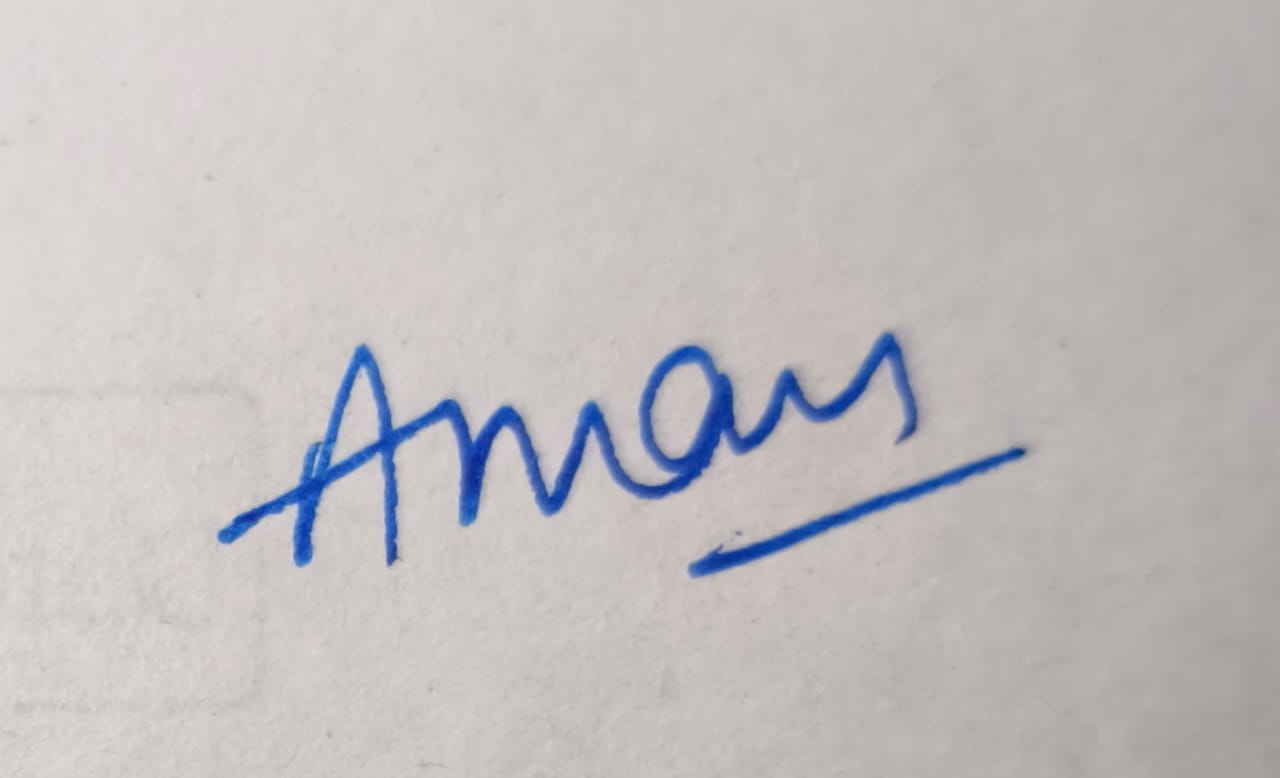
**DATA SCIENCE  
PROJECT REPORT**(Project Semester January-April 2025) **Exploratory Data Analysis using Python**Submitted by  
Aman  
Registration No: 12309783  
Program: Bachelor of Technology Section: K23FD  
Course Code: INT375  
  
Under the Guidance of  
Mrs. Baljinder Kaur (U. Id: 27952) Assistant Professor  
Discipline of CSE/IT  
Lovely School of Computer Science and Engineering  
Lovely Professional University, Phagwara

# CERTIFICATE :-

This is to certify that Aman bearing Registration no. 12309783 has completed INT375 project titled, “Exploratory Data Analysis of Death Suicide using Python” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.  
  
Signature and Name of the Supervisor  
Designation of the Supervisor  
School of Computer Science and Engineering  
Lovely Professional University  
Phagwara, Punjab.  
Date: 13 April, 2025

# DECLARATION :-

I, Aman, student of Bachelor of Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.  
  
Date: 13 April, 2025  
  
Signature:   
  
Registration No.: 12309783  
  
Name: Aman

# ACKNOWLEDGEMENT :-

I would like to express my deepest gratitude to Mrs. Baljinder Kaur, my esteemed project guide, for her constant support, expert guidance, and valuable feedback throughout this project. Her motivation and encouragement enabled me to complete this project successfully and learn extensively through every step.

I am also profoundly thankful to Lovely Professional University for creating an academic environment that fosters learning and innovation. The facilities, libraries, and online resources offered by the university have been indispensable for my research and project development.

I would also like to thank all the professors and faculty members of the Discipline of CSE/IT, whose teachings laid the foundation for my technical knowledge and understanding.

I am sincerely grateful to my classmates and friends for their constructive discussions, motivation, and moral support, which kept me going even during challenging phases of this project.

Lastly, I would like to extend my heartfelt thanks to my family, whose patience, love, and support were the backbone of my success. Without their constant encouragement, completing this project would not have been possible.

This project report is a reflection of collective inspiration, support, and learning from various individuals and institutions, and I am genuinely appreciative of all the contributions that made this journey fruitful and fulfilling.

# TABLE OF CONTENTS :-

1. Introduction

2. Source of Dataset

3. EDA Process

4. Analysis on Dataset

5. Conclusion

6. Future Scope

7. References

**1. INTRODUCTION :-**

In today's fast-paced digital world, the volume of data generated every second is enormous and growing rapidly. This abundance of information holds immense potential to drive innovation, inform decision-making, and improve societal outcomes. However, to harness the power of data effectively, it is essential to first understand its structure, detect underlying patterns, and identify meaningful insights. This crucial first step in the data analysis journey is known as Exploratory Data Analysis (EDA).

EDA involves summarizing the main characteristics of a dataset, often using visual methods and statistical techniques. It allows data scientists to detect anomalies, test hypotheses, and uncover the hidden structure within the data. By providing a clear and insightful overview of the data, EDA sets the stage for more sophisticated analysis, modeling, and predictive tasks.

In this project, we focused on performing an EDA on a sensitive and socially important dataset related to deaths by suicide in the United States between 1999 and 2016. The dataset contains detailed information categorized by gender, race, age group, state, and year, providing a multi-dimensional view of suicide statistics across the nation. The primary aim was to extract meaningful insights that could contribute to a better understanding of the trends and factors associated with suicides, ultimately helping to inform prevention strategies and public health initiatives.

Through systematic data cleaning, statistical summarization, and powerful visualizations, we delve deep into the patterns hidden within the suicide data. Our analysis not only highlights high-risk groups and regions but also emphasizes the urgent need for proactive mental health interventions. The findings are intended to serve as a valuable resource for policymakers, healthcare professionals, and researchers working towards reducing suicide rates and improving mental well-being in society.

# 2. SOURCE OF DATASET :-

The dataset used for this Exploratory Data Analysis project was sourced from a highly trustworthy and reputable organization: the Centers for Disease Control and Prevention (CDC) in the United States. The CDC is a leading national public health institute that collects a vast range of health-related data, including detailed records of causes of death across the country. Because the data is compiled and verified by healthcare professionals and official government channels, it is considered extremely reliable, accurate, and comprehensive.

The specific dataset we used relates to suicide deaths in the United States between the years 1999 and 2016. It includes important information such as:

* The **year** of death,
* The **state** in which the death occurred,
* The **gender** of the person,
* The **race** of the person,
* The **age group** of the person,
* And the **crude death rate**, which represents the number of deaths per 100,000 people in each group.

