Concepts And Technologies

Of AI Assessment – 1

Student ID: 2358900

Student Name: Subekshya Dhamala

Group: L5CG22

Module Leader: Mr. Siman Giri

Tutor: Mr. Ronit Shreshta

Abstract

This report investigates the data about suicides in India. I cleaned the data, looked at overall statistics and studied individual and combined factors. I also made charts to show important details visually. Doing all these things helped me better understand what the data is like.

Table of Contents

Table of Figure

1.Sucide in India

2.Introduction

This study examines a dataset regarding suicides in India in depth. The data comes from the well-known open dataset platform Kaggle and covers a lengthy time from 2001 to 2012. It's a great tool for learning about the complex aspects of suicide in India. The primary goal of the research is to identify significant trends and patterns in a massive dataset. The goal is to offer valuable perspectives for analysts, researchers, and policymakers. With this knowledge, they may make educated judgements and carry out focused actions in mental health programming.

3. Data Cleaning and Summary Statistics

3.1. Data cleaning

The data preparation code which has been provided, the most common values, referred to as the "mode," are used to fill in the missing values in the "Year," "Gender," and "Age_group" columns. After that, the code examines the dataset for any further missing values and displays them. Furthermore, the default value of 0 is assigned to any missing values in the 'Total' column. This is done on an assumption that a missing total value can be properly replaced with this default value. Consequently, these data preprocessing steps contribute to a more complete and analytically usable dataset. The entire procedure

ensures that the dataset is handled effectively, with missing values restored using suitable techniques.

```
State         0
Year          0
Type_code     0
Type          0
Gender        0
Age_group     0
Total         6
dtype: int64
```

Figure 1 before removing missing values:

```
State       0
Year        0
Type_code   0
Type        0
Gender      0
Age_group   0
Total       0
dtype: int64
Total Number of Missing values after filling missing values: 0
```

Figure 2 Data after removing missing values.

3.2 Summary Statistics

We classified the columns as 'int64,' 'float64,' or 'object' types in the summary statistics analysis by using the select_dtypes function to separate the data into numerical and categorical categories. We used the 'describe ()' function to focus on the numerical columns, namely 'int64' and 'float64." This process provided crucial statistical metrics like mean, median, minimum, maximum, and standard deviation. The normal range and distribution of values within the numerical features of our dataset are effectively evaluated using these measures.

When evaluating categorical data, we started by showing each dataset column's unique values. The most frequent value (mode) for each of these categorical columns was then shown by using the select_dtypes function to clearly select columns getting categorical data types (object).

