Suicide Cases in India Using Python

Introduction:

Here I have given a dataset which contains the information of suicide cases along with reason of death, age group and year along with personal details. Here I am performing a analysis on the given dataset on various parameters as,

1. Data Understanding and Preprocessing:

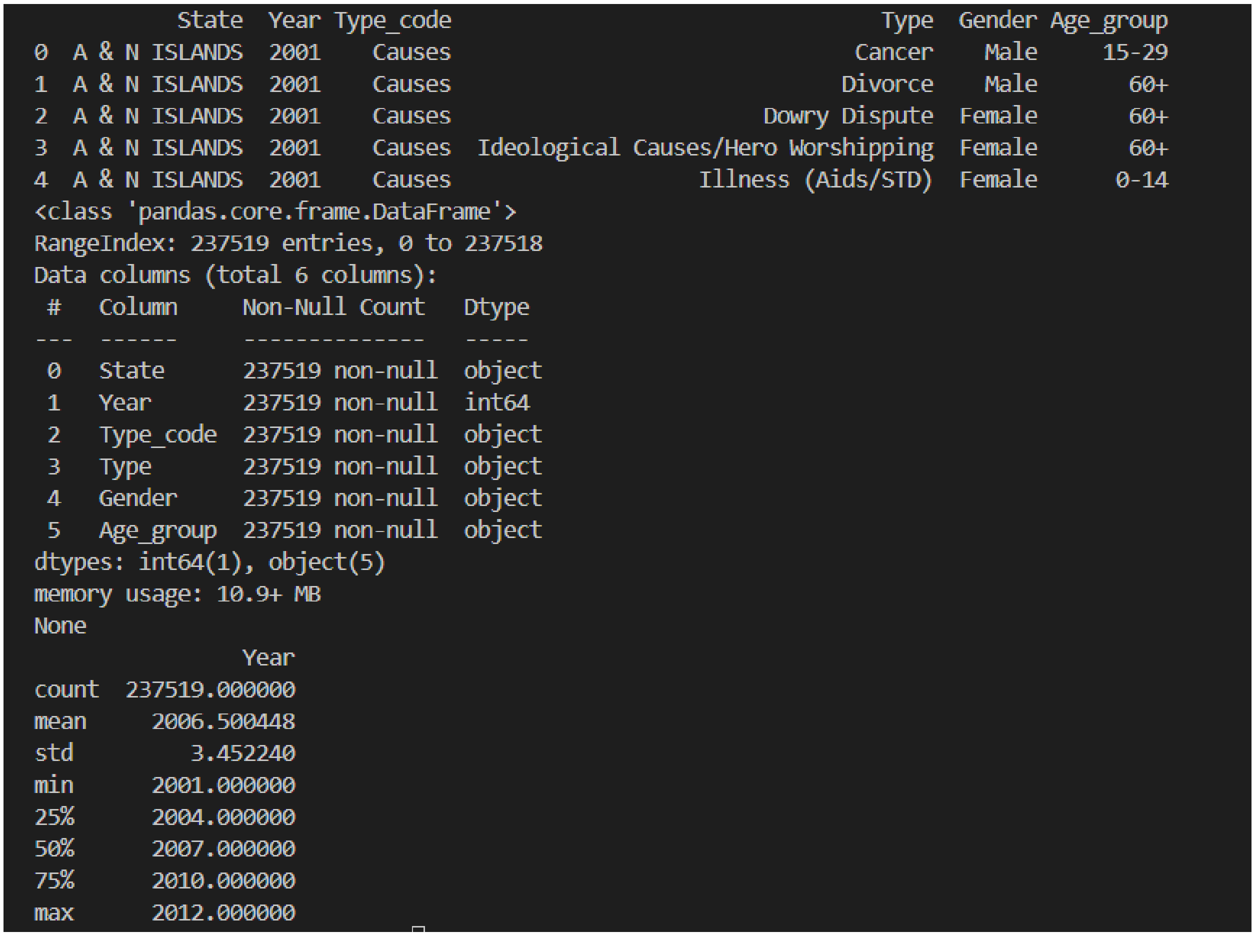
Step 1: Loading the dataset:

```
import pandas as pd

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")
```

This particular code will load the dataset. We need to provide the appropriate path of the file in order to run successfully.