This dataset was made publicly available to researchers and analysts through **Kaggle**, a widely used online platform for data science competitions and data sharing. Kaggle allows users to access datasets for free, ensuring that they are accessible to students, researchers, and data scientists globally.

We selected this dataset because:

* It comes from an **official, government-backed source**, ensuring the quality and authenticity of the data.
* It covers a **broad timeline of 18 years**, making it possible to observe long-term trends and patterns.
* It includes **rich demographic details**, allowing deeper analysis across different sections of the population.

Before beginning the analysis, the dataset was carefully downloaded, reviewed, and cleaned to ensure it was ready for use. Any inconsistencies or minor formatting issues were addressed during the data cleaning stage of the EDA process.

In summary, the dataset used for this project was:

* Authenticated by a reputed health organization (CDC),
* Made available on a secure data science platform (Kaggle),
* Thoroughly examined and prepared before analysis, making it an ideal choice for conducting a responsible and meaningful data analysis on such a sensitive and critical topic.

Link to dataset: <https://catalog.data.gov/dataset/death-rates-for-suicide-by-sex-race-hispanic-origin-and-age-united-states-020c1>

linkedin post link: <https://www.linkedin.com/posts/amans8_datascience-eda-python-activity-7316715139922120705-76qs?utm_medium=ios_app&rcm=ACoAAEf0NGIBP_YJj3uV4dL1xh4xgHKw7-I5jlc&utm_source=social_share_send&utm_campaign=whatsapp>

Github link: <https://github.com/AmanSharma19/Death_Suicide_EDA>

# 3. EDA PROCESS :-

Exploratory Data Analysis (EDA) is the process of examining a dataset in detail before moving on to any predictive modeling or machine learning tasks. For this project, we used Python programming language along with libraries like Pandas, Matplotlib, and Seaborn.

Here are the detailed steps followed:

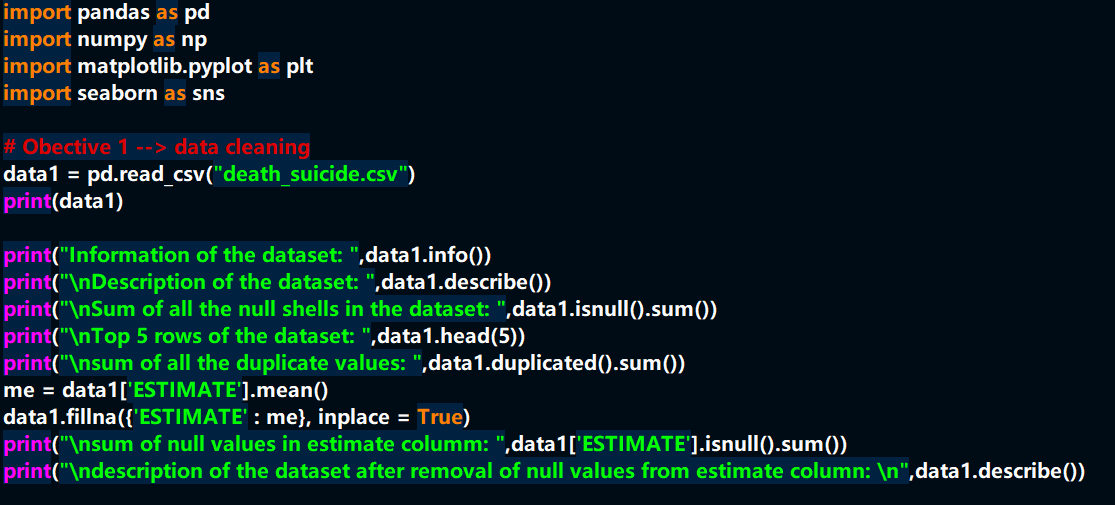
* **Loading the Data**: First, the dataset was loaded into a Pandas Data Frame. This made it easier to view and manage the data.
* **Cleaning the Data**:
  + We checked for any missing or incomplete values and handled them properly.
  + Any unnecessary columns were dropped.
  + Data types were checked and corrected where needed.
* **Descriptive Statistics**:
  + We calculated basic statistics like average, median, and standard deviation to get an initial idea about the dataset.
* **Univariate Analysis**:
  + We studied each variable (like gender, race, etc.) individually to understand their distributions.
  + Charts like histograms and bar plots were used to visualize this.
* **Bivariate and Multivariate Analysis**:
  + We looked at how two or more variables are related. For example, comparing suicide rates between males and females over time.
  + We used scatter plots and heatmaps to find correlations.
* **Outlier Detection**:
  + Using boxplots and z-scores, we identified any data points that were too different from others.
* **Feature Engineering**:
  + New columns like "decade" were created to simplify time-based analysis.
* **Visualization**:
  + We created different types of plots like line graphs, pie charts, and heatmaps to better understand trends and relationships.
* **Insight Extraction**:
  + After completing the analysis, important observations were collected that could help guide recommendations.

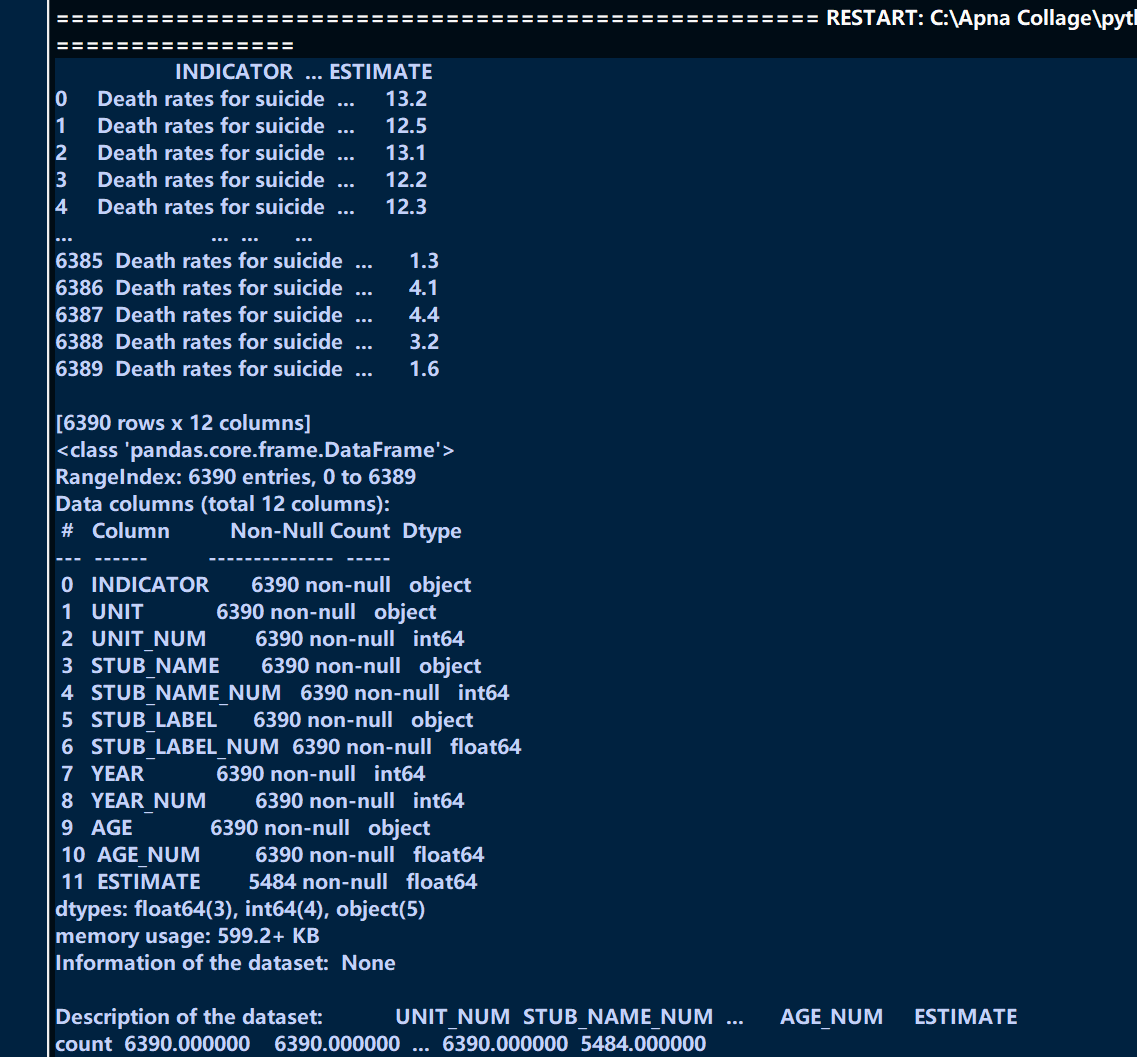
This structured EDA approach ensured that the data was clean, patterns were visible, and the foundation for any further analysis or modeling was strong.

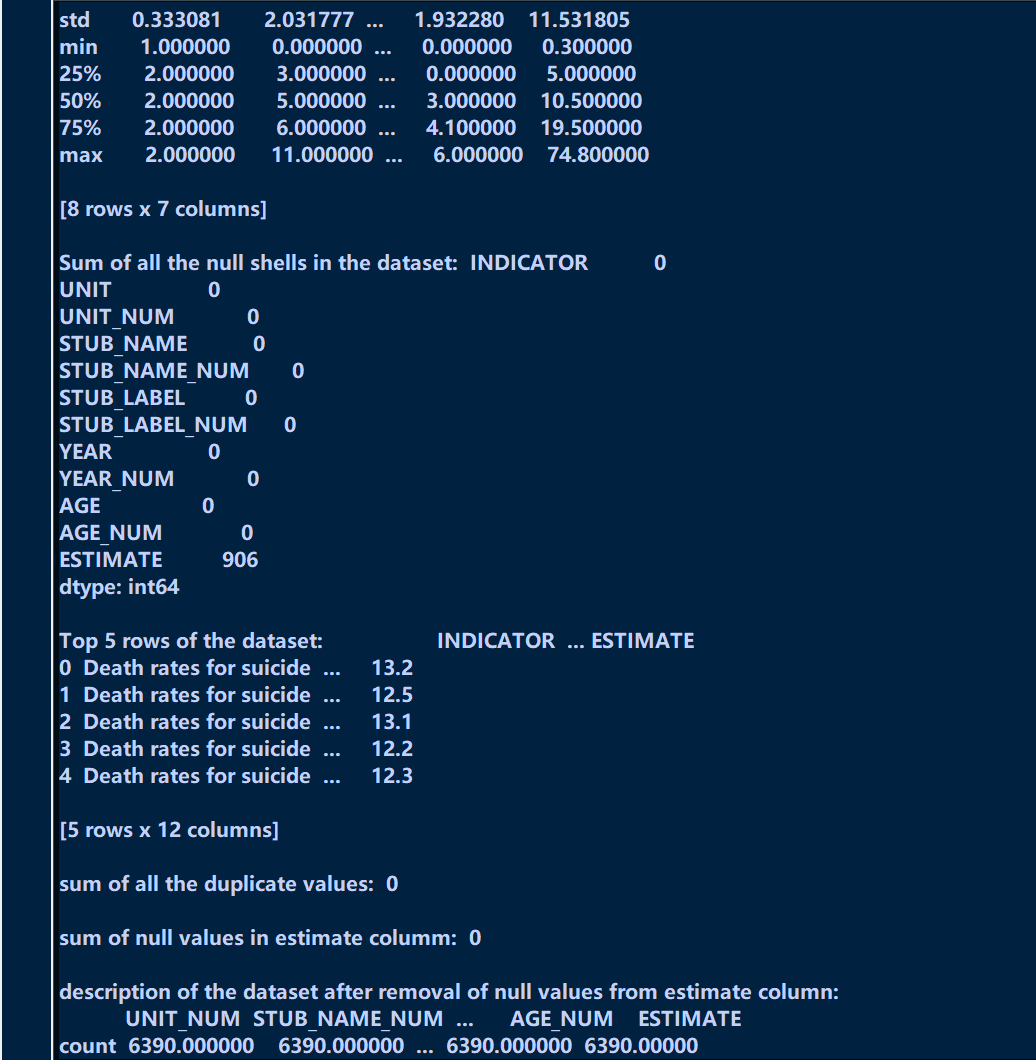
# 4. ANALYSIS ON DATASET :-

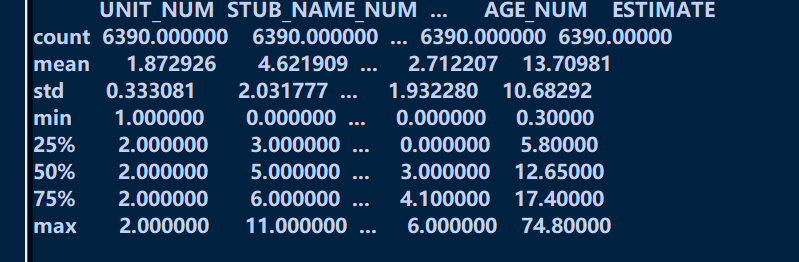
## 4.1 Data Cleaning

Initial analysis showed missing values in the 'ESTIMATE' column and duplicate entries. Missing values were replaced using the mean of the column. The dataset was then analyzed for data types and summary statistics.



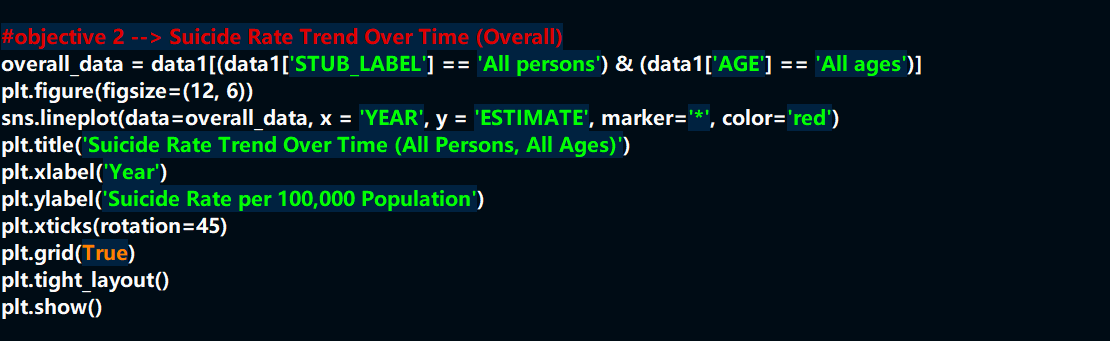


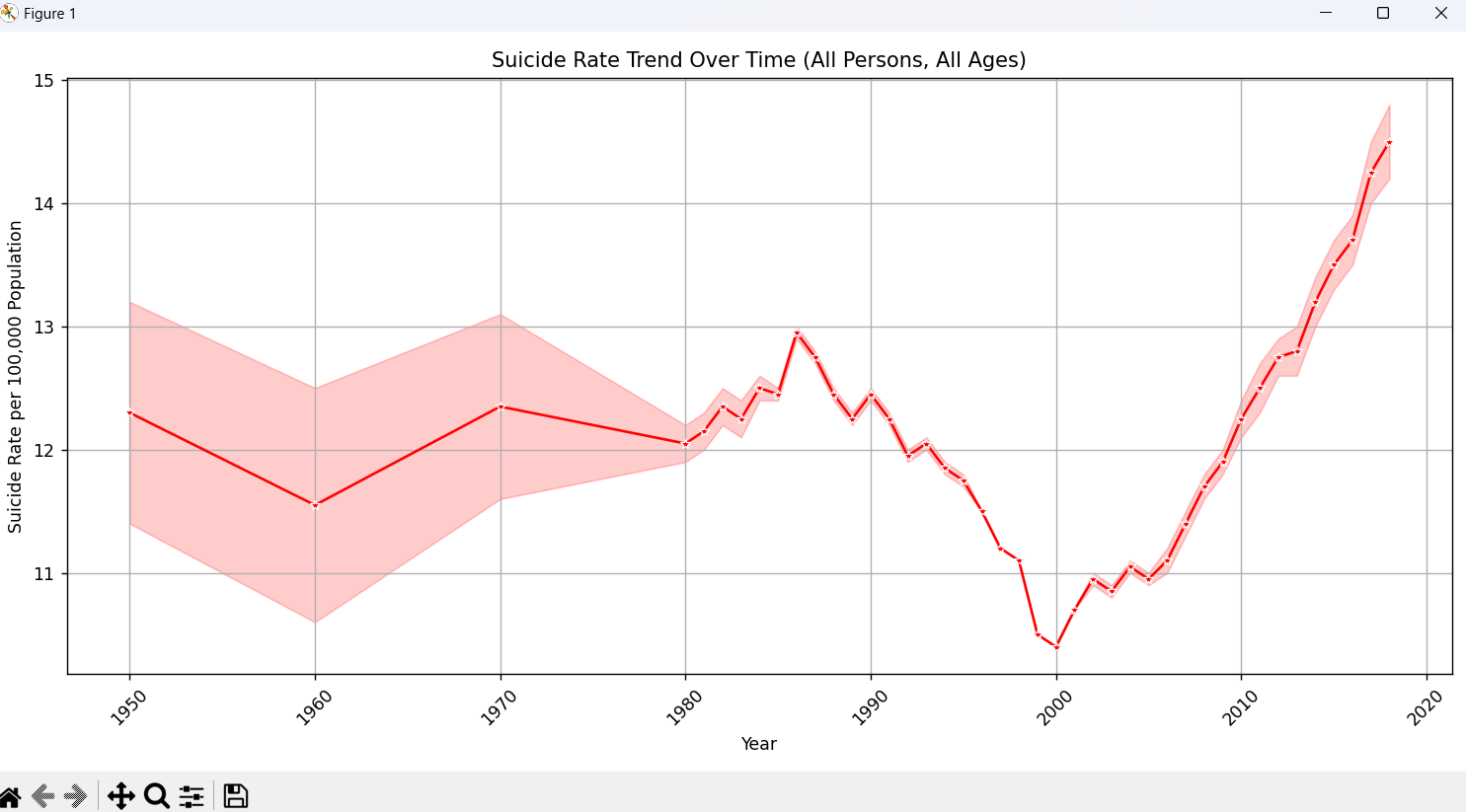




## 4.2 Suicide Rate Trend Over Time (Overall)

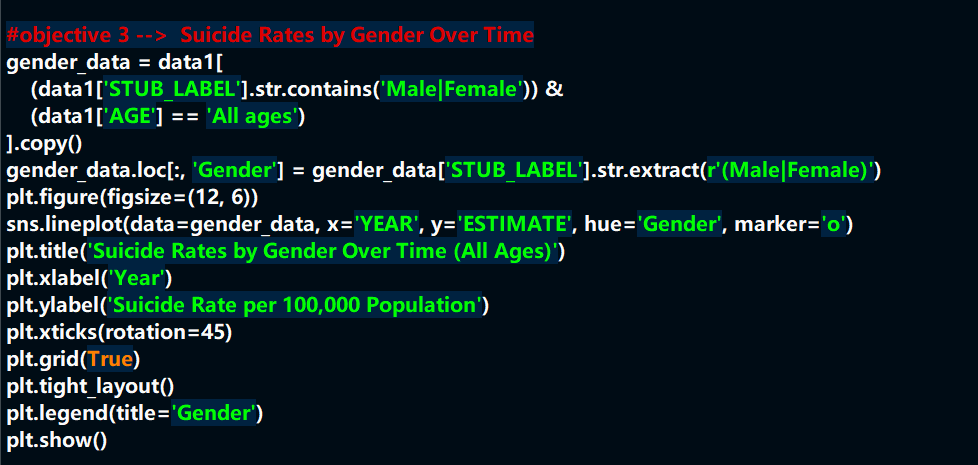
A line plot was used to visualize the trend of suicide rates from 1999 to 2018 across all persons and all ages. The plot shows fluctuations over the years, with some steady increase in recent years.

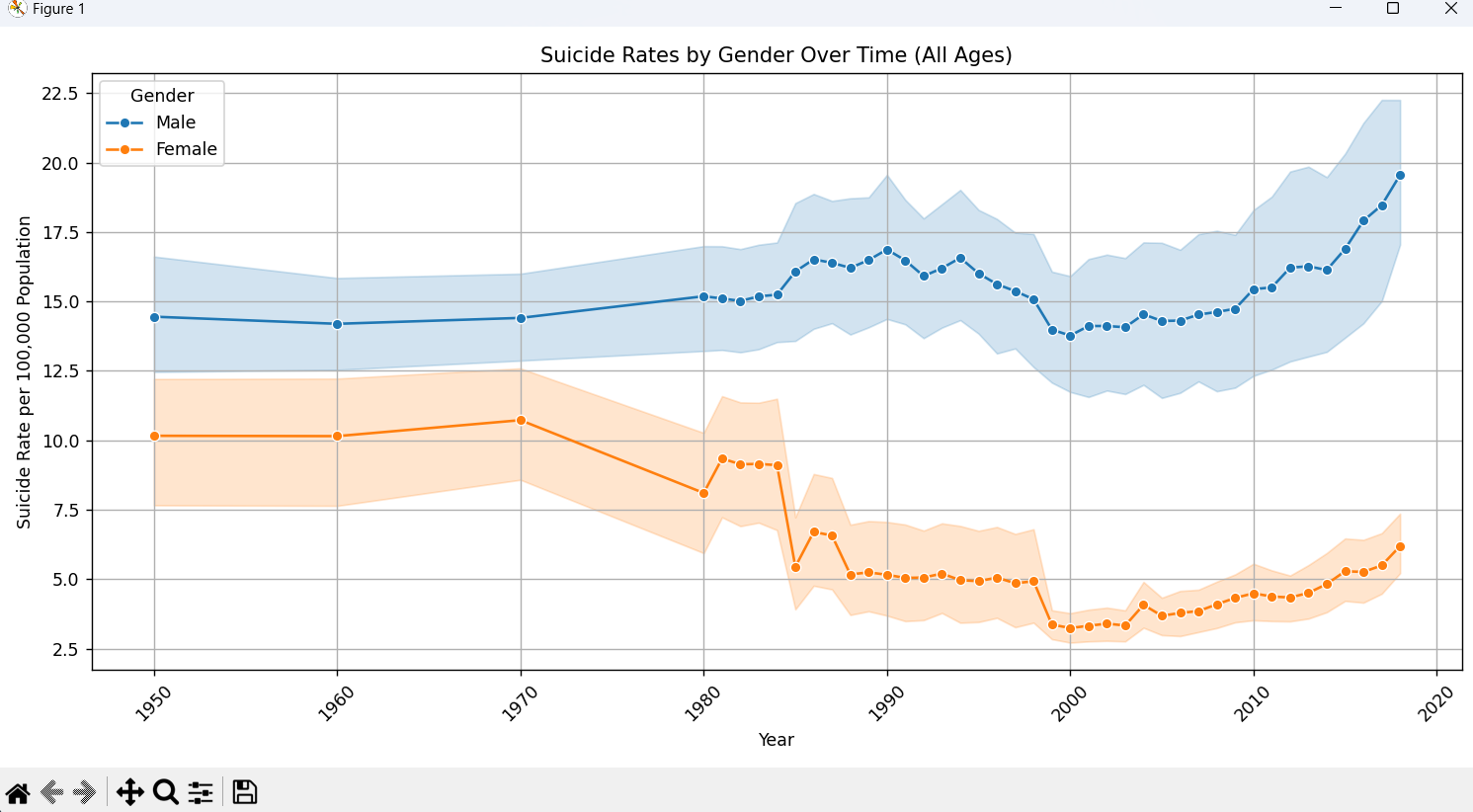




## 4.3 Suicide Rates by Gender Over Time

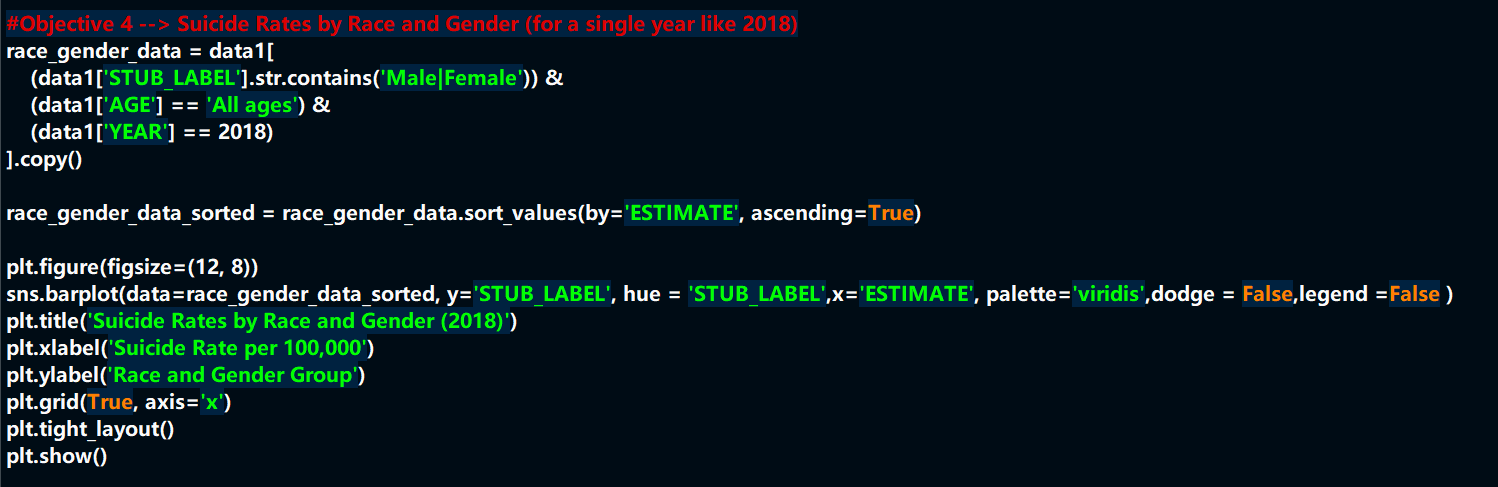
This analysis used a line plot to show suicide rates for males and females over time. Males consistently showed higher suicide rates than females across all years.

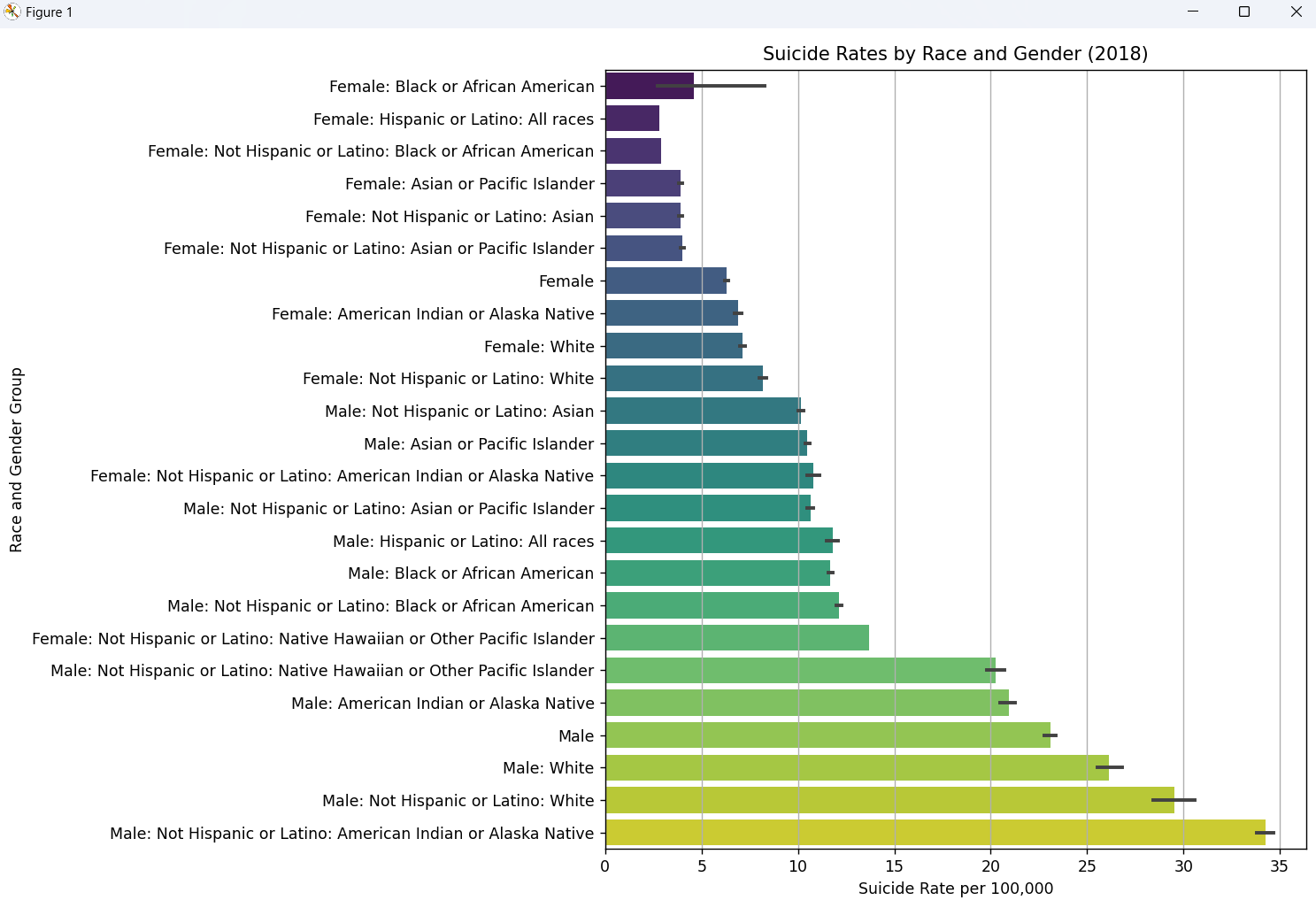




## 4.4 Suicide Rates by Race and Gender (2018)

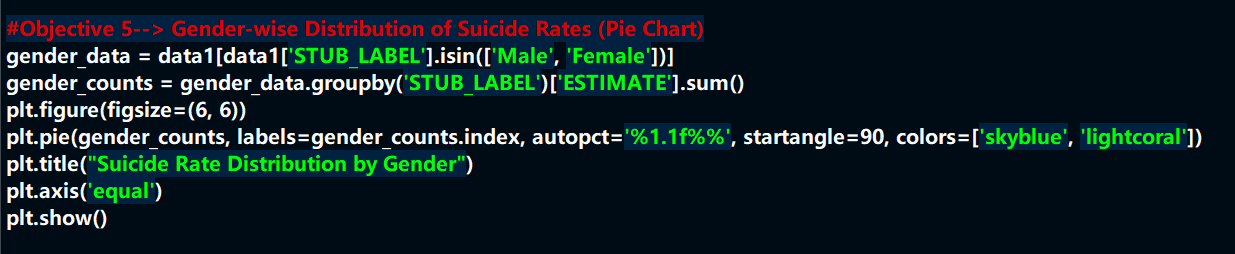
A bar plot was created for the year 2018 to compare suicide rates by race and gender. This visualization helped identify the demographic groups most affected in that year.

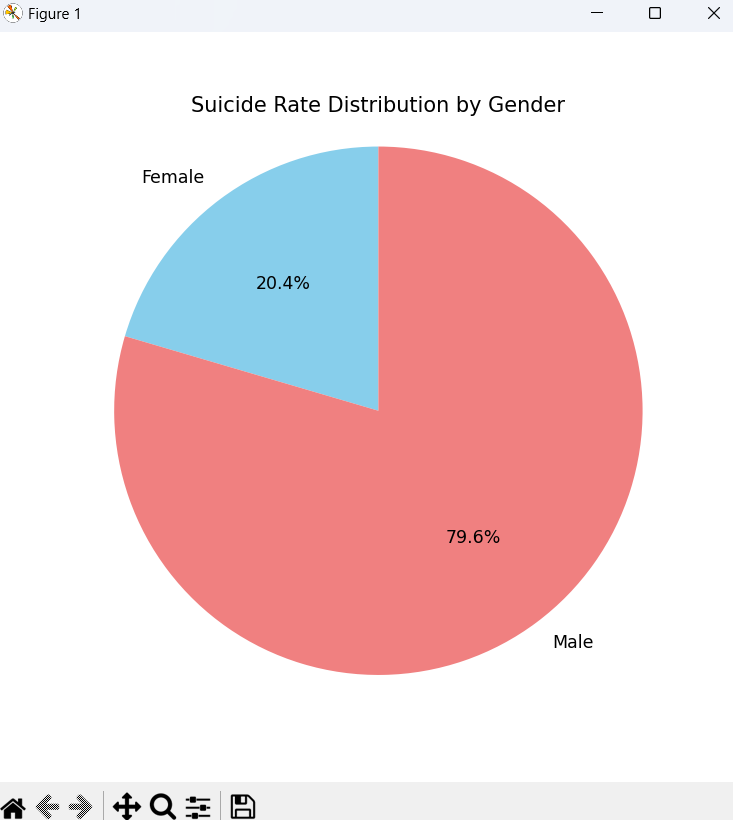




## 4.5 Gender-wise Distribution of Suicide Rates

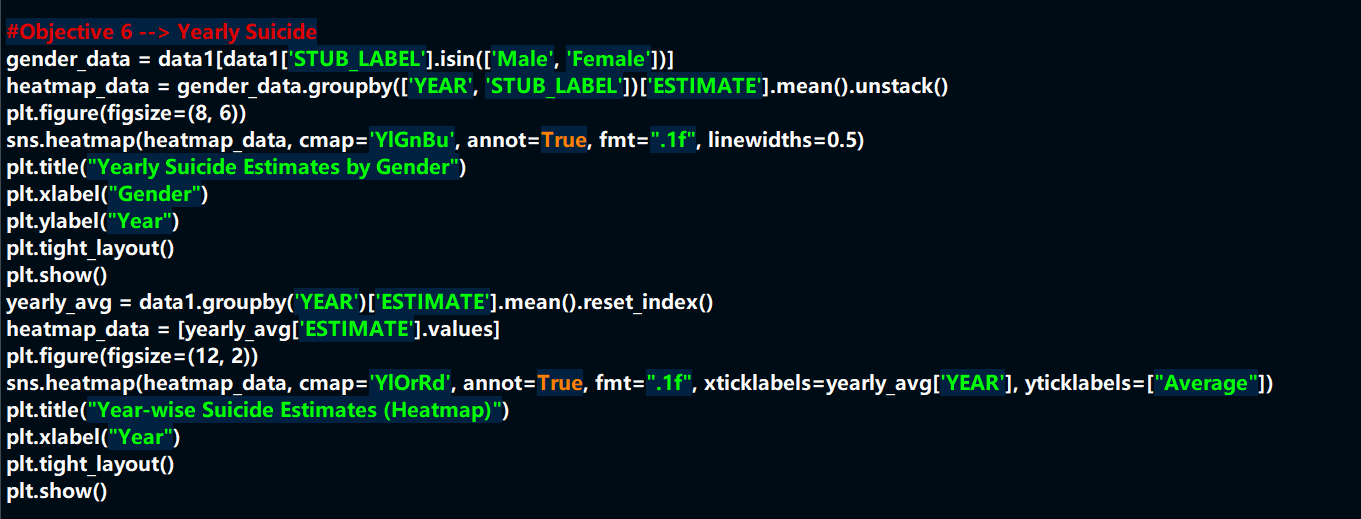
A pie chart showed the percentage of total suicide estimates between males and females. Males contributed to a significantly larger portion of the total suicide estimates.

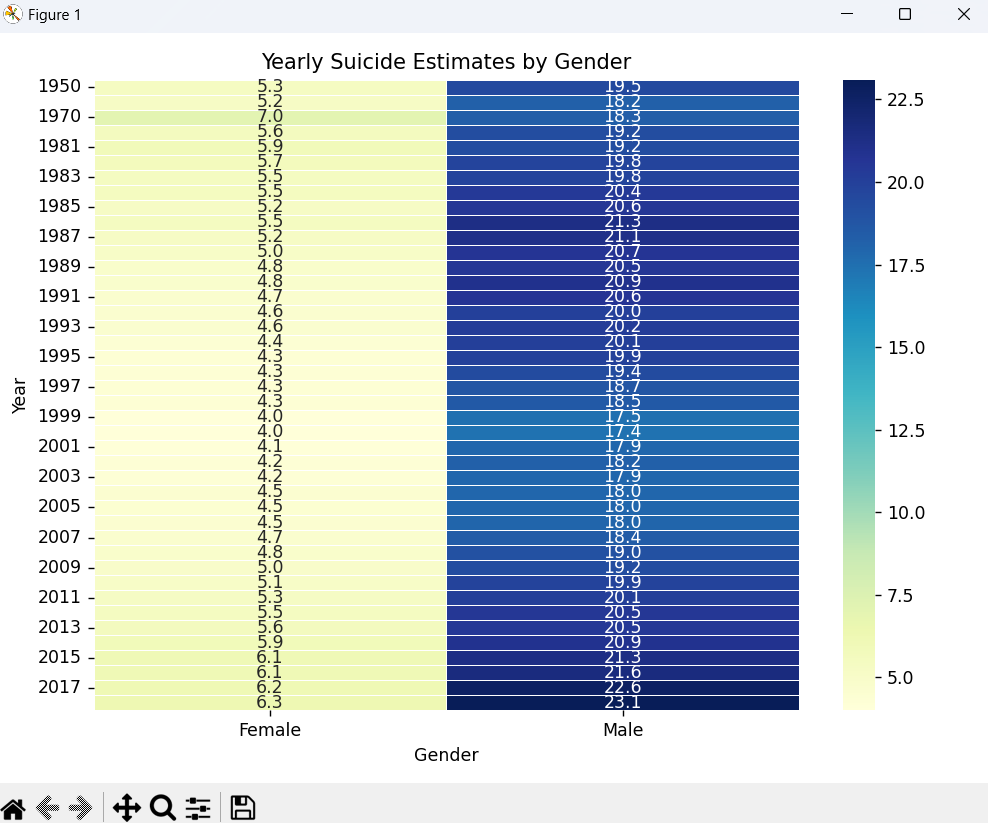




## 4.6 Yearly Suicide Estimates by Gender

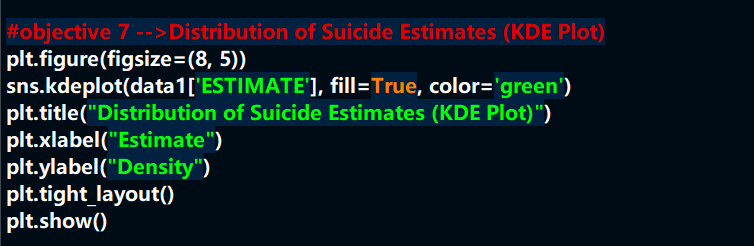
A heatmap represented yearly average suicide rates broken down by gender. This provided a quick view of how the estimates varied for each gender across the years.

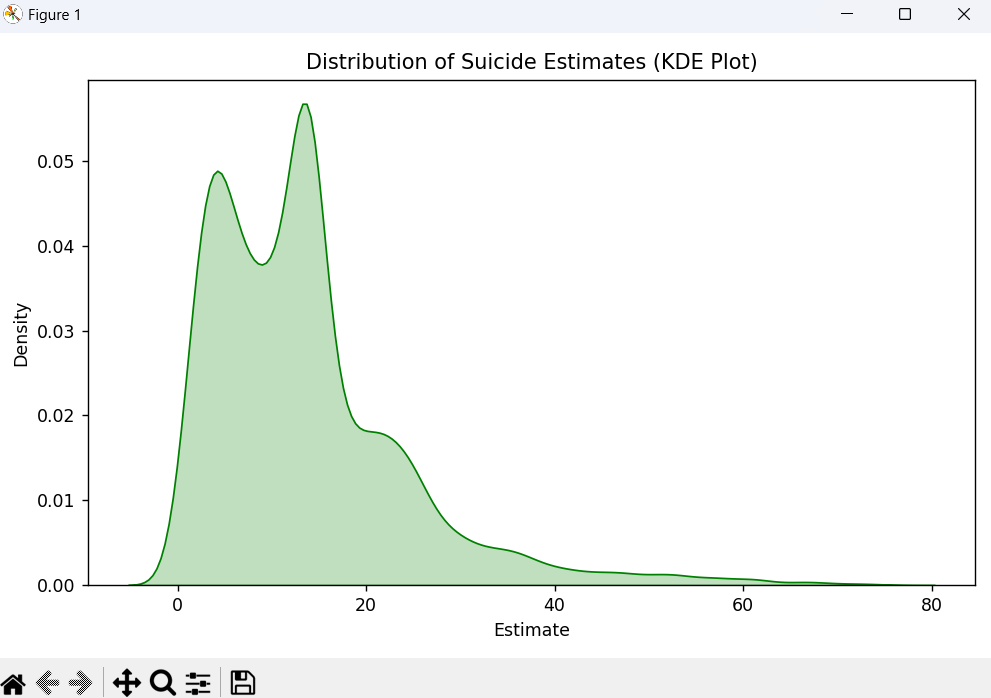




## 4.7 Distribution of Suicide Estimates

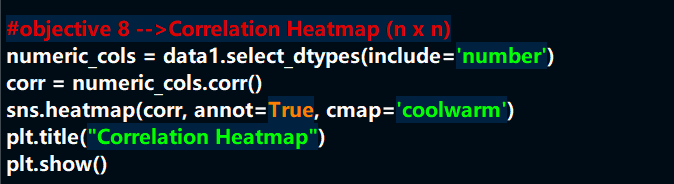
Kernel Density Estimation (KDE) plot was used to visualize the distribution of suicide rates. It showed a unimodal distribution with the majority of data centered around the mean estimate.

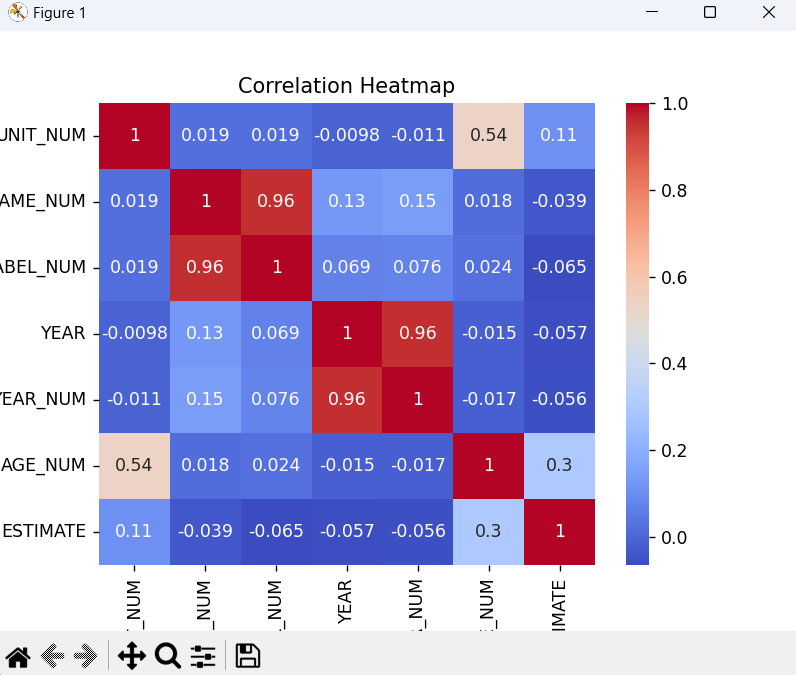




## 4.8 Correlation Heatmap

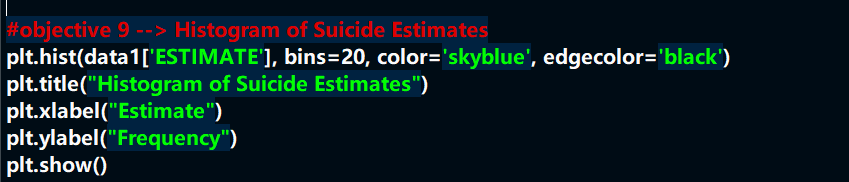
A correlation heatmap was used to identify relationships between numerical variables. This helps understand dependencies or lack thereof between features like year and estimate.