4. Data Visualization and Exploration

4.1. Univariate Analysis:

4.1.1 chart 1

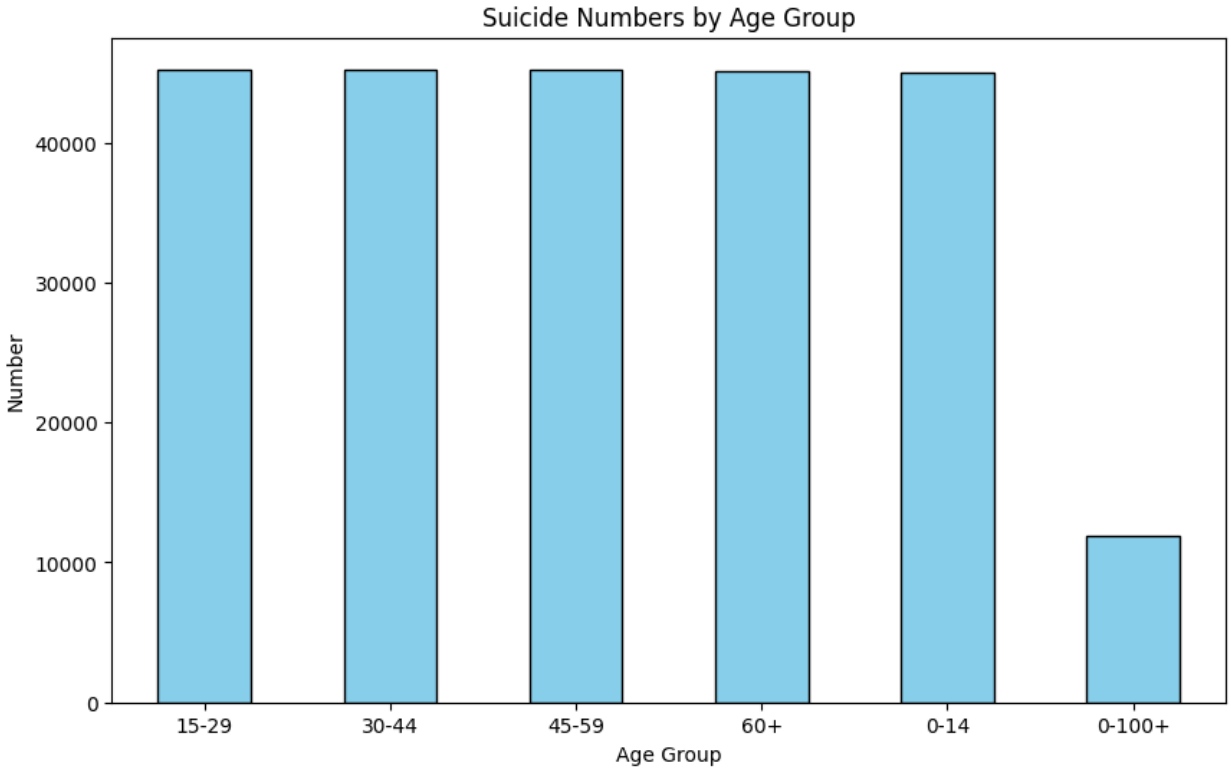

Figure 3 :Suicide number by age group

A bar graph displaying the overall suicide rate distribution by age group. According to the "Suicide Numbers by Age Group" bar chart, the age categories of 15–29, 30-44, 45–59, and 60+ have the highest suicide rates, having each group exceeding 40,000 cases. The age groups 0-14 and 0-100+ show a decrease. The graph highlights the importance of suicide as the main cause of death which makes mental health resources readily available to those in need of care.

4.1.2. Chart 2

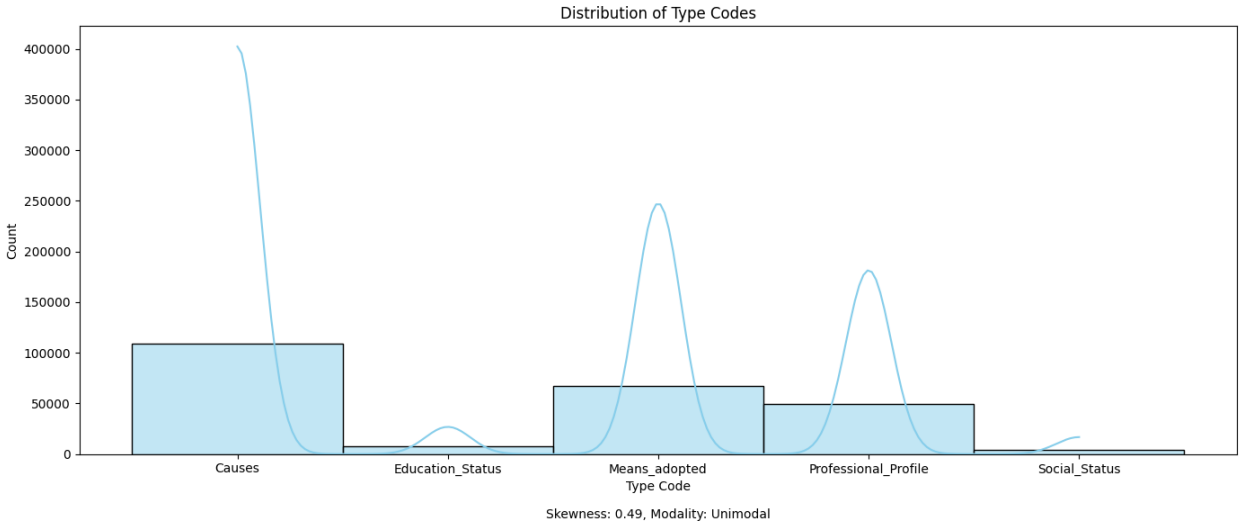

Figure 4:Distribution of type code

A visual representation of the causes, education status, means adopted, professional profile, and social status is offered by the "Distribution of Type Codes" chart. The biggest number of causes is notable and is followed by a decrease in education status and an increase in means adopted. A unimodal distribution can be seen by the chart's skewness of 0.49, which highlights a clear pattern. With categories on the x-axis and counts ranging up to 400,000 on the y-axis, this representation offers valuable insights into the prevalence of various types.

4.1.3. Chart 3

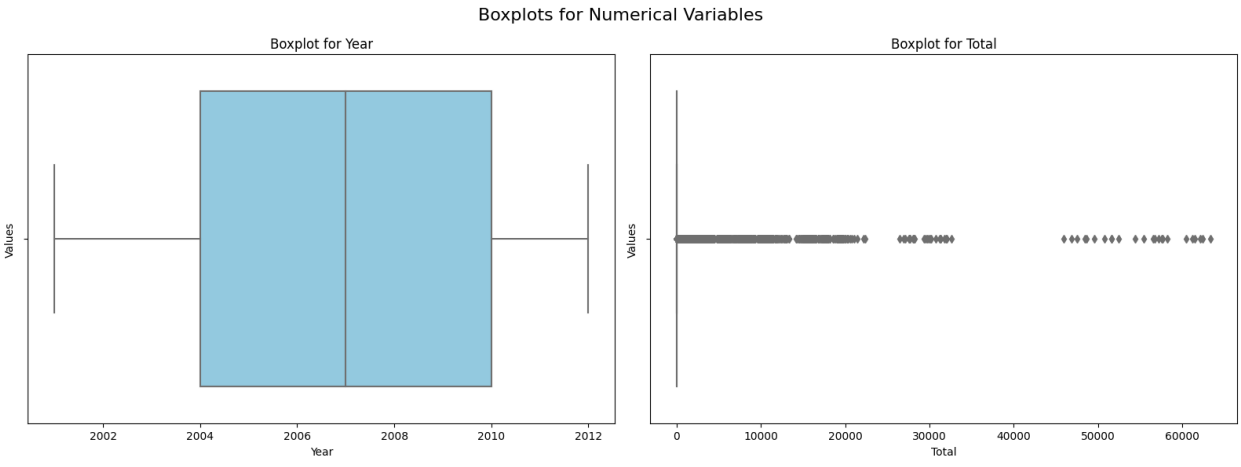

Figure 5: Boxplots for Numeric variables

Two boxplots are included in the charts. The one on the left, titled "Boxplot for Year," displays data with no outliers from 2002 to 2012. The height of the box doesn't change, indicating a stable range. The one on the right, "Boxplot for Total," shows information from 0 to 60,000. A few extremely high values can be seen as dots to the right but the majority of numbers of clusters close to zero.

4.2. Bivariate Analysis:

4.2.1.Chart1

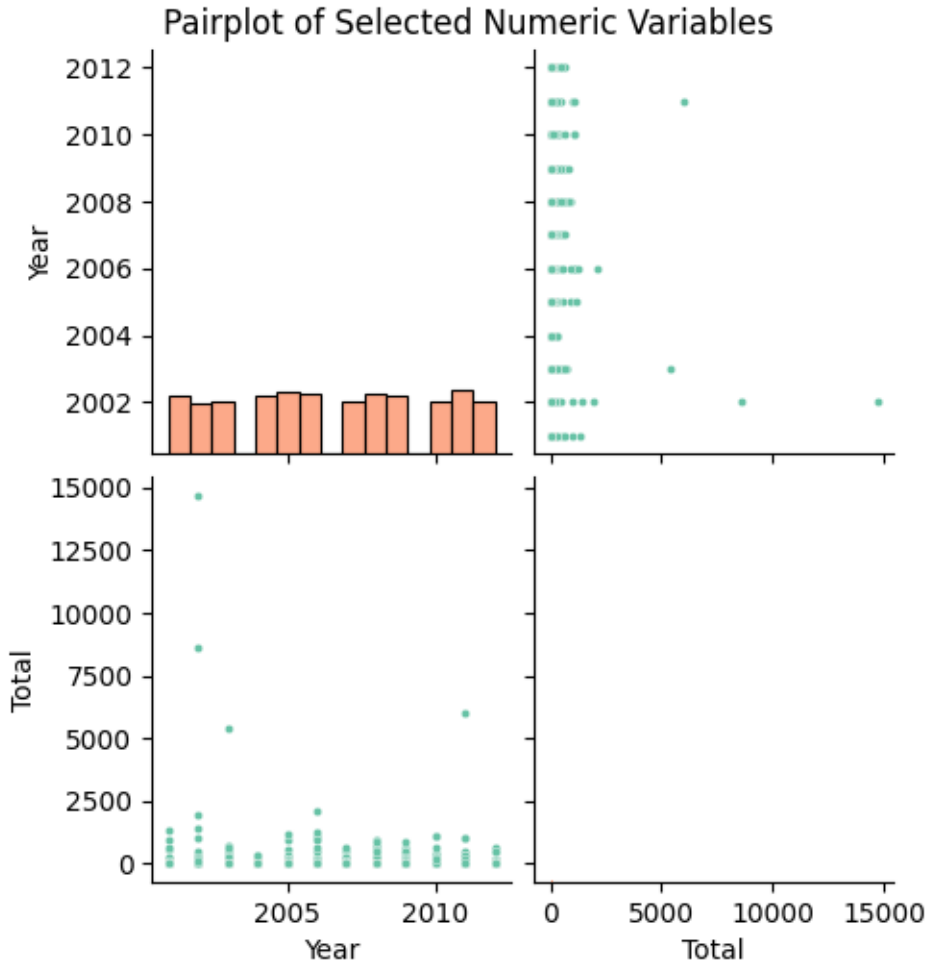

Figure 6 Pair plot of numeric variable

To produce "Pair plot of Selected Numeric Variables," use the pairplot function from Seaborn. This function shares each numerical variable between rows and columns to visualize pairwise relationships in a dataset. The distribution of the data in each column can be seen in the horizontal charts. The following three components make up the graph: a scatter plot showing the total values, a histogram showing the distribution of data over the years, and a second scatter plot showing the relationship between the

total values and the years. Individual data points are denoted by green dots in the graphs, and exact value references are provided by numerical scales on both axes.

