UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

5DATA006C.1 DATA VISUALIZATION AND COMMUNICATION

COURSEWORK

Lecturer Name: Mr. Fouzul Hassan

Name: M S Noor

UOW ID: w2052142

IIT No: 20231154

Contents

Research Question 3

Data Preparation 4

Exploratory Data Analysis 8

Data Storytelling 14

Research Question

The research question I have chosen is "**How do suicide rates vary by gender and age groups across different countries?** The question is highly relevant as it addresses a critical issue. Suicide remains as a leading cause of death globally, and by identifying high risk groups can help allocate effectively, making this a significant public priority. Insights gained from this analysis can assist policymakers evolve and take relevant actions to prevent specific groups and cultural context. Understanding the variations in suicide rates by demographic factors such as gender, age groups, countries is crucial for creating prevention strategies.

This research question is important, because suicide rates are influenced by so many factors such as psychological, biological, social and cultural. Cross country analysis shows how these factors play a role In contributing variations in suicide rates.

This research can help reduce the inequalities in mental health care by indicating the contrast between suicide rates among different community groups. The research provides valuable insights into the global impact of suicide and assists to make new prevention strategies to address the issue.

Research question source: https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016/data

Data Preparation

I made some changes for the dataset in excel as well, because the age group column had range age type (14-34). To have a tidy dataset u cannot have multi values in a cell. To make the dataset tidy, I split the age range to two columns and found the mean age. I round off the mean age to the nearest whole number to make it easier to analyze, because it's not practical to have decimal numbers in age.

Converting an object data type into an integer

The gdp_for_year column data contains as an object data type, because it mostly contains string values. To do mathematical analyzations we need statistical data types such as integer and float. So, changing the data type will also prevent from unexpected errors from incorrect data types.

```
[ ]  #COnverting object into an integer
     #converts data type of gdp_for_year column to an integer
     DataViz[' gdp_for_year  '] = DataViz[' gdp_for_year  '].str.replace(',', '').fillna(0).astype(int)
```

```
DataViz.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   country            1000 non-null   object 
 1   year               1000 non-null   int64  
 2   sex                1000 non-null   object 
 3   population         1000 non-null   int64  
 4   suicides/100k pop  1000 non-null   float64
 5   HDI for year       216 non-null    float64
 6    gdp_for_year      1000 non-null   int64  
 7   gdp_per_capita     1000 non-null   int64  
 8   generation         1000 non-null   object 
 9   suicides_no        1000 non-null   int64  
 10  mean_age           1000 non-null   int64  
dtypes: float64(2), int64(6), object(3)
memory usage: 86.1+ KB
```

Removing rows with 0 values

I removed the rows which contain zero values in suicide_no column, because the primary goal of my analysis is to understand the factors and patterns related to suicide. So having rows with zero suicidal numbers is not relevant. Having rows with 0 suicides will skew the calculations and graphs while not having an accurate dataset.

```
[ ] # Remove rows with 0 values in 'suicides_no' column
    DataViz = DataViz[DataViz['suicides_no'] != 0]

    # Optionally reset the index if needed
    DataViz = DataViz.reset_index(drop=True)
```

```
DataViz
```

	country	year	sex	population	suicides/100k pop	HDI for year	gdp_for_year	gdp_per_capita	generation	suicides_no	mean_age
0	Albania	1987	male	312900	6.71	NaN	2156624900	796	Generation X	21	20
1	Albania	1987	male	308000	5.19	NaN	2156624900	796	Silent	16	45
2	Albania	1987	female	289700	4.83	NaN	2156624900	796	Generation X	14	20
3	Albania	1987	male	21800	4.59	NaN	2156624900	796	G.I. Generation	1	85
4	Albania	1987	male	274300	3.28	NaN	2156624900	796	Boomers	9	30
...	...	...	...	...	...	...	...	...	...	...	...
615	Armenia	1992	female	290100	0.34	NaN	1272577456	385	Generation X	1	20
616	Armenia	1993	female	53100	11.30	NaN	1201313201	357	G.I. Generation	6	85
617	Armenia	1993	male	226400	8.83	NaN	1201313201	357	Silent	20	65
618	Armenia	1993	male	28400	7.04	NaN	1201313201	357	G.I. Generation	2	85
619	Armenia	1993	male	371900	6.72	NaN	1201313201	357	Boomers	25	45

Handling the missing values in the dataset

Having missing values in a dataset is also a main reason for an untidy dataset. Having a dataset with missing values will significantly impact on the analysis of the dataset. Ignoring the missing values will create inaccurate, unreliable results. This will lead to incorrect decision makings. With having missing it will make hard to make accurate predictions.

```
[ ] #Hnadling all the missing values in HDI for year column
    DataViz['HDI for year'].fillna(DataViz['HDI for year'].mean(), inplace=True)
```

```
<ipython-input-15-fd8b7f7ad30f>:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation i

  DataViz['HDI for year'].fillna(DataViz['HDI for year'].mean(), inplace=True)
```

Before handling missing values

```
[ ] #data cleaning part
    #number of missing values for dataset variables
    Dataviz_missingvalues = DataViz.isnull().sum()
    Dataviz_missingvalues
```

	0
country	0
year	0
sex	0
population	0
suicides/100k pop	0
HDI for year	457
gdp_for_year	0
gdp_per_capita	0
generation	0
suicides_no	0
mean_age	0

dtype: int64

After handling missing values

```
#data cleaning part
#number of missing values for dataset variables
Dataviz_missingvalues = DataViz.isnull().sum()
Dataviz_missingvalues
```

	0
country	0
year	0
sex	0
population	0
suicides/100k pop	0
HDI for year	0
gdp_for_year	0
gdp_per_capita	0
generation	0
suicides_no	0
mean_age	0

dtype: int64

Dataset after handling missing values

	country	year	sex	population	suicides/100k pop	HDI for year	gdp_for_year	gdp_per_capita	generation	suicides_no	mean_age
0	Albania	1987	male	312900	6.71	0.747252	2156624900	796	Generation X	21	20
1	Albania	1987	male	308000	5.19	0.747252	2156624900	796	Silent	16	45
2	Albania	1987	female	289700	4.83	0.747252	2156624900	796	Generation X	14	20
3	Albania	1987	male	21800	4.59	0.747252	2156624900	796	G.I. Generation	1	85
4	Albania	1987	male	274300	3.28	0.747252	2156624900	796	Boomers	9	30
...	...	...	...	...	...	...	...	...	...	...	...
615	Armenia	1992	female	290100	0.34	0.747252	1272577456	385	Generation X	1	20
616	Armenia	1993	female	53100	11.30	0.747252	1201313201	357	G.I. Generation	6	85
617	Armenia	1993	male	226400	8.83	0.747252	1201313201	357	Silent	20	65
618	Armenia	1993	male	28400	7.04	0.747252	1201313201	357	G.I. Generation	2	85
619	Armenia	1993	male	371900	6.72	0.747252	1201313201	357	Boomers	25	45

