Mechanical and Industrial Engineering Department IE 6600: Computation and Visualization.
Spring 2025

**Death rates for suicide, by sex, race, Hispanic origin, and age: United States**

Group 2

Members: Praneeta Mahentesh, Bhavya Kapadia, Soumya Gosala

# Introduction:

Suicide rates are a growing concern worldwide, posing significant challenges for public health, mental well-being, and social stability. Understanding the trends and underlying factors behind suicide is crucial for developing effective prevention strategies and mental health interventions. This report analyzes historical suicide rates across different age groups, genders, and racial/ethnic backgrounds to uncover key patterns and correlations. By leveraging structured data analysis techniques, this study aims to provide data-driven insights that can inform policy decisions, resource allocation, and targeted support programs to reduce suicide risks and promote mental health awareness.

# Data Processing and Analysis

## 1. Data Acquisition:

The "Death_rates_for_suicide_by_sex_race_Hispanic_origin_and_age_United_States.csv"([Link to dataset](#)) dataset, contains 6,390 records spanning from 1950 to 2020, providing insights into suicide rates across different demographic groups in the United States. It includes key attributes that allow for an in-depth examination of suicide trends by sex, race, age group, and time period.

This dataset contains:

- INDICATOR: Specifies the type of data recorded (e.g., "Death rates for suicide").
- UNIT: The unit of measurement, given as deaths per 100,000 residents.
- YEAR: The time period for each recorded suicide rate.
- AGE: Age group categorization (e.g., "All ages" or specific age ranges).
- SEX & RACE: Captures demographic differences in suicide rates.
- ESTIMATE: The recorded suicide rate per 100,000 residents.
- FLAG: Marks specific data points that may require additional interpretation.

Purpose of the Dataset
The dataset serves as a comprehensive resource for analyzing historical and demographic patterns in suicide rates. By evaluating the trends in suicide rates across different age groups, genders, and racial backgrounds, this analysis can:
- Identify demographic groups at higher risk.
- Reveal long-term trends and patterns in suicide rates.
- Support data-driven policymaking and targeted mental health interventions.

This dataset forms the foundation for exploratory analysis and statistical modeling, helping to uncover meaningful insights that can guide public health initiatives and suicide prevention strategies.

## 2. Data Inspection and Cleaning:

2.1 Initial Data Inspection

Before performing any analysis, it is essential to inspect the dataset for redundant columns, missing values, and inconsistencies. The dataset originally contained 13 columns, including categorical and numerical identifiers. A preliminary check using df.info() and df.head() provided insights into the dataset structure, confirming the presence of unnecessary numerical representations of categorical variables.

2.2 Cleaning Process

1. Dropping Redundant Columns

To streamline the dataset and avoid duplicate information, we removed redundant numerical columns that replicate categorical data. The following columns were dropped:
INDICATOR – Descriptive label with no unique analytical value.
UNIT & UNIT_NUM – Represents measurement units, which are constant across the dataset.
STUB_NAME_NUM & STUB_LABEL_NUM – Numeric versions of categorical variables.
YEAR_NUM & AGE_NUM – Numeric versions of YEAR and AGE.
The cleaned dataset retained only relevant fields, making it more efficient for analysis.

2. Renaming Columns for Clarity

To improve readability and ensure consistency, key columns were renamed:
STUB_NAME → Category
STUB_LABEL → Subcategory
ESTIMATE → Suicide_Rate
This renaming enhances data clarity, making it easier to interpret during analysis.

```python
# Dropping redundant numerical columns that duplicate categorical information
columns_to_drop = ['INDICATOR', 'UNIT', 'UNIT_NUM', 'STUB_NAME_NUM', 'STUB_LABEL_NUM', 'YEAR_NUM', 'AGE_NUM']
df_cleaned = df.drop(columns=columns_to_drop)

# Renaming columns for clarity
df_cleaned.rename(columns={
    'STUB_NAME': 'Category',
    'STUB_LABEL': 'Subcategory',
    'ESTIMATE': 'Suicide_Rate'
}, inplace=True)

# Display updated structure
df_cleaned.info()
df_cleaned.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6390 entries, 0 to 6389
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Category      6390 non-null   object
 1   Subcategory   6390 non-null   object
 2   YEAR          6390 non-null   int64
 3   AGE           6390 non-null   object
 4   Suicide_Rate  5484 non-null   float64
 5   FLAG          906 non-null    object
dtypes: float64(1), int64(1), object(4)
memory usage: 299.7+ KB
```

| | Category | Subcategory | YEAR | AGE | Suicide_Rate | FLAG |
|---|---|---|---|---|---|---|
| 0 | Total | All persons | 1950 | All ages | 13.2 | NaN |
| 1 | Total | All persons | 1960 | All ages | 12.5 | NaN |
| 2 | Total | All persons | 1970 | All ages | 13.1 | NaN |
| 3 | Total | All persons | 1980 | All ages | 12.2 | NaN |
| 4 | Total | All persons | 1981 | All ages | 12.3 | NaN |

2.3 Handling Missing Values
A crucial step in data preparation is addressing missing values, particularly in the Suicide_Rate column:
The df.isnull().sum() function revealed missing values in Suicide_Rate and FLAG columns.
Since FLAG represents data annotations rather than numerical values, it was not imputed.

Two approaches were considered for handling missing values in Suicide_Rate:
- Removing rows with missing values (df_cleaned.dropna(subset=['Suicide_Rate'], inplace=True)) – This ensures only complete records are analyzed.

- Imputing missing values with the median (df_cleaned['Suicide_Rate'].fillna(df_cleaned['Suicide_Rate'].median(), inplace=True)) – This helps retain data while minimizing bias.

For this analysis, removal of missing values was chosen to maintain data integrity.

```python
# Check missing values
df_cleaned.isnull().sum()

# Handling missing values in Suicide_Rate (either remove or fill)
df_cleaned.dropna(subset=['Suicide_Rate'], inplace=True)  # Option 1: Remove rows with missing rates
# df_cleaned['Suicide_Rate'].fillna(df_cleaned['Suicide_Rate'].median(), inplace=True)  # Option 2: Fill with median

# Verify missing values are handled
df_cleaned.isnull().sum()
```

|  | 0 |
| --- | --- |
| Category | 0 |
| Subcategory | 0 |
| YEAR | 0 |
| AGE | 0 |
| Suicide_Rate | 0 |
| FLAG | 5484 |

**dtype:** int64

1. Final Data Verification

A final check confirmed that all missing values in Suicide_Rate were handled, and the dataset was now structured for further exploration analysis.

# Exploratory Data Analysis (EDA)

### 1. Suicide Rate Trend Over Time

A line plot of the average suicide rate over time (1950-2019) provides insight into the historical evolution of suicide rates across different decades.
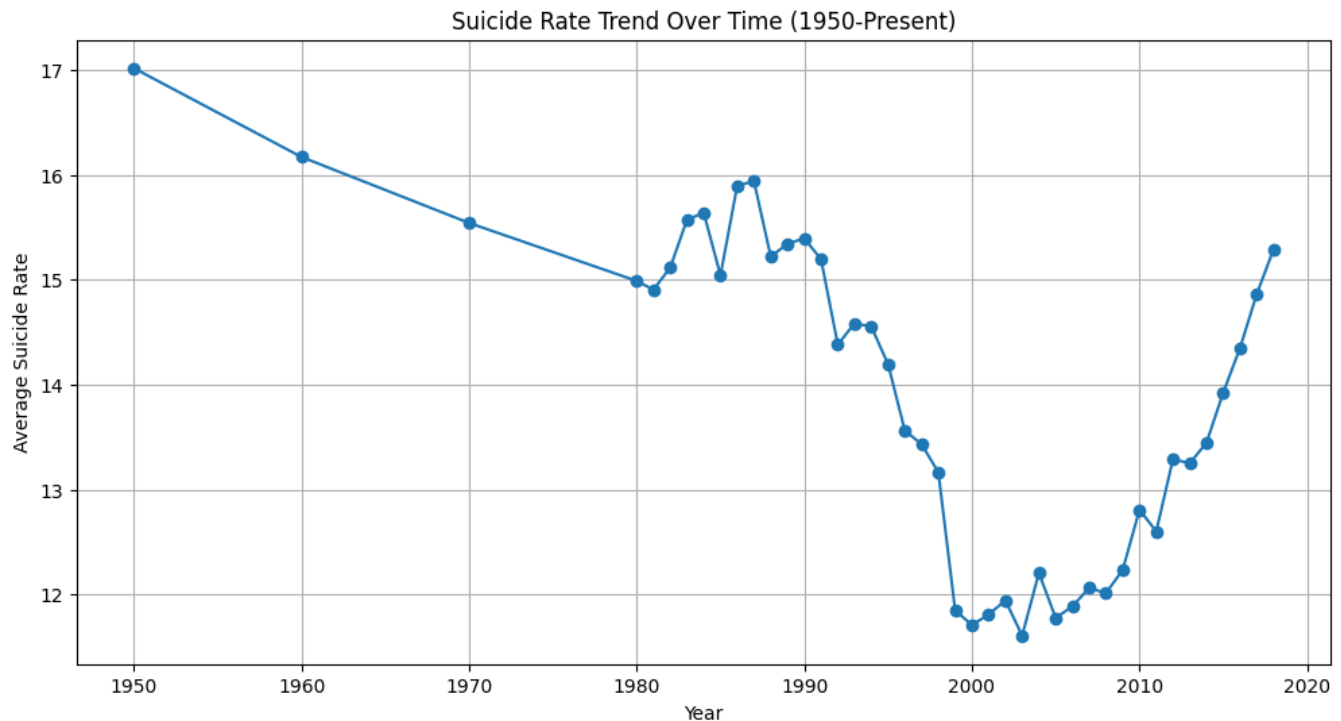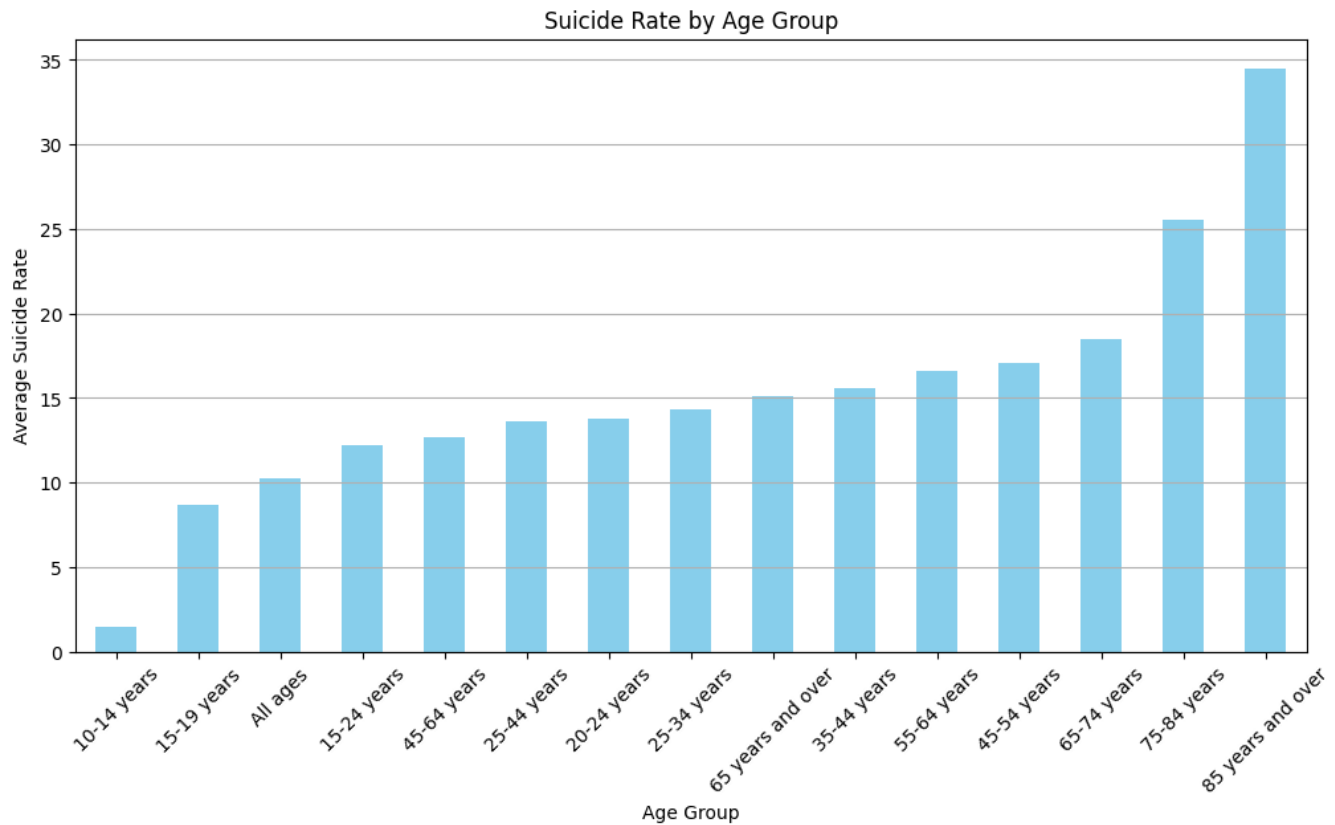


**Key Observations**

- 1950s-1970s: The suicide rate gradually declined, moving from 17 per 100,000 in 1950 to around 15 per 100,000 by 1975.
- 1980s-1990s: A sudden rise in suicide rates was observed, peaking at nearly 16.5 per 100,000.
- 2000s: A sharp decline occurred, reaching its lowest recorded point (below 12 per 100,000) around the early 2000s.
- 2010s-2019: The suicide rate has shown a continuous rise, reaching close to 15.5 per 100,000 by 2019.

This analysis suggests that suicide rates have experienced significant fluctuations due to various socioeconomic, policy, and mental health-related factors over time.

## 2. Suicide Rate by Age Group

Analyzing suicide rates across different age groups helps identify the most vulnerable demographics. The bar chart below visualizes the average suicide rate per 100,000 residents for each age group.


Suicide Rate by Age Group

**Key Observations**
- Youth (10-14 years): Suicide rates are lowest in this age group.
- Adolescents and Young Adults (15-24 years): A gradual increase is observed, indicating rising risks during early adulthood.
- Middle-Aged Adults (25-64 years): Suicide rates remain relatively stable, fluctuating between 12-16 per 100,000 residents.
- Older Adults (65+ years): The rates increase significantly, peaking in the 85+ age group at over 35 per 100,000 residents.

Higher suicide rates among elderly populations may be attributed to factors such as social isolation, health deterioration, and financial stress, making them particularly vulnerable. The gradual increase in suicide rates from adolescence to middle adulthood underscores the need for targeted mental health interventions that support young and middle-aged adults during critical life transitions. Additionally, the sharp rise in suicide rates among individuals aged 75 and older highlights the urgent need for improved elderly care programs, accessible mental health services, and awareness initiatives to address the unique challenges faced by aging populations.

## 3. Suicide Rate by Gender

The pie chart above represents the percentage of suicides by gender over all recorded years in the dataset. The visualization highlights a stark contrast in suicide rates between males and females.

Percentage of Suicides by Gender (Overall Years)
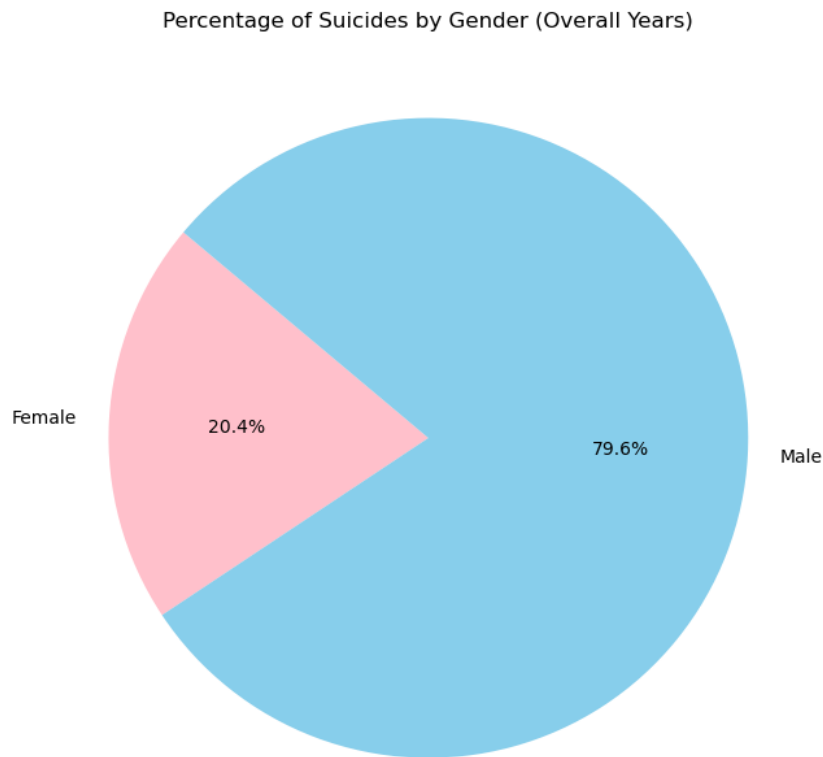


**Key Observations**
- Males account for nearly 80% of suicides, while females make up approximately 20%.
- This trend is consistent across multiple time periods and aligns with global patterns in suicide rates.
- The significantly higher suicide rate among males suggests underlying societal, psychological, and behavioral factors contributing to this disparity.

Research indicates that males are less likely to seek mental health support compared to females, largely due to societal expectations around masculinity and emotional expression, which may contribute to their higher suicide rates. Additionally, differences in suicide methods play a role, as males tend to use more lethal means, resulting in higher fatality rates, while women, despite experiencing higher rates of depression, have lower suicide completion rates. These disparities highlight the need for gender-specific prevention strategies, including male-focused mental health interventions, awareness campaigns, and efforts to reduce stigma around help-seeking behaviors.

## 4. Suicide Rate by Race/Ethnicity

The bar chart below presents the average suicide rates across different racial and ethnic groups, categorized by gender. This analysis helps identify disparities among racial demographics, which can inform targeted mental health interventions and policy decisions.



Average Suicide Rate by Race/Ethnicity
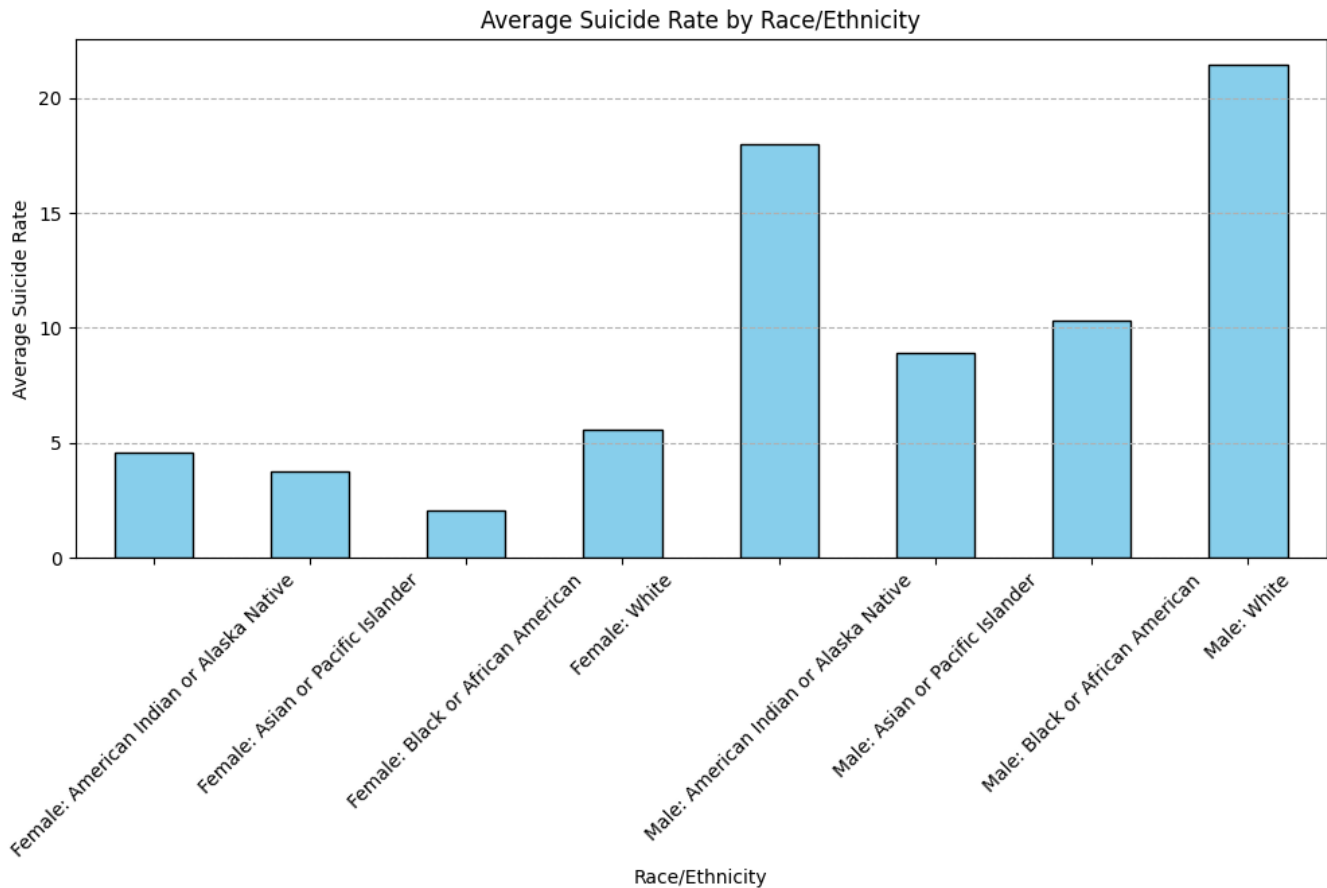
**Key Observations**
- White males exhibit the highest suicide rate, followed by American Indian or Alaska Native males.
- Among females, suicide rates are relatively lower across all racial groups, with White and American Indian/Alaska Native women experiencing the highest rates.
- Black or African American, Asian, and Pacific Islander populations have comparatively lower suicide rates.
- Across all racial groups, males have significantly higher suicide rates than females.

## 5. Correlation Between Year and Suicide Rate by Age Group

The scatter plot above visualizes the relationship between year and suicide rate, segmented by age group. Each colored point represents a specific age category, allowing for a detailed examination of age-based trends over time.
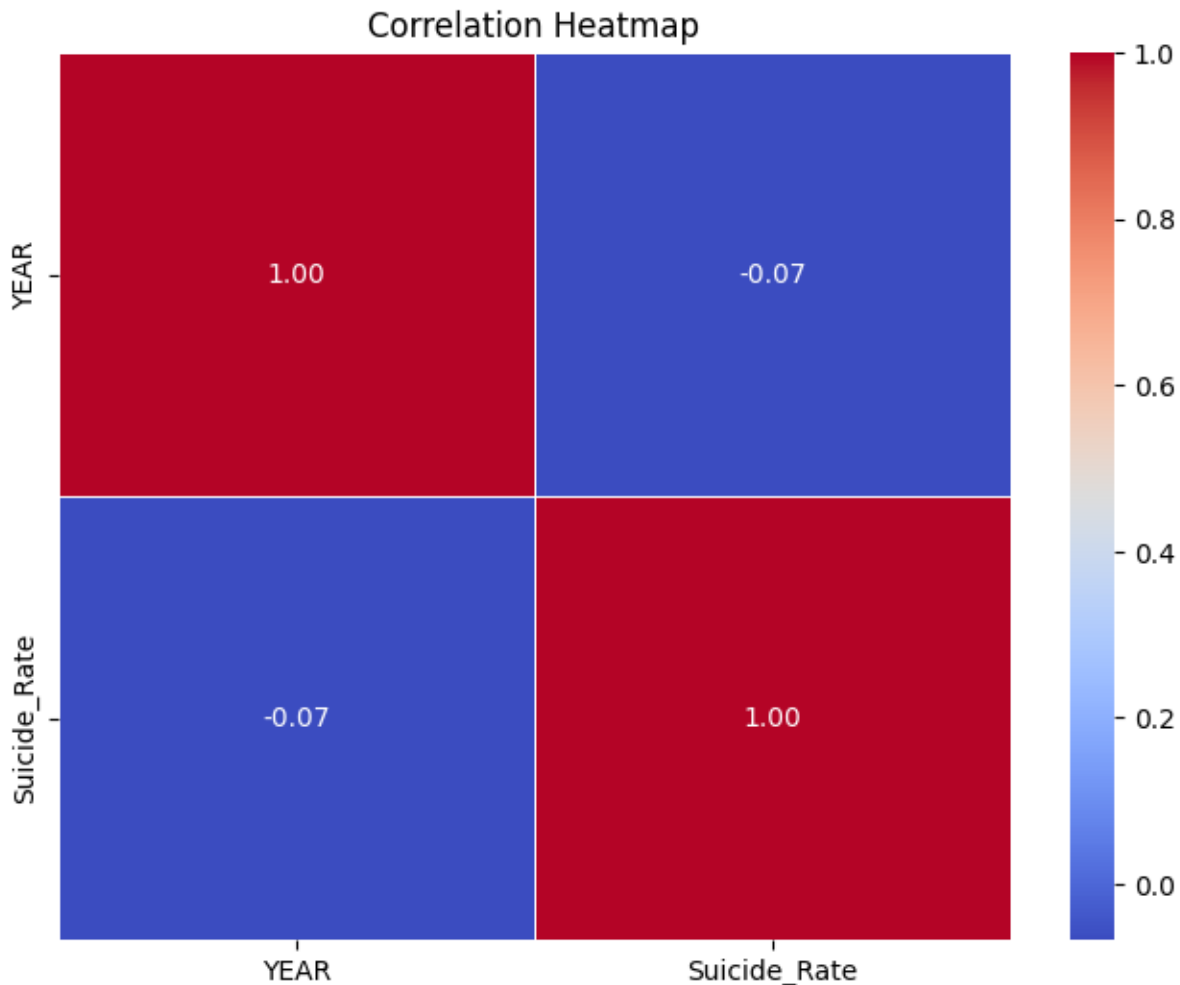


Correlation between Year and Suicide Rate by Age Group
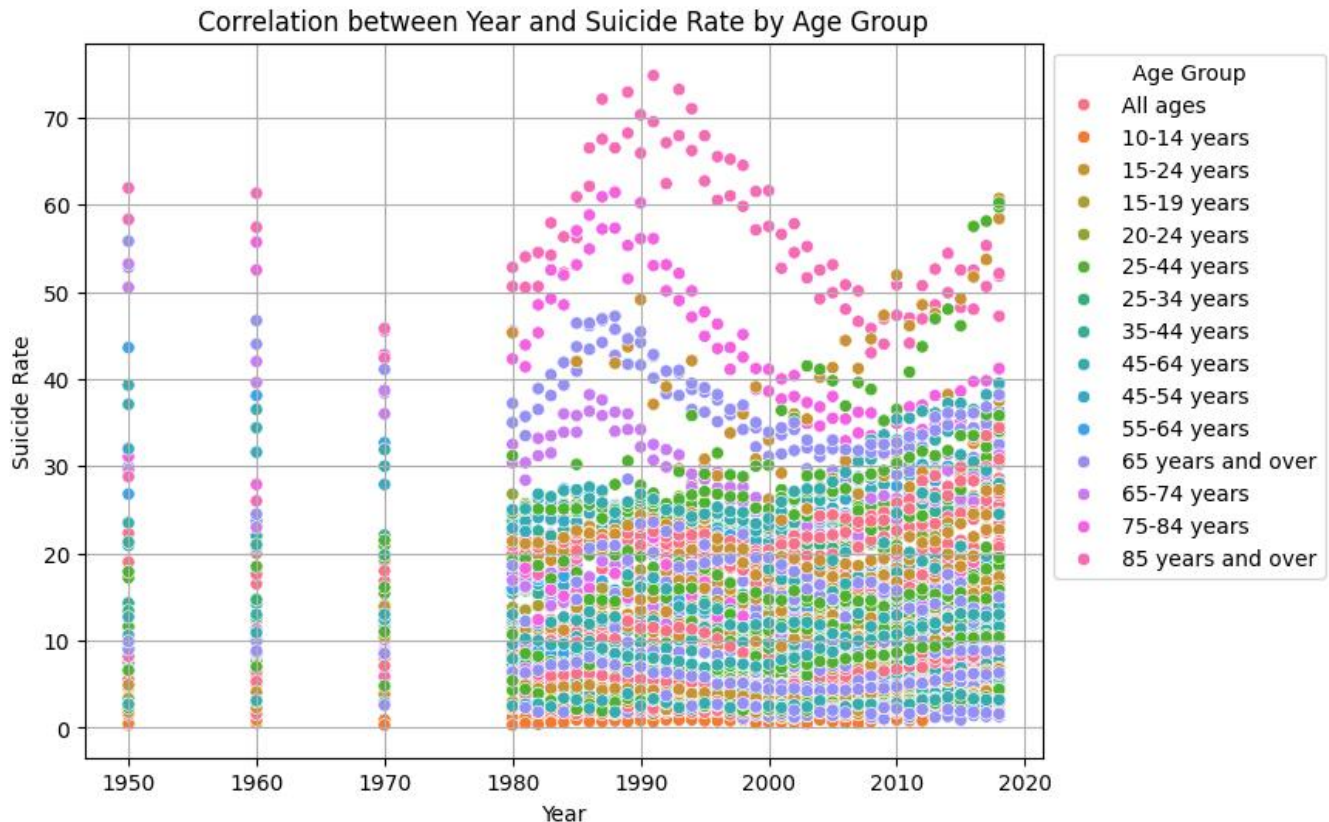
**Key Observations**
- Elderly populations (85+ years) consistently show higher suicide rates compared to younger age groups.
- Fluctuations are more prominent in the mid-to-late 20th century, particularly in the 1970s to 1990s, indicating possible socioeconomic or policy-driven factors influencing suicide rates.
- A noticeable decline around the early 2000s is observed across all age groups.
- Since 2010, suicide rates have been rising again, particularly in the older adult categories.

The rise in suicide rates during the 1980s and 1990s may be linked to economic recessions, social shifts, and limited mental health awareness, while the decline post-2000 could be due to advancements in mental health care and crisis interventions. However, the increase post-2010 suggests that modern stressors like financial instability and social isolation require further attention. Elderly individuals (75+ years) consistently show the highest suicide rates, highlighting the need for

specialized mental health support, while younger age groups (10-24 years) are experiencing a concerning rise post-2000. These trends call for targeted interventions—elderly-focused mental health programs and youth-centered initiatives. Additionally, further research on socio-political influences can enhance long-term suicide prevention policies. This analysis underscores the importance of historical and demographic context in mental health research.

## 6. Correlation Analysis: Heatmap Interpretation



Correlation Heatmap

The heatmap above represents the correlation between numerical variables in the dataset, particularly YEAR and Suicide_Rate. Correlation values range from -1 to 1, where:

- 1.0 represents a strong positive correlation (variables increase together).
- -1.0 represents a strong negative correlation (one variable increases as the other decreases).
- 0.0 indicates no correlation between variables.

**Key Observations**

- The correlation between YEAR and Suicide Rate is -0.07, which is very weakly negative.
- This suggests no significant relationship between time and suicide rate across all age groups in a

simple linear sense.
- The lack of strong correlation indicates that suicide rates do not follow a simple increasing or decreasing trend over time and are likely influenced by multiple external factors.

# Conclusion

This report examined suicide rates in the United States, analyzing their variations across demographics, time periods, and socio-economic factors. Through data acquisition, cleaning, exploratory analysis, and statistical modeling, several critical insights emerged.

Final Analysis-
- Temporal Trends: Suicide rates have fluctuated over the decades, showing a decline from the 1950s to early 2000s, followed by a sharp increase post-2010, indicating the influence of external socio-political and economic conditions rather than a consistent trend.
- Age-Based Differences: Suicide rates increase with age, peaking among individuals 85 and older, while younger populations maintain lower rates but show a rising trend, particularly among adolescents and young adults.
- Gender Disparities: Males account for nearly 80% of suicides, significantly outpacing females. This highlights the need for targeted mental health interventions, focusing on help-seeking behaviors and social support for men.
- Racial and Ethnic Patterns: White males and American Indian/Alaska Native populations experience the highest suicide rates, while African American and Asian/Pacific Islander groups show lower rates, possibly due to cultural resilience or underreporting. These disparities emphasize the importance of culturally sensitive mental health programs.
- Correlation Insights: A weak correlation between time and suicide rates suggests that suicide trends are not simply increasing or decreasing in a linear manner. Instead, they are shaped by multiple factors, including age, gender, race, economic conditions, healthcare policies, and societal attitudes toward mental health.

Suicide is a complex, multi-dimensional issue that requires a comprehensive, data-driven approach to prevention. While certain age groups, genders, and racial communities are more vulnerable, suicide prevention must be holistic and inclusive. By leveraging historical data, demographic insights, and correlation analysis, this study provides a foundation for targeted mental health strategies and policy reforms.

Ultimately, collaborative efforts among healthcare providers, policymakers, and community organizations are crucial to reducing suicide rates and fostering a healthier, more supportive society.