Department of Computer Engineering
College of Engineering
Polytechnic University of the Philippines Sta. Mesa

CMPE 40163: Exploratory Data Analysis

Exploratory Data Analysis of WHO Suicide Statistics

Submitted by:

Merque, John Ric C.
BSCOE 3-2

Submitted to:

EDCEL B. ARTIFICIO

I. **Data Dictionary**

WHO Suicide Statistics

Basic historical (1979-2016) data by country, year and demographic groups

Suicide is a significant public health issue, with approximately 700,000 people dying by suicide each year worldwide according to the World Health Organization (WHO). For each suicide, there are an estimated 20 suicide attempts. It ranks among the leading causes of death, affecting diverse demographics across different regions. Interestingly, suicide rates tend to exhibit considerable differences based on factors such as year, age, gender, and geographic location and these perspectives this exploratory data analysis plan aims to look at.

Here is the outline of data variables and high-level information about the dataset:

```
Data Dictionary
Number of rows: 43776
Number of columns: 6
Column 'country': 0 null values
Column 'year': 0 null values
Column 'sex': 0 null values
Column 'age': 0 null values
Column 'suicides_no': 2256 null values
Column 'population': 5460 null values
```

```
root
 |-- country: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- sex: string (nullable = true)
 |-- age: string (nullable = true)
 |-- suicides_no: integer (nullable = true)
 |-- population: integer (nullable = true)
```

Column	Data Type	Description	Null Values
country	String	The name of the country where the data was recorded	0
year	Integer	The year the data was recorded	0
sex	String	The sex of the individuals (e.g., male, female)	0
age	String	The age group of the individuals	0
suicides_no	Integer	The number of suicides recorded	2256
population	Integer	The population of the specific demographic group in the year	5460

II. Summary Statistics

```
summary_stats = data_df.describe()
summary_stats.show()
```

```
+-------+--------+------------------+------+-----------+-----------------+------------------+
|summary| country|              year|   sex|        age|      suicides_no|        population|
+-------+--------+------------------+------+-----------+-----------------+------------------+
|  count|   43776|             43776| 43776|      43776|            41520|             38316|
|   mean|    NULL|1998.5024671052631|  NULL|       NULL|193.3153901734104|1664091.1353742562|
| stddev|    NULL|10.338711176746282|  NULL|       NULL|800.5899259349637|3647231.2274873867|
|    min| Albania|              1979|female|15-24 years|                0|               259|
|    max|Zimbabwe|              2016|  male|  75+ years|            22338|          43805214|
+-------+--------+------------------+------+-----------+-----------------+------------------+
```

Here's a breakdown of each summary statistic provided in the table:

Count:

country: 43776
year: 43776
sex: 43776
age: 43776
suicides_no: 41520
population: 38316

This indicates the number of non-null entries for each column. There are some missing values in the suicides_no and population columns since their counts are less than the total count.

Mean:

suicides_no: 193.32
population: 1,664,091.14

On average, there are about 193.32 suicides per recorded data point, and the average population for the age groups is about 1,664,091.

Standard Deviation:

suicides_no: 800.59
population: 3,647,231.23

This measures the amount of variation or dispersion in the dataset. There is significant variability in the number of suicides and population across the data points.

Min:

year: 1979
age: 5-14 years
suicides_no: 0
population: 259

The earliest year in the dataset is 1979. The minimum number of suicides recorded is 0. The smallest population recorded for an age group is 259.

Max:

year: 2016
age: 75+ years
suicides_no: 22,338
population: 43,805,214

The latest year in the dataset is 2016. The maximum number of suicides recorded in a single entry is 22,338. The largest population recorded for an age group is 43,805,214.

Data Interpretation

The data spans from 1979 to 2016, providing a wide range of years for analysis across different countries. Both male and female data are included, along with various age groups ranging from 15-24 years to 75+ years There is significant variability in the number of suicides reported, with some entries recording no suicides and others recording very high numbers. The population for the age groups also varies widely, indicating diverse data sources or differences in demographic sizes.

Quartiles and Percentiles

```
# Define the columns of interest
columns_of_interest = ['suicides_no', 'population']
probabilities = [0.25, 0.5, 0.75, 0.1, 0.9]
relative_error = 0.01  # 1% error

# Compute quartiles and percentiles for each column
quantiles = {col: data_df.approxQuantile(col, probabilities, relative_error) for col in columns_of_interest}

quantiles
```

```
{'suicides_no': [1.0, 13.0, 86.0, 0.0, 333.0],
 'population': [78000.0, 367000.0, 1246900.0, 13374.0, 4212476.0]}
```

The provided output shows the approximate quartiles and percentiles for the `suicides_no` and `population` columns in the dataset.