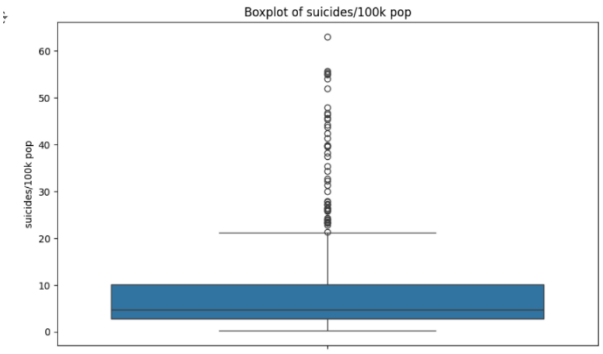
320 rows × 11 columns

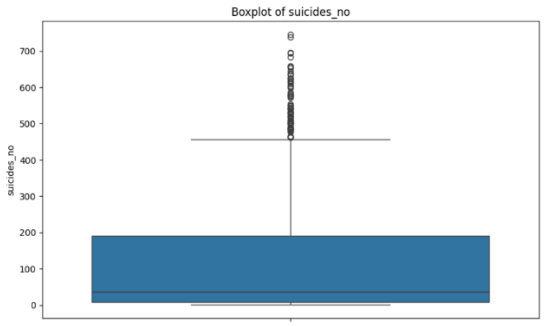
Identifying outliers and removing

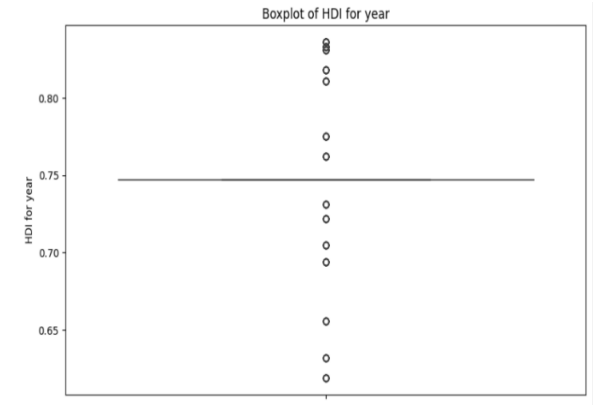
Outliers are data points which way different compared to other data points in the dataset. Outliers can occur due to data entry mistakes, measurement errors, genuine extreme values. Outliers can mislead conclusions about the variability and relationships. This can affect the overall performance of the model. In my dataset there were outliers in 3 variables.

```python
# Identify outliers using the IQR method
def identify_outliers(DataViz):
    Q1 = DataViz.quantile(0.25)
    Q3 = DataViz.quantile(0.75)
    IQR = Q3 - Q1
    outliers = ((DataViz < (Q1 - 1.5 * IQR)) | (DataViz > (Q3 + 1.5 * IQR))).sum()
    return outliers

# Apply the function to the dataset
outliers = identify_outliers(DataViz.select_dtypes(include=['float64', 'int64']))

# Print the number of outliers in each variable
print("Number of outliers in each variable:")
print(outliers)

# Review the outliers by visualizing them using boxplots
def visualize_outliers(DataViz, column):
    plt.figure(figsize=(10, 6))
    sns.boxplot(data=DataViz, y=column)
    plt.title(f'Boxplot of {column}')
    plt.show()

# Visualize outliers for relevant columns
columns_to_review = ['suicides/100k pop', 'HDI for year', 'suicides_no']
for column in columns_to_review:
    visualize_outliers(DataViz, column)
```

```
Number of outliers in each variable:
year                 0
population           0
suicides/100k pop    51
HDI for year         163
 gdp_for_year        0
gdp_per_capita       0
suicides_no          75
mean_age             0
dtype: int64
```



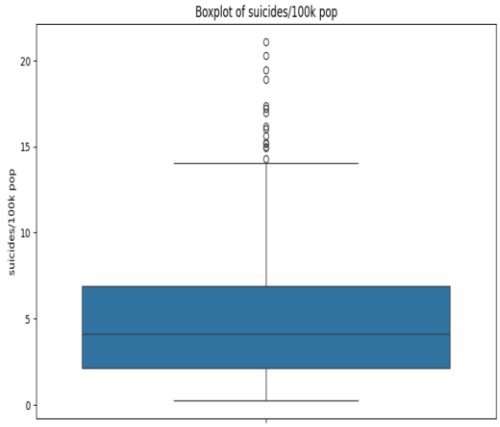


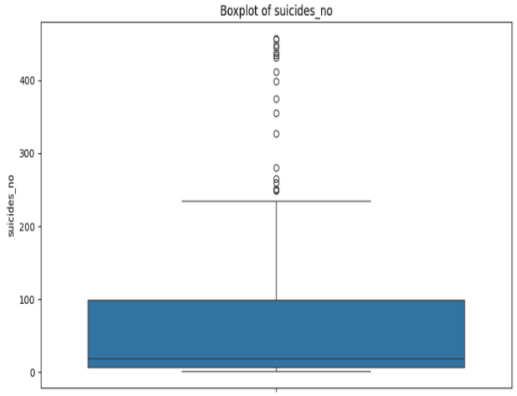


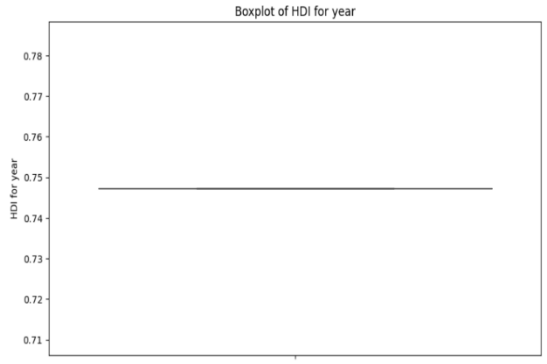
Removing outliers will increase the overall quality of the dataset. Even though I removed outliers from the variables, still there's some outliers within the variables. Because those outliers has extreme genuine values.

```
[ ] def remove_outliers(DataViz, columns):
        Q1 = DataViz[columns].quantile(0.25)
        Q3 = DataViz[columns].quantile(0.75)
        IQR = Q3 - Q1
        df_no_outliers = DataViz[~((DataViz[columns] < (Q1 - 1.5 * IQR)) | (DataViz[columns] > (Q3 + 1.5 * IQR))).any(axis=1)]
        return df_no_outliers

    # Columns to remove outliers from
    columns_to_remove_outliers = ['suicides/100k pop', 'HDI for year', 'suicides_no']

    # Print the number of data points before and after removing outliers
    print(f"Number of data points before removing outliers: {len(DataViz)}")


    # Remove outliers from the dataset
    DataViz = remove_outliers(DataViz, columns_to_remove_outliers)

    print(f"Number of data points after removing outliers: {len(DataViz)}")
```

```
Number of data points before removing outliers: 620
Number of data points after removing outliers: 383
```



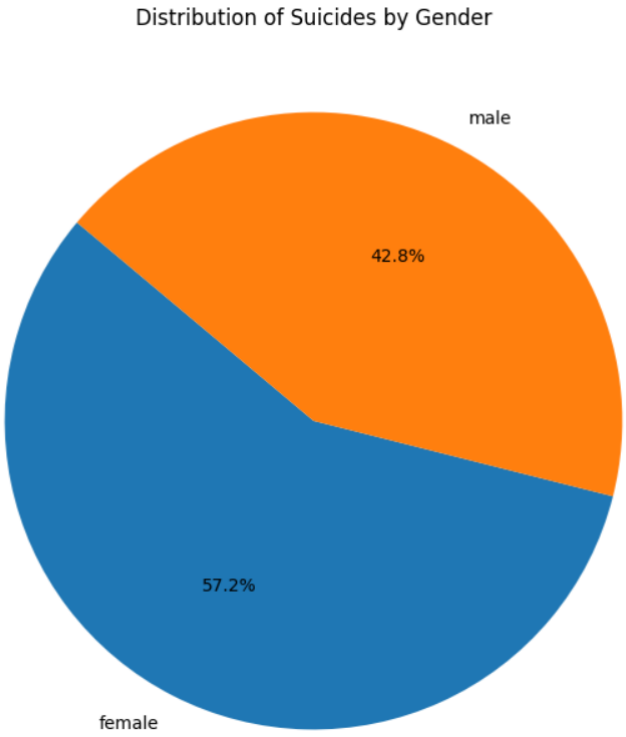




Exploratory Data Analysis

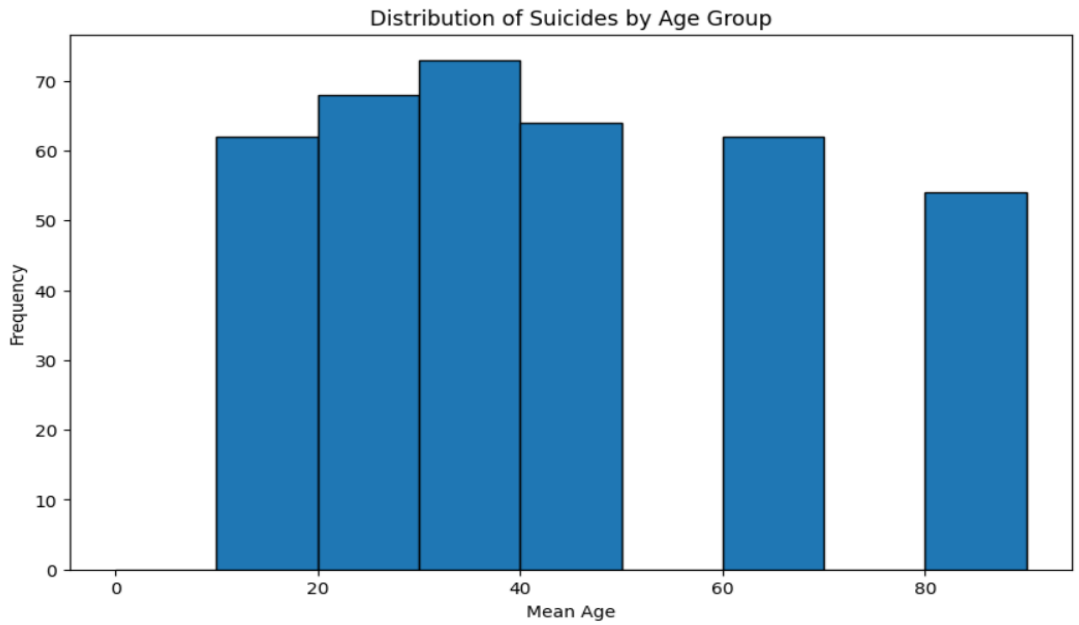
Univariant Analysis

Distribution of suicide by gender (Pie chart)



The above pie chart shows the distribution of suicides by gender according to the dataset. This analysis is a univariate analysis, because it focus only on one variable(gender). This EDA graph aims to understand the suicides occurred across genders. By observing the pie chart we can identify which gender has higher suicide percentage. This insight is valuable to understand to investigate and target the reason behind this intervention. This gender distribution helps to seize the trends and imbalance of suicide occurrence between male and females.

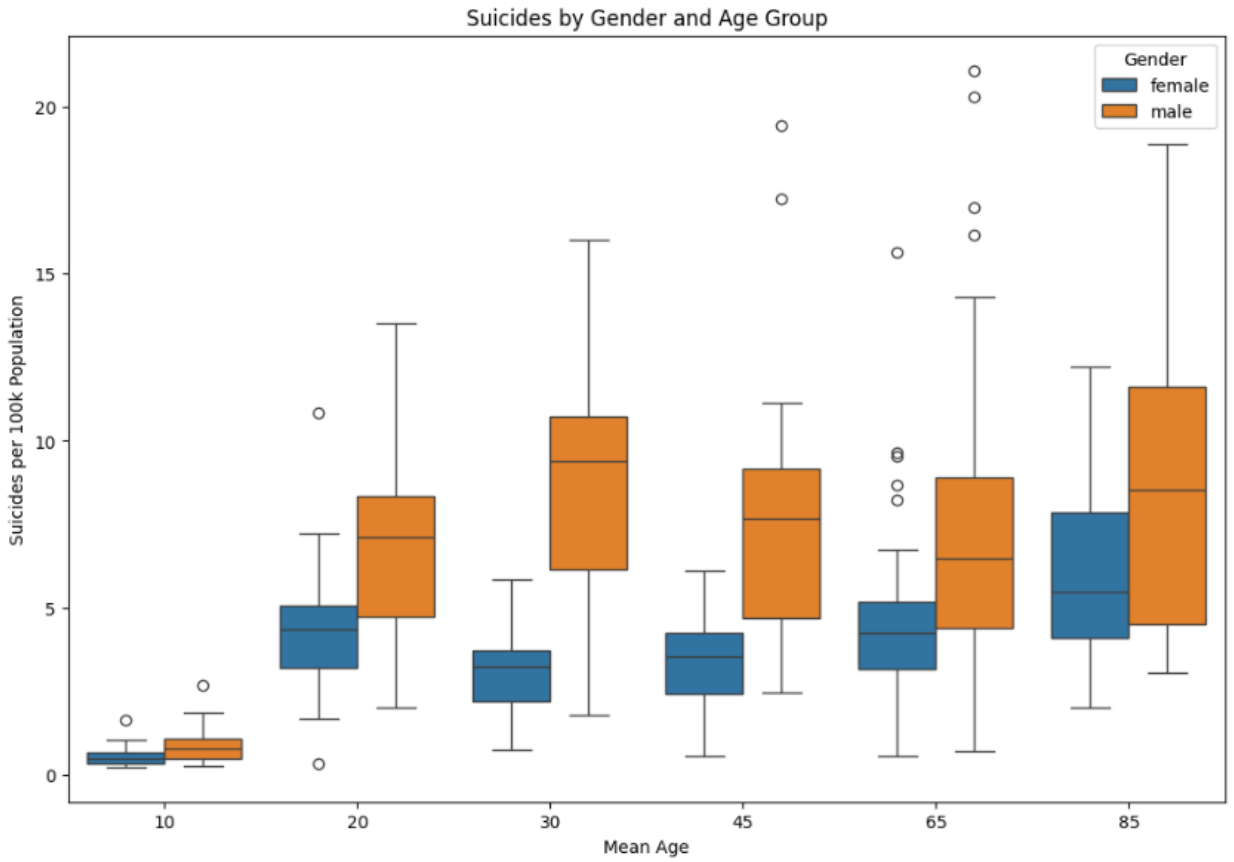
Distribution of suicides by Age group (Histogram)



The above histogram graph also indicates a univariate analysis according to the research question. The EDA graph represents the frequency of suicides with each age range by insighting to age related pattern of suicides. We can identify in which age range has the highest frequencies of suicides. We can see that this distribution is positively skewed. The age range of 30-40 has high risk of having suicides according to the above EDA graph. This insight can prompt further for research in factors pitched in to increase this risk and take necessary actions.

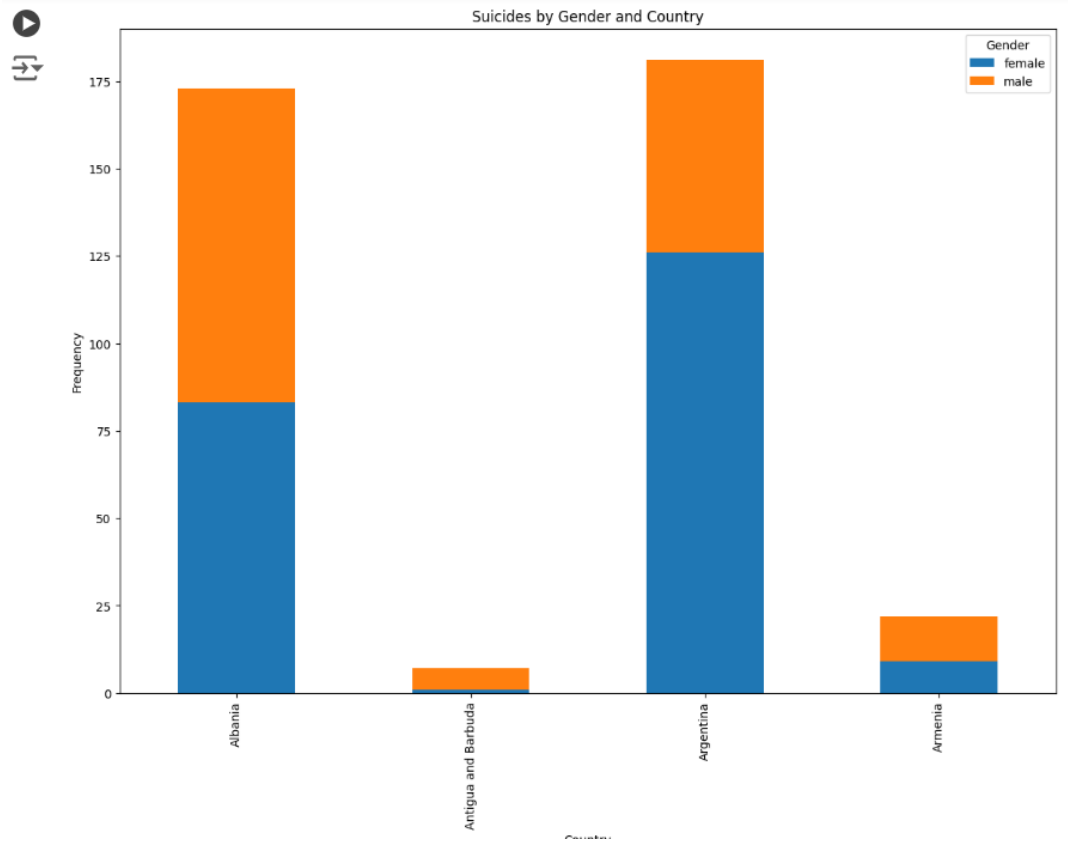
Multivariant Analysis

Suicides by gender and age groups (boxplot)



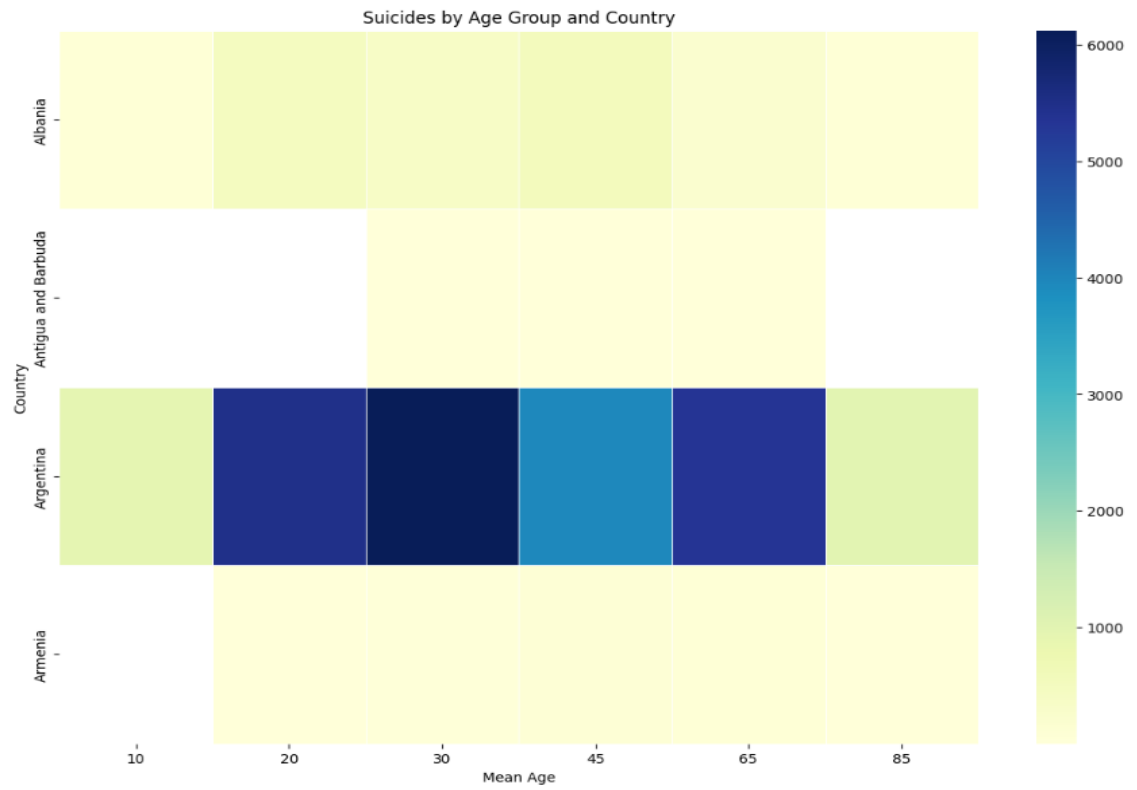
The above diagram is a clear example of a multivariate analysis, because it examines the relationship between three variables. This analysis go beyond the univariate and bivariate analysis. This EDA boxplot visualize the distribution between gender and mean age. This boxplot allows us to make comparisons of suicides rates across different age groups within each gender and vise versa. The plot shows that the males has the high risk of suicide in gender because the increase is steeper for males. As a limitation the graph shows only a summary of the distribution but doesn't show individual data points to provide additional insights.

Suicides by gender and country (stacked bar chart)



The above stacked bar chart also a multivariate analysis EDA graph. This graph explores the relationship between gender, country and suicide frequency simultaneously. The graph allows to do comparisons of suicides across countries within each gender and vise verca. We can identify the countries which involved in higher or lower suicide rates by the height of the bars across genders. We can observe the variations in gender differences across countries. Some countries shows a huge gap between genders across countries compared to others, indicating societal and cultural factors influencing suicidal patterns.

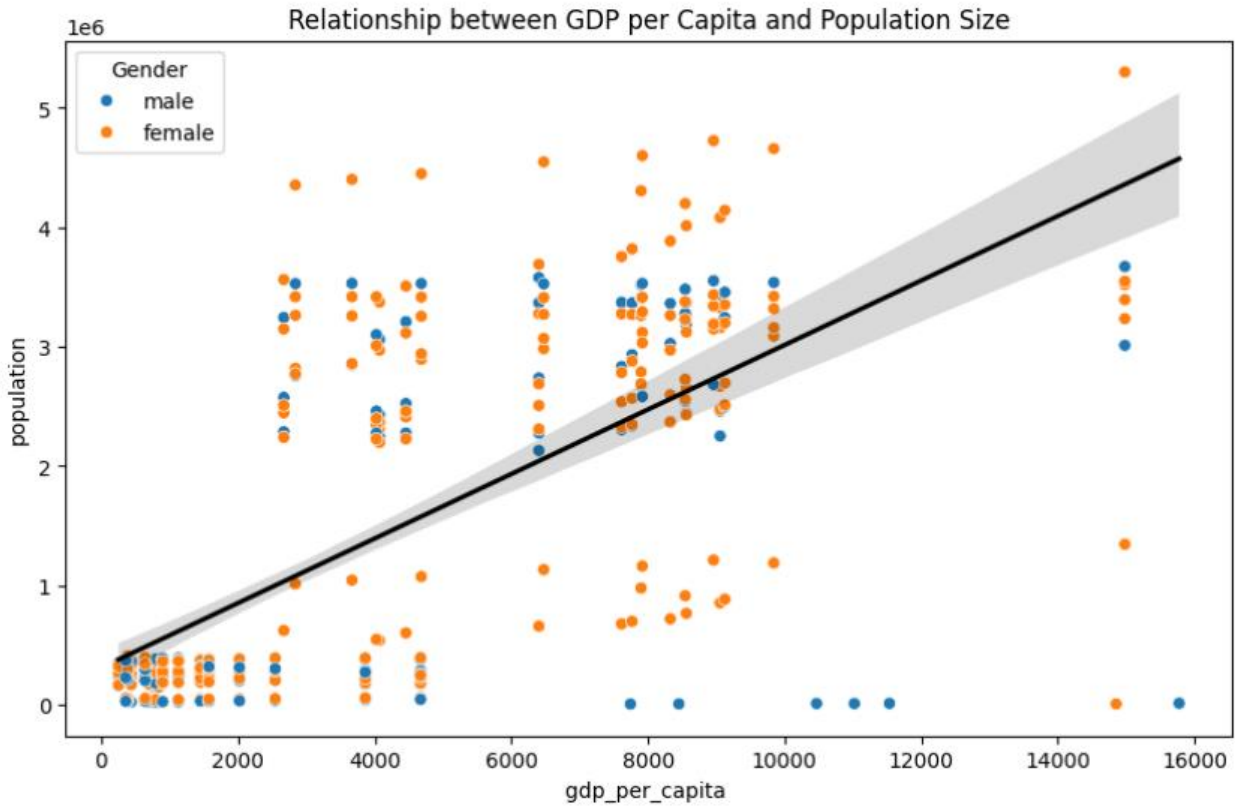
Suicides by age group and country (heatmap)



The heatmap which indicates above is also a multivariate relationship analysis by using color intensity to appear suicide count for each country and age group. We can identify countries and age groups higher and lower suicide rates by observing color intensity of the cells in the heatmap. The darker colored cells shows the age groups that are in danger to suicide within specific countries. This analysis will help us to prevent the suicides happening in different segments.

