishes the properties for categorical variables. A bar plot for the suicides per 100K population against the age group is drawn using R studio. The code for the bar plot is given below.



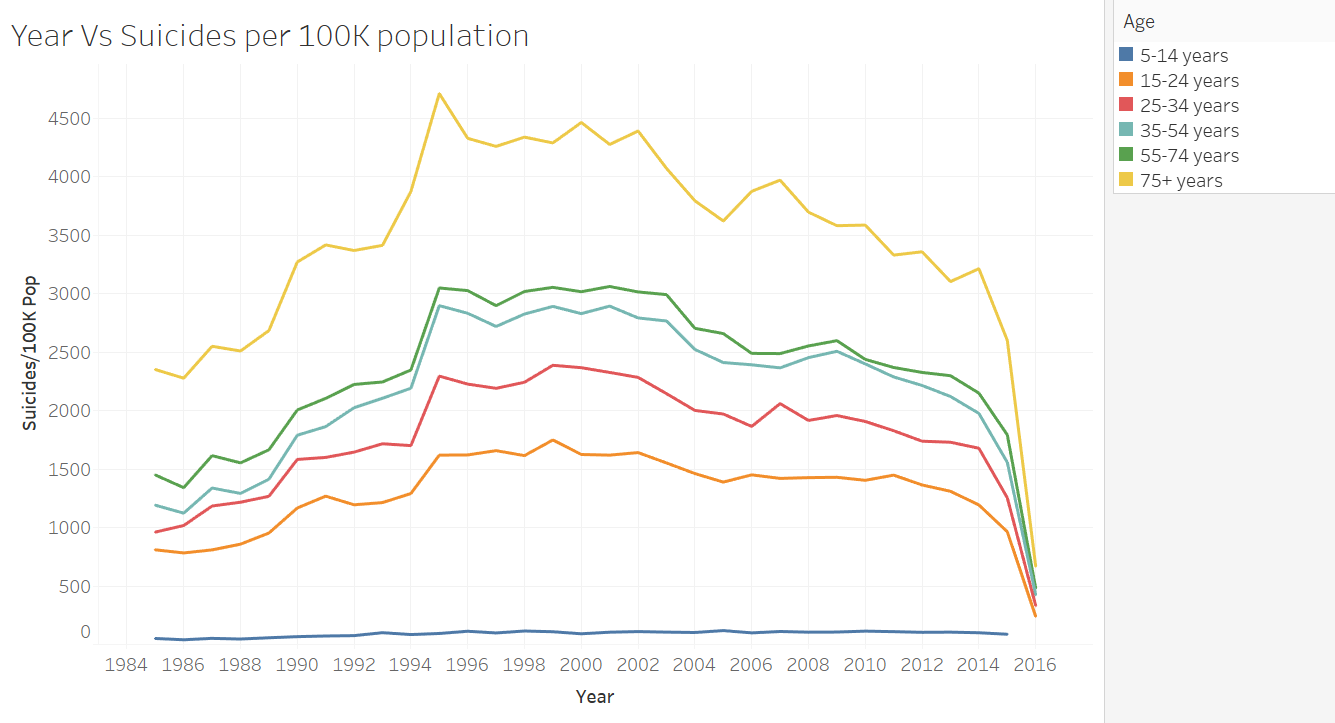


From the above bar graph, it can be inferred that the number of suicides is highest for the age group greater than 75 years and least for the age group between 5 and 14 years.

1. **Line chart**

The below line chart displays the change in suicides per 100K population over the years 1985 to 2016 for each group. This plot is created using Tableau.

The number of suicides is higher in the year 1995 comparative to other years for all the groups. Also, the number of suicides decreased rapidly in 2016.D

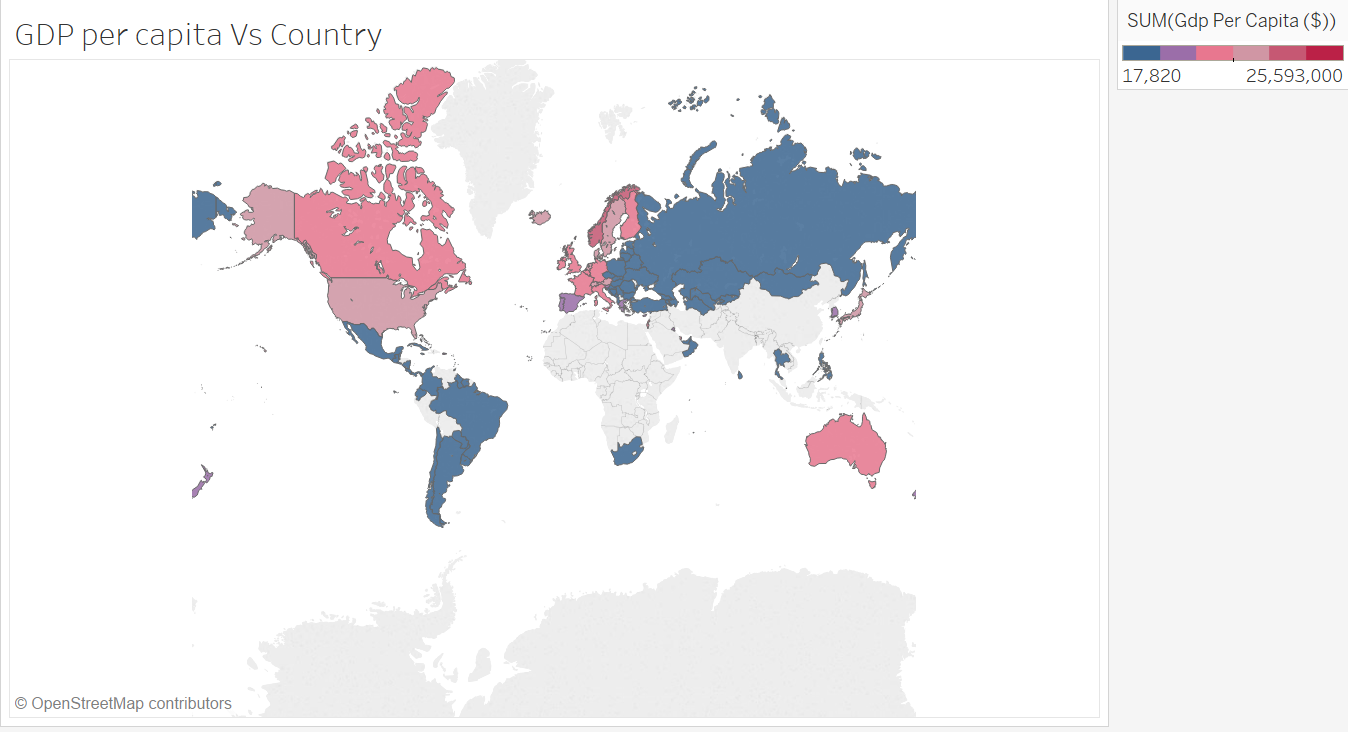


1. **Heat Map**

A heat map is a representation of data in the form of a map or a diagram in which the data values are represented as colors.

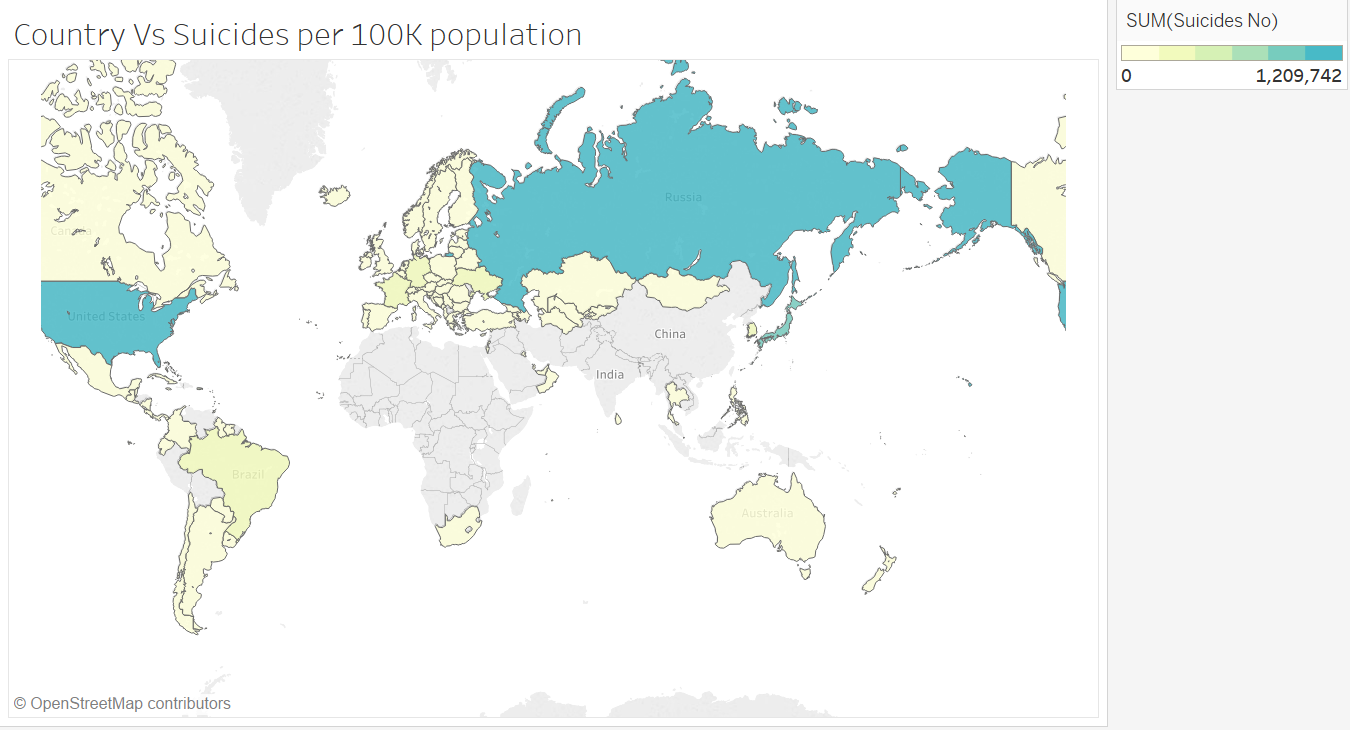
* 1. **GDP per Capita**

The below graph is created in Tableau. The variation in color is based on the GDP per capita for different countries. A lower GDP value is indicated with blue color, and higher GDP is represented in red color. From the graph, it can be inferred that Canada has the highest GDP per capita and few Asia and South America countries have very low GDP per capita



* 1. **Suicides per 100K population**

The below graph represents the variation in the number of suicides per 100K population across the world with yellow representing the least number of suicides and blue representing the highest number of suicides. The number of suicides is highest in the USA and Russia.



**Conclusion**

It can be concluded from all the visualizations that some suicides per 100K population are highest in the USA for males aged above 75 years. Also, it can be inferred that Russia has the highest GDP per capita and the highest number of suicides.

Therefore, visualizations help analyze and the data better and conclude.

# References

|  |  |
| --- | --- |
| [1] | W. H. Organization, National suicide prevention strategies Progress, examples and indicators, World Health Organization, 2018. |
| [2] | Kaggle, "Suicide Rates Overview 1985 to 2016," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016. [Accessed 7 March 2019]. |