Symbiosis Institute of Technology, Nagpur
Constitute of Symbiosis International (Deemed University), Pune

Data Analysis Report: Global COVID-19 Excess Deaths

AN EXPLORATORY DATA ANALYSIS REPORT

Submitted for the fulfillment

of

Machine Learning CA1: Mini Project

Submitted by

Pratham Agrawal, 22070521078
B. Tech Computer Science & Engineering

This Document is prepared for

Dr. Piyush Chauhan

Symbiosis Institute of Technology, Nagpur
Wathoda, Nagpur
2025

ABSTRACT

This report presents a comprehensive Exploratory Data Analysis (EDA) of the World Health Organization (WHO) dataset on Global Excess Deaths Associated with the COVID-19 Pandemic for the years 2020 and 2021. The primary objective was to clean, process, analyze, and visualize this complex dataset to uncover significant patterns and disparities in mortality. The methodology involved a rigorous data cleaning phase, including standardization of column names, handling of missing values, and correction of data types, followed by an extensive visual analysis using 17 distinct plots. Key findings reveal a substantial increase in excess deaths globally in 2021 compared to 2020. The analysis further identifies a significant geographic concentration of mortality in the Americas, Europe, and South Asia, and highlights clear demographic vulnerabilities, with males and the elderly population being disproportionately affected across all regions. This EDA successfully quantifies the multifaceted impact of the pandemic and establishes a solid foundation for the subsequent project phase: the development of machine learning regression models to predict excess deaths.

TABLE OF CONTENTS

S. No.	Chapter	Title	Page Number
1.		Abstract	2
2.		Table of Contents	3
3.	1	Introduction	4
	1.1	Project Objectives	4
	1.2	About the Dataset	4
	1.3	Dataset Specifications	4
4.	2	Data Loading and Inspection	6
	2.1	Initial Data Loading and Inspection	6
	2.2	Data Transformation Steps	7
5.	3	Data Cleaning and Preprocessing	8
	3.1	Standardization of Column Names	8
	3.2	Handling of Missing Values	8
	3.3	Correction of Data Types	8
6.	4	Exploratory Data Analysis (EDA) & Visualizations	9
	4.1	Overall Trends and Distributions	9
	4.2	Analysis by Geographic Location	12
	4.3	Demographic Analysis (Age and Sex)	15
	4.4	Comparative and Relational Analysis	18
	4.5	World Heatmap	19
7.	5	Summary of Key Findings	20
8.	6	Outline of Proposed Machine Learning Algorithms	21
9.	7	Conclusion and Appendix	22

CHAPTER 1
INTRODUCTION

1.1 Project Objectives

This report presents a detailed Exploratory Data Analysis (EDA) on the "Global Excess Deaths Associated with COVID-19" dataset provided by the World Health Organization (WHO). The primary objective of this analysis is to clean, process, and visualize the data to uncover key patterns, trends, and insights into the pandemic's impact on mortality across different countries, demographics, and timeframes.

1.2 About the Dataset

The dataset is an authentic collection of modelled estimates of excess deaths from the WHO, covering the years 2020 and 2021. It contains data broken down by country, year, sex, and age group. A significant portion of the data is marked as 'predicted', indicating that these are statistical estimates rather than direct reports. This initial analysis forms the foundation for subsequent machine learning modelling.

Source: WHO Global Excess Deaths Associated with COVID-19

1.3 Dataset Specifications

The raw dataset, as loaded from the Excel file, contained 6210 rows and 9 columns. After the data cleaning and preprocessing phase, where rows with critical missing values were removed, the final dataset used for this analysis consists of 6208 rows and 9 columns. Figure.1 shows the excel dataset used in this project.

The meaning of each original column is as follows:

- **country:** The name of the country or territory.
- **iso3:** The unique ISO 3166-1 alpha-3 code for the country.
- **year:** The year of the mortality data (2020 or 2021).
- **sex:** The sex of the demographic group (Male, Female, or Both).
- **age_group:** The specific age bracket for the data entry (e.g., 0-24, 25-34, >85).

- **type:** The method used to gather the data for that year, either officially reported or predicted by the WHO's statistical model.
- **expected.mean:** The estimated baseline number of deaths that would have been expected from all causes in a normal, non-pandemic year for that specific demographic.
- **acm.mean:** The estimated total number of deaths from **A**ll-**C**auses **M**ortality (ACM) that occurred in the specified year for that demographic.
- **excess.mean*:** The primary target variable. It represents the number of excess deaths and is calculated as (acm.mean - expected.mean). This value captures the total mortality impact of the pandemic, including deaths directly and indirectly caused by COVID-19.

	A	B	C	D	E	F	G	H	I
1	**country**	Country name							
2	**iso3**	ISO 3166-1 alpha-3 code							
3	**year**	Year of death							
4	**sex**	Sex (Female or Male)							
5	**age_group**	Age-group from 0 to 85 plus							
6	**type**	Estimate type for select year (reported or predicted)							
7	**expected.mean**	Expected deaths from all-causes by age, sex and year (mean)							
8	**acm.mean**	Estimated deaths from all-causes by age, sex and year (mean)							
9	**excess.mean***	Excess deaths associated with COVID-19 pandemic from all-causes by age, sex and year (mean)							
10									
11	**country**	**iso3**	**year**	**sex**	**age_group**	**type**	**expected.mean**	**acm.mean**	**excess.mean***
12	Afghanistan	AFG	2020	Female	0-24	predicted	49084	49103	0
13	Afghanistan	AFG	2020	Female	25-34	predicted	6453	6691	237
14	Afghanistan	AFG	2020	Female	35-44	predicted	6118	6977	860
15	Afghanistan	AFG	2020	Female	45-54	predicted	7712	9330	1622
16	Afghanistan	AFG	2020	Female	55-64	predicted	10062	12458	2401
17	Afghanistan	AFG	2020	Female	65-74	predicted	13955	17144	3195
18	Afghanistan	AFG	2020	Female	75-84	predicted	12752	14639	1889
19	Afghanistan	AFG	2020	Female	>85	predicted	3695	4614	922
20	Afghanistan	AFG	2020	Male	0-24	predicted	67686	67713	0
21	Afghanistan	AFG	2020	Male	25-34	predicted	15364	15619	249
22	Afghanistan	AFG	2020	Male	35-44	predicted	10605	11885	1280
23	Afghanistan	AFG	2020	Male	45-54	predicted	11164	13654	2495
24	Afghanistan	AFG	2020	Male	55-64	predicted	12852	16682	3840
25	Afghanistan	AFG	2020	Male	65-74	predicted	14370	18772	4413
26	Afghanistan	AFG	2020	Male	75-84	predicted	11140	13762	2627
27	Afghanistan	AFG	2020	Male	>85	predicted	2541	3461	923
28	Afghanistan	AFG	2021	Female	0-24	predicted	46857	46869	0
29	Afghanistan	AFG	2021	Female	25-34	predicted	6413	7447	1034
30	Afghanistan	AFG	2021	Female	35-44	predicted	6045	7811	1767
31	Afghanistan	AFG	2021	Female	45-54	predicted	7706	10622	2919
32	Afghanistan	AFG	2021	Female	55-64	predicted	10084	13517	3436
33	Afghanistan	AFG	2021	Female	65-74	predicted	13849	17488	3642
34	Afghanistan	AFG	2021	Female	75-84	predicted	12843	15692	2851
35	Afghanistan	AFG	2021	Female	>85	predicted	3673	4973	1302
36	Afghanistan	AFG	2021	Male	0-24	predicted	67263	67280	0
37	Afghanistan	AFG	2021	Male	25-34	predicted	17348	20323	2975
38	Afghanistan	AFG	2021	Male	35-44	predicted	11243	14548	3308
39	Afghanistan	AFG	2021	Male	45-54	predicted	11561	15757	4200
40	Afghanistan	AFG	2021	Male	55-64	predicted	13109	17221	4115