Bivariant Analysis

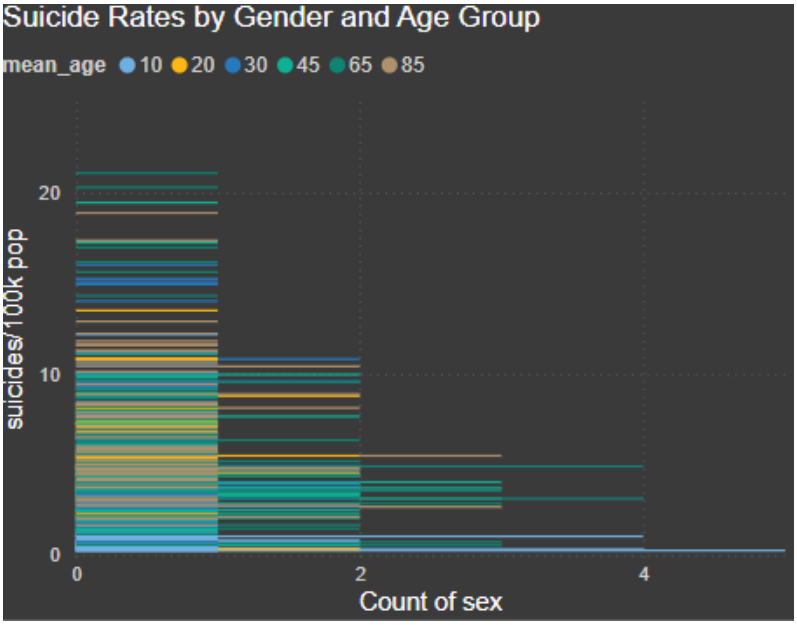
Relationship between GDP per capita and population (scatter plot)



The above scatter plot diagram illustrate bivariate relationship between GDP per capita and population. This is a statistical method which can use to analyze two variables. This aims to learn whether there is a correlation between the variables and understand the strength of the relationship. The trend line in the graph shows a positive slope and correlation means when one variable increases the other variable tends to increase,which means the countries with higher GDP has higher population.

Data Storytelling

Suicide remains as a critical global issue, and analyzing its trend across, age groups, genders and countries is crucial for designing constructive interventions. Let's hunt through these patterns using the worldwide data.

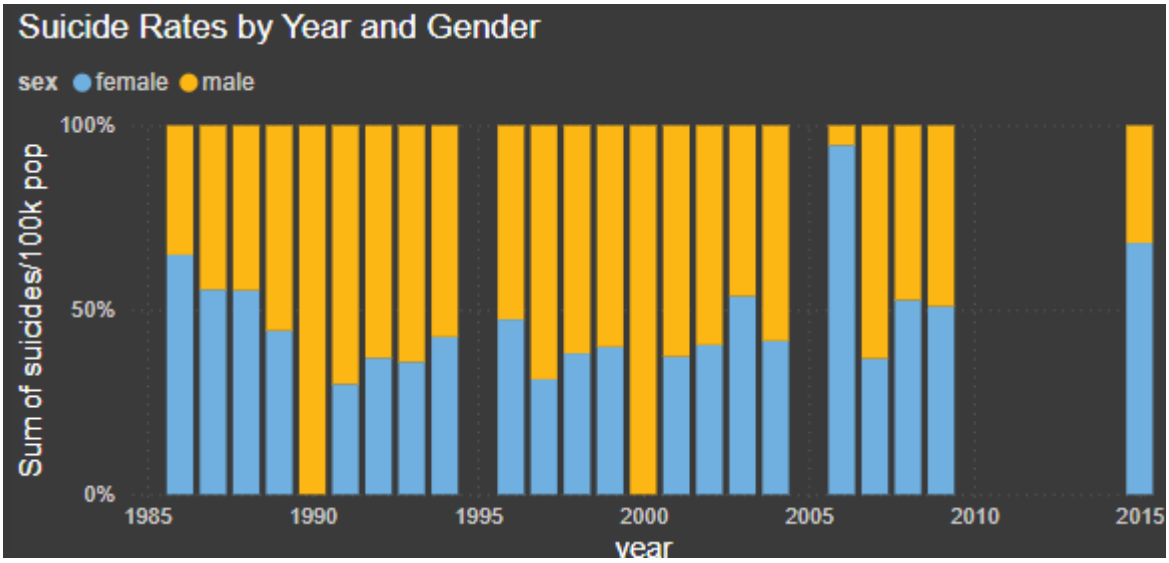


The above stacked bar chart indicates the average suicide rates per 100,000 population, classified by gender and further broken down by age group.

The X axis in the chart appears as gender and for Y axis indicates the average suicide rate. Colors in the visuals represent the various age groups.

The chart shows a significant imbalance in suicide rates between genders, with males displays a higher suicide rate across all age groups. Middle-aged males indicate the highest suicide rates among males. These suggests that the middle-aged males may face significant amount of stress and challenges which leads to higher suicide rates. The data not emphasizing the targeted psychological interventions and support systems for males, mainly for middle-aged males. By understanding these patterns, we can assists in developing suicidal prevention strategies to particular age groups and genders. It's heartbreaking to see such high rates, and it emphasizes the importance of reaching out supporting those who are struggling.

These differences are not fixed, they evolve over time. Let's explore how these trends changed over the years...

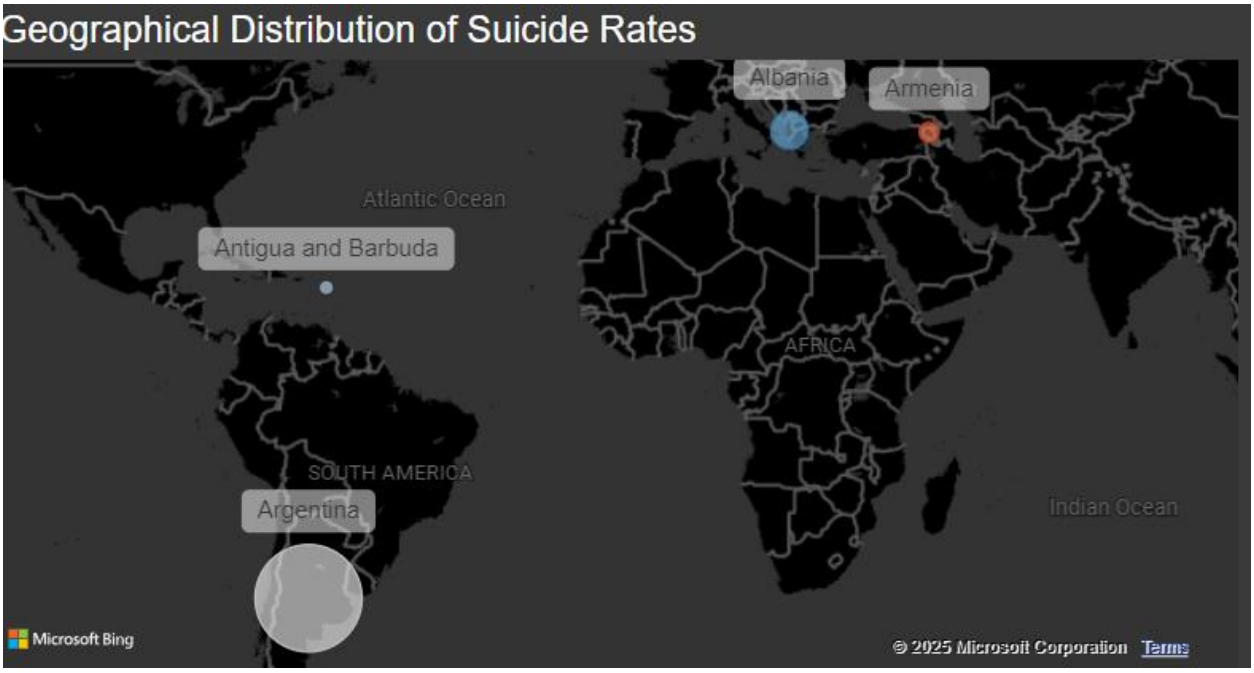


The above stacked column chart indicates suicide rates over time, categorized by gender.

The X axis represents the years and Y axis represents the suicide rates. Colors represent the two genders.

The chart accompanies the trends in suicide rates over the years by both genders. It is apparant that male suicide rates have been consistently higher than female rates. However, both genders show fluctuations in suicide rates over the years. This highlights the importance of monitoring and intervention efforts to address the altering trends in suicide rates. By understanding these patterns, policymakers and health officials can develop timely and effective strategies to turn down suicide rates. These fluctuations remind us that behind each data point is a human life and preventing these tragedies is critical.

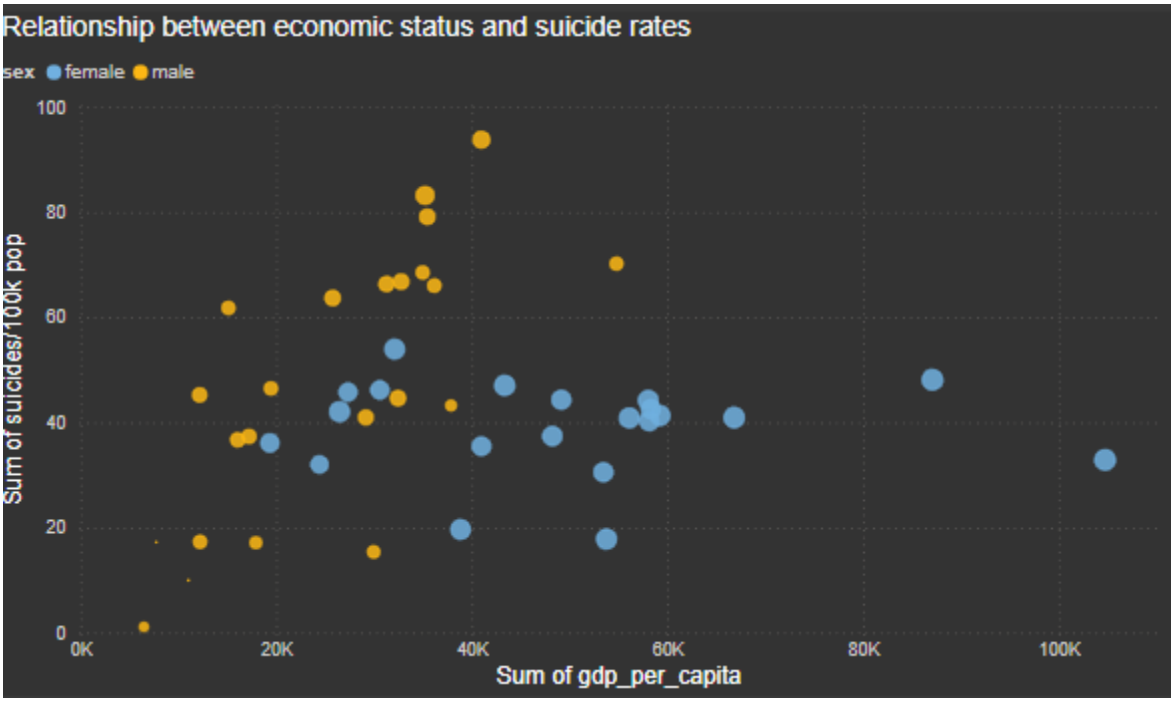
To examine deeper insights, we'll now visualize the geographical distribution of suicide rates to identify global hotspots…



This map visualization shows the geographical distribution of suicide rates worldwide. The size of the datapoints represent the total number of suicides across countries.

The map highlights the geographical distribution of suicide rates across different countries. Latin Americas and Eastern Europe shows a higher suicide rate, indicating that the economical, cultural and social factors has affected highly to these regions. The size of the data points represents the total number of suicides, providing an extensive view of the global suicide landscape. By understanding these patterns can help in formulating region specific preventive measures effectively to the areas which has high suicidal rates. It's sensible to see the global impact of suicide, and we must be united and give effort to address this issue across borders.