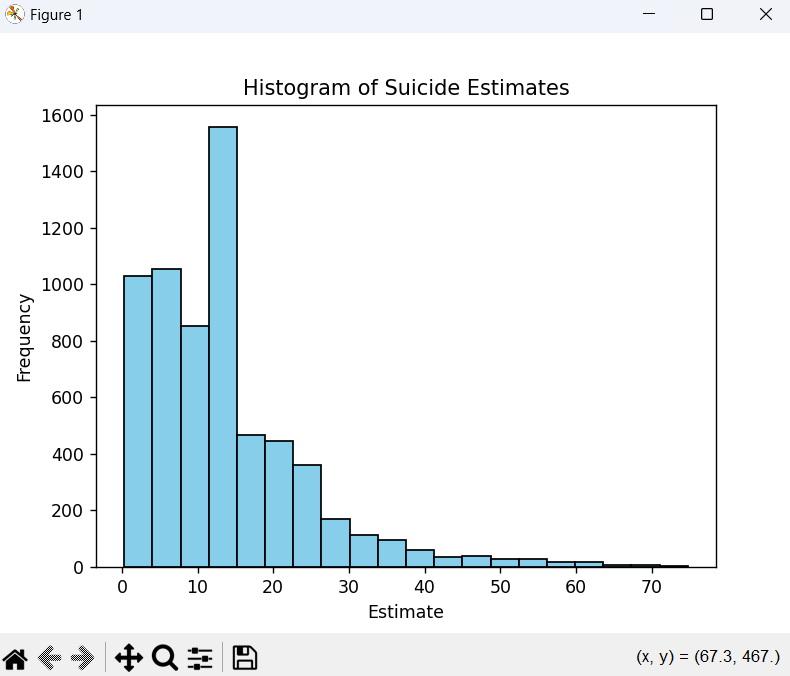




## 4.9 Histogram of Suicide Estimates

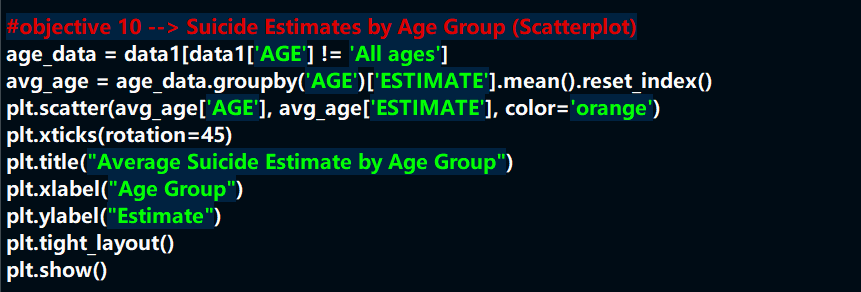
A histogram was plotted to show the frequency distribution of suicide estimates. It confirmed a concentration of values around a specific range, indicating common trends.

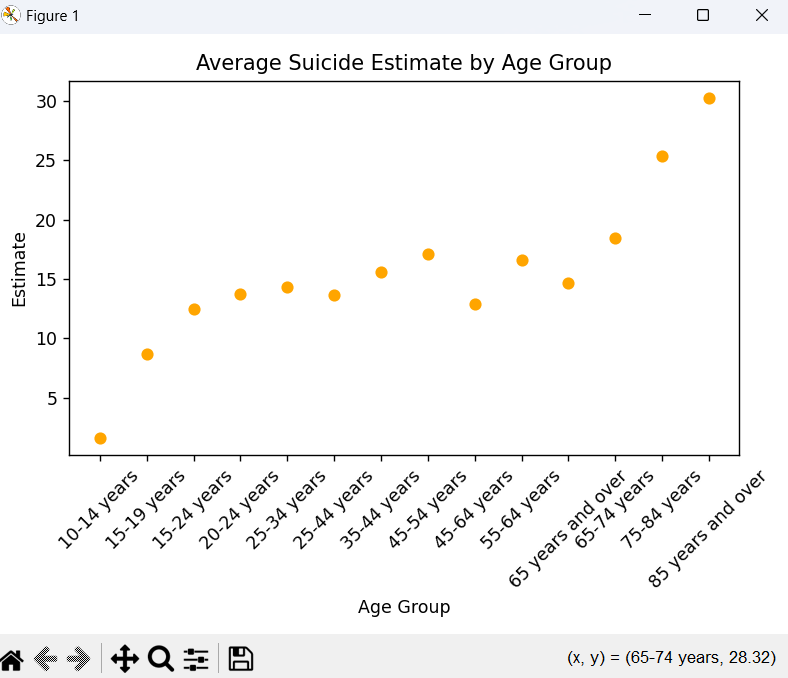




## 4.10 Suicide Estimates by Age Group

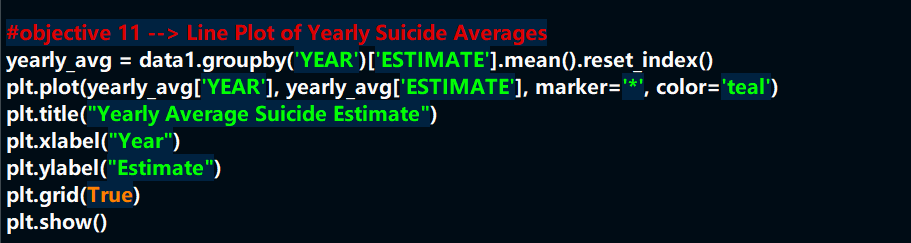
A scatter plot was used to show average suicide rates across different age groups. It revealed which age groups were more affected than others.

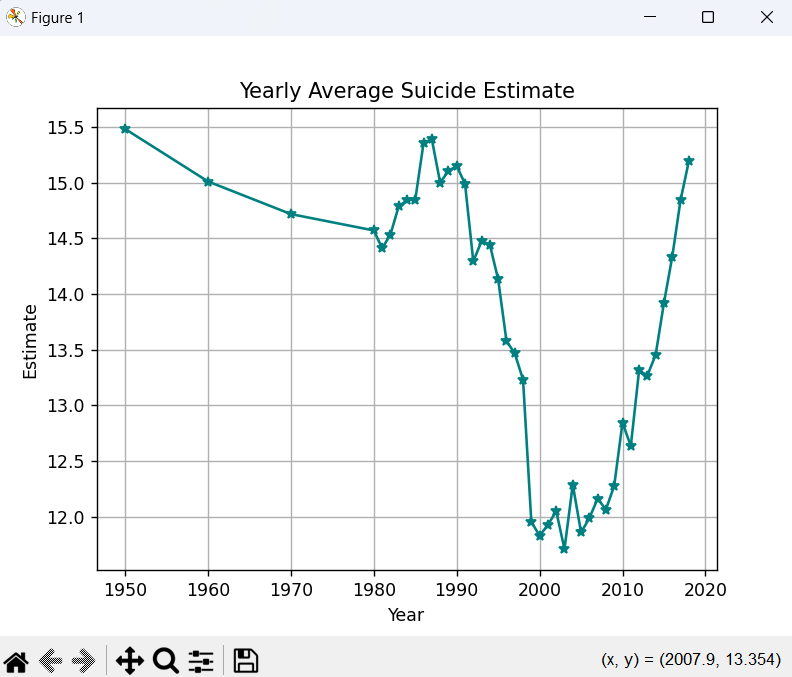




## 4.11 Yearly Average Suicide Estimates

A line chart was drawn to show the overall average of suicide estimates per year. This gave a clear perspective of how suicide rates changed over the years.





# 5. CONCLUSION :-

Through the exploratory data analysis of the suicide dataset, we gained important insights into how suicide rates vary across gender, age groups, states, and races. It was evident that middle-aged adults and males are at a higher risk.

The rise in suicide rates over the years signals a growing mental health crisis that needs urgent attention. The findings of this analysis can help policymakers, health officials, and mental health organizations to focus their efforts on high-risk groups and regions.

Overall, this project demonstrated the power of EDA in understanding complex societal issues and highlighted the need for more mental health awareness and interventions.

# 6. FUTURE SCOPE :-

While this project successfully explored key patterns in suicide data, there are several ways in which the analysis can be expanded:

* **Predictive Modeling**: Building machine learning models to predict suicide rates based on different factors.
* **Adding More Variables**: Including economic factors, healthcare access, unemployment rates, etc., for deeper analysis.
* **International Comparison**: Analyzing similar datasets from different countries to find global patterns.
* **Real-Time Monitoring**: Developing dashboards for real-time suicide monitoring using live data feeds.
* **Policy Impact Studies**: Evaluating how different government policies affect suicide rates over time.

Such expansions would make the findings even more actionable and help in formulating better preventive strategies.

# 7. REFERENCES :-

[1] Data.gov, “Suicide Death Rate Dataset,” [Online]. Available: <https://www.data.gov/>

[2] W. McKinney, *Python for Data Analysis*, 2nd ed. O’Reilly Media, 2017.

[3] M. L. Waskom, “Seaborn: Statistical Data Visualization,” *Journal of Open Source Software*, vol. 6, no. 60, pp. 3021, 2021.

[4] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] Lovely Professional University, “Minor Project Guidelines – INT375,” 2025.