Statistics for Data Science (UE19CS203)

**Project Report**

EDA of Suicide Rate

|  |  |  |
| --- | --- | --- |
| Name | SRN | Section |
| Serena A Gomez | PES2UG19CS372 | F |
| Sagarika M Chavan | PES2UG19CS347 | F |
| Shanoo Raghav | PES2UG19CS375 | F |
| Roshni Govind | PES2UG19CS339 | F |

## Abstract

Use the input dataset obtained from multiple data source; that comprise of country-wise population data & corresponding suicide rate, along with their socio-economic indices for the years spanning 1985-2016 – to perform Exploratory Data Analysis using the Jupyter Notebook. This project involves – data pre-processing, data-cleaning, data-preparation, data-analysis, and hypothesis-testing. Our analysis revealed the following –

* The average suicide-rates is 11.82 per 100K population; suicides rates are highest for the following three countries - Lithuania, The Russian Federation, and Sri Lanka; and the lowest for the following countries - Saint Kitts & Nevis, Dominica, and Jamaica.
* The Males population were more likely to commit suicide than Females.
* People in the age group ‘55+ years’ were most likely to commit suicide,
* The suicide-rate is higher for countries with high HDI; whereas, it is higher for countries with low GDP. Although there is a direct correlation between GDP & HDI
* Similarly, the suicide-rate increases as the HDI increases; whereas it decreases as the GDP increases.

## Introduction

The objective of this project is to analyse the relationship between Suicide rates in countries with different socio-economic characteristics such as - Gross Domestic Product (GDP) per capita, Human Development Index (HDI) - across multiple years. The project used the exploratory analysis techniques to find answers the following questions:

1. Countries with the highest & lowest suicide rate?
2. Who amongst the men and women commit more suicides?
3. People of what Age are more susceptible to commit suicides?
4. Effects of GDP & HDI on the suicide rates.

## Dataset

The input dataset is a CSV file, with 32 years (1985-2016) of country-wise population data, along with its GDP-per-capita, suicide-rate, and human-development-index. It contains over 25,000 records, with 13 features that describe the socio-economic spectrum of 101 countries (out of 193 countries in the world) – specifically the suicide rates. The primary features of the include - categorical data (*country*, *sex*, *age*, *country-year*, *gdp\_for\_year* and *generation*), numeric discrete data (*year*), and numeric continuous data (*suicides\_no*, *population*, *gdp\_per\_capita ($)*, *suicides/100k pop*, and *HDI for year*).

## Pre-processing or Data cleaning

The input dataset was analysed for its inaccuracies, incompleteness, redundancies, and/or validity. Based on visual data inspection, we noticed the following:

* 70% of missing data in the *HDI for year* feature.
* Redundant data in the ‘*country-year*’, ‘*country*’ and ‘*year*’; wherein, the ‘*country*’ and ‘*year*’ are exactly the same as ‘*country-year*’.

The data quality was improved by performing the following steps:

* Removed the redundant ‘*country-year*’ feature from the input dataset.
* Divided the input dataset into 2 data-frames
  1. Data-frame-1 (Socio-economic data with suicide-rates, GPD & HDI) - this data-frame will have all the features, however will not include the rows with blank or null value in the *HDI for year* feature. This data-frame will have only 30% of the original data.
  2. Data-frame-2 (Country-wise population data with suicide-rates & GDP) – this data-frame will not have the *HDI for year* feature.
* Remove outliers, in both the data-frames
  1. Data-frame-1 - Draw a boxplot for ‘*HDI for year*’, we did not see any outliers; However, there very few outliers in the boxplot for ‘*suicides/100k pop*’ feature; in order to remove these outliers, we dropped all the rows that had a value greater than (mean + 6 \* standard deviation) in the ‘*suicides/100k pop*’ feature.
  2. Data-frame-2 - Draw a boxplot for ‘*suicides/100k pop*’ and we saw outliers. In order to remove these outliers, we dropped all the rows that had a value greater than (mean + 6 \* standard deviation) in the ‘*suicides/100k pop*’ feature.

This **data cleaning** step helped us to identify and remove - errors, omissions, and redundant data. This is going to help us maximize efficiencies in the next steps - by reducing all kinds of noise, duplications and inconsistencies.

## Exploratory Data Analysis

**Data preparation:**

The data-frames was augmented with the normalized data, for the following features - ‘*suicides\_no*’, ‘*population’*, ‘*suicides/100k pop*’, ‘*gdp\_per\_capita ($)*’. As part of the exploratory analysis – we used the normalized data features.

**Data correlation:**

We used the Data-frame-2 to plot a heatmap to see the correlation between our various numeric variables.

Chart, bar chart

Description automatically generated

From the above chart, we inferred that there is a strong positive correlation between *population* and *suicides\_no* with a correlation factor of 0.62. In addition, there is a very minor negative correlation between *suicides/100k pop* and *year* (with a correlation factor of -0.04), which can be ignored.

**Data analysis:**

1. Suicide rates by Country: We analysed the data to find the countries that reported the highest suicide-rate. Hence, we plotted a graph to see the number of *suicides/100k* versus *population* of each country in the years 1985-2016 using the Data-Frame-2.