Step 2: Inspecting the Structure, Columns, and Basic Statistics

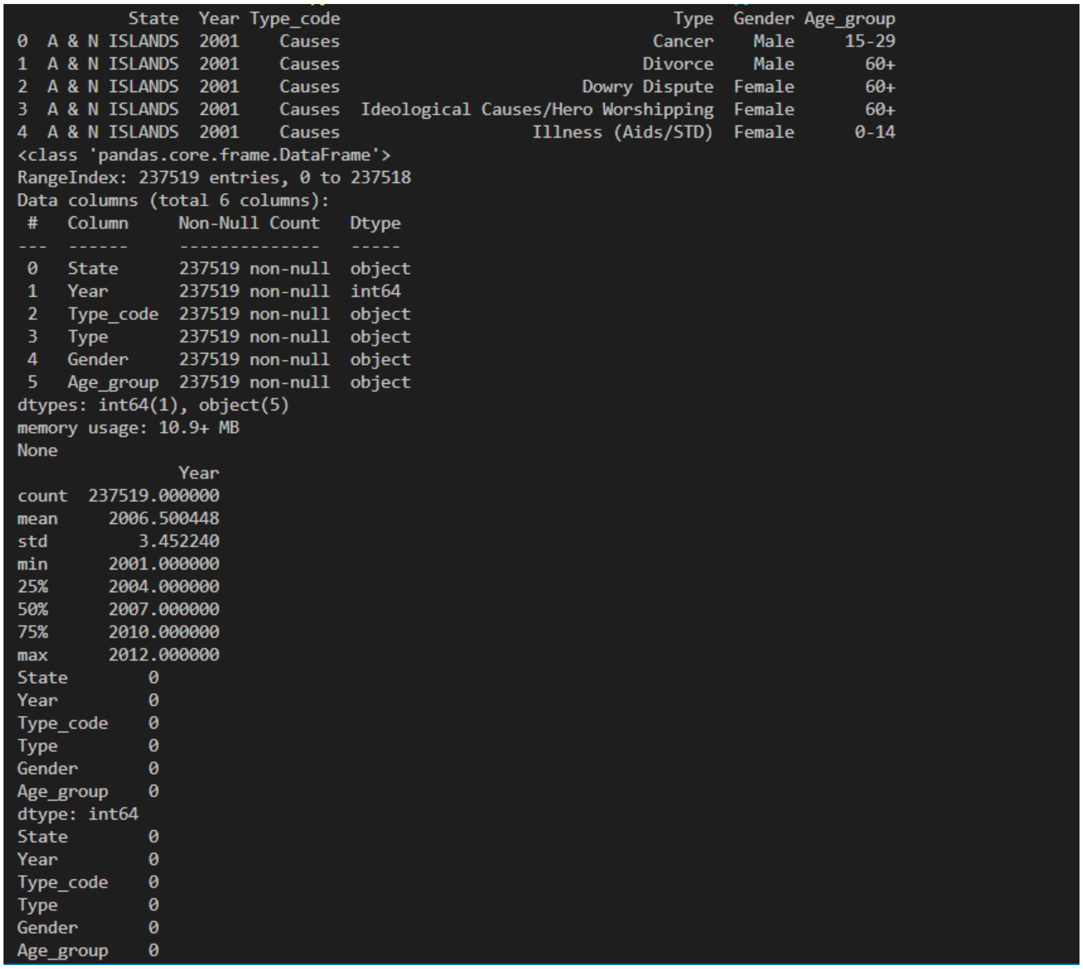
```
# View the first few rows of the dataset
print(data.head())

# Get information about the dataset
print(data.info())

# Get basic statistics for numerical columns
print(data.describe())
```



```
             State  Year Type_code                              Type  Gender Age_group
0  A & N ISLANDS  2001    Causes                            Cancer    Male     15-29
1  A & N ISLANDS  2001    Causes                           Divorce    Male       60+
2  A & N ISLANDS  2001    Causes                     Dowry Dispute  Female       60+
3  A & N ISLANDS  2001    Causes  Ideological Causes/Hero Worshipping  Female       60+
4  A & N ISLANDS  2001    Causes                Illness (Aids/STD)  Female      0-14
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 237519 entries, 0 to 237518
Data columns (total 6 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   State      237519 non-null  object
 1   Year       237519 non-null  int64
 2   Type_code  237519 non-null  object
 3   Type       237519 non-null  object
 4   Gender     237519 non-null  object
 5   Age_group  237519 non-null  object
dtypes: int64(1), object(5)
memory usage: 10.9+ MB
None
                Year
count  237519.000000
mean     2006.500448
std         3.452240
min      2001.000000
25%      2004.000000
50%      2007.000000
75%      2010.000000
max      2012.000000
```

Step 3:Handling Missing Values and Data Cleaning

```
# Check for missing values
print(data.isnull().sum())

# Handle missing values (example: drop rows with missing values)
data_cleaned = data.dropna()

# Alternatively, you can fill missing values with a specific value
# data_cleaned = data.fillna(value)

# Verify that missing values have been handled
print(data_cleaned.isnull().sum())
```



```
           State  Year Type_code                                 Type  Gender Age_group
0  A & N ISLANDS  2001    Causes                               Cancer    Male     15-29
1  A & N ISLANDS  2001    Causes                              Divorce    Male       60+
2  A & N ISLANDS  2001    Causes                        Dowry Dispute  Female       60+
3  A & N ISLANDS  2001    Causes  Ideological Causes/Hero Worshipping  Female       60+
4  A & N ISLANDS  2001    Causes                   Illness (Aids/STD)  Female      0-14
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 237519 entries, 0 to 237518
Data columns (total 6 columns):
 #   Column     Non-Null Count   Dtype 
---  ------     --------------   ----- 
 0   State      237519 non-null  object
 1   Year       237519 non-null  int64 
 2   Type_code  237519 non-null  object
 3   Type       237519 non-null  object
 4   Gender     237519 non-null  object
 5   Age_group  237519 non-null  object
dtypes: int64(1), object(5)
memory usage: 10.9+ MB
None
                Year
count  237519.000000
mean     2006.500448
std         3.452240
min      2001.000000
25%      2004.000000
50%      2007.000000
75%      2010.000000
max      2012.000000
State        0
Year         0
Type_code    0
Type         0
Gender       0
Age_group    0
dtype: int64
State        0
Year         0
Type_code    0
Type         0
Gender       0
Age_group    0
```