4.2.2. Chart 5

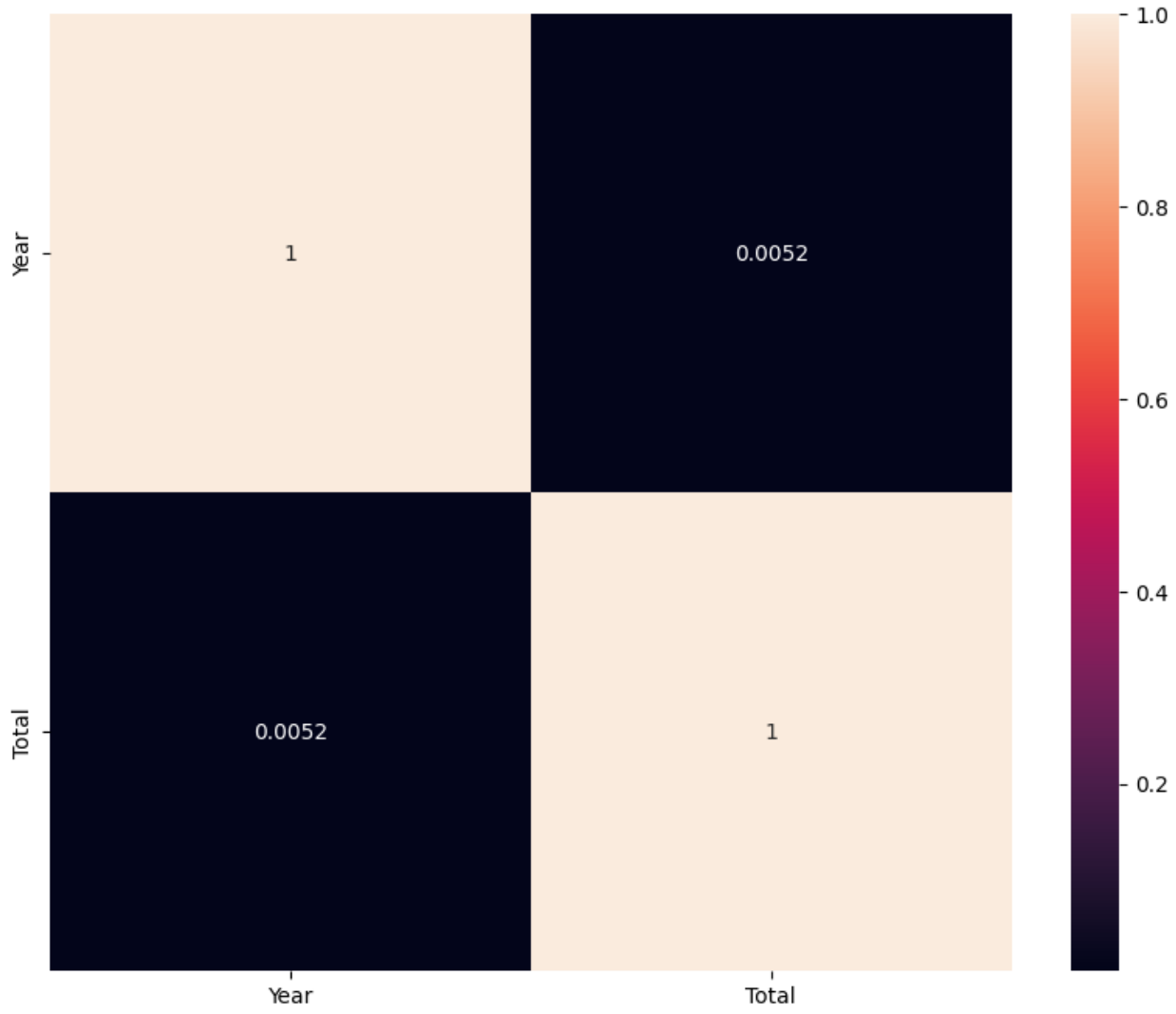

Figure 7:Heatmap

The correlation matrix for numeric columns in a pandas Data Frame is computed and shown by this code. It helps the finding of variable connections by producing a heatmap with color-coded values utilizing the Seaborn library. Marks are used to display the plot and provide information about the direction and strength of linear correlations. To find patterns and relationships in a dataset, this code is useful for exploratory data analysis.

5. Conclusion:

In conclusion, we looked carefully at the Kaggle database of Suicide in India, which includes the years 2001–2012. After cleaning the data, we examined the numbers to look for unexpected trends. The histogram showed the distribution of various case types, while the bar graph indicated which age groups had the highest rates of suicide. By visually connecting the "Year" and "Total," the correlation heatmap and pair plot revealed how things change over time. For those making decisions about mental health treatments, these visualizations provide insightful information. Our findings highlight important trends, such as the age groups that are most affected, which add to our increasing comprehension of the complicated structure of suicide in India.