For `suicides_no`:

0.25 Quantile (Q1): 1.0
(25% of the entries have 1 or fewer suicides.)

0.50 Quantile (Median/Q2): 13.0
(50% of the entries have 13 or fewer suicides.)

0.75 Quantile (Q3): 86.0
(75% of the entries have 86 or fewer suicides.)

0.10 Percentile (P10): 0.0
(10% of the entries have 0 suicides.)

0.90 Percentile (P90): 333.0
(90% of the entries have 333 or fewer suicides.)

This indicates that the distribution of suicides is skewed, with a significant number of entries having relatively low suicide counts, while a smaller number of entries have very high counts.

For `population`:

0.25 Quantile (Q1): 78,000.0
25% of the entries have a population of 78,000 or fewer.

0.50 Quantile (Median/Q2): 367,000.0
50% of the entries have a population of 367,000 or fewer.

0.75 Quantile (Q3): 1,246,900.0
75% of the entries have a population of 1,246,900 or fewer.

0.10 Percentile (P10): 13,374.0
10% of the entries have a population of 13,374 or fewer.

0.90 Percentile (P90): 4,212,476.0
90% of the entries have a population of 4,212,476 or fewer.

This indicates a wide variation in the population sizes across the dataset, with some entries representing very small populations and others representing very large populations.

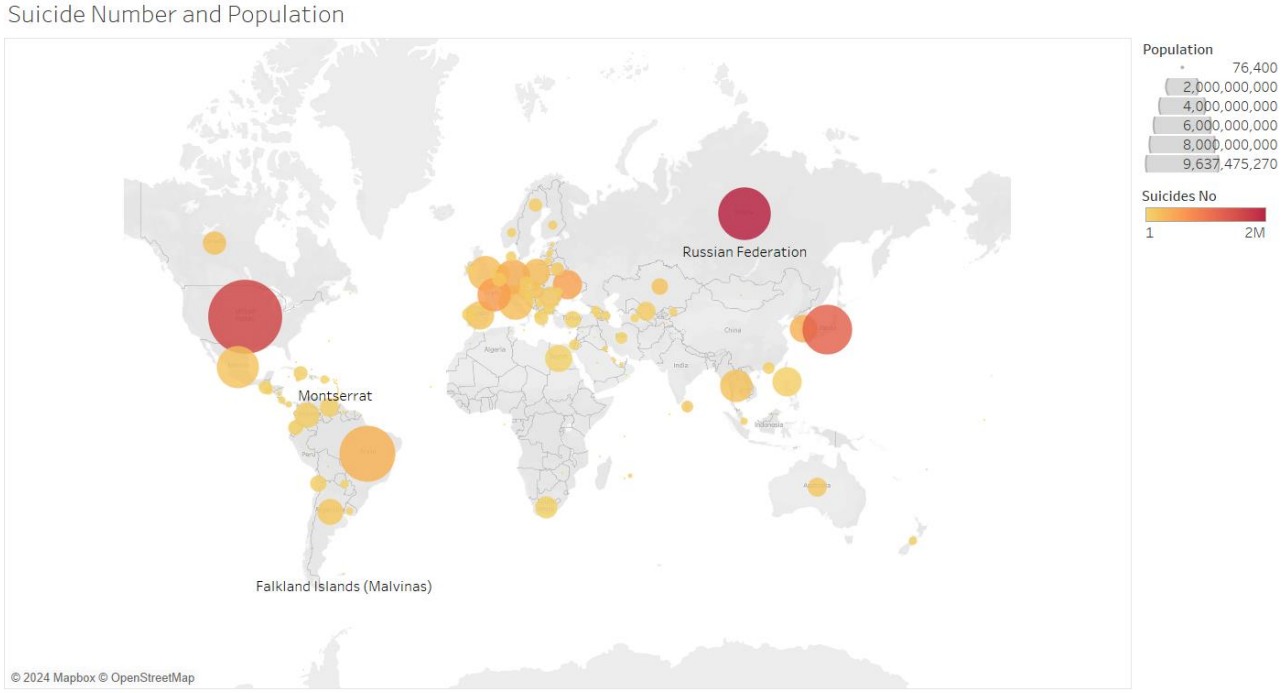
Data Interpretation

The `suicides_no` column has a highly skewed distribution, with most entries having relatively low suicide numbers, but a few having very high numbers. The `population` column shows significant variability, with both small and very large population sizes represented in the dataset. These quartiles and percentiles provide insights into the distribution of the data, helping to understand the spread and central tendency of the suicide numbers and population sizes.