Chart, histogram

Description automatically generated

From the above graph, we inferred that the top three countries with the highest suicides rates in the years 1985-2016 were *Lithuania*, *The Russian Federation* and *Sri Lanka*; and the bottom three countries with the lowest suicide rates in these years are *Saint Kitts and Nevis*, *Dominica* and *Jamaica*.

In addition, we inferred that the global mean of suicide rates in the years 1985-2016, was 11.82 per 100K population.

2. Suicide rates by Population: We analysed the data to understand the suicidal tendencies in the male and female population, across all countries over the years (1985-2016). We plotted a graph to visualize the data using the Data-Frame-2.

Chart, histogram

Description automatically generated

From the above graph, we inferred that males were more likely to commit suicide than females.

In addition, we inferred that the suicide rates of females were highest in the year 1995 and that of males was highest in the years 1995 and 1997.

3. Suicide rates by Generation: We analysed the data to understand, which generation committed more suicide in the years 1985-2016, across all countries using the Data-Frame-2.

Chart, pie chart

Description automatically generatedIn the following plot, the generations identify the following:

* GI Generation – 1901-27
* Silent Generation – 1925-45
* Baby Boomers – 1946-64
* Generation X – 1965-79
* Millennials – 1975-94
* Generation Z- 1995-2012

From the above graph, we inferred that the *Silent generation* commit the greatest number of suicides.

Although, the plot shows that *Generation Z* committed the least number of suicides, on visual analysis of the data – we found that there was very little data in the dataset to conclude. Hence, we infer that the *Millennials* committed the least number of suicides.

4. Suicide rates by Age: We analysed the data to understand, the suicide trend of the population by age groups - over the years 1985-2016, across all countries using the Data-Frame-2.

Chart, line chart

Description automatically generated

From the above graph, we inferred that the people in the age group ‘*75+ years*’ were most likely to commit suicide, and the next in the list is in the age group of ‘*55-74 years*’.

5. Suicide rates by HDI: Using the normalized data in the Data-frame-1, we analysed the relationship between the Human Development Index, and Number of suicides by population. We plotted a graph to see the effect of *HDI for year* on *suicides/100k population*.

Chart, scatter chart

Description automatically generated

From the above graph, we inferred that the suicide rate is lower for countries with low *HDI for year*; and increases as the *HDI for year* increases.

6. Suicide rates by GDP per capita: Using the normalized data in the Data-frame-1, we analysed the relationship between the Number of suicides and the *GDP per capita*. We plotted a graph to see the effect of *GDP per capita* on *suicides/100k population*.

Chart, scatter chart

Description automatically generated

From the above graph, we inferred that the *number of suicides* is higher for the countries with lower *GDP per capita*; and it reduces as the *GDP per capita* increases.

7. HDI versus GDP: Finally, we analysed the relationship between the *GDP per capita* and the *Human Development Index* (HDI) of all the countries. We plotted a graph *HDI for year* and *GDP per capita*.

Chart, scatter chart

Description automatically generated

On expected lines, the above graph, indicates that *HDI for year* is higher for countries with higher *GDP per capita*; and the *GDP per capita* increases exponentially with the *HDI for year*.

## Hypothesis Testing

We did a hypothesis testing using the following hypothesis:

* The suicide rate in Ukraine was high,
* The average suicide rate in Ukraine is more than the global average.

The global average of the suicide rate is computed as 11.82. Hence, the hypothesis is as follows:

Null hypothesis: x¯ < 11.82  
Alternate hypothesis: x¯ >= 11.82  
α = 0.05

We took a sample of 200 observations from the dataset, for *country* = “Ukraine”.

The sample mean is computed using the mean of *suicides/100k pop* of this sample.

The Z test was performed, and our null hypothesis was rejected as shown below:



Hence, we concluded that the suicide rates in Ukraine was more than the global mean suicide rate (11.82).

## Results and Discussion

The result of the exploratory data analysis is as follows:

1. Countries with the highest & lowest suicide rate?
   * The top three countries with the highest suicides rates in the years 1985-2016 were:
     + Lithuania,
     + The Russian Federation, and
     + Sri Lanka;
   * The bottom three countries with the lowest suicide rates in these years are:
     + Saint Kitts and Nevis,
     + Dominica, and
     + Jamaica.
   * The global mean of suicide rates in the years 1985-2016, was 11.82 per 100K population
2. Who amongst the men and women commit more suicides?
   * The Males population were more likely to commit suicide than Females.
   * The suicide rates of females were highest in the year 1995 and that of males was highest in the years 1995 and 1997
3. People of what Age are more susceptible to commit suicides?
   * People in the age group ‘75+ years’ were most likely to commit suicide,
   * Next in the list, is people in the age group of ‘55-74 years’.
4. Effects of GDP & HDI on the suicide rates.
   * HDI is higher for countries with higher GDP; and the GDP increases exponentially with the HDI.
   * The number of suicides is lower for countries with low HDI; and it increases as the HDI increases
   * On the other hand, the number of suicides is higher for the countries with lower GDP; and it decreases as the GDP increases.