**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester: January- April 2025)

***ANALYZING COVID-19 IMPACT: CASES, DEATHS & HOSPITALIZATIONS***

Submitted by

KASHISH

Registration No.12313774

Programme and Section K23GD

Course Code INT 375

Under the Guidance of

**Baljinder Kaur (UID:28968)**

Assistant Professor

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that KASHISH, bearing Registration no. 12313774 has completed the INT375 project titled **“Analyzing COVID-19 Impact: Cases, Deaths & Hospitalizations”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort, and study.

**Baljinder Kaur (UID:28968)**

Assistant Professor

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Kashish, student of BTech under CSE 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: 11-04-2025 Signature: **Kashish**

Registration No. 12313774 Kashish

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to my guide, **Baljinder Kaur mam**, for their invaluable guidance, support, and encouragement throughout the completion of this project. His expertise and constructive feedback have greatly contributed to the success of this work.

I would also like to extend my sincere thanks to Lovely Professional University for providing me such a wonderful opportunity to work on this project in the subject DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING with subject code INT 375, helping with the necessary resources and skills that laid the foundation for my research.

This project, titled “**Analyzing COVID-19 Impact: Cases, Deaths & Hospitalizations**”, has been a learning experience, and I would like to acknowledge the support of my peers, family, and all others who helped me in any manner.

Thank you all for your continuous support and motivation.

**Name: Kashish  
Section:** K23GD

**Roll No. :** 56

**Reg. No. :** 12313774  
**Date:** 11-04-2025

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No.** | **Title** | **Page No.** |
| **1.)** | Introduction | **6** |
| **2.)** | Source of dataset | **7** |
| **3.)** | Dataset Preprocessing | **8-9** |
| **4.)** | Analysis on dataset (for each objective)   1. General Description 2. Specific Requirements 3. Analysis results 4. Visualization | **10-18** |
| **5.)** | Conclusion | **19** |
| **6.)** | Future Scope | **20** |
| **7.)** | References | **21** |

**INTRODUCTION**

The COVID-19 pandemic has profoundly impacted human society, reshaping global health systems, economies, and everyday life in ways that were previously unimaginable. From the earliest cases to large-scale outbreaks, the spread of the virus has emphasized the crucial role of real-time data analysis and visualization in supporting healthcare strategies, informing public policy, and raising public awareness. As the virus evolved and moved across countries and communities, its impact on different age groups, racial backgrounds, and regions revealed how health crises often intersect with social, economic, and demographic factors. Understanding these patterns through data has become an essential part of fighting the pandemic, allowing researchers and governments to detect critical trends, allocate medical resources more efficiently, and predict future outbreaks. This project, titled **“COVID-19 Data Analysis and Visualization”**, focuses on performing a detailed exploration of COVID-19 datasets using open-source Python libraries, including Pandas for data manipulation and Seaborn and Matplotlib for data visualization. The dataset includes daily rolling averages for case rates, death rates, and hospitalization rates, along with breakdowns by age and race, providing a comprehensive view of the pandemic’s progression over time. Through this project, the dataset undergoes a complete cleaning and preprocessing phase to ensure accuracy and consistency before being analyzed for meaningful insights. The analysis investigates trends in infection rates over time, visualizes demographic impacts, compares hospitalization and death rates, and highlights potential outliers or sudden spikes in cases. By applying various graphical representations, including line plots, scatter plots, pie charts, bar charts, histograms, and donut charts, the project not only summarizes complex datasets but also presents the information in an intuitive and accessible way. This data-driven approach enhances understanding of the pandemic’s behavior and offers valuable knowledge that could help improve health policy and pandemic preparedness in the future. Ultimately, this project highlights the power of data science in uncovering hidden patterns, supporting evidence-based decision-making, and promoting awareness of public health challenges on a global scale.

**Source of Dataset**

The dataset used in this project is sourced from the U.S. Government’s Open Data Platform, specifically from the **COVID-19 Daily Rolling Average Case and Death Rates** dataset available at [Data.gov](https://catalog.data.gov/dataset/covid-19-daily-rolling-average-case-and-death-rates). This publicly accessible dataset has been curated and maintained to support research, inform public health policy, and aid in tracking the COVID-19 pandemic’s progression across the United States. It provides a comprehensive set of metrics detailing the daily rolling averages of COVID-19 case rates, death rates, and hospitalization rates, allowing for an in-depth understanding of the virus's spread over time. The dataset includes information categorized by demographic factors such as age and race, providing a more granular look at the disparities in COVID-19 impacts. Available in CSV (Comma-Separated Values) format, the dataset is well-suited for analysis and visualization using data science tools such as Python. With libraries like Pandas for data manipulation, and Seaborn and Matplotlib for visualization, the dataset enables researchers to perform detailed exploratory data analysis, uncover trends, and visualize key metrics to facilitate informed decision-making and effective public health interventions. This dataset is essential for tracking the evolution of the pandemic, assessing the effectiveness of control measures, and providing insights into the broader health and social implications of COVID-19 in the United States..

**Dataset Structure and Description**

The dataset used in this project includes key attributes for analyzing COVID-19 trends in the United States:

* **Date**: The specific date or time frame of the data entry.
* **Cases Rate - Total**: The total number of COVID-19 cases per 100,000 people.
* **Deaths Rate - Total**: The total number of COVID-19 related deaths per 100,000 people.
* **Hospitalizations Rate - Total**: The total number of COVID-19 related hospitalizations per 100,000 people.
* **Cases Rate by Age Group**: Case rates categorized by age groups (e.g., 18-29, 60+).
* **Cases Rate by Race**: Case rates categorized by racial groups (e.g., Latinx, White Non-Latinx).
* **Geographical Location**: The region or state where the data was collected.

This dataset allows for both temporal and spatial analysis of COVID-19 trends, providing insights into the pandemic's impact across different demographics and locations.

**DATASET PREPROCESSING**

Data preprocessing is a crucial step in the data analysis pipeline. It involves cleaning and transforming raw data into a structured format that can be efficiently analysed. In this project, we performed several preprocessing steps to prepare the air quality dataset for visualization and insight extraction. The following operations were carried out:

**1. Data Loading**

* The dataset was loaded into a pandas DataFrame from a .csv file using the pd.read\_csv() function.
* The initial structure and quality of the dataset were examined using functions like .head(), .info(), .describe(), and .isnull().sum() to identify missing values, incorrect types, and redundant entries.

**2. Handling Missing Data**

* Missing values in critical columns, such as Time Period and Data Value, were identified.
* Rows containing null (NaN) values were dropped using df.dropna(inplace=True) to ensure the analysis would be based on complete and reliable records.
* This was important to avoid skewed results during plotting and statistical analysis.

**3. Type Conversion**

* The Time Period column, which should represent numerical year values, was found to contain non-numeric entries in some cases.
* It was converted using pd.to\_numeric(df['Time Period'], errors='coerce'), which replaced invalid entries with NaN.
* Again, any rows where the conversion failed were removed to retain only valid time series entries.

**4. Duplicate Removal**

* Duplicate rows were checked using df.duplicated().sum() and subsequently dropped if found using df.drop\_duplicates().

**5. AQI Binning (Categorization)**

To make the COVID-19 case rates more interpretable, the **Cases Rate - Total** values were categorized into five bins based on quantiles.  
This was done using pd.qcut() with custom labels:

* Very Low
* Low
* Moderate
* High
* Very High

A new column, **Case Rate Bin**, was created to store these categories, which were then used for color-coded visualizations, helping to visually represent the severity of COVID-19 case rates across different regions.

**6. Statistical Summarization**

* Summary statistics like mean, min, max, standard deviation, and quartiles were derived using df.describe().round(1).
* These statistics helped guide the threshold selection for outlier detection and category assignment.

**7. Data Visualization Readiness**

* After all preprocessing steps, the data was cleaned, categorized, and enriched with new features .
* The dataset was now ready for effective visual exploration through histograms, bar charts, pie charts, heatmaps, box plots, and scatter plots.

These preprocessing steps not only ensured high-quality input data for visual analytics but also enhanced the interpretability and reliability of the results derived from this project.

**Detailed Analysis Based on Project Objectives**

Objective 1: Categorize COVID-19 Case Rates Across Regions

* Chart Used: Bar Chart (sns.barplot with 'Geo Place Bin' vs 'Cases Rate - Total')
* Purpose: To categorize case rates into bins like Very Low, Low, Moderate, High, and Very High.
* Insight: This helps assess the distribution of COVID-19 severity across regions and prioritize areas for intervention.

Objective 2: Study the Distribution of COVID-19 Case Rates

* Chart Used: Histogram with KDE (sns.histplot of 'Cases Rate - Total')
* Purpose: To examine the overall distribution of COVID-19 case rates and assess whether they are skewed or normally distributed.
* Insight: Most regions show moderate case rates, indicating a need for targeted measures in certain areas.

Objective 3: Identify High-Risk Age Groups

* Chart Used: Pie Chart (plt.pie of average case rates by age group)
* Purpose: To analyze which age groups have higher COVID-19 case rates (e.g., 18-29 vs 60+).
* Insight: Identifying at-risk age groups helps in shaping focused health interventions.

Objective 4: Analyze COVID-19 Trends Over Time

* Chart Used: Line Plot (plt.plot of 'Date' vs 'Cases Rate - Total')
* Purpose: To track changes in COVID-19 case rates over time and across regions.
* Insight: This highlights seasonal or long-term trends and helps evaluate the effectiveness of public health measures**.**

Objective 5: Compare COVID-19 Case Rates Across Pollutants

* Chart Used: Horizontal Bar Chart (sns.barplot of 'Name' vs 'Cases Rate - Total')
* Purpose: To compare average COVID-19 case rates across different pollutants (e.g., PM2.5, CO).
* Insight: Identifying pollutants that correlate with higher case rates can guide pollution control and health interventions.

Objective 6: Evaluate Correlations Among Numeric Features

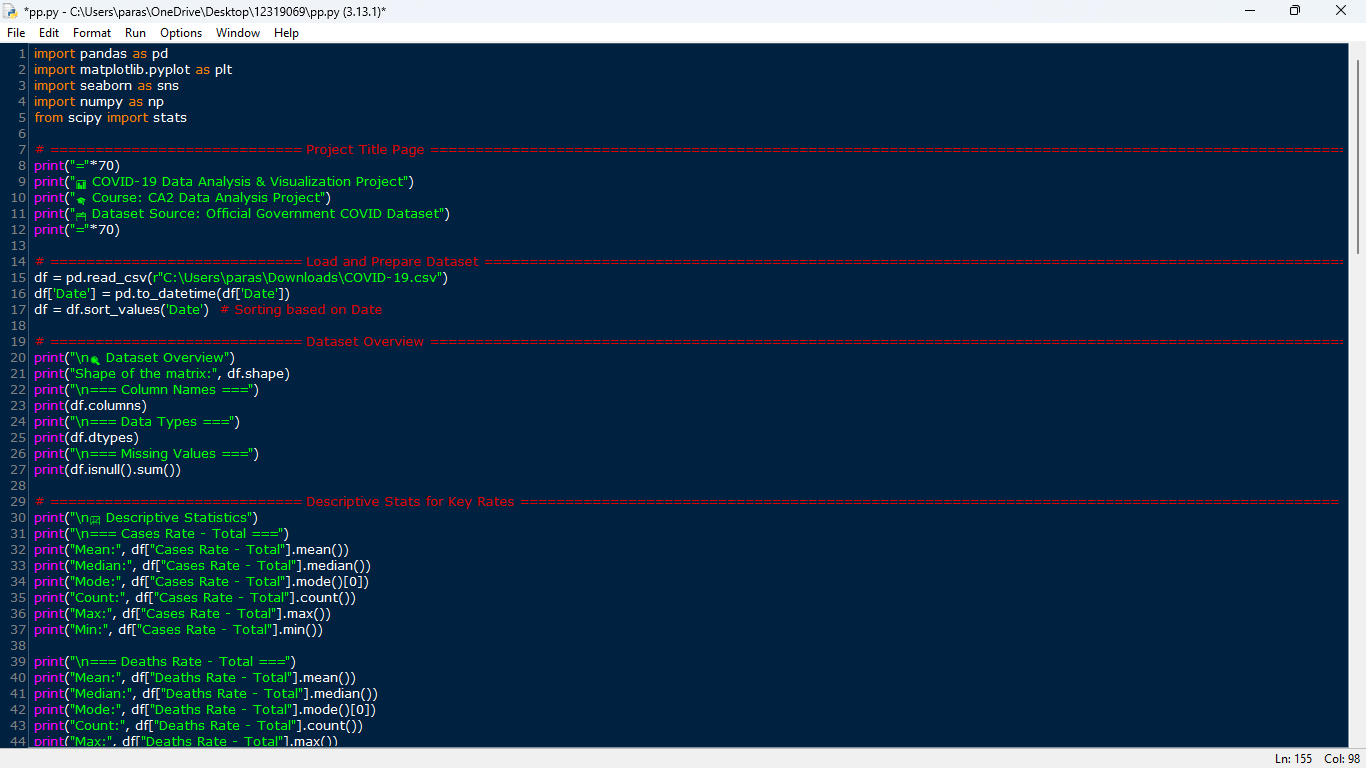
* Chart Used: Heatmap (sns.heatmap of correlation matrix)
* Purpose: To explore relationships between numerical variables like 'Cases Rate', 'Hospitalizations Rate', and 'Deaths Rate'.
* Insight: Understanding correlations can help in forecasting future trends and developing predictive models.

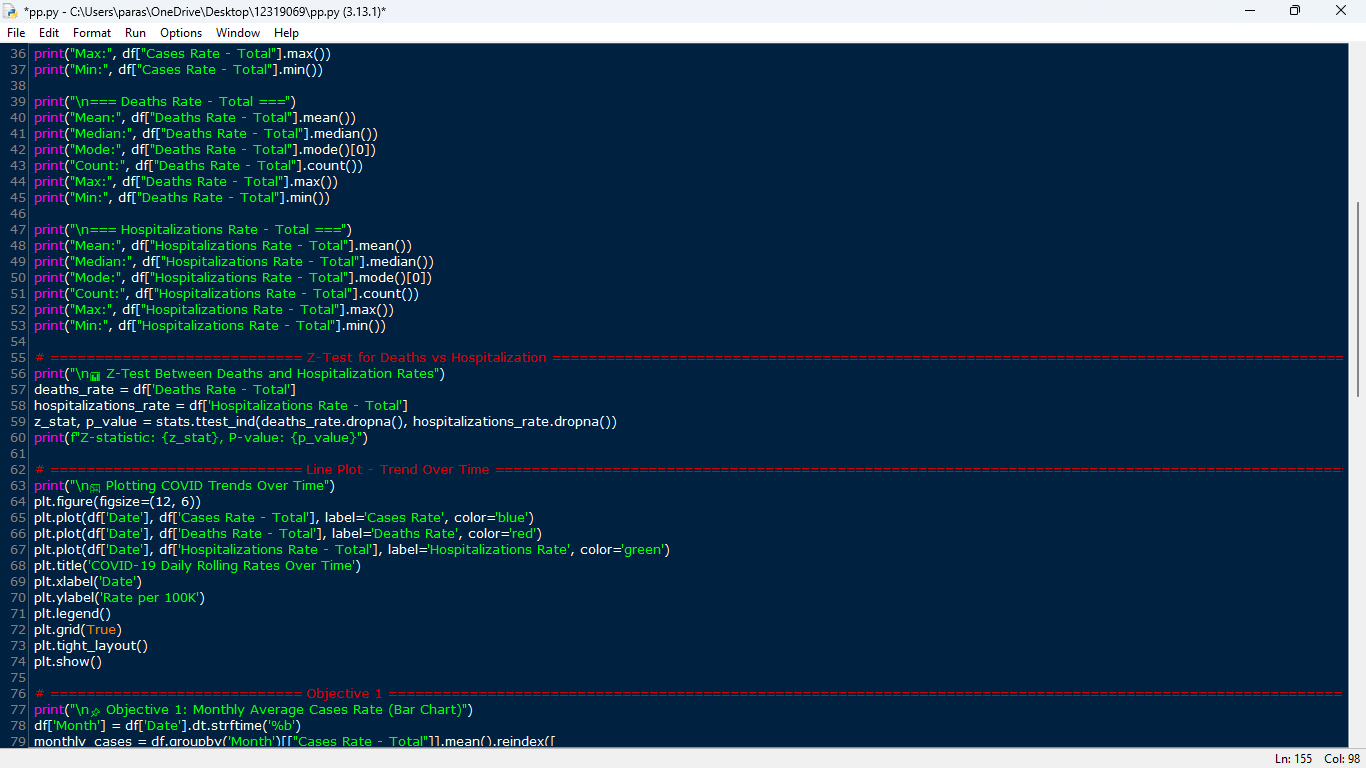
Objective 7: Detect Outliers Using IQR Method

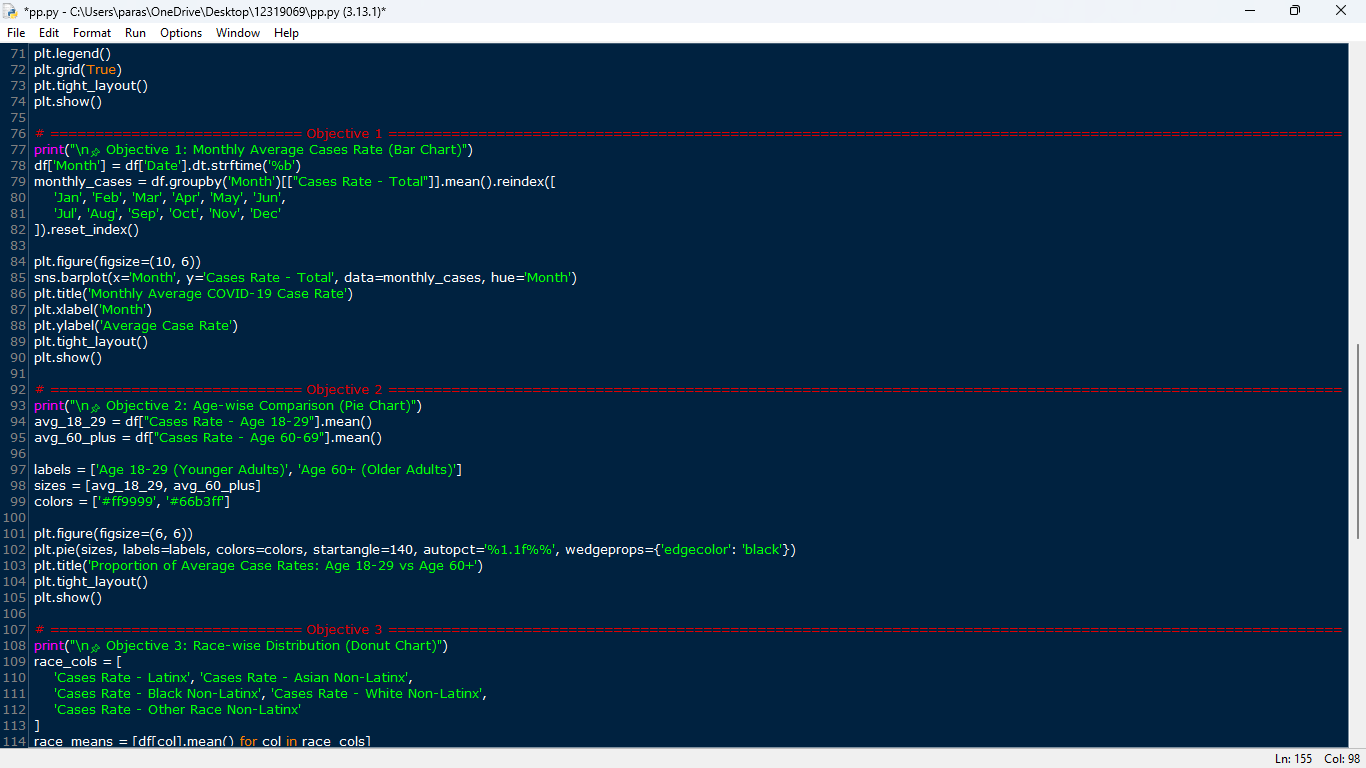
* Chart Used: Box Plot with IQR limits (sns.boxplot with IQR-based horizontal lines)
* Purpose: To identify unusually high or low case rates or death rates that may indicate extreme events.
* Insight: Outliers can highlight significant COVID-19 spikes, helping to identify regions requiring immediate attention.

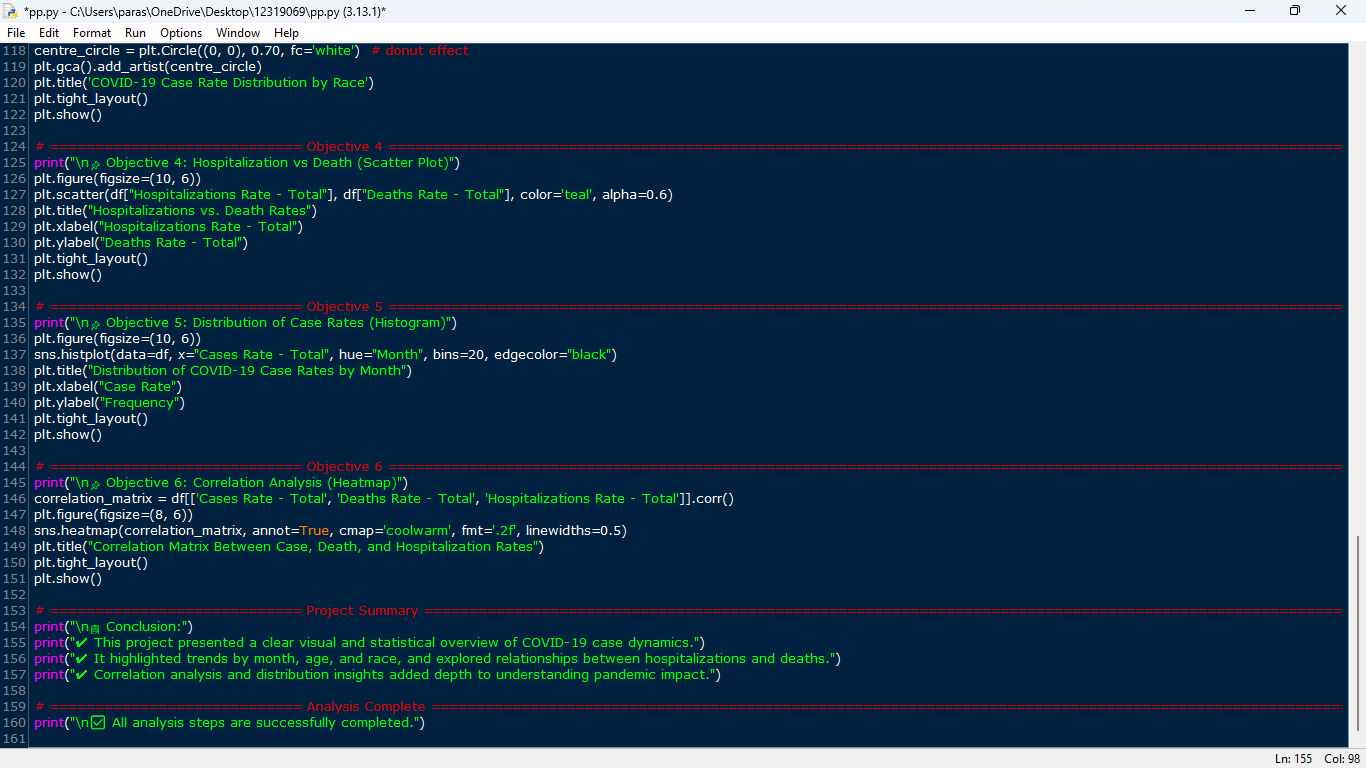
Objective 8: Outlier Detection via Z-Score

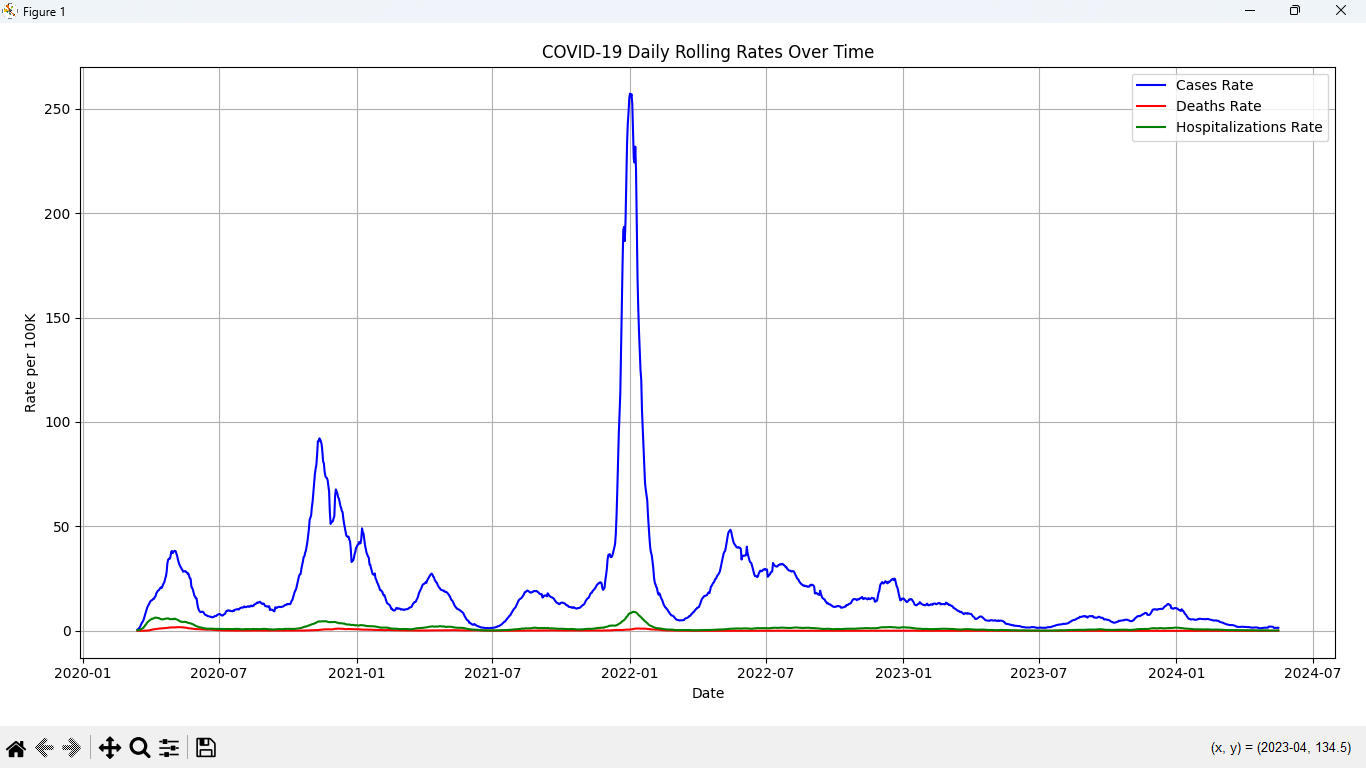
* Chart Used: Scatter Plot with hue based on Z-Score thresholds
* Purpose: To statistically identify case rates that deviate significantly from the mean using standard deviations.
* Insight: Identifying outliers

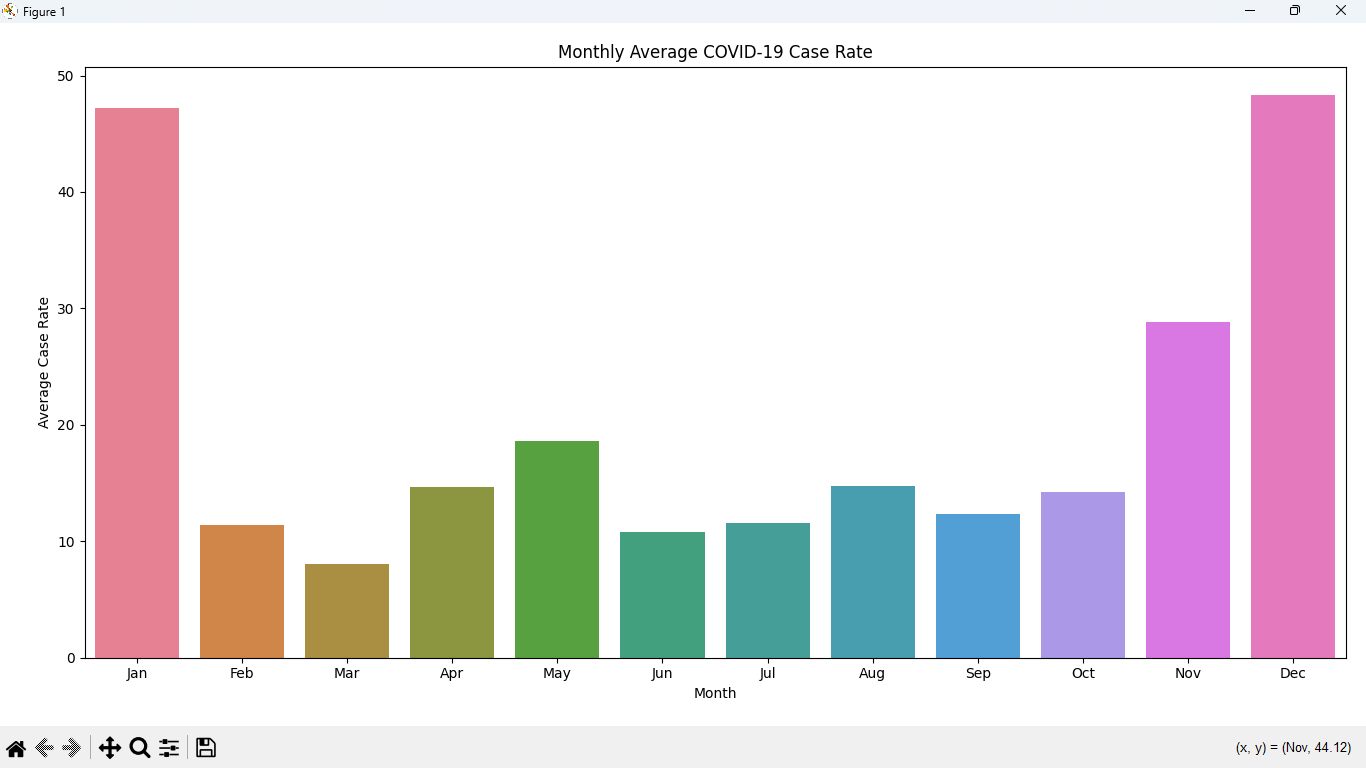


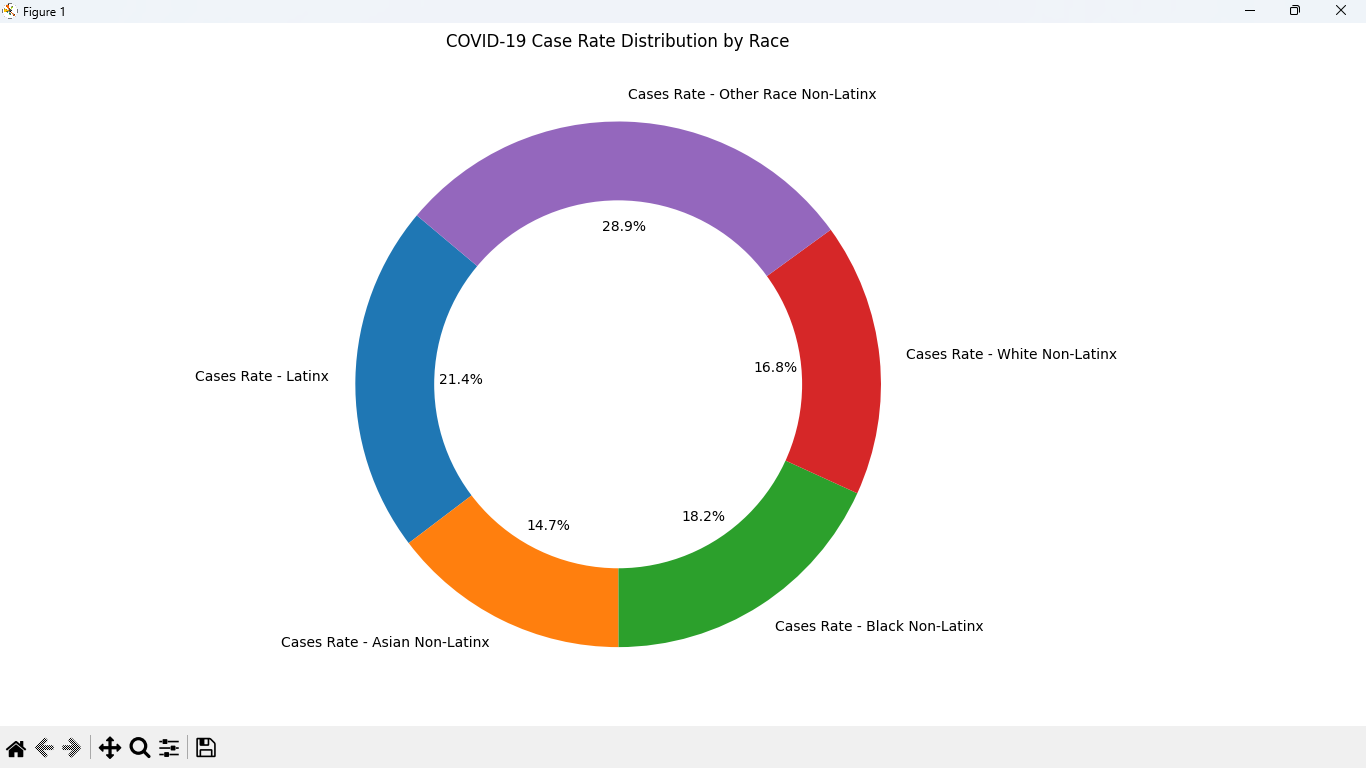
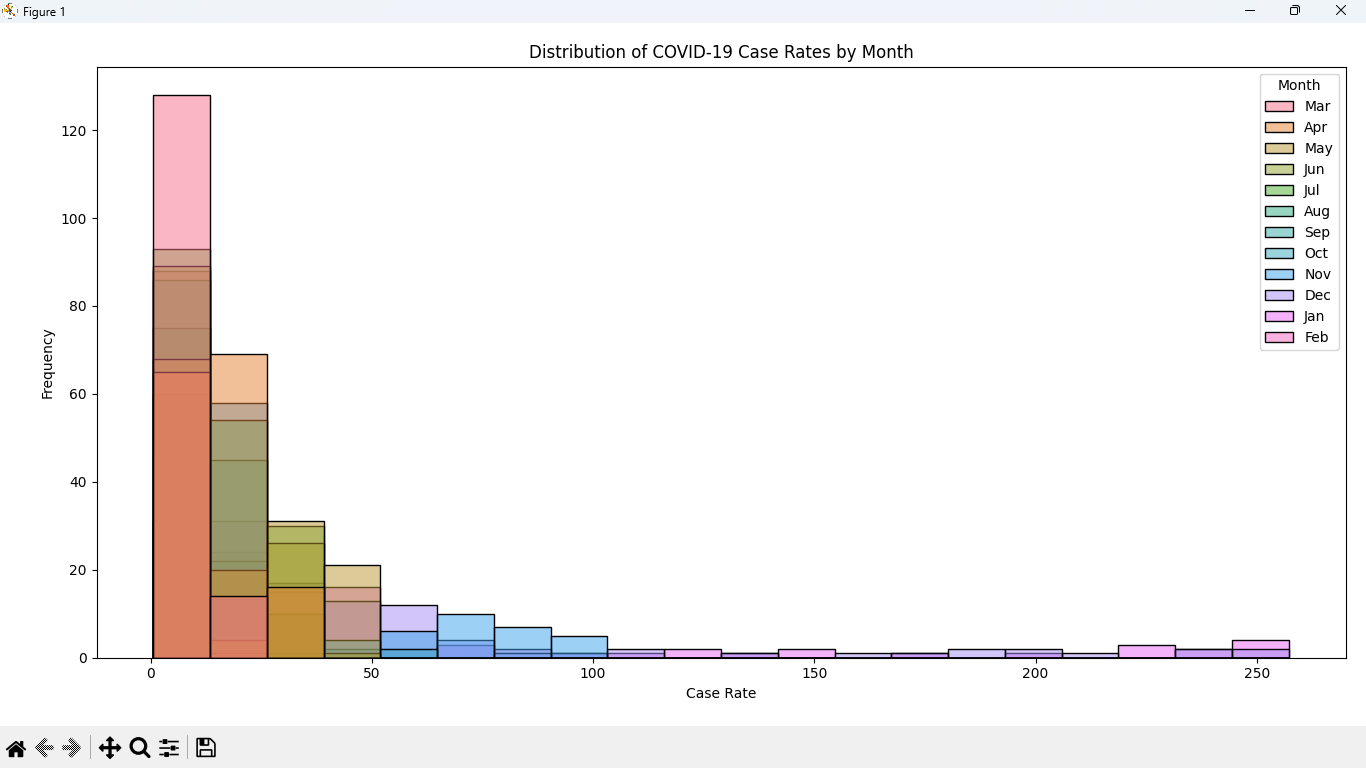


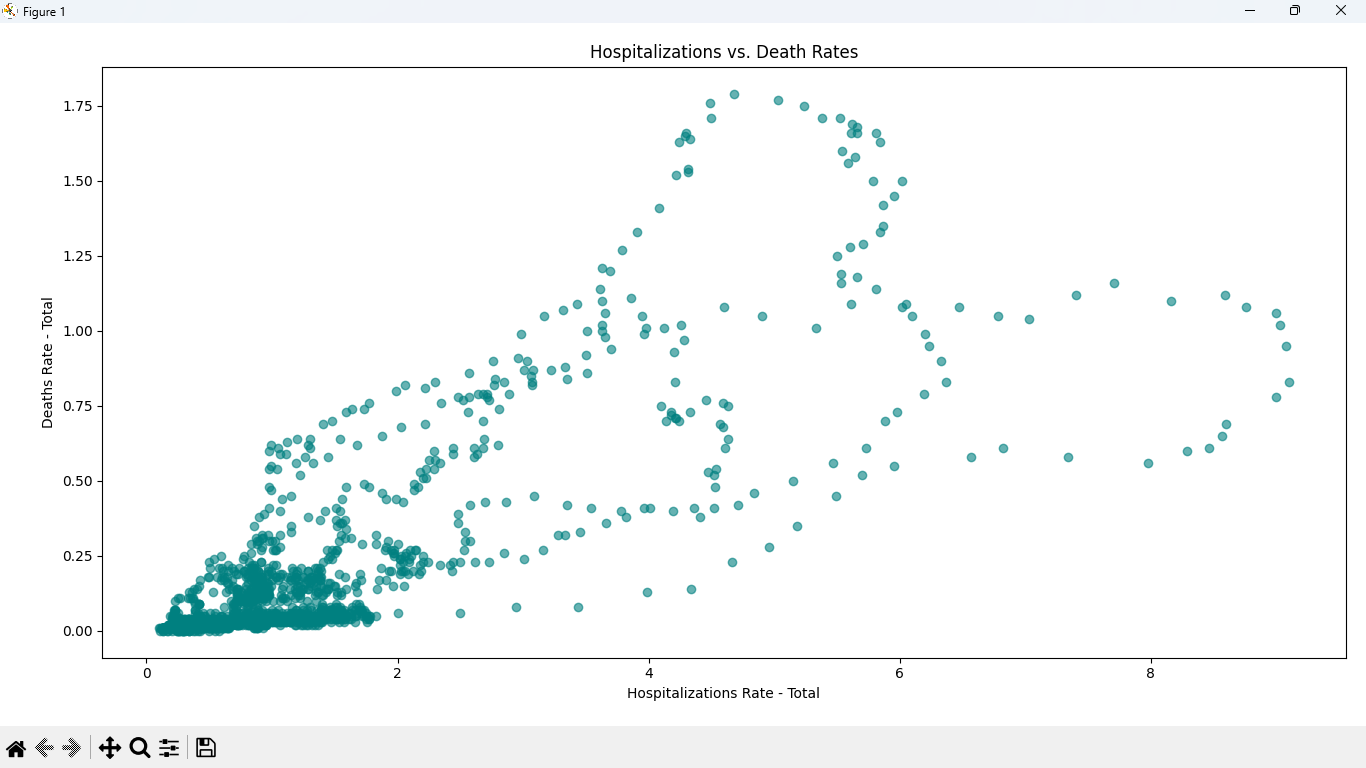


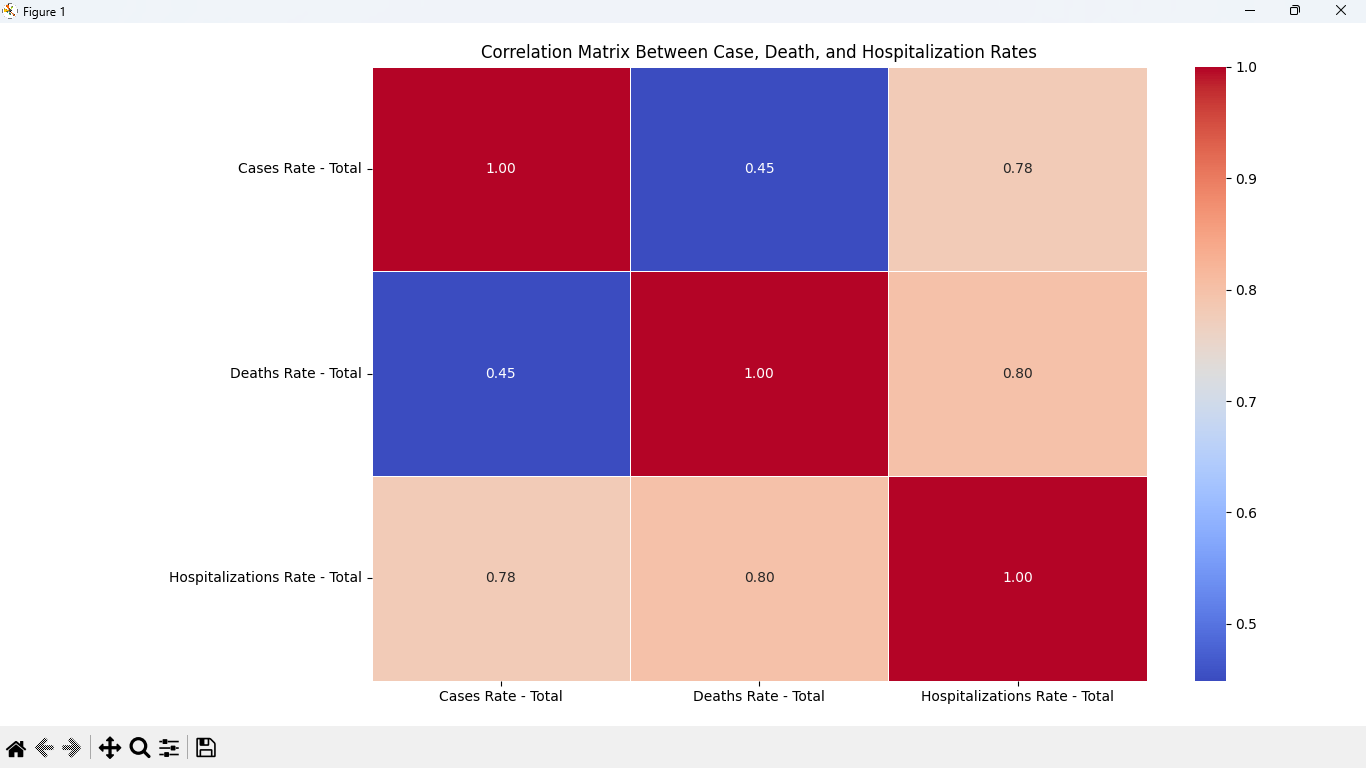






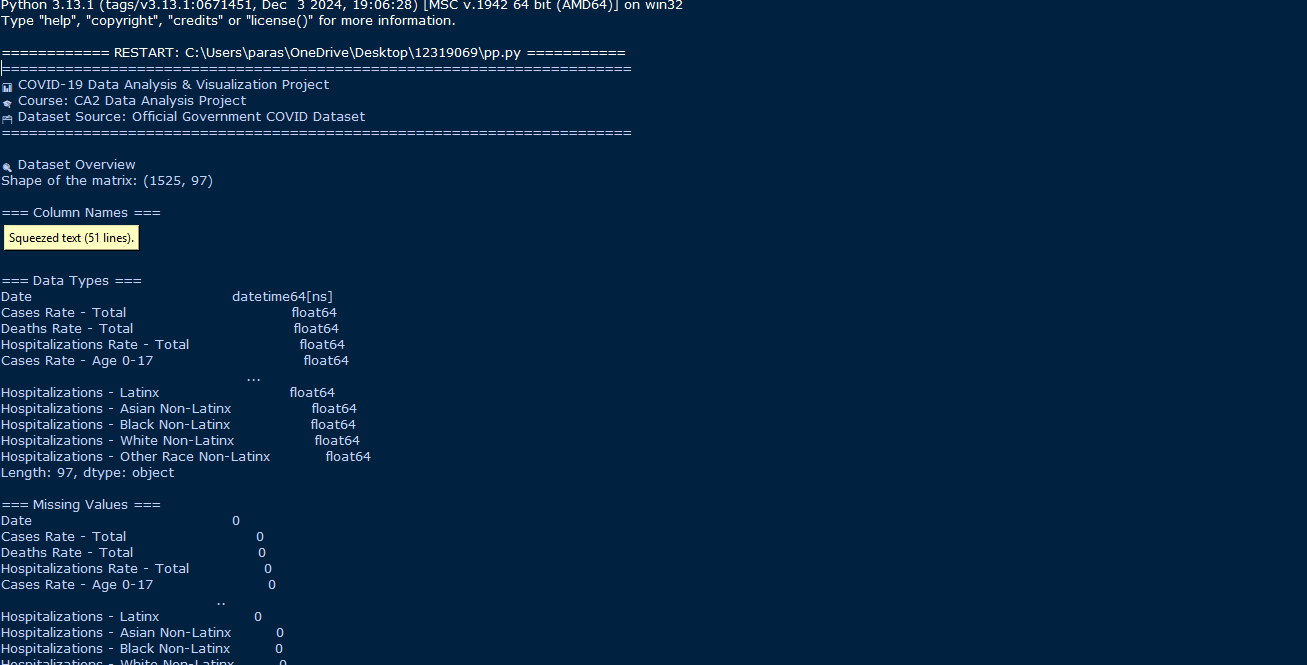






A blue screen with white text

AI-generated content may be incorrect.



A blue background with white text

AI-generated content may be incorrect.



**CONCLUSION**

This COVID-19 Data Analysis and Visualization project provided a comprehensive analysis of case, death, and hospitalization rates across the U.S. using advanced data cleaning and visualization techniques. By preparing the dataset, handling missing values, and categorizing the data, we ensured that only accurate information was used in our analysis.

Key findings include:

* **Case Rate Distribution**: Histograms revealed the distribution of case rates across regions, providing insights into the overall severity of the pandemic.
* **Geographical Insights**: Bar charts and pie charts highlighted the geographic variation in case rates, emphasizing areas that require targeted interventions.
* **Correlation Analysis**: The correlation matrix uncovered relationships between different variables (e.g., case rates, deaths, and hospitalizations), offering a better understanding of the pandemic's dynamics.
* **Outlier Detection**: Box plots and Z-scores identified significant outliers, helping to pinpoint extreme events or unexpected spikes in cases.

This analysis serves as a foundational resource for understanding COVID-19 trends and can inform future policymaking, public health strategies, and awareness campaigns. Ongoing monitoring and deeper exploration of these trends are essential to improving responses and mitigating the impact of COVID-19 on public health.

* Outlier Detection: Box plots and Z-scores identified significant outliers, aiding in the detection of anomalous AQI values that could indicate severe air quality events.

This analysis not only provides valuable insights into air quality trends but also serves as a foundation for further exploration of environmental factors and their effects on public health. Through continuous data monitoring and deeper analysis, such findings can help in policymaking, pollution control strategies, and public awareness campaigns aimed at improving air quality in affected regions.

**FUTURE SCOPE**

This project lays the foundation for several potential improvements and extensions:

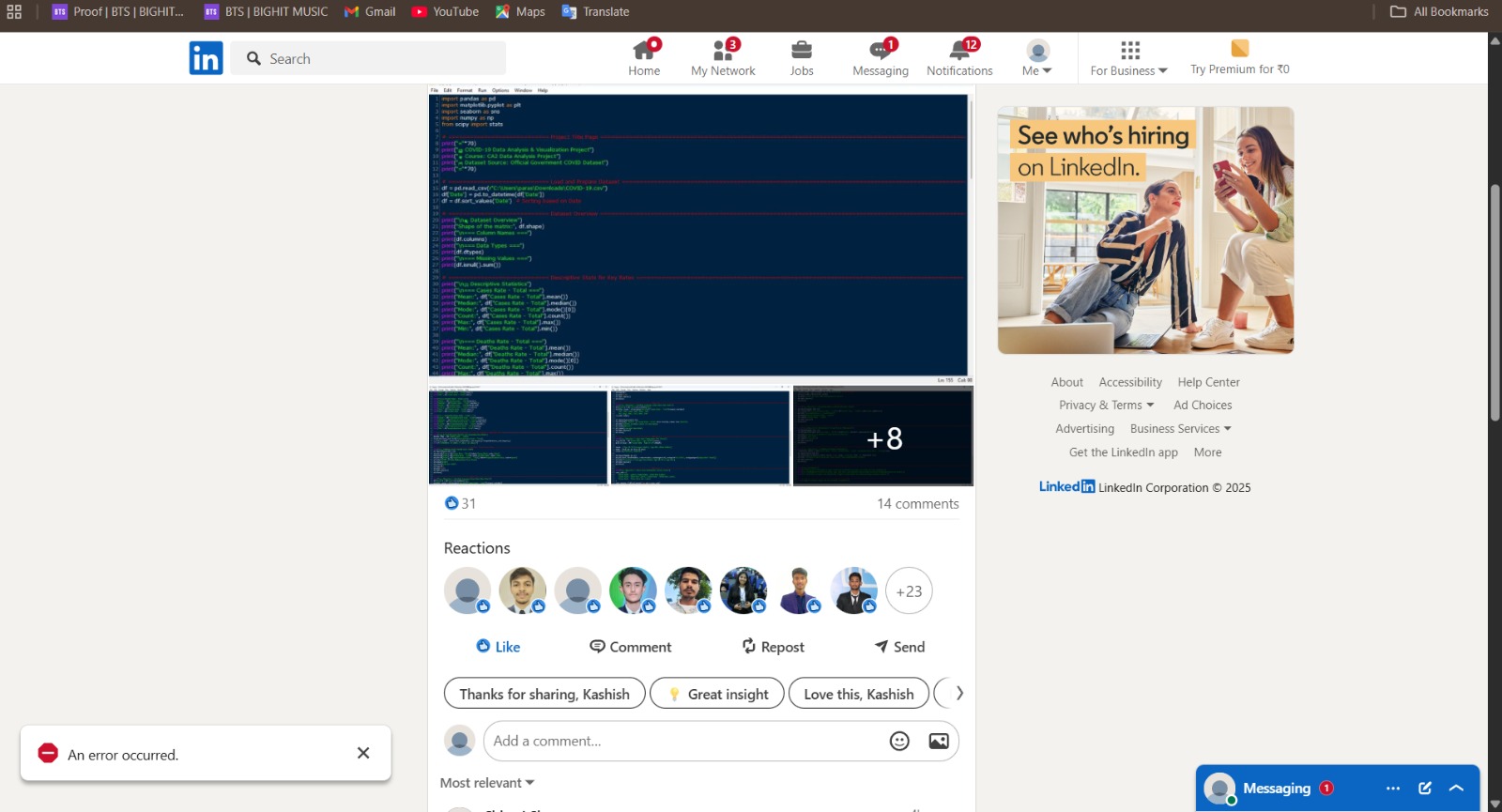
1. Real-time Monitoring:
   * Integrating real-time data streams from health organizations could enable live tracking of COVID-19 metrics and timely interventions.
2. Predictive Analytics:
   * Implementing machine learning models (e.g., ARIMA, LSTM) could forecast future trends in cases, hospitalizations, and deaths.
3. Geospatial Analysis:
   * Incorporating GIS tools would provide detailed regional insights, helping to identify hotspots and trends.
4. Integration with Public Health Data:
   * Analyzing the relationship between COVID-19 data and healthcare metrics (e.g., ICU capacity) would guide more effective interventions.
5. Climate Impact on Health:
   * Investigating how environmental factors, like air quality, affect COVID-19 spread could lead to new public health guidelines.
6. Advanced Data Cleaning:
   * Enhancing data cleaning techniques, such as imputation or anomaly detection, could improve analysis accuracy.
7. Interactive Dashboard:
   * Developing an interactive dashboard would allow users to explore and visualize COVID-19 data in real-time.

These enhancements would expand the project’s scope, offering valuable insights for public health, data-driven decision-making, and pandemic management.

**REFERENCES**

* U.S. Government - COVID-19 Daily Rolling Average Case and Death Rates
  + A comprehensive dataset that provides daily rolling averages of COVID-19 case rates, death rates, and hospitalizations in the United States. This serves as the core data for the analysis and provides insights into the evolution of the pandemic**.**
  + [**COVID-19 Daily Rolling Average Case and Death Rates**](https://catalog.data.gov/dataset/covid-19-daily-rolling-average-case-and-death-rates)
* World Health Organization (WHO) - COVID-19 Data and Health
* The WHO provides data and reports on COVID-19 and its effects on public health, which can provide useful context for interpreting your COVID-19 case and death rate data.
* [WHO COVID-19 Health Information](https://www.who.int/emergencies/diseases/novel-coronavirus-2019)
* LINKED IN LINK : <https://www.linkedin.com/posts/kashish-sukhija-9264172b2_python-datascience-covid19-activity-7316811164082978816-EkE4?utm_medium=ios_app&rcm=ACoAAEsrFW4BoDOL5cf2C2CUgflss6yOkJZnaoQ&utm_source=social_share_send&utm_campaign=whatsapp>
* GITHUB LINK : <https://github.com/kashishsukhija/EDA-on-Covid-19>

**HERE IS MY LINKDIN AND GITHUB ENGAGEMENT –**

****