III. **Visualization**

Tableau Symbol Map

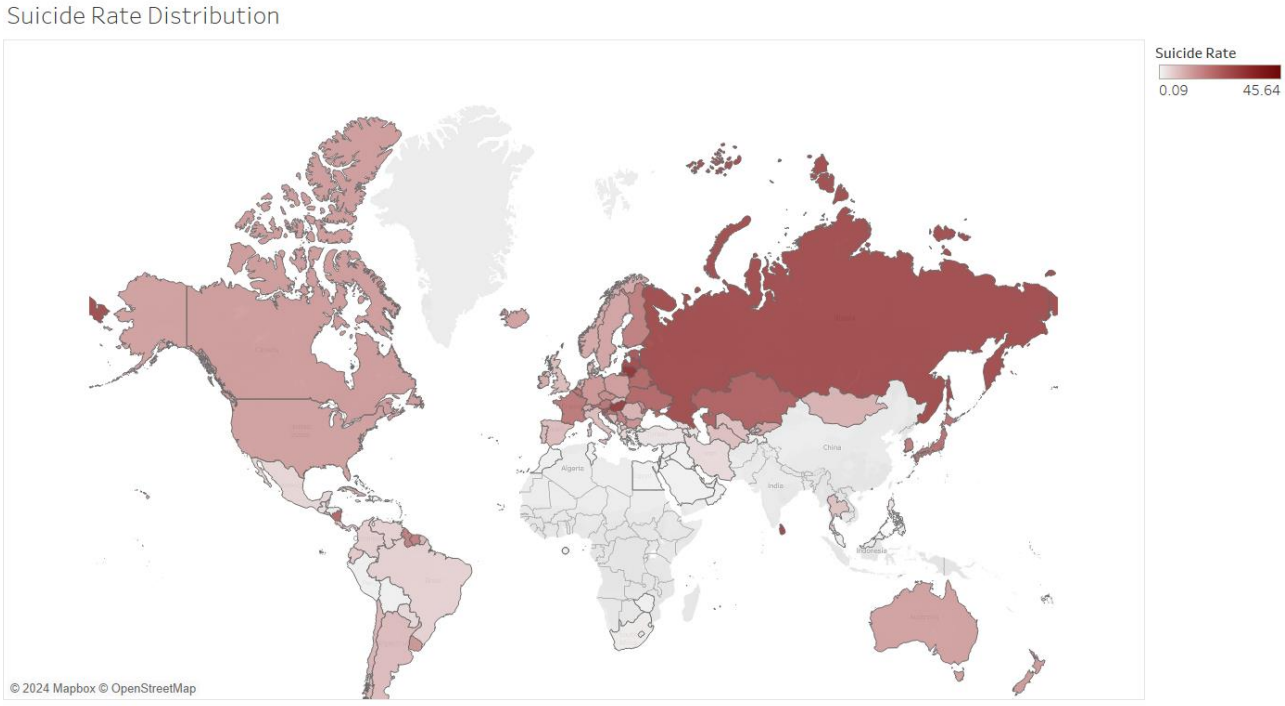
To visualize suicide number per population across countries and highlight minimum and maximum suicide rate:



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Suicides No. Size shows sum of Population. The marks are labeled by Country. Details are shown for Country.