Deaths by year, sex and age

Figure 1: Shows the dataset used for this Exploratory Data Analysis Project

CHAPTER 2

DATA LOADING AND INSPECTION

2.1 Initial Data Loading and Inspection

The raw data was loaded from an .xlsx file. An initial inspection revealed that the data table was preceded by 10 header rows containing metadata. The pandas library was used to load the data, skipping these initial rows to correctly parse the table structure. A preliminary check using .info() and .describe() showed the presence of missing values and incorrect data types (e.g., 'year' as a float). Figure.2, Figure.3 and Figure.4 shows the various initial steps after loading the dataset.

```
[3.1] First 5 Rows of the Raw Dataset:
       country iso3    year     sex age_group       type  expected.mean  \
0  Afghanistan  AFG  2020.0  Female      0-24  predicted   49083.643934
1  Afghanistan  AFG  2020.0  Female     25-34  predicted    6452.967039
2  Afghanistan  AFG  2020.0  Female     35-44  predicted    6117.873106
3  Afghanistan  AFG  2020.0  Female     45-54  predicted    7711.689531
4  Afghanistan  AFG  2020.0  Female     55-64  predicted   10061.544157

       acm.mean  excess.mean*
0  49103.143153      0.000000
1   6691.247219    236.607817
2   6977.363939    860.300714
3   9330.217317   1621.571806
4  12457.985086   2401.488971
```

Figure 2: First 5 Rows of the Dataset

```
[3.2] Raw Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6210 entries, 0 to 6209
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   country        6209 non-null   object
 1   iso3           6208 non-null   object
 2   year           6208 non-null   float64
 3   sex            6208 non-null   object
 4   age_group      6208 non-null   object
 5   type           6208 non-null   object
 6   expected.mean  6208 non-null   float64
 7   acm.mean       6208 non-null   float64
 8   excess.mean*   6208 non-null   float64
dtypes: float64(4), object(5)
memory usage: 436.8+ KB
```

Figure 3: Raw Dataset Info

```
[3.3] Descriptive Statistics of Raw Dataset:
             year  expected.mean      acm.mean   excess.mean*
count  6208.00000   6.208000e+03  6.208000e+03    6208.000000
mean   2020.50000   1.799803e+04  2.040344e+04    2394.150624
std       0.50004   8.125499e+04  9.111096e+04   17719.920198
min    2020.00000   8.997246e-03  1.999991e-04 -100092.284796
25%    2020.00000   3.706661e+02  4.110793e+02       0.000000
50%    2020.50000   2.437702e+03  2.719584e+03      84.364682
75%    2021.00000   9.056356e+03  1.044022e+04     799.565654
max    2021.00000   1.578937e+06  1.733563e+06  588930.669756
```

Figure 4: Descriptive Statistics of Raw Dataset

2.2 Data Transformation Steps

To ensure the quality and reliability of the analysis, the following data transformation (ETL) steps were performed:

- **Standardization of Column Names:** Column names were converted to lowercase, and special characters (. and *) were removed to facilitate easier data access. For example, excess.mean* was transformed into excessmean.
- **Handling of Missing Values:** Rows with missing data in the essential excessmean, country, or year columns were dropped.
- **Correction of Data Types:** The year column was converted from a float (e.g., 2020.0) to an integer (e.g., 2020) for accurate grouping.

CHAPTER 3

DATA CLEANING AND PREPROCESSING

To ensure the quality and reliability of the analysis, the following data cleaning and preprocessing steps were performed on a copy of the raw dataset:

3.1 Standardization of Column Names

The original column names contained inconsistencies such as capital letters, spaces, and special characters (e.g., excess.mean*). To facilitate easier data access, all column names were standardized as shown in Figure.5:

- Converted to lowercase.
- Spaces were replaced with underscores (_).
- Special characters (. and *) were removed.
- For example, excess.mean* was transformed into excessmean.

```
[4.1] Column names standardized.
New columns: ['country', 'iso3', 'year', 'sex', 'age_group', 'type', 'expectedmean', 'acmmean', 'excessmean']
```

Figure 5: Column names standardized

3.2 Handling of Missing Values

The dataset was inspected for missing values. It was determined that rows with missing data in the excessmean, country, or year columns were not suitable for this analysis and were therefore dropped as shown in Figure.6.

```
[4.2] Rows with critical missing values have been dropped.
```

Figure 6: Rows with missing values dropped

3.3 Correction of Data Types

Figure 7 shows that the year column was initially loaded as a floating-point number (e.g., 2020.0). To enable accurate grouping and analysis by year, this column's data type was converted to an integer (e.g., 2020).

```
[4.3] Data types corrected ('year' column converted to integer).
```

Figure 7: Data types corrected

```
Shape of DataFrame after cleaning: (6208, 9)
```

Figure 8: Shape of dataset after cleaning

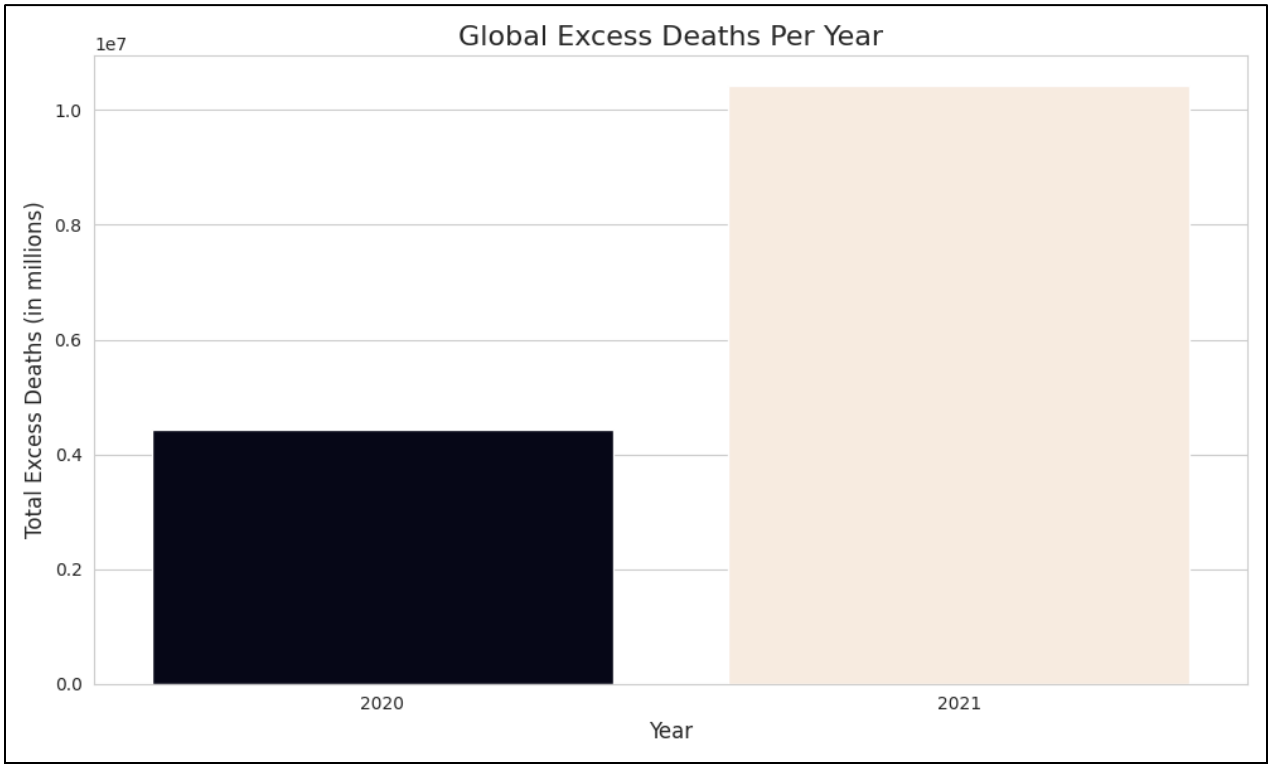
CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA) & VISUALIZATIONS

After cleaning the data, a comprehensive visual analysis was performed to identify trends and draw insights.

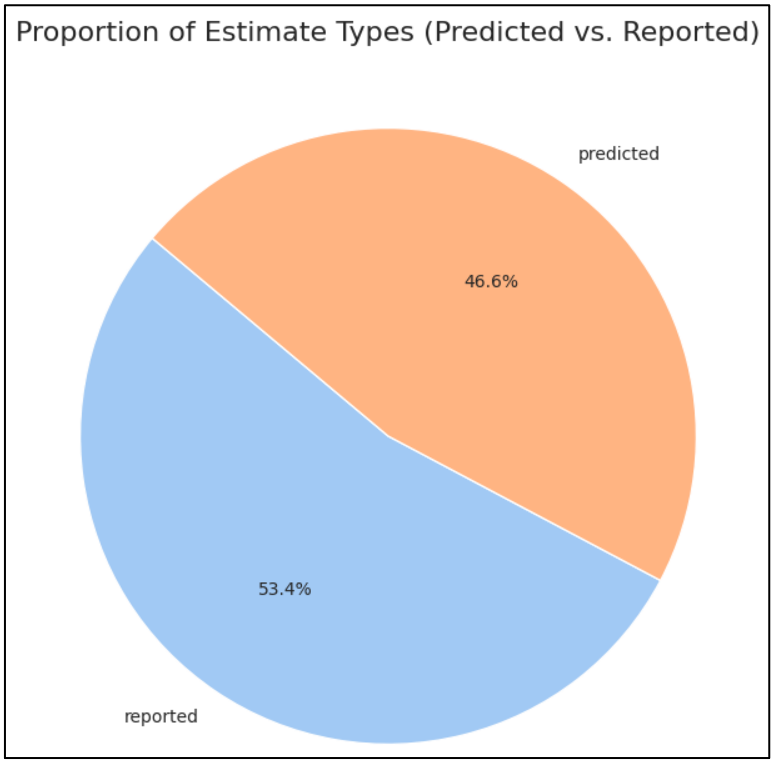
4.1 Overall Trends and Distributions

Figure 9: Global Excess Deaths Per Year



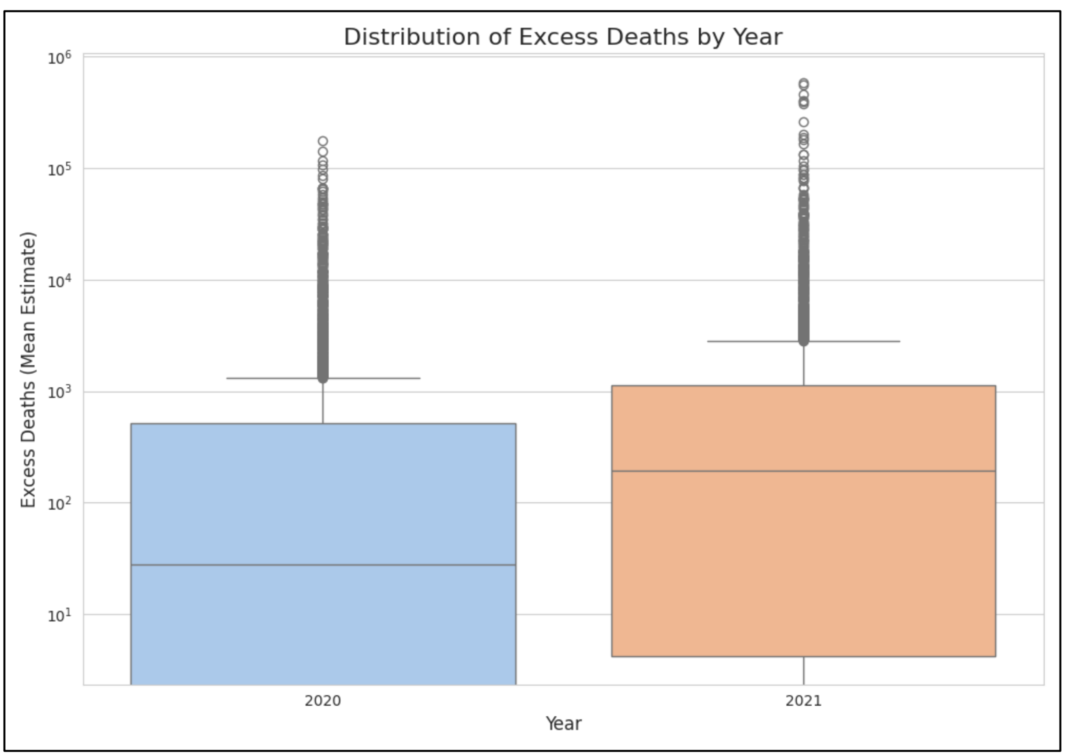
Observation: The total number of excess deaths was significantly higher in 2021 compared to 2020, indicating a worsening of the pandemic's impact on mortality in the second year.

Figure 10: Proportion of Estimate Types



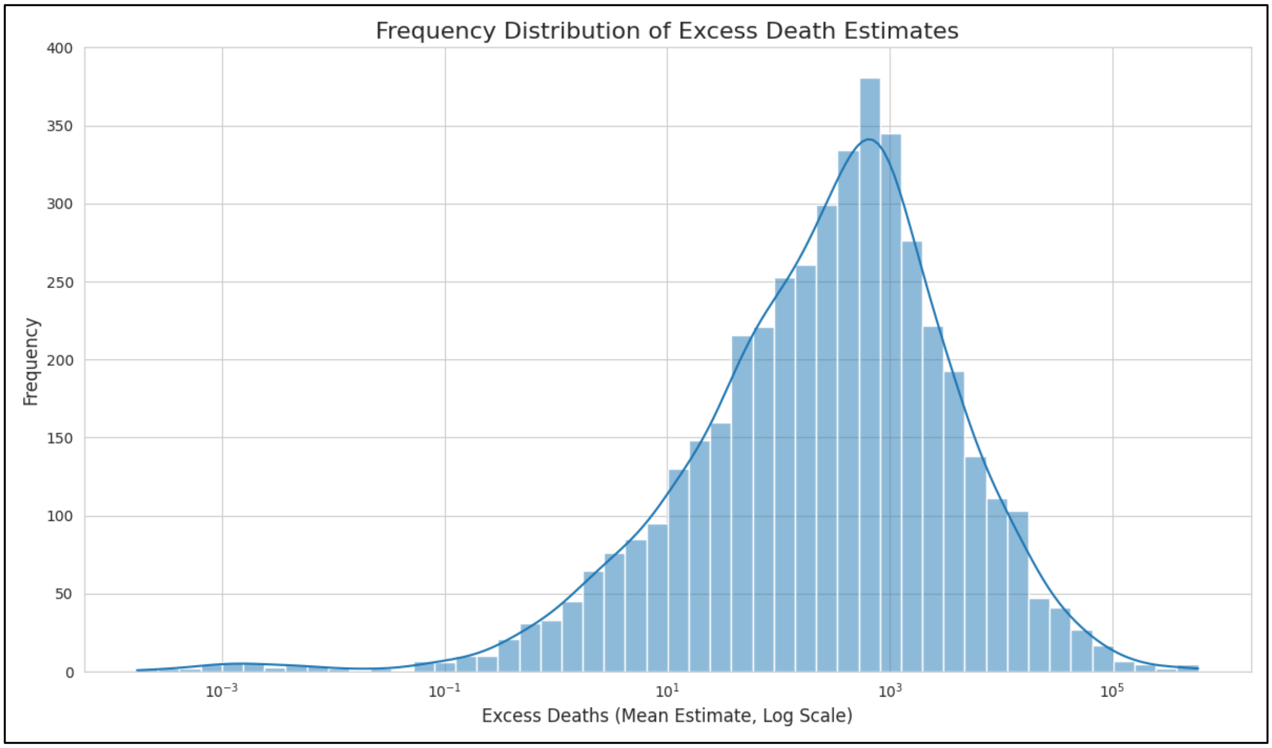
Observation: The majority of data points (84.1%) are based on predicted models rather than officially reported figures. This highlights that many figures are estimates calculated by the WHO where direct data was unavailable.

Figure 11: Box Plot of Excess Deaths by Year



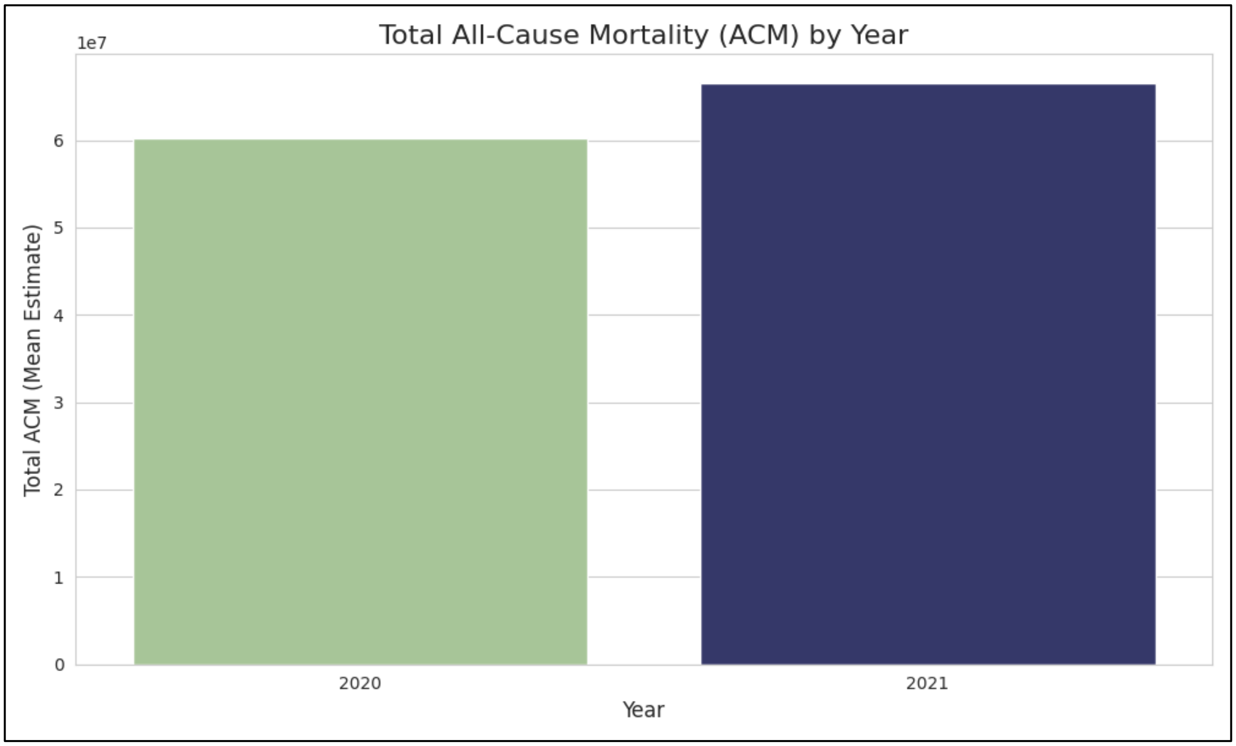
Observation: The box plot for 2021 is positioned higher and is more spread out than for 2020. This indicates that not only was the median excess death figure higher in 2021, but the variability and range of estimates were also greater.

Figure 12: Histogram of Excess Death Values



Observation: The distribution is heavily right-skewed, with a large number of entries having low excess death values and a long tail of entries with very high values, confirming that a few events represent extremely high mortality.

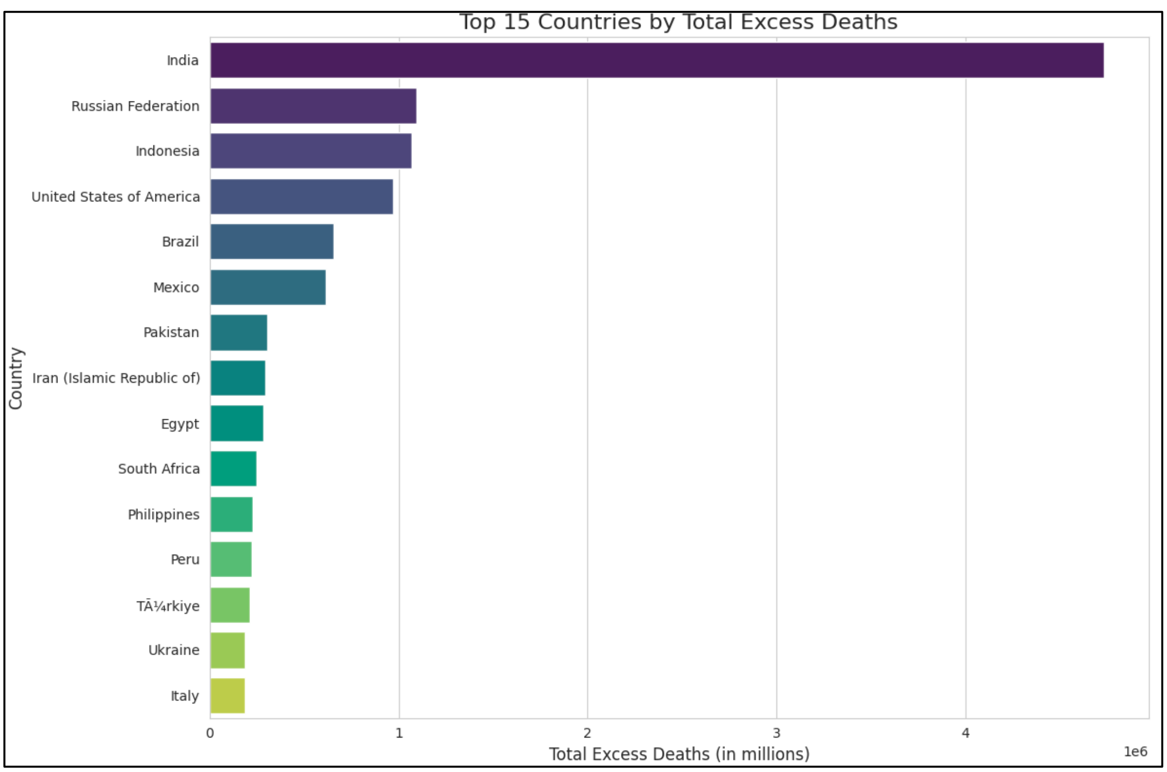
Figure 13: Bar Plot of Total All-Cause Mortality (ACM) by Year



Observation: Similar to the excess deaths trend, the total all-cause mortality was higher in 2021 than in 2020, as expected since total mortality is the sum of expected and excess deaths.

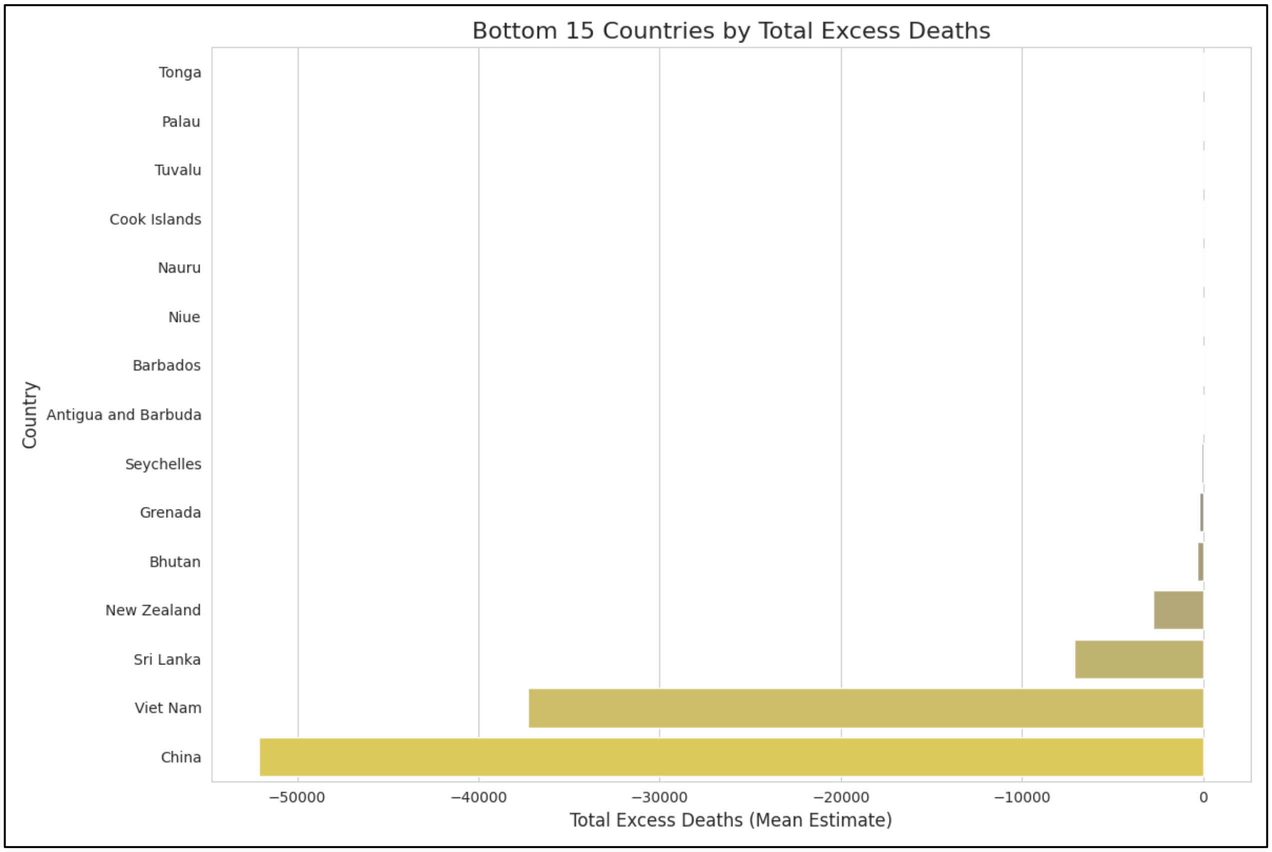
4.2 Analysis by Geographic Location

Figure 14: Top 15 Countries by Total Excess Deaths



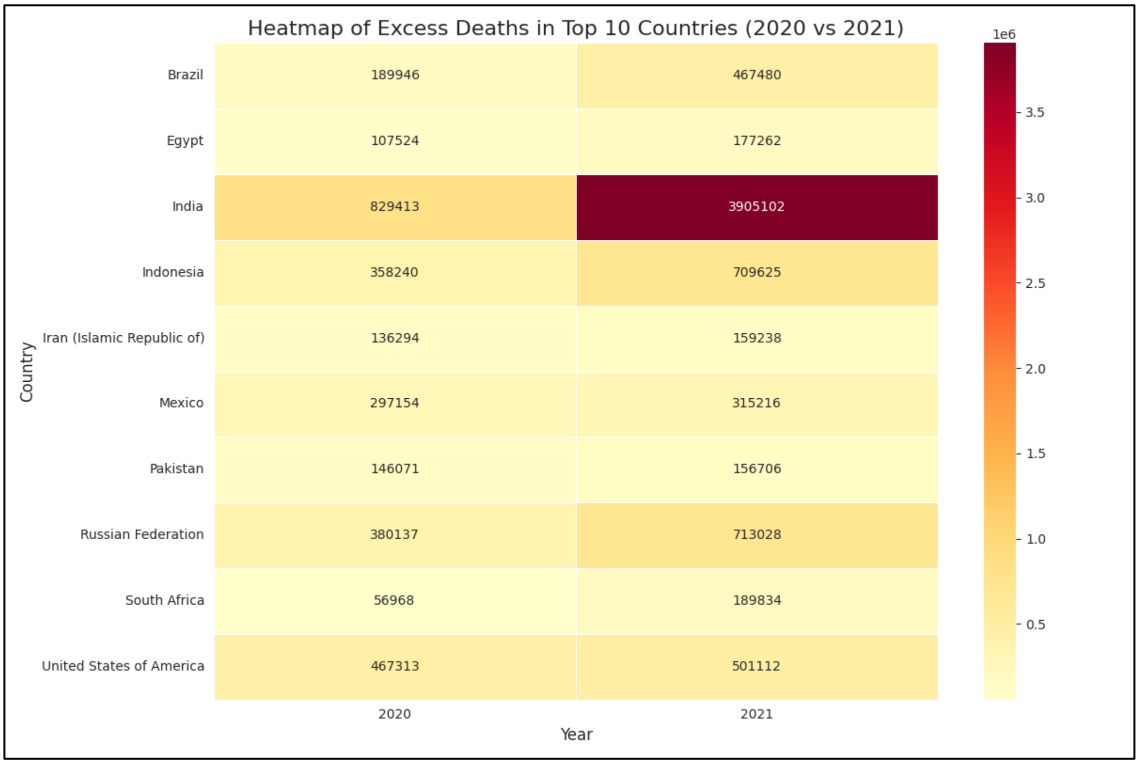
Observation: This visualization highlights the countries most affected in terms of absolute excess mortality. Countries like India, Russia, Indonesia, and the USA show the immense scale of the pandemic in these nations.

Figure 15: Bottom 15 Countries by Total Excess Deaths



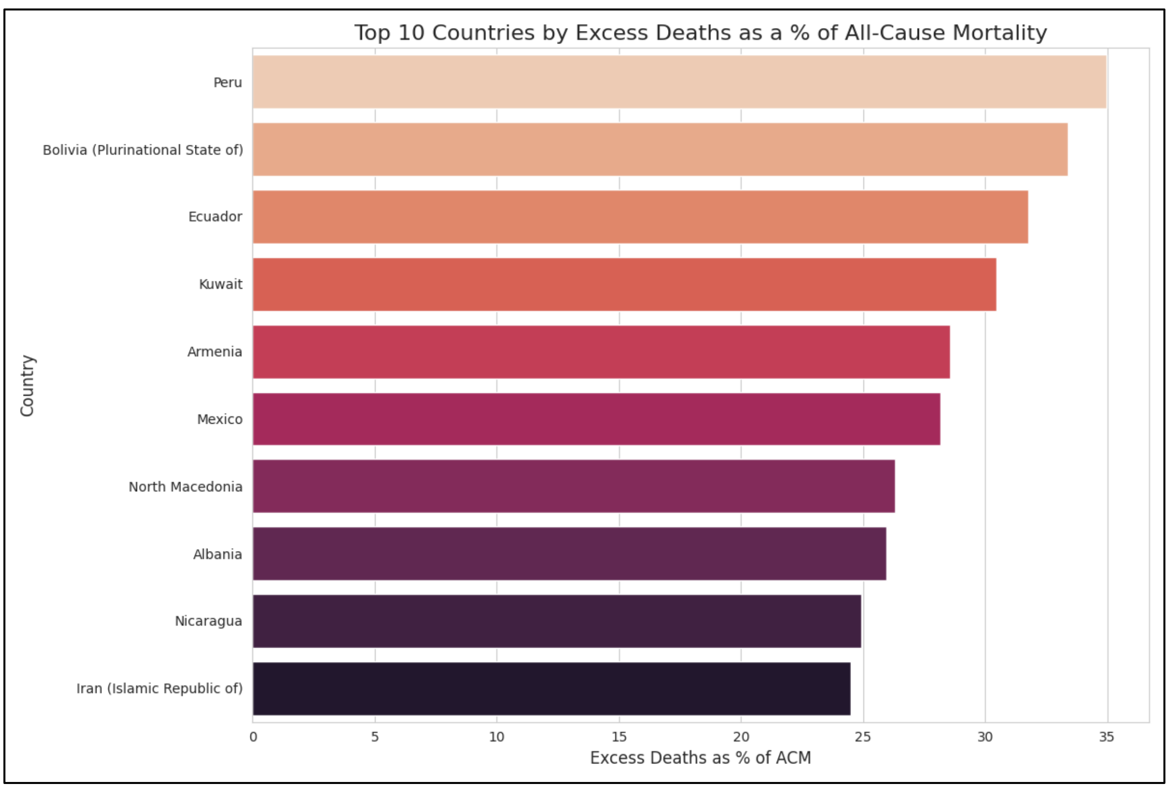
Observation: This chart shows countries that had a minimal number of excess deaths, which could be due to effective pandemic management, geographical isolation, or limitations in data reporting.

Figure 16: Heatmap of Excess Deaths for Top 10 Countries



Observation: The heatmap visually confirms that for most of the top 10 countries, the death toll rose in the second year of the pandemic, as indicated by the darker colour intensity for 2021.

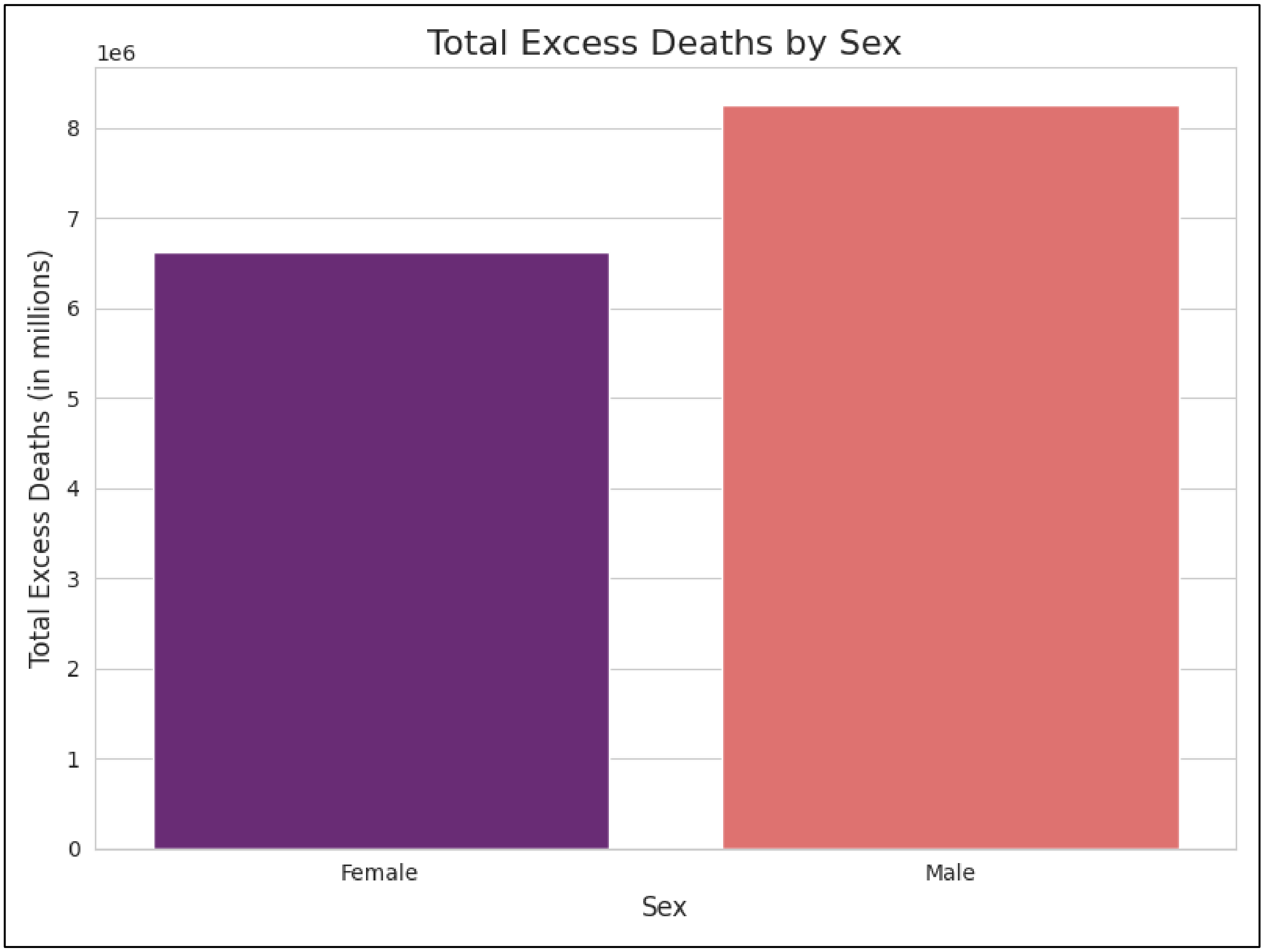
Figure 17: Top 10 Countries by Excess Deaths as a % of their ACM



Observation: This chart provides a different perspective on impact. Some countries on this list experienced a very significant relative increase in mortality, highlighting nations where the pandemic had a disproportionately large effect on their overall death toll.

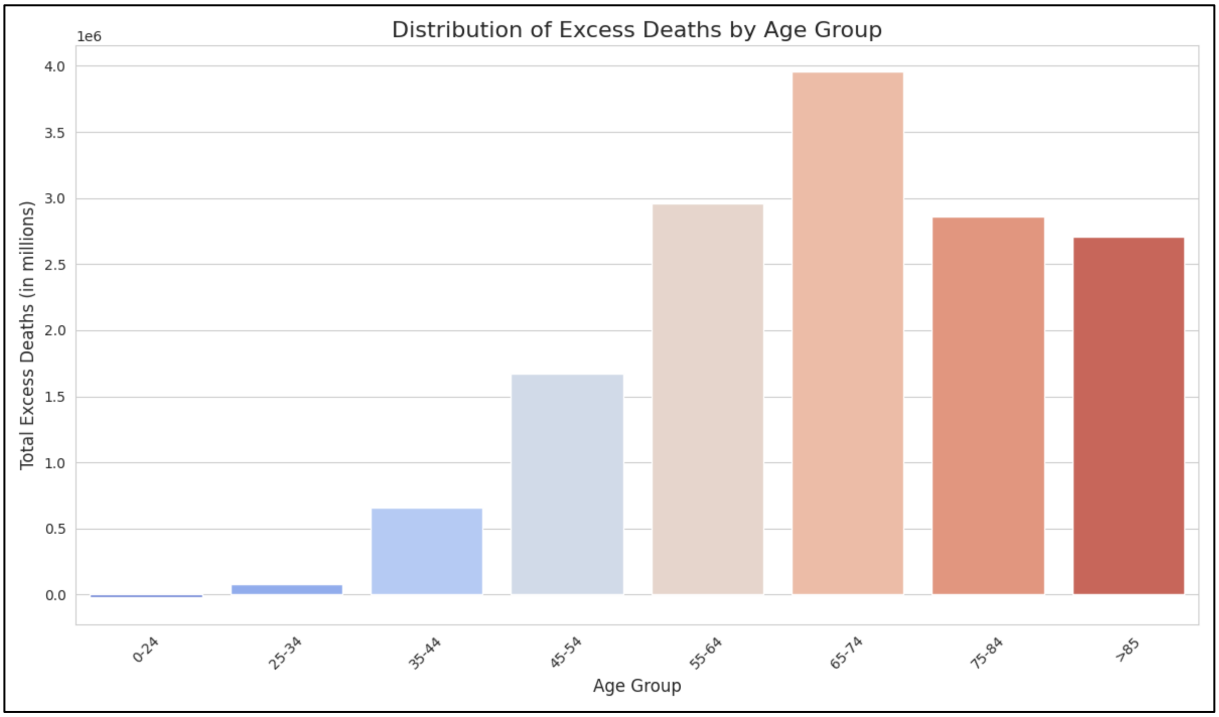
4.3 Demographic Analysis (Age and Sex)

Figure 18: Comparison of Excess Deaths by Sex



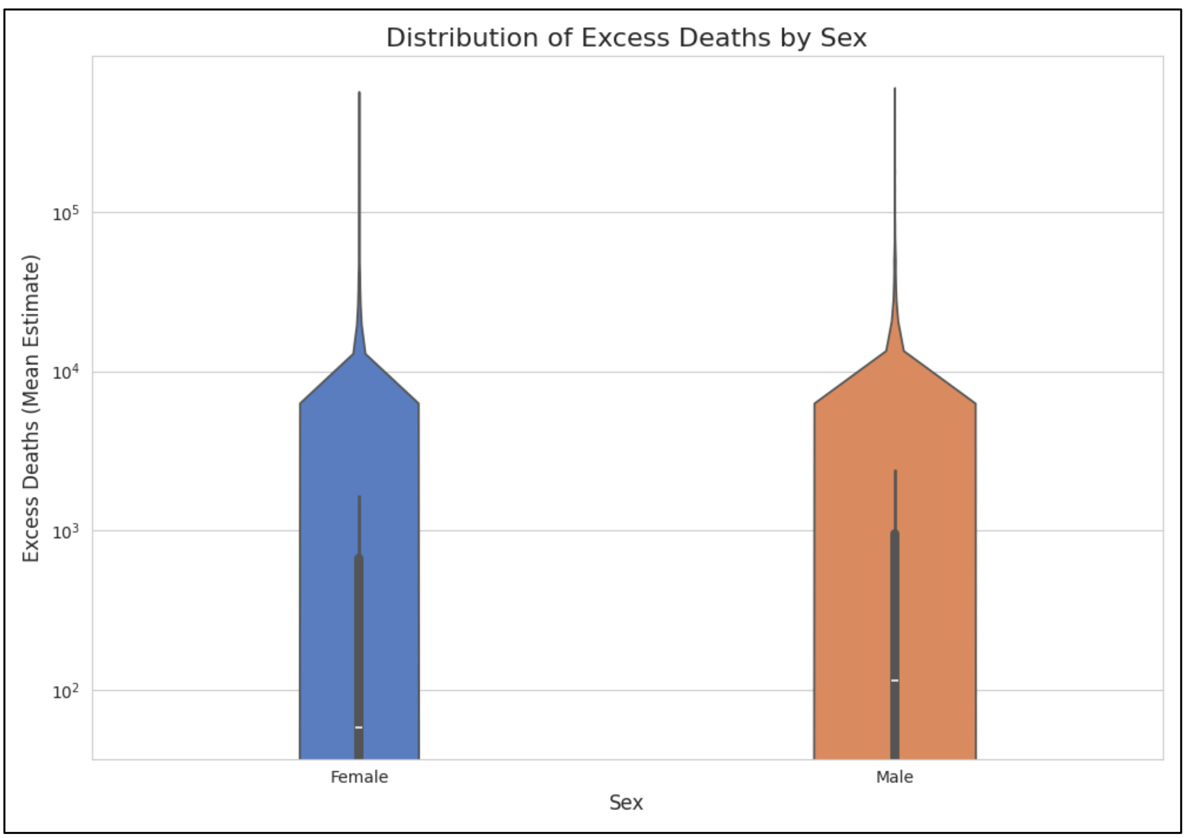
Observation: A higher number of excess deaths were recorded for males than for females globally, suggesting a gender disparity in the pandemic's impact.

Figure 19: Distribution of Excess Deaths by Age Group



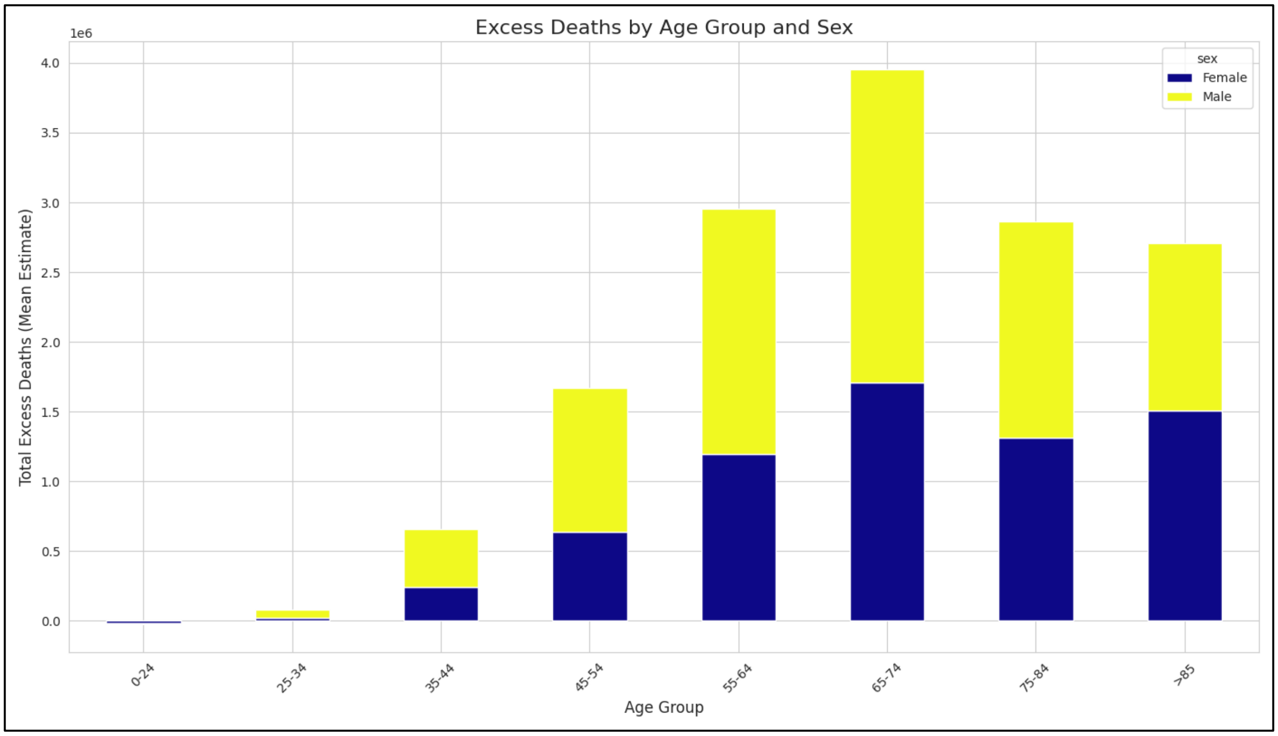
Observation: There is a clear trend showing that excess deaths increase significantly with age. The older age groups, particularly >65, account for the vast majority of excess deaths.

Figure 20: Violin Plot of Excess Deaths by Sex



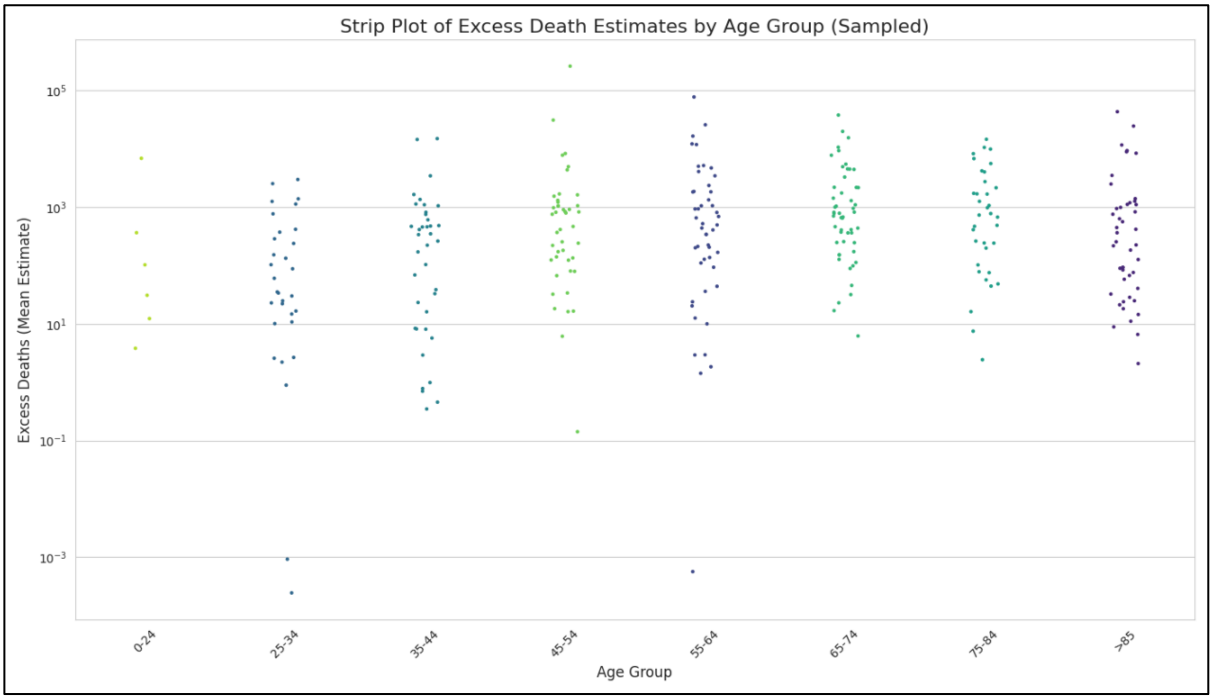
Observation: The violin plot for males is wider at higher values compared to the plot for females, showing that the density of higher-end estimates is greater for the male population.

Figure 21: Stacked Bar Chart of Deaths by Age Group and Sex



Observation: In almost every age group, the portion of the bar representing males is larger than that for females, confirming the gender disparity across different ages. The disparity appears most pronounced in the middle and older age groups.

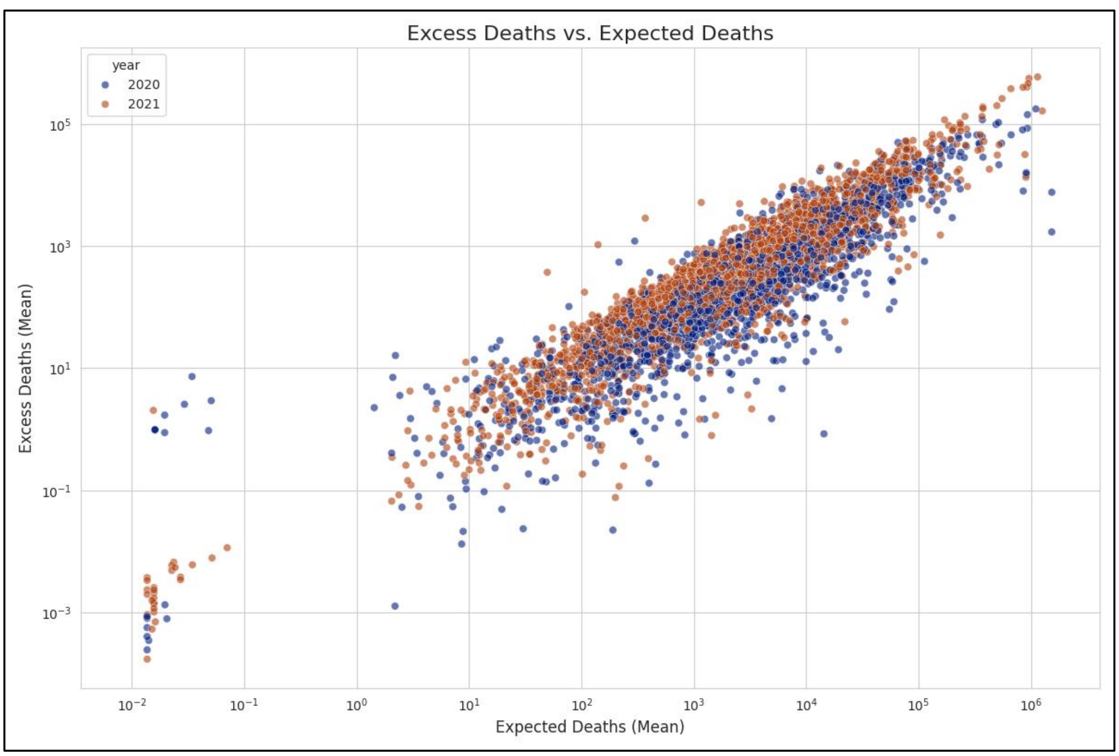
Figure 22: Strip Plot for Excess Deaths by Age Group



Observation: The swarm plot visually confirms the trend of increasing excess deaths with age. The density of points shifts upwards as age increases, and it also shows the wide range of estimates within each age category.

4.4 Demographic Analysis (Age and Sex)