To explore impact of wealth and economic stability, Lets Analyze the relationship between GDP per capita and suicide rates…

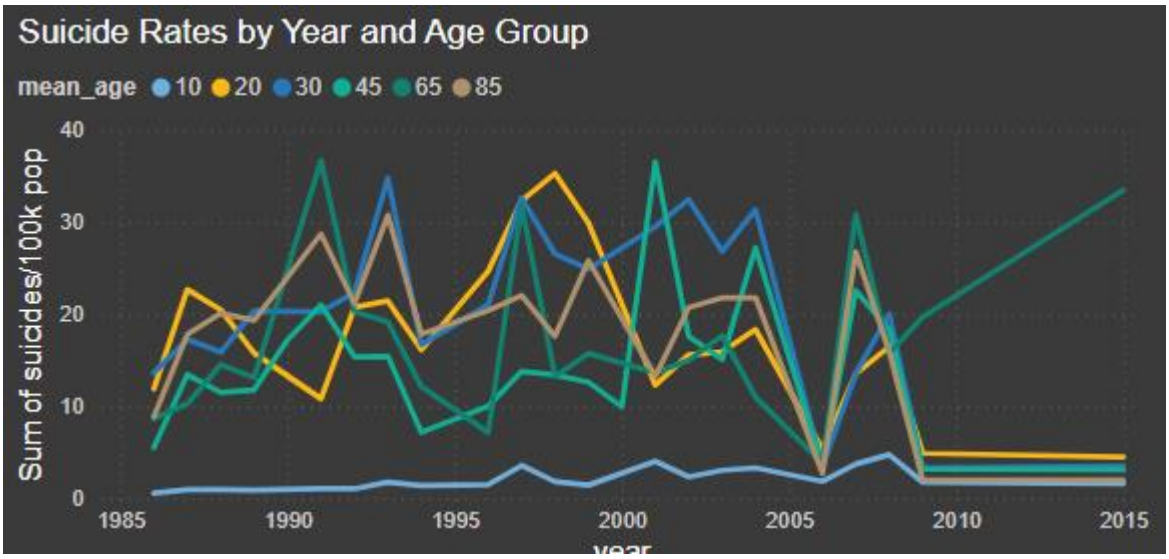


The above scatterplot indicates the relationship between GDPs per capita and suicide rates, data points sized by population and colored by gender.

The X axis represents the GDP per capita, while the Y axis represents the suicide rates.

The plot indicates that there's no correlation between a country's wealth and suicide rates. This suggests that not only economic factors determine suicide rates. Factors such as psychological, social and cultural also influence these rates. This visualization highlights the complications of suicide rates and the need of many approaches for prevention. It is sharply defined that wealth does not equate to mental well-being. This shows that every society should prioritize mental health.

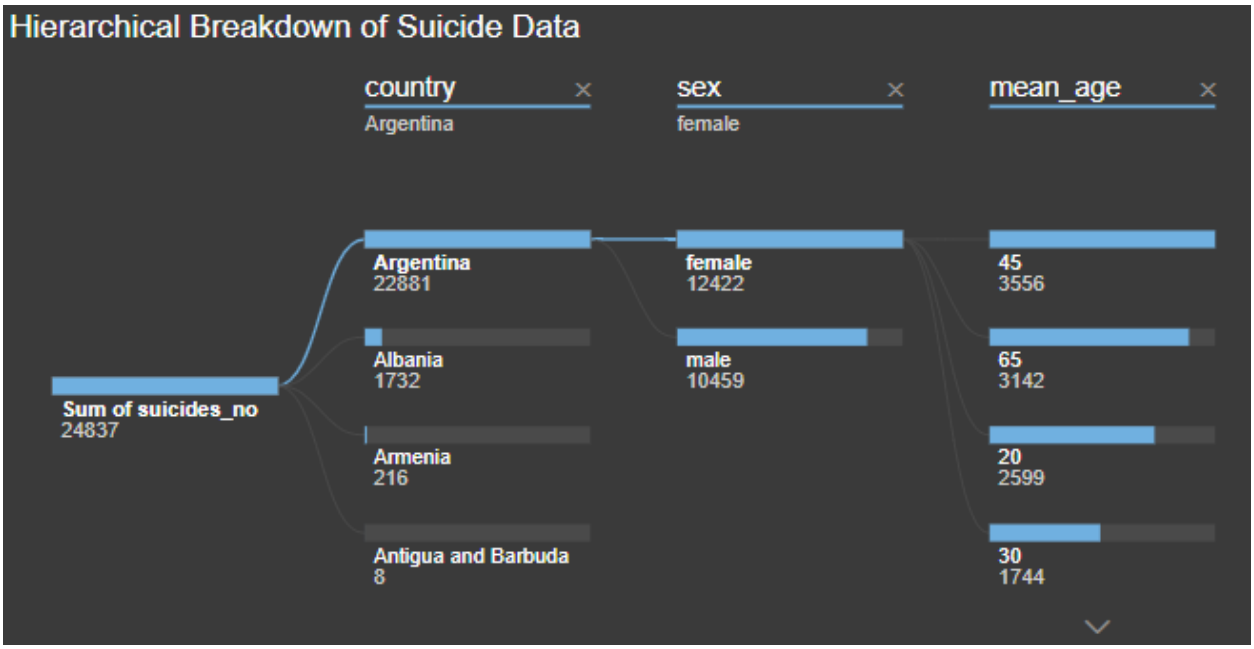
To identify endangered age groups, we'll analyze how suicide rates vary across age demographics over the years…



The line chart shows the trend analysis for suicide across age groups over time. The X axis represents the years, and the Y axis represents the suicide rate.

The chart indicates that the middle-aged individuals have higher suicide rates compared to younger and older aged groups. This suggests that the middle-aged individuals facing many challenges and stressors. This visualizations helps to identify the patterns and take necessary actions to prevent these tragedies. The lines in the chart represents real lives, and each peak is a knock on the door to provide support.

By using a hierarchical breakdown, we can drill down into the data to discover the particular demographics at risk…



The above decomposition tree diagram reveals the hierarchical breakdown of suicide data by country, gender and age group. The tree breakdowns by the number of suicides and branching out by country, gender and age groups.

The diagram reveals that certain countries have higher number of suicides among males compared to females. It also highlights the age groups most affected in a county. This breakdown helps to identify the specific demographics and the regions that require targeted interventions. By understanding the breakdown policymakers and health officials can develop more prevention strategies. Each branch of the tree represents a life lost, and it emphasizes the urgent need for prevention of suicides.

In summary, the analysis reveals significant inequality in suicide rates across genders age and regions. Male middle-aged have higher suicidal rates indicating different factors. Higher rates in regions show the influence in culture. Continuous monitoring and interventions are crucial for productive preventions. Each diagram and pattern highlight the urgency of compassionate action to save a life.

Dashboard with the infographics