Tableau Map

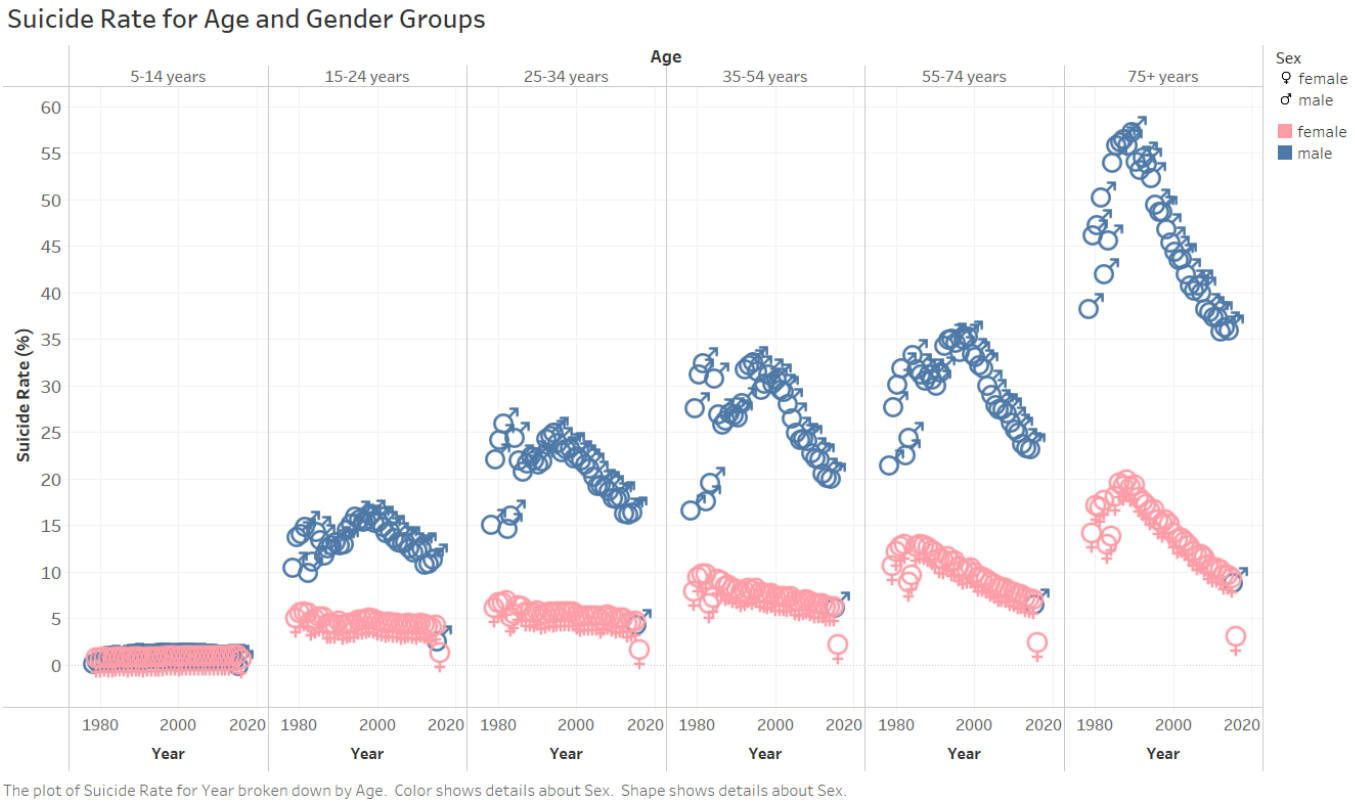
To visualize suicide rate distribution across countries:



Map based on Longitude (generated) and Latitude (generated). Color shows Suicide Rate. Details are shown for Country.

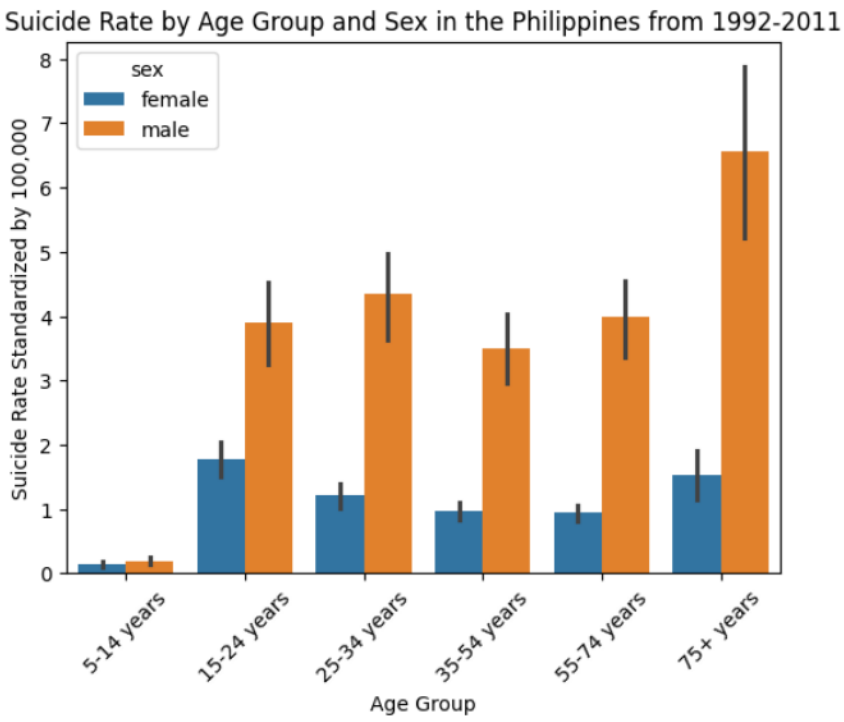
Tableau Scatter Plot

To visualize suicide rate distribution across sex per age group per across the years:



Seaborn Bar Pot

To visualize gender distribution across age groups of suicide rate in the Philippines:



Seaborn Line Plot

To visualize Philippine suicide trends by age group throughout the years: