Advanced Data Visualization
Experiment - 2

Harshey Kaur Soi
2021300057
BE COMPS A - BATCH G

Aim:

Analyze socio-economic dataset using various advanced charts like Word Cloud, Box Plot, Whisker Plot, Regression Plot.

Description of Dataset:

1. The dataset contains 27,820 entries and 12 columns. Here's a brief description of each column:

    1. **country**: The name of the country where the data was recorded.
    2. **year**: The year the data was recorded.
    3. **sex**: The gender of the individuals (male/female).
    4. **age**: The age group of the individuals (e.g., 15-24 years).
    5. **suicides_no**: The number of suicides in the given country for that specific year, gender, and age group.
    6. **population**: The population of the specified age group in the given country for that year and gender.
    7. **suicides/100k pop**: The number of suicides per 100,000 people.
    8. **country-year**: A combined identifier of the country and year.
    9. **HDI for year**: The Human Development Index for that year (available for only some records).
    10. **gdp_for_year ($)**: The total GDP of the country for that year (as a string, needs conversion for analysis).
    11. **gdp_per_capita ($)**: The GDP per capita for that year.
    12. **generation**: The generation classification of the individuals (e.g., Generation X, Silent, etc.).

2. Dataset Characteristics:

    - **Data Source**: Likely derived from global health or economic databases, focusing on suicide rates across different countries, age groups, and genders over time.
    - **Temporal Coverage**: 1985 to 2016.
    - **Geographical Coverage**: 101 countries.

- **Main Variables of Interest**: Suicide numbers, suicide rates per 100k population, GDP, population, and Human Development Index.

3. <u>Overall Socio-Economic Impact:</u>

1. **Public Health Policy**: By identifying vulnerable demographics (age groups, genders, countries), governments can develop targeted public health campaigns and allocate resources to areas most in need.
2. **Economic Interventions**: Understanding the relationship between economic factors and suicide rates can guide economic policies that mitigate risks, such as providing unemployment benefits, promoting job security, and reducing income inequality.
3. **Social Services**: Insights from the dataset can justify investments in social services, mental health resources, and community support networks, particularly for groups identified as high-risk.
4. **Global Awareness and International Collaboration**: Countries with lower suicide rates might offer lessons for those with higher rates. International collaborations can be formed to share successful intervention strategies, enhancing global well-being.

4. <u>Usage</u>

This dataset is valuable for socio-economic research, particularly in analyzing how demographic factors, economic conditions, and development indices influence suicide rates across different countries and over time.

Advanced Graphs and Analysis

Let's create some advanced visualizations like word cloud, box plot, whisker plot, and regression plot.

1. **Word Cloud**: To visualize the frequency of different countries.
2. **Box Plot**: To show the distribution of suicides per 100k population across different age groups.
3. **Whisker Plot**: To visualize the distribution of suicides per 100k population by gender.
4. **Regression Plot**: To visualize the relationship between GDP per capita and suicides per 100k population.

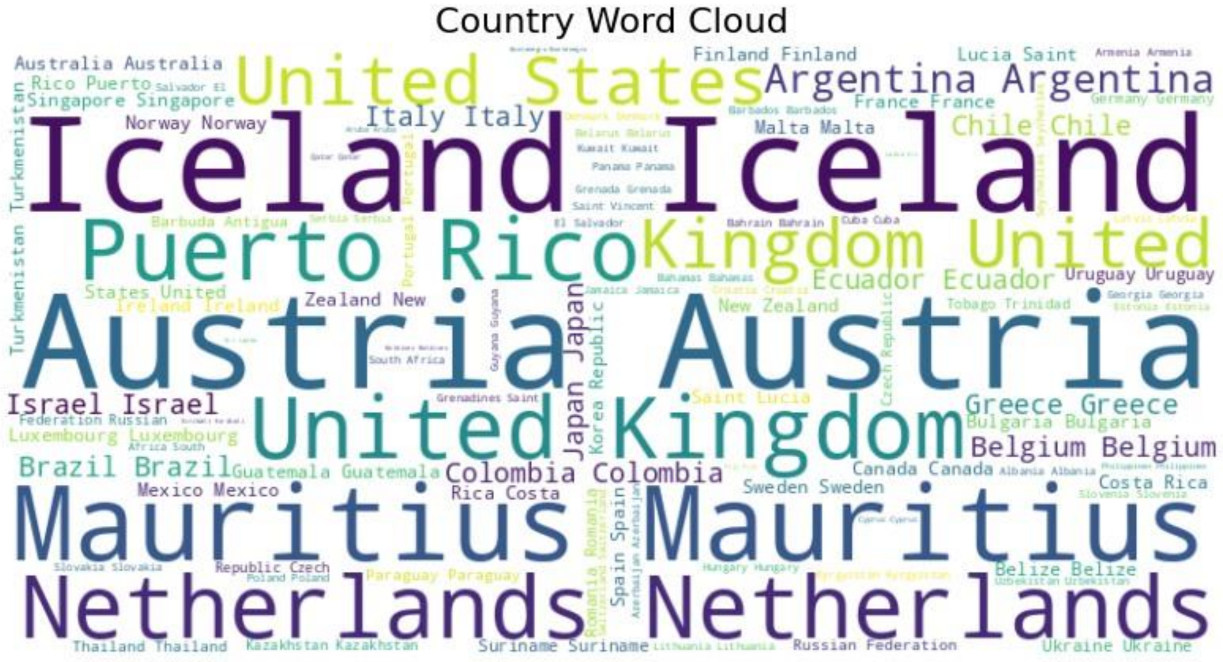
Charts and Analysis:

1)Word Cloud

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Generate a word cloud for the 'country' column
wordcloud = WordCloud(width=800, height=400,
background_color='white').generate(' '.join(data['country']))


# Display the word cloud
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Country Word Cloud', fontsize=16)
plt.show()
```



The word cloud visualizes the frequency of countries in the dataset. Countries that appear more frequently in the dataset have larger and bolder text in the word cloud. This helps to quickly identify which countries contribute the most data to the dataset.

Analysis:

The word cloud visualization shows the frequency of different countries in the dataset. Larger and bolder text indicates countries with more data entries. This can help quickly identify which countries contribute the most to the dataset, and may also point towards regions with potentially higher suicide rates. Understanding country representation is crucial for conducting region-specific analysis.

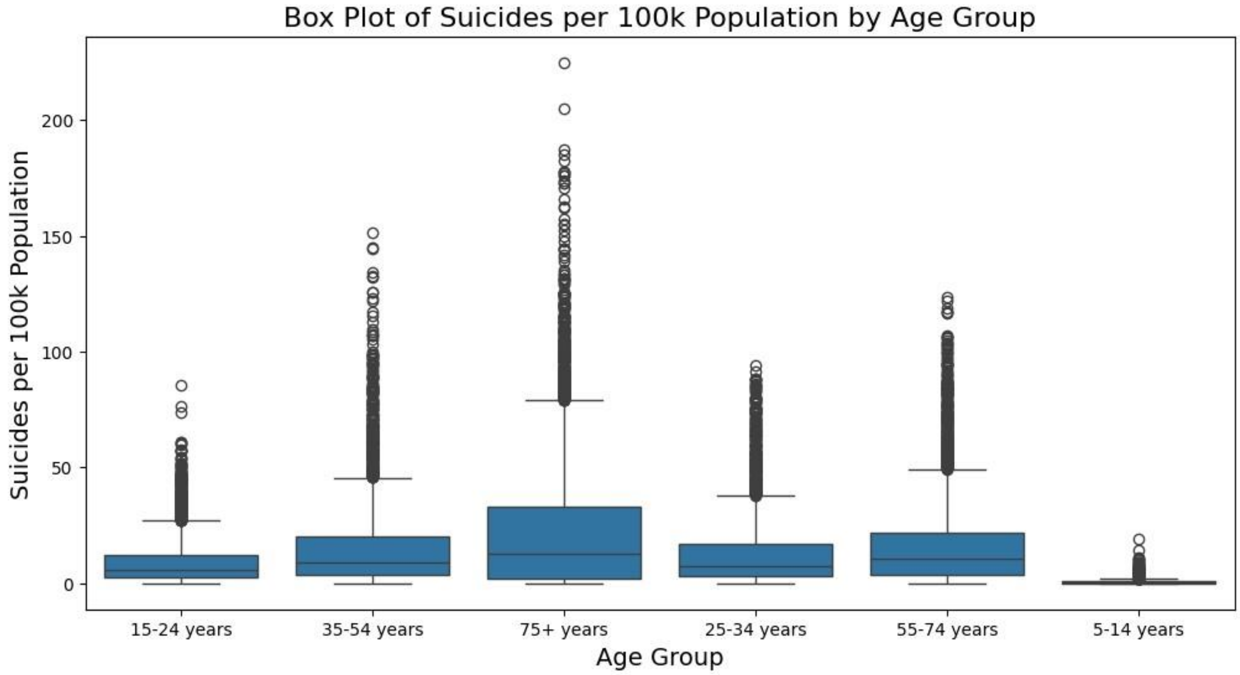
Insight: The word cloud helps identify the countries most represented in the dataset. By focusing on these countries, researchers can conduct more in-depth analyses of regions where suicide rates are particularly high or low, enabling targeted interventions.

Socio-Economic Relevance: Identifying countries with higher data representation allows policymakers to study specific socio-economic conditions (e.g., poverty, unemployment, mental health services) in those regions, leading to more effective policy decisions.

2)Box Plot by Age Group

```
import seaborn as sns


# Create a box plot for suicides per 100k population across different age
groups
plt.figure(figsize=(12, 6))
sns.boxplot(x='age', y='suicides/100k pop', data=data)
plt.title('Box Plot of Suicides per 100k Population by Age Group',
fontsize=16)
plt.xlabel('Age Group', fontsize=14)
plt.ylabel('Suicides per 100k Population', fontsize=14)
plt.show()
```



The box plot visualizes the distribution of suicides per 100,000 people across different age groups. Key observations include:

- **Median Values**: The median number of suicides per 100k population varies across age groups, with higher medians observed in older age groups (e.g., 75+ years).
- **Interquartile Range (IQR)**: The IQR, representing the middle 50% of the data, is wider for older age groups, indicating greater variability in suicides per 100k population.
- **Outliers**: Outliers are present in most age groups, especially in younger and older groups, indicating some extreme cases of high suicide rates.

Insight: This plot highlights the distribution of suicides across different age groups, showing which age groups are more vulnerable.
Socio-Economic Relevance: Age-specific patterns of suicides can guide the allocation of resources and the design of age-appropriate mental health services. For instance, if older age groups exhibit higher suicide rates, governments might invest more in elderly care, social inclusion programs, and healthcare services aimed at older populations.

Analysis:

The box plot shows the distribution of suicides per 100,000 population across different age groups. Key insights include:

- Older age groups generally show a higher median suicide rate.
- The variability in suicide rates (indicated by the interquartile range) is higher in older age groups.
- Outliers indicate extreme cases, which could signal areas needing further investigation.

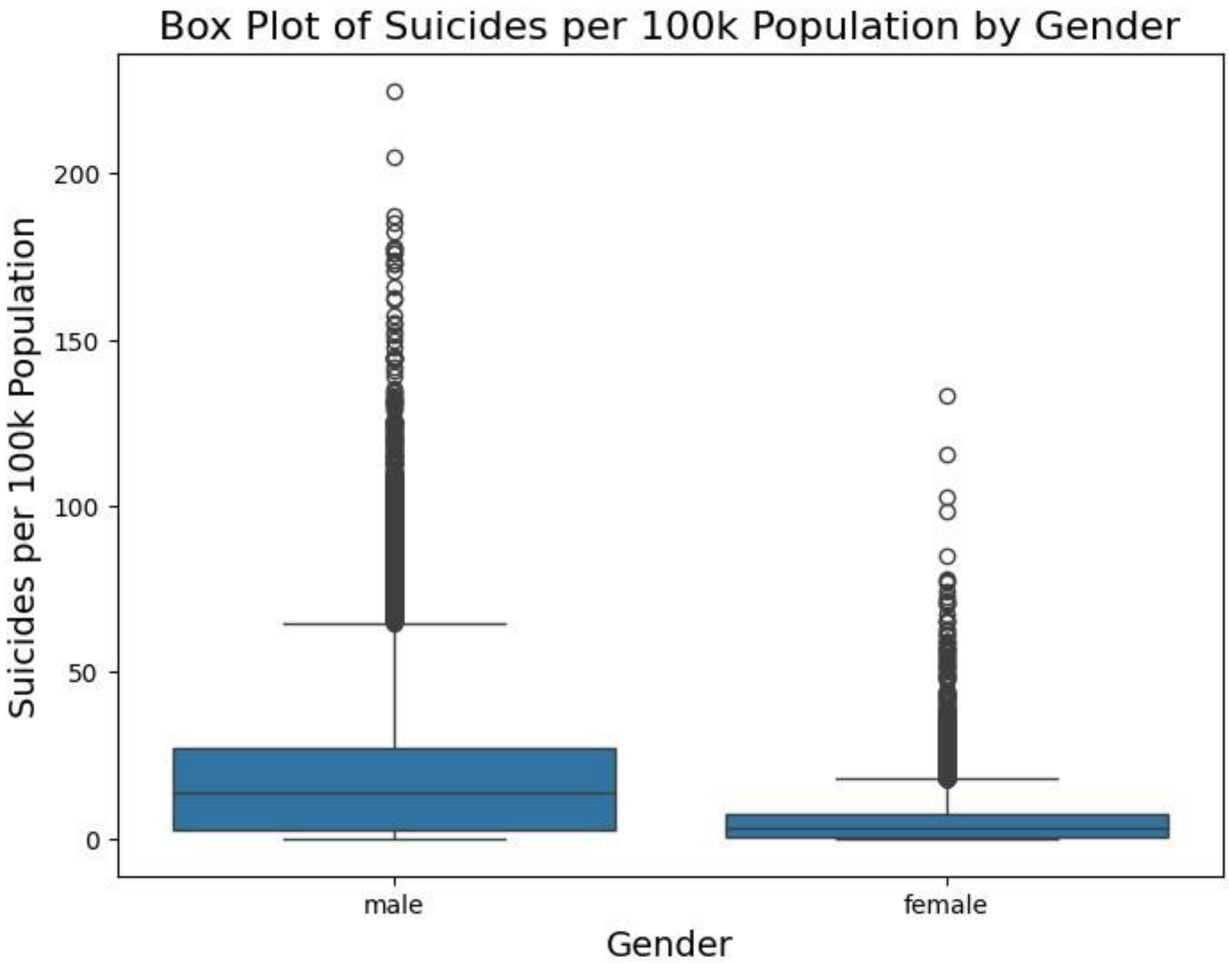
This plot helps identify which age groups are most at risk, guiding age-targeted interventions.

3)Whisker Plot by Gender

```
import seaborn as sns
import matplotlib.pyplot as plt


# Create a box plot (whisker plot) for suicides per 100k population by
gender
```

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='sex', y='suicides/100k pop', data=data)
plt.title('Box Plot of Suicides per 100k Population by Gender',
fontsize=16)
plt.xlabel('Gender', fontsize=14)
plt.ylabel('Suicides per 100k Population', fontsize=14)
plt.show()
```



This plot will show the distribution of suicides per 100k population for males and females. You can analyze whether there is a significant difference between the genders.

Analysis:

This plot displays the distribution of suicides per 100k population by gender. Key observations might include:

- Differences in median suicide rates between men and women.
- Variability within each gender group, highlighting the presence of outliers.
- Understanding these gender-based differences is essential for crafting gender-sensitive mental health programs.

Insight: The gender-based distribution of suicides per 100k population shows whether one gender is more susceptible to suicides than the other.

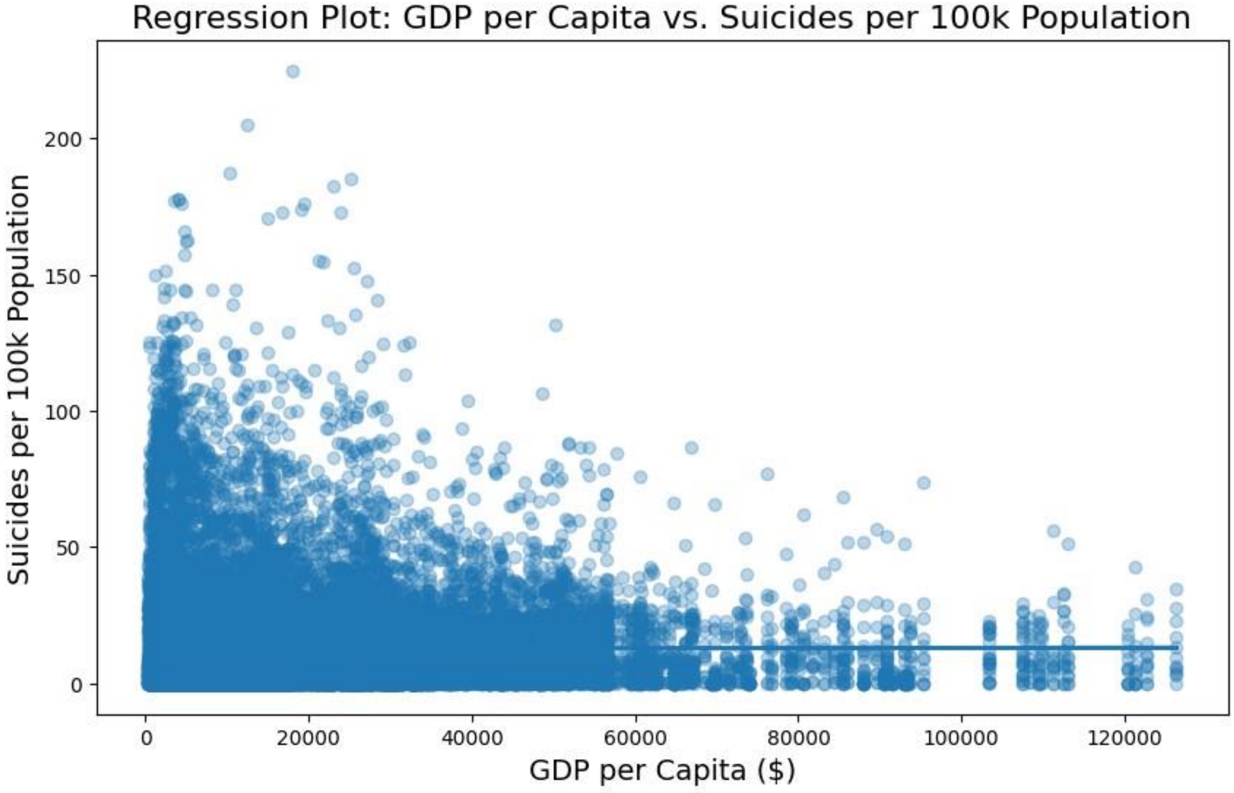
Socio-Economic Relevance: Understanding gender differences in suicide rates can help in creating gender-sensitive policies and interventions. For example, if men are found to have higher suicide rates, this might indicate the need for targeted mental health programs, addressing issues such as toxic masculinity, job stress, and social isolation that disproportionately affect men.

4)Regression Plot: GDP per Capita vs. Suicides per 100k Population

```
# Remove leading and trailing spaces from column names
data.columns = data.columns.str.strip()

# Convert GDP for year to numeric after removing commas
data['gdp_for_year ($)'] = data['gdp_for_year ($)'].replace({'\$': '', ',': ''}, regex=True).astype(float)

# Create a regression plot
plt.figure(figsize=(10, 6))
sns.regplot(x='gdp_per_capita ($)', y='suicides/100k pop', data=data, scatter_kws={'alpha':0.3})
plt.title('Regression Plot: GDP per Capita vs. Suicides per 100k Population', fontsize=16)
plt.xlabel('GDP per Capita ($)', fontsize=14)
plt.ylabel('Suicides per 100k Population', fontsize=14)
plt.show()
```



The regression plot will help you understand the relationship between GDP per capita and the number of suicides per 100k population. Look for the trend line's slope and the scatter of data points to infer whether wealthier countries have higher or lower suicide rates.

Insight: The regression plot explores the relationship between a country's economic status (measured by GDP per capita) and its suicide rates.

Socio-Economic Relevance: This plot can reveal whether wealthier countries (higher GDP per capita) have lower or higher suicide rates. If there is a strong correlation, it can indicate that economic factors such as unemployment, inequality, or financial insecurity significantly influence mental health. This information is critical for economic policy, as it underscores the importance of economic stability and equitable distribution of wealth in reducing suicide rates.

Analysis:

This plot examines the relationship between a country's economic status (GDP per capita) and its suicide rate. A key question is whether there is a correlation between economic well-being and suicide rates:

- A downward trend could suggest that higher economic well-being (higher GDP per capita) is associated with lower suicide rates.
- Conversely, an upward trend could indicate that wealthier countries face higher suicide rates, potentially due to factors like increased social pressure or inequality.

This plot can reveal how economic conditions influence mental health, informing economic policies aimed at suicide prevention.

Conclusion:

These visualizations offer a multi-faceted view of the suicide data, highlighting key socio-economic factors like age, gender, and economic status. Each graph provides a unique perspective, contributing to a comprehensive understanding of the data. By analyzing these plots, policymakers and researchers can identify vulnerable groups and develop targeted strategies to reduce suicide rates globally.