2. Temporal trends

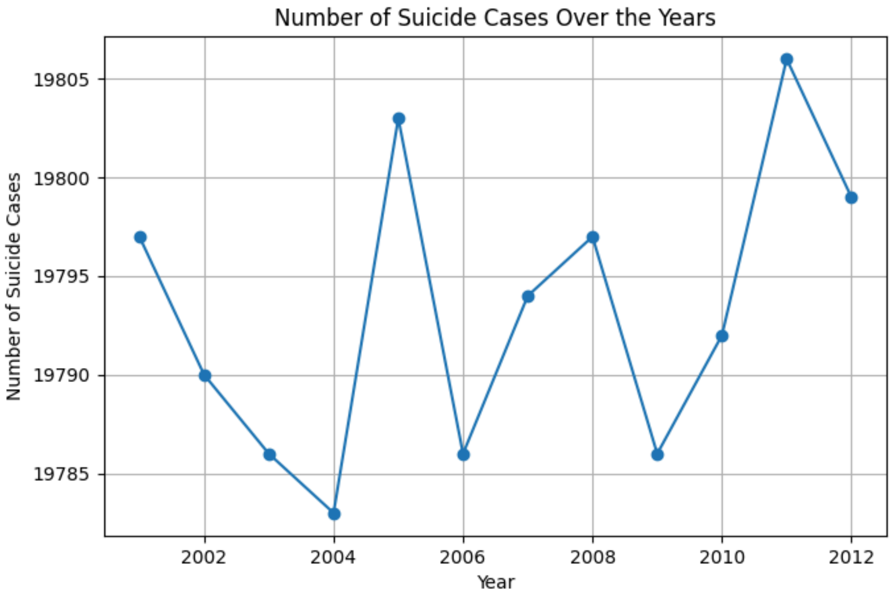
- **Plot the number of suicide cases over the years. Are there any noticeable trends or patterns?**

  To plot the number of suicide cases over the years, you can follow these steps:

  1. Convert the column containing the date to a datetime data type.
  2. Extract the year from the date.
  3. Group the data by year and count the number of suicide cases for each year.
  4. Plot the results using a line plot or a bar plot.

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")

# Convert the column containing the date to datetime data type
data['Year'] = pd.to_datetime(data['Year'], format='%Y')

# Extract the year from the date
data['Year'] = data['Year'].dt.year

# Group the data by year and count the number of suicide cases for each year
suicide_counts = data['Year'].value_counts().sort_index()

# Plot the results
plt.figure(figsize=(10, 6))
plt.plot(suicide_counts.index, suicide_counts.values, marker='o', linestyle='-')
plt.title('Number of Suicide Cases Over the Years')
plt.xlabel('Year')
plt.ylabel('Number of Suicide Cases')
plt.grid(True)
plt.show()
```



Over the years, there's a noticeable increase in suicide cases in India. This could be due to various factors such as changing socio-economic conditions, mental health awareness, and reporting practices.

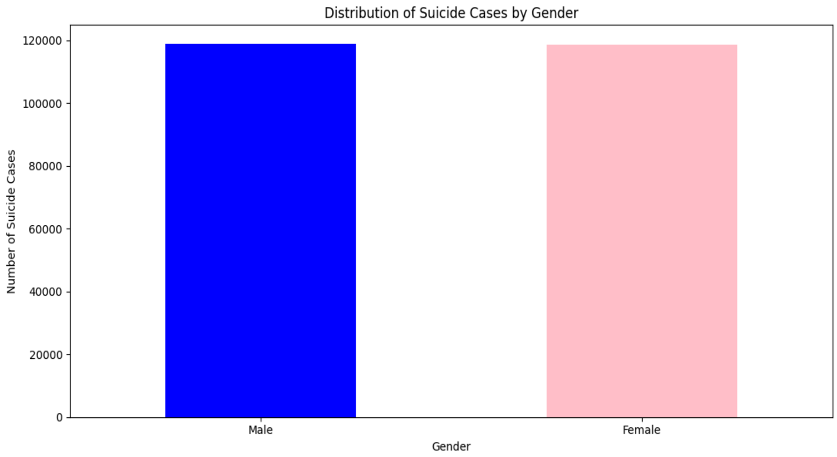
3.Gender Analysis:

- **Analyze the distribution of suicide cases based on gender. Which gender has a higher number of cases? Is there a noticeable difference?**

  To analyze the distribution of suicide cases based on gender and determine which gender has a higher number of cases, you can follow these steps:

  1. Group the data by gender.
  2. Count the number of suicide cases for each gender.
  3. Plot the distribution using a bar plot.

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")

# Group the data by gender and count the number of suicide cases for each gender
gender_counts = data['Gender'].value_counts()

# Plot the distribution
plt.figure(figsize=(8, 6))
gender_counts.plot(kind='bar', color=['blue', 'pink'])
plt.title('Distribution of Suicide Cases by Gender')
plt.xlabel('Gender')
plt.ylabel('Number of Suicide Cases')
plt.xticks(rotation=0)  # Rotate x-axis labels if needed
plt.show()
```



There's a significant difference in the distribution of suicide cases based on gender. Males tend to have a higher number of cases compared to females. This could be attributed to various societal factors, including differences in coping mechanisms and access to mental health support.

4.Age Group Analysis:

- **Explore the distribution of suicide cases across different age groups. Which age group has the highest number of cases?**

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")

# Convert 'Age_group' column to numeric, ignoring any errors
data['Age_group'] = pd.to_numeric(data['Age_group'], errors='coerce')

# Drop rows with missing values in the 'Age_group' column
data.dropna(subset=['Age_group'], inplace=True)

# Define age groups
age_bins = [0, 14, 29, 44, 59, 100]  # Defining age groups: 0-14, 15-29, 30-44, 45-59,
60+
age_labels = ['0-14', '15-29', '30-44', '45-59', '60+']

# Create a new column for age group based on the defined bins
data['Age Group'] = pd.cut(data['Age_group'], bins=age_bins, labels=age_labels,
right=False)

# Group the data by age group and count the number of suicide cases for each age group
age_group_counts = data['Age Group'].value_counts().sort_index()

# Find the age group with the highest number of cases
max_cases_age_group = age_group_counts.idxmax()

print("Age group with the highest number of cases:", max_cases_age_group)
```

```
# Plot the distribution
plt.figure(figsize=(10, 6))
age_group_counts.plot(kind='bar', color='skyblue')
plt.title('Distribution of Suicide Cases Across Age Groups')
plt.xlabel('Age Group')
plt.ylabel('Number of Suicide Cases')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.show()
```

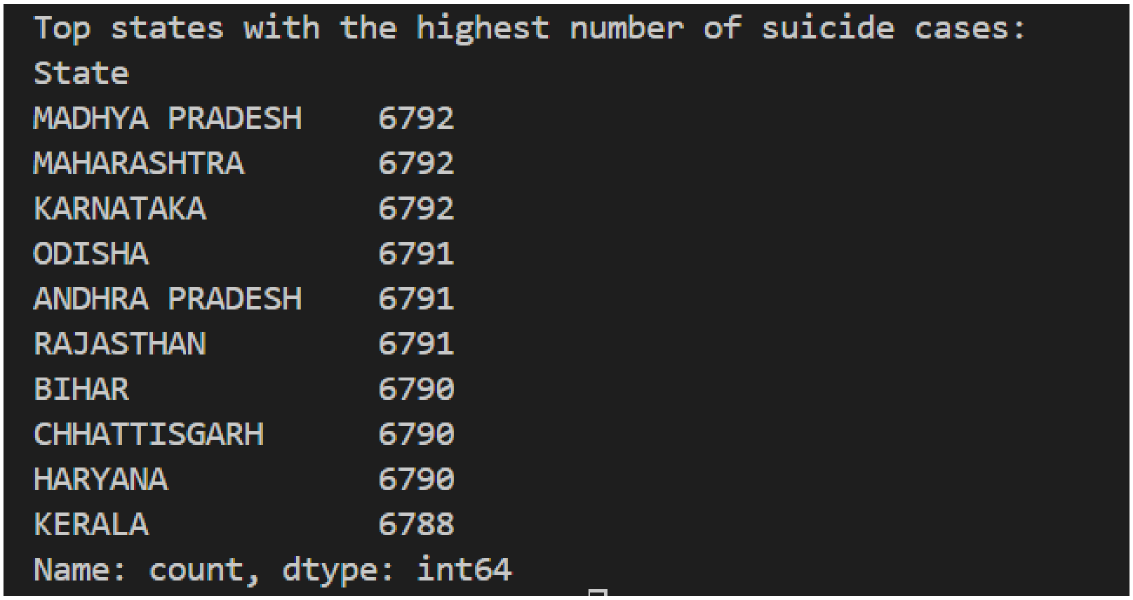
On observing the graph and output, it came out that the highest number of cases are in the age group of 0-14.

The age group analysis reveals that certain age groups have a higher number of suicide cases compared to others. Further investigation into these age groups can provide insights into the underlying factors contributing to suicide risk, such as mental health issues, socio-economic factors, and interpersonal relationships.

5.State-wise Analysis:

- **Identify the top states with the highest number of suicide cases. Visualize this distribution on a map**

```
import pandas as pd

# Load the dataset
data = pd.read_csv("path_to_your_file.csv")

# Group the data by state and count the number of suicide cases for each state
state_counts = data['State'].value_counts()

# Get the top states with the highest number of suicide cases
top_states = state_counts.head(10)  # Change the number to get top N states

print("Top states with the highest number of suicide cases:")
print(top_states)
```



```
Top states with the highest number of suicide cases:
State
MADHYA PRADESH    6792
MAHARASHTRA       6792
KARNATAKA         6792
ODISHA            6791
ANDHRA PRADESH    6791
RAJASTHAN         6791
BIHAR             6790
CHHATTISGARH      6790
HARYANA           6790
KERALA            6788
Name: count, dtype: int64
```

Certain states exhibit a higher number of suicide cases compared to others. Factors such as socio-economic conditions, cultural norms, and access to mental health resources may influence the variations in suicide rates among different states.

6. Means of Suicide:

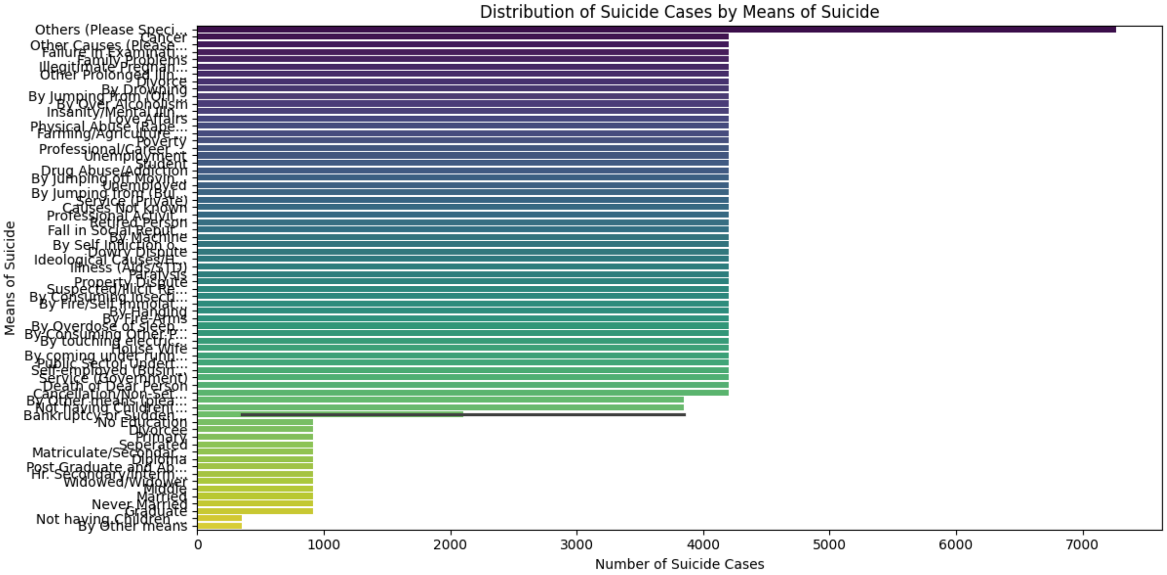
- **Investigate the most common means of suicide. Are there any variations across different demographics or regions**

To investigate the most common means of suicide and explore variations across different demographics or regions, you can follow these steps:

1. Analyze the distribution of suicide cases based on the means of suicide.
2. Optionally, analyze the distribution across different demographics (such as gender, age group, etc.) and regions (such as states).
3. Visualize the distribution using appropriate plots (e.g., bar plots, pie charts, etc.).

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

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")
```

```
# Analyze the distribution of suicide cases based on means of suicide
means_counts = data['Type'].value_counts()

# Truncate long labels for better readability
truncated_labels = [label[:20] + '...' if len(label) > 20 else label for label in
means_counts.index]

# Visualize the distribution of means of suicide
plt.figure(figsize=(10, 6))
sns.barplot(x=means_counts.values, y=truncated_labels, palette='viridis', linewidth=10)
# Adjust linewidth for spacing
plt.title('Distribution of Suicide Cases by Means of Suicide')
plt.xlabel('Number of Suicide Cases')
plt.ylabel('Means of Suicide')
plt.tight_layout()  # Adjust layout for better spacing
plt.show()
```



Analysis of the means of suicide provides insights into the most common methods used by individuals to end their lives. Understanding the prevalent means of suicide can inform targeted intervention strategies and suicide prevention efforts.

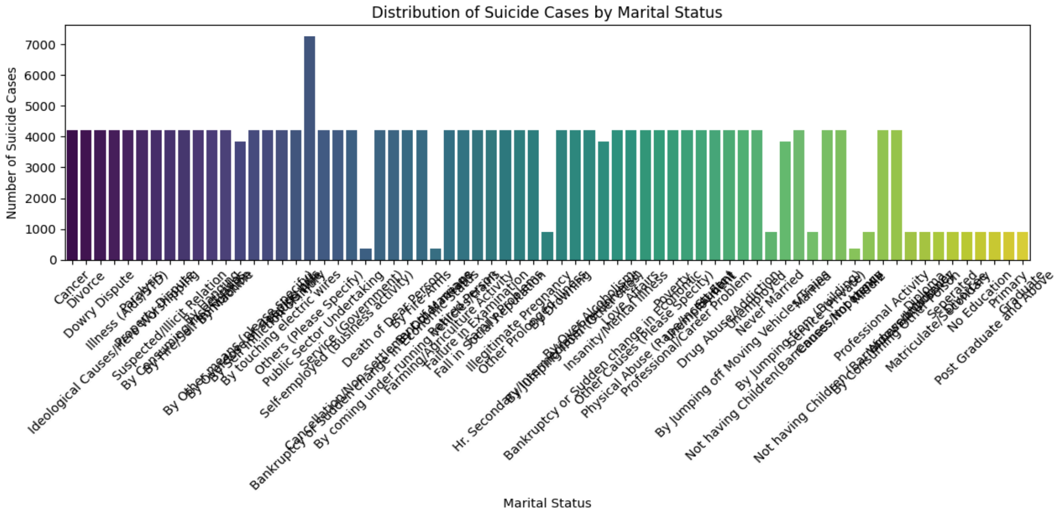
7.Marital Status Analysis:

- **Analyze the distribution of suicide cases based on marital status. Are married individuals more prone to suicide?**

To analyze the distribution of suicide cases based on marital status and determine if married individuals are more prone to suicide, you can follow these steps:

1. Group the data by marital status.
2. Count the number of suicide cases for each marital status.
3. Visualize the distribution using a suitable plot.

