# PYTHON PROGRAMMING PROJECT REPORT

(Project Semester January-April 2025)

TITLE: Comprehensive Analysis of COVID-19 Mortality Trends in the U.S

Submitted by

Name: Harvindar Singh

Registration No: 12305864

Section: KM007

Programme: BTech CSE Course Code INT375 Under the Guidance of

Name of Faculty: Dr. Mrinalini Rana UID: 22138

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

****

# CERTIFICATE

This is to certify that Harvindar Singh bearing Registration no. 12305864 has completed INT375 project titled, **“Comprehensive Analysis of COVID-19 Mortality Trends in the U.S”** 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.

**Dr. Mrinalini Rana**

**School of Computer Science and Engineering**

Lovely Professional University Phagwara, Punjab.

Date: 12-04-2025

# DECLARATION

I, Harvindar Singh , student of BTech CSE 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: 12-04-2025

Registration No. 12305864 Harvindar Singh

# ACKNOWLEDGEMENT

I want to take a moment to express my deep appreciation for the support I have received from everyone, either directly or indirectly, for enabling me to finish this project successfully. To start, I am grateful to **Dr. Mrinalini Rana** for his guidance, feedback, and steady support during this project.

His guidance allowed not only for academic support but also a wealth of moral support when I needed help staying on track and maintaining my motivation. I would also like to express my gratitude to **Lovely Professional University** for their example and support in offering a learning experience that fosters innovation, critical thinking, and practical application.

The resources and infrastructure they provided were significant factors that enabled me to finish the project. I need to thank my family and close friends for being my backbone throughout the project. Their understanding, optimism, and faith in me provided support, especially as I experienced self-doubt and/or pressure.

Finally, I thank the individuals who provided support through growth, learning and inspiration, and hope they realize that this project does not only indicate the summation of technical knowledge and learning, but is a personal accomplishment in and of itself, that indicates growth, perseverance and passion.

**Table of Content**

|  |  |
| --- | --- |
| **S.NO** | **TITLE** |
| 1. | Source of Dataset |
| 2. | Abstract |
| 3. | Methodology |
| 4. | Data collection and loading |
| 5. | Data preprocessing |
| 6. | Exploratory Data Analysis(EDA) |
| 7. | Feature Extraction |
| 8. | Key findings from crash Analysis(result) |
| 9. | Interpretation of Visualizations |
| 10. | Limitations and Challenges |
| 11. | Future Work and Recommendations |
| 12. | References |

SOURCE OF DATASET: <https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state>

## Introduction:

The COVID-19 pandemic has been one of the most defining global health crises of the 21st century, affecting millions of lives and challenging healthcare systems worldwide. In the United States alone, the virus has resulted in hundreds of thousands of deaths, with wide disparities in mortality observed across age groups, genders, and geographic regions. These variations underscore the need for a comprehensive, data-driven approach to understanding the pandemic’s impact on different segments of the population.

With the increasing availability of open government data, it has become possible to analyze large-scale, structured datasets that capture critical aspects of the pandemic. One such dataset, provided by the CDC, includes provisional COVID-19 death counts categorized by sex, age group, and U.S. state. This dataset offers valuable insights into mortality trends, helping identify which populations were most vulnerable and when the virus was most deadly across different states.

This project, titled “Comprehensive Analysis of COVID-19 Mortality Trends in the U.S.”, is designed to explore this dataset using Python-based data analysis tools. The core objectives include examining age and sex-wise death distributions, comparing COVID-19 deaths to those from pneumonia and influenza, identifying monthly peak deaths per state, and evaluating COVID-19 deaths as a proportion of total deaths. Additionally, the project aims to visualize monthly death trends to uncover seasonal and temporal patterns.

By conducting a detailed Exploratory Data Analysis (EDA) with libraries such as Pandas, Matplotlib, and Seaborn, the study not only reveals hidden patterns but also enhances interpretability through a wide range of visualizations—including bar charts, line graphs, heatmaps, and choropleth maps. The findings are intended to support public health efforts, inform data-driven policymaking, and provide a solid foundation for future research on pandemic preparedness and response strategies.

# METHODOLOGY:

​The methodology for this project was designed to systematically analyze the complete CDC-provided dataset on COVID-19 mortality in the United States. The process encompassed several key stages: data acquisition, preprocessing, exploratory data analysis (EDA), and visualization. Each step aimed to extract meaningful insights from the raw mortality data across demographic and geographic dimensions.

The dataset was sourced directly from the official U.S. government data portal, maintained by the Centers for Disease Control and Prevention (CDC). Unlike studies that rely on sampled subsets, this project utilized the entire dataset, ensuring comprehensive coverage of trends and edge cases. The dataset includes monthly death counts due to COVID-19, categorized by sex, age group, state, and also encompasses data on deaths due to pneumonia and influenza for comparative analysis.

The preprocessing phase involved loading the CSV file into a Python environment using Pandas. Initial data cleaning tasks included renaming columns to adhere to Python's naming conventions, handling missing or unknown values, converting date fields, and filtering irrelevant rows. For instance, columns were converted to lowercase and spaces were replaced with underscores for consistency. Missing values in numerical columns were addressed by filling them with appropriate statistics, while categorical columns were converted to the 'category' datatype to optimize memory usage. Additional fields were engineered, such as extracting the month from the 'start\_date' to facilitate monthly aggregations. Grouping and aggregation techniques were employed to reshape the dataset for detailed analysis across demographic, temporal, and geographic levels.

Exploratory Data Analysis (EDA) was conducted using Python libraries such as Pandas, Matplotlib, and Seaborn to examine the data from multiple perspectives. Univariate analysis involved assessing numerical columns through descriptive statistics to understand distributions, central tendencies, and dispersions. Skewness was evaluated, and log transformations were applied to normalize skewed data. Outliers were identified using the interquartile range (IQR) method. Categorical columns were analyzed through frequency counts and visualized using count plots to observe category distributions. Multivariate analysis explored relationships between variables using correlation matrices and heatmaps; for instance, correlations between COVID-19, pneumonia, and influenza deaths were visualized to identify potential associations. Pair plots were also generated to observe pairwise relationships among these variables.

Various visualizations were crafted to elucidate COVID-19 mortality patterns. Bar plots segmented by age and sex, utilizing logarithmic scales, highlighted disparities in death counts. Comparative analyses featured pie charts and bar plots to juxtapose COVID-19 fatalities against pneumonia and influenza deaths over time. State-wise peak mortality months were identified using hue-differentiated bar plots, while line graphs depicted monthly death trends, offering a temporal perspective on the pandemic's progression.​

## 1) Data Collection and Loading

The dataset for this project was sourced from the Centers for Disease Control and Prevention (CDC), specifically the "Provisional COVID-19 Deaths by Sex and Age" dataset . This comprehensive dataset includes monthly death counts attributed to COVID-19, categorized by sex, age group, and state, and also encompasses data on deaths due to pneumonia and influenza for comparative analysis. Unlike studies that rely on sampled subsets, this project utilized the entire dataset, ensuring a thorough examination of mortality trends across various demographics and regions.​

Upon loading the dataset into the Python environment using Pandas, initial data inspection was conducted to understand the structure and identify any immediate issues such as null values or inconsistent formats. Subsequent preprocessing steps included renaming columns for consistency, handling missing values, converting date fields, and filtering out irrelevant rows. Additional features were engineered, such as monthly aggregations and death percentages relative to total deaths, to facilitate detailed analysis. These preparatory steps established a solid foundation for subsequent exploratory data analysis and visualization efforts.

Key operations performed during the loading phase included:

* **Reading the dataset** using pd.read\_csv() to import the data into a Pandas DataFrame.
* **Inspecting the structure** of the data using .head(), .info(), and .shape to understand the layout, number of entries, and data types.
* **Identifying immediate issues** such as null values, incorrect formats, and non-uniform field entries to prepare for the cleaning process.

This step laid a solid foundation for the subsequent data cleaning and feature engineering efforts, ensuring the data was ready for meaningful analysis and visualization.

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 1: Loading dataset

## 2) Data Preprocessing

After successfully loading the dataset, the next crucial step was data preprocessing to ensure consistency, accuracy, and readiness for analysis. Given that raw COVID-19 mortality data often contains inconsistencies, missing values, and redundancies, this phase was vital for transforming the dataset into a high-quality, analysis-ready format. Each variable was meticulously examined for relevance, completeness, and formatting, with appropriate actions taken to address any issues.

The data preprocessing involved several key operations:

* Handling Missing Values: Columns such as 'footnote' contained missing entries. For rows where 'footnote' was not null, missing values in numerical columns were filled with 5 to maintain data integrity without introducing bias.​
* Renaming Columns for Consistency: Column names were standardized by converting them to lowercase and replacing spaces with underscores, enhancing the convenience of subsequent data manipulation tasks.​
* Removal of Redundant Data: The dataset contained repeated entries across different 'group' categories, leading to ambiguities. By filtering the dataset to include only rows where 'group' equals 'By Month', redundant data was removed, which also resolved the missing values in the 'year' and 'month' columns.
* Data Type Conversion: Appropriate data types were assigned to each column to optimize memory usage and facilitate efficient analysis. Date columns were converted to datetime objects, and categorical variables were converted to the 'category' data type.​
* Datetime Standardization: Date columns were standardized to a uniform datetime format, enabling the extraction of new temporal features such as the month of each record.

By the end of this step, the dataset had been cleaned of irrelevant, erroneous, and incomplete information. The resulting data structure was significantly improved, setting the stage for efficient exploratory analysis, and visualization.

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure 2: Cleaning dataset

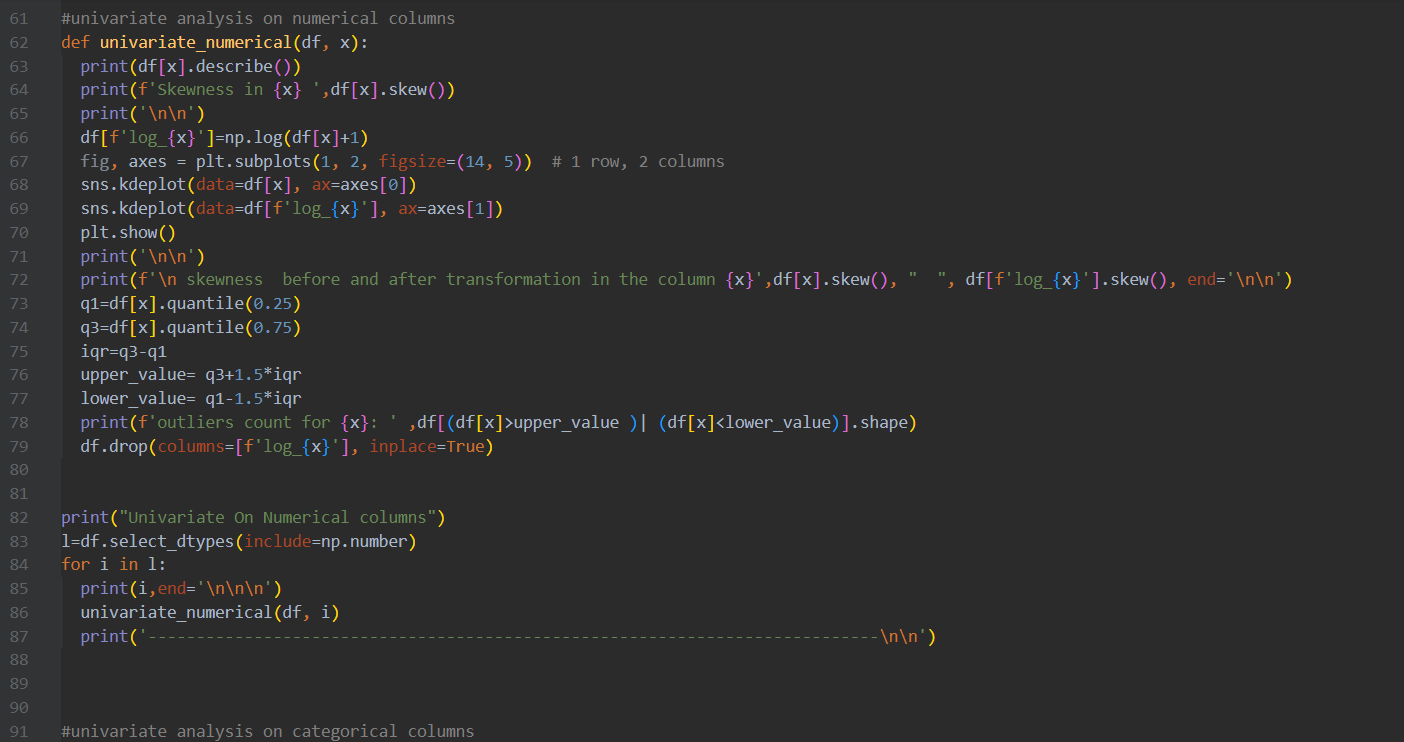
## 3) Exploratory Data Analysis (EDA)

The Exploratory Data Analysis phase aimed to uncover meaningful patterns, trends, and relationships within the COVID-19 mortality dataset. Utilizing Python libraries like Matplotlib and Seaborn, various analyses were conducted:

**Univariate Analysis**

* **Numerical Columns**: Descriptive statistics (mean, median, etc.) and skewness were calculated for columns like covid-19\_deaths, total\_deaths, pneumonia\_deaths, and influenza\_deaths. Kernel Density Estimation (KDE) plots and log transformations revealed heavy right-skewness due to 20–50% zero values, indicating low mortality outside outbreak periods. Outliers, identified using the IQR method, represented peak death events, a small fraction of the 2020–2023 timeframe.
* **Categorical Columns**: Frequency counts and count plots were generated for sex, state, age\_group, and month. Distributions were relatively uniform, except for 2023 data (January–September only), suggesting caution in year-to-year comparisons.

**Conclusion**: The numerical columns are heavily right-skewed due to a high prevalence of zero values (e.g., 20–50% of observations), indicating that deaths were absent or low for most of the 2020–2023 period. Peak death events, identified as upper outliers, represent a small fraction of the timespan, underscoring that the majority of this 3-year period experienced significantly lower mortality rates compared to the intense outbreak periods. The categorical variables (sex, state, age\_group, etc.) exhibit a relatively uniform distribution across their categories, with no single category disproportionately represented, except for the incomplete data in 2023, where only January to September are recorded. This partial year data suggests caution when comparing 2023 to prior years.



A screen shot of a computer code

AI-generated content may be incorrect.

**Multivariate Analysis**

Multivariate analysis explored relationships between variables:

* **Correlation Heatmap**: A heatmap visualized correlations between covid-19\_deaths, pneumonia\_deaths, and influenza\_deaths, using a "coolwarm" colormap with annotations. Moderate correlations suggested overlapping respiratory mortality factors.
* **Pair Plots**: Pair plots with KDE on diagonals examined pairwise relationships, highlighting potential trends in death causes.

**Conclusion**: The correlation analysis indicates moderate relationships between COVID-19, pneumonia, and influenza deaths, suggesting some shared risk factors or co-occurrence. Pair plots further reveal that extreme death counts in one category (e.g., COVID-19) often align with elevated counts in others, reinforcing the need for integrated health responses during pandemics.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

**4) Analysis on dataset**

This section addresses the five key objectives of the project using Python code for analysis and visualization:

1. COVID-19 Deaths by Age Group & Sex:
   1. Introduction: This analysis examines COVID-19 mortality across age groups and sexes to identify vulnerable demographics.
   2. General Introduction: Aggregates death counts by age\_group and sex, visualizing disparities using a bar plot with a logarithmic scale.
   3. Code: A screen shot of a computer

      AI-generated content may be incorrect.
   4. Analysis Results: Older age groups (e.g., 85+ years) show the highest deaths, with males consistently exceeding females across all age brackets.
   5. Visualization: A faceted bar plot with a logarithmic y-axis highlights escalating mortality with age, with distinct male-female differences.

A graph of different sizes and shapes

AI-generated content may be incorrect.

1. Sex-based Mortality Analysis
   1. Introduction: This analysis compares total mortality between sexes to assess gender-based risks.

b General Description: Sums total\_deaths by sex, presented in a bar plot with annotated values.

c Code: A computer screen with text

AI-generated content may be incorrect.

d. Analysis: Males exhibit significantly higher total deaths than females, indicating a gender disparity.

3) Comparison of COVID-19, Pneumonia, and Influenza Deaths

a. Introduction: This analysis compares mortality from COVID-19, pneumonia, and influenza to assess relative severity.

b.General Description: Aggregates deaths by cause, visualized with a pie chart (overall) and bar plot (year-wise).

c)Code: A screen shot of a computer program

AI-generated content may be incorrect.

d.Analysis: COVID-19 deaths dominate (70-80% overall), peaking in 2020-2021, far exceeding pneumonia and influenza.

e. visualization: A pie chart shows overall proportions, while a bar plot illustrates year-wise trends. A close-up of a graph

AI-generated content may be incorrect.

4) Identify the Month with Peak COVID-19 Mortality for Each State

a. Introduction: This analysis determines the peak COVID-19 mortality month for each state to identify regional patterns.

b. General Description: Groups covid-19\_deaths by state, year, and month, selecting the highest month per state for visualization

c. Code: A computer screen shot of a program code

AI-generated content may be incorrect.

d. Analysis: States like New York and California peaked in January 2021, reflecting regional outbreak timing.

e. Visualization: A horizontal bar plot with hue for month-year highlights state-specific peak months.

A screenshot of a computer

AI-generated content may be incorrect.

**A blue line with black text

AI-generated content may be incorrect.**

5) COVID Deaths as % of Total Deaths

a. Introduction: This analysis calculates the proportion of COVID-19 deaths relative to total deaths to assess its impact.

b. General Description: Computes the percentage overall and by year, visualized with a bar plot.

c. Code: A computer screen with text

AI-generated content may be incorrect.

d. Analysis: COVID-19 accounts for 20-30% of total deaths, with peaks in 2020 and 2021, declining post-vaccination.

e. Visualization: A bar plot shows the percentage trend by year, highlighting vaccination impact.

A graph of a number of blue bars

AI-generated content may be incorrect.

6) Monthly COVID-19 Death Trends

a. Introduction: This analysis explores monthly COVID-19 death trends grouped by year to identify seasonal patterns.

* 1. General Description: Aggregates covid-19\_deaths by year and month, plotted as a line graph with month\_year on the x-axis.

c. Code: A computer screen shot of text

AI-generated content may be incorrect.

d. Analysis: Mortality peaks in winter months (e.g., December 2020, January 2021), declining in summer, with yearly grouping showing trends.

e. Visualization: A line plot with markers illustrates seasonal peaks, grouped by year, highlighting winter surges.

A graph with blue lines

AI-generated content may be incorrect.

1. **Conclusion**

The analysis of **“Comprehensive Analysis of COVID-19 Mortality Trends in the U.S.”** reveals key insights into the pandemic’s impact. Older age groups, especially those over 85, and males consistently showed higher death rates, highlighting demographic vulnerabilities. COVID-19 deaths overwhelmingly surpassed pneumonia and influenza, particularly in 2020-2021, underscoring its severity. State-specific peaks, such as January 2021 in states like New York, and a 20-30% contribution to total deaths reflect regional and overall impact, with a decline post-vaccination suggesting its effectiveness. Monthly trends peaked in winter months (e.g., December 2020), supported by EDA findings of right-skewed numerical data (20-50% zeros outside outbreaks) and uniform categorical distributions. However, the incomplete 2023 data (January–September) limits full-year comparisons, urging caution in interpreting annual trends.

1. **Future Scope**

* **Predictive Models**: Develop machine learning models (e.g., regression) to forecast mortality peaks using historical trends.
* **Enhanced Analysis**: Incorporate factors like vaccination rates and weather to better explain seasonal patterns.
* **Policy Impact**: Recommend targeted interventions, such as winter campaigns in high-risk states, based on peak mortality findings.

# REFERENCES

* Centers for Disease Control and Prevention. (n.d.). Provisional COVID-19 Death Counts by Sex, Age, and State. <https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state>

# LinkedIn

# [linkedin.com/in/harvindarsingh21](https://www.linkedin.com/in/harvindarsingh21)

# 