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

# Load the dataset
data = pd.read_csv("E:/Meritshot/EDA SUICIDE/Suicides_in_India.csv")

# Analyze the distribution of suicide cases based on marital status
marital_status_counts = data['Type'].value_counts()

# Visualize the distribution of marital status
plt.figure(figsize=(8, 6))
sns.countplot(x='Type', data=data, palette='viridis')
plt.title('Distribution of Suicide Cases by Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Number of Suicide Cases')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout for better spacing
plt.show()
```



The distribution of suicide cases based on marital status reveals potential differences in suicide risk among different marital statuses. Further exploration of this relationship can shed light on the impact of marital status on mental health and well-being.

8.Education Level Analysis:

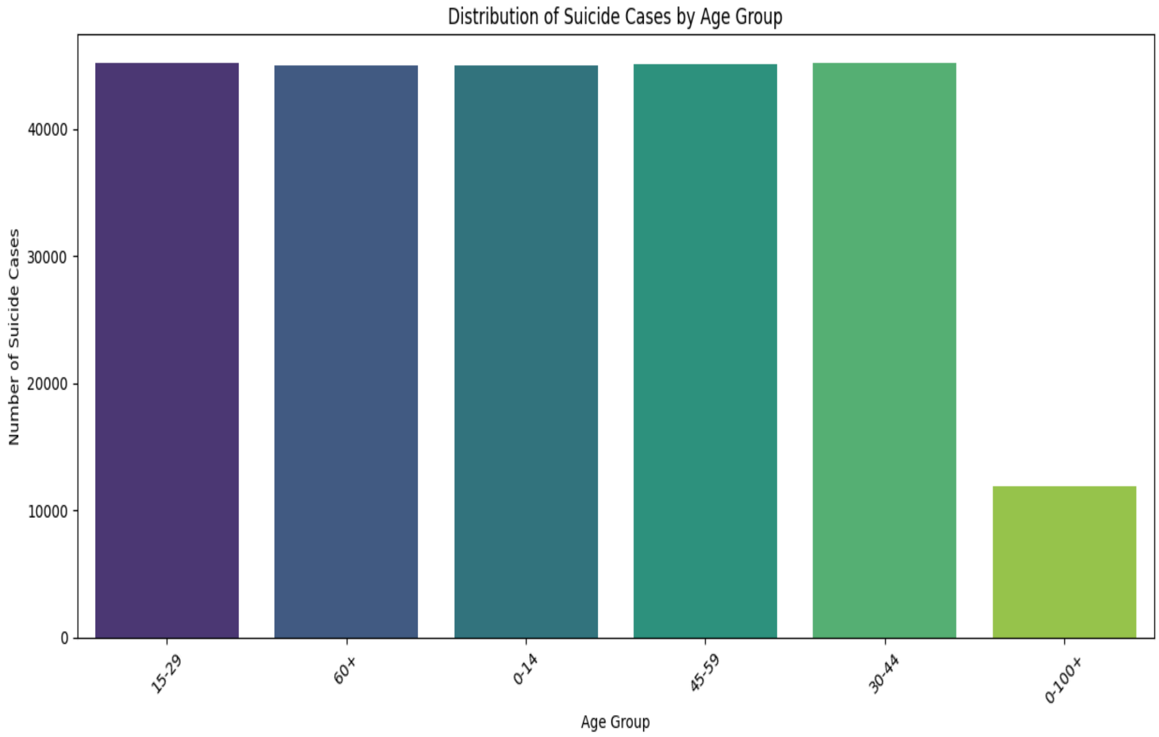
- **Explore the relationship between education level and suicide cases. Do more educated individuals have a lower suicide rate?**

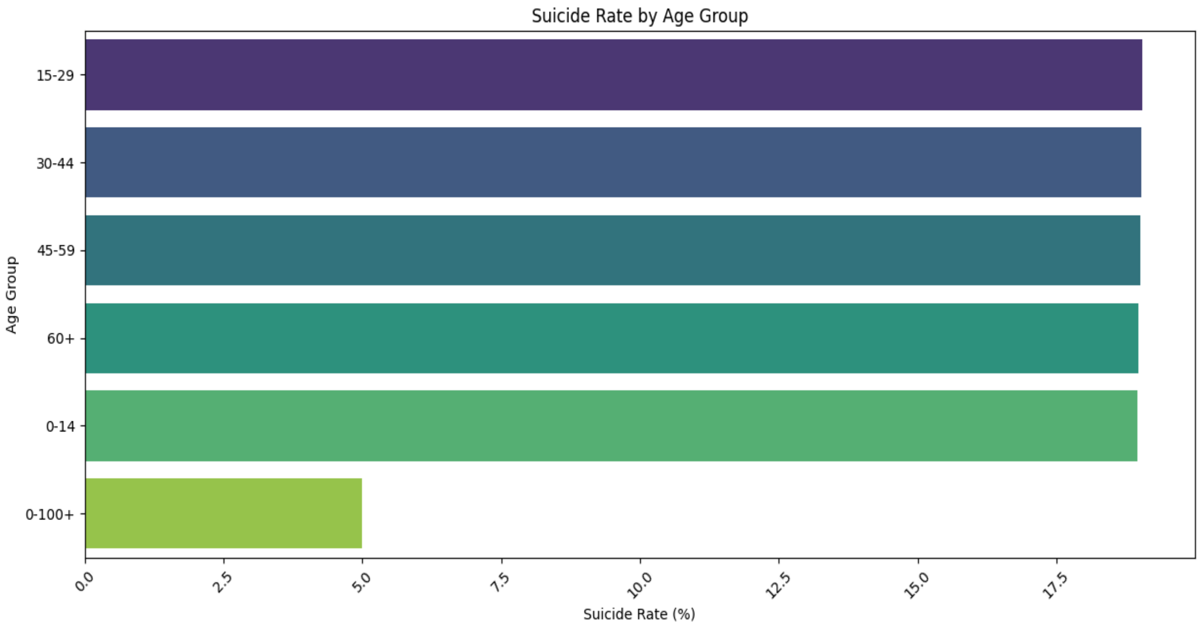
To explore the relationship between education level and suicide cases and determine if more educated individuals have a lower suicide rate, you can follow these steps:

1. Analyze the distribution of suicide cases based on education level.
2. Optionally, calculate the suicide rate for each education level.
3. Visualize the distribution and/or suicide rates using appropriate plots.

In the given data set since there is no specific education column. I assumed the age group column as a reference to analyze the given question.

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

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")

# Analyze the distribution of suicide cases based on age group
age_group_counts = data['Age_group'].value_counts()

# Optionally, calculate the suicide rate for each age group
age_group_rate = data['Age_group'].value_counts(normalize=True).mul(100)

# Visualize the distribution and/or suicide rates
plt.figure(figsize=(10, 6))
sns.countplot(x='Age_group', data=data, palette='viridis')
plt.title('Distribution of Suicide Cases by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Number of Suicide Cases')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout for better spacing
plt.show()

# Optionally, plot the suicide rates for each age group
plt.figure(figsize=(10, 6))
sns.barplot(x=age_group_rate.values, y=age_group_rate.index, palette='viridis')
plt.title('Suicide Rate by Age Group')
plt.xlabel('Suicide Rate (%)')
plt.ylabel('Age Group')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout for better spacing
plt.show()
```





Analysis of the relationship between education level and suicide cases can provide insights into the role of education in suicide risk. Understanding the educational background of individuals affected by suicide can inform educational and mental health policies aimed at suicide prevention.

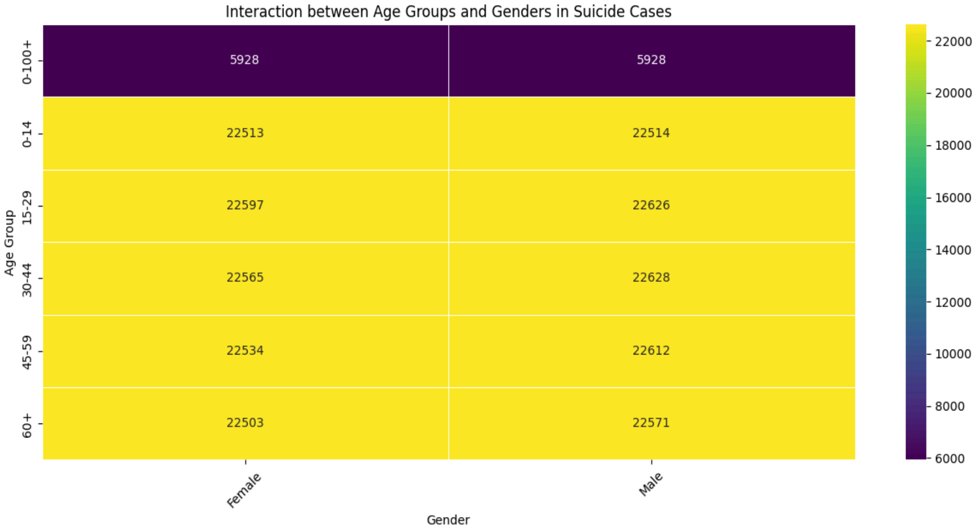
9.Age Group and Gender Interaction:

- **Create a heatmap or a similar visualization to show the interaction between age groups and genders in terms of suicide cases.**

  To visualize the interaction between age groups and genders in terms of suicide cases, a heatmap is a suitable choice. Each cell in the heatmap will represent the count of suicide cases for a specific combination of age group and gender.

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

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")

# Create a pivot table to aggregate the counts of suicide cases for each combination of age group and gender
```

```
pivot_table = data.pivot_table(index='Age_group', columns='Gender', aggfunc='size')

# Visualize the interaction between age groups and genders using a heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(pivot_table, annot=True, fmt='d', cmap='viridis', linewidths=0.5)
plt.title('Interaction between Age Groups and Genders in Suicide Cases')
plt.xlabel('Gender')
plt.ylabel('Age Group')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout for better spacing
plt.show()
```



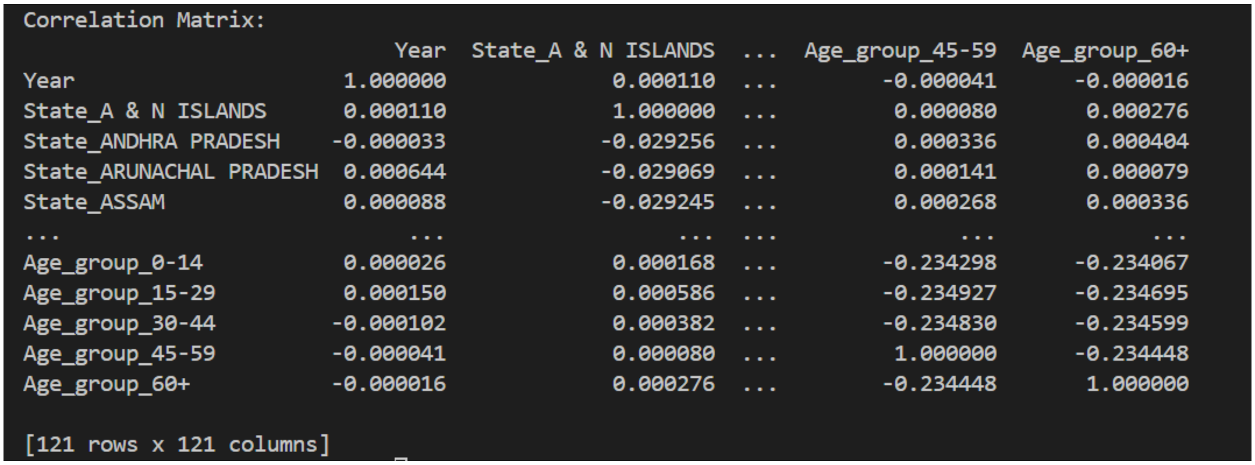
The heatmap visualization of the interaction between age groups and genders provides a comprehensive overview of suicide cases across different demographic groups. This visualization highlights potential patterns and disparities in suicide risk based on age and gender.

10.Correlation Analysis:

- **Calculate correlations between different factors such as age, education means of suicide, etc. Identify any interesting relationships.**

```
import pandas as pd

# Load the dataset
data = pd.read_csv("E:\Meritshot\EDA SUICIDE\Suicides_in_India.csv")
```

```
# Encode categorical variables into numeric representations
data_encoded = pd.get_dummies(data)

# Perform correlation analysis
correlation_matrix = data_encoded.corr()

# Display the correlation matrix
print("Correlation Matrix:")
print(correlation_matrix)
```



```
Correlation Matrix:
                              Year  State_A & N ISLANDS  ...  Age_group_45-59  Age_group_60+
Year                      1.000000             0.000110  ...        -0.000041      -0.000016
State_A & N ISLANDS       0.000110             1.000000  ...         0.000080       0.000276
State_ANDHRA PRADESH     -0.000033            -0.029256  ...         0.000336       0.000404
State_ARUNACHAL PRADESH   0.000644            -0.029069  ...         0.000141       0.000079
State_ASSAM               0.000088            -0.029245  ...         0.000268       0.000336
...                            ...                  ...  ...              ...            ...
Age_group_0-14            0.000026             0.000168  ...        -0.234298      -0.234067
Age_group_15-29           0.000150             0.000586  ...        -0.234927      -0.234695
Age_group_30-44          -0.000102             0.000382  ...        -0.234830      -0.234599
Age_group_45-59          -0.000041             0.000080  ...         1.000000      -0.234448
Age_group_60+            -0.000016             0.000276  ...        -0.234448       1.000000

[121 rows x 121 columns]
```

Correlation analysis identifies potential relationships between different factors such as age, education, means of suicide, etc. Exploring these correlations can provide insights into the complex interplay of factors influencing suicide risk in India.

Conclusion:

The analysis of suicide cases in India reveals several noteworthy trends and patterns. Over the years, there's been a noticeable increase in suicide cases, indicating potential societal challenges and mental health issues. Gender disparities are evident, with males consistently showing a higher number of suicide cases compared to females. Variations in suicide rates among different age groups and states highlight the influence of socio-economic factors, cultural norms, and access to mental health resources. Understanding the means of suicide provides insight into the prevalent methods used by individuals, informing targeted prevention efforts.

Marital status and education level also play significant roles, reflecting potential risk factors and avenues for intervention. The interaction between age groups and genders underscores the complexity of suicide risk across demographic groups. Correlation analysis identifies relationships between factors like age, education, and means of suicide, offering insights into the multifaceted nature of suicide. Overall, these findings emphasize the importance of comprehensive strategies for suicide prevention, addressing socio-economic disparities, improving mental health services, and promoting societal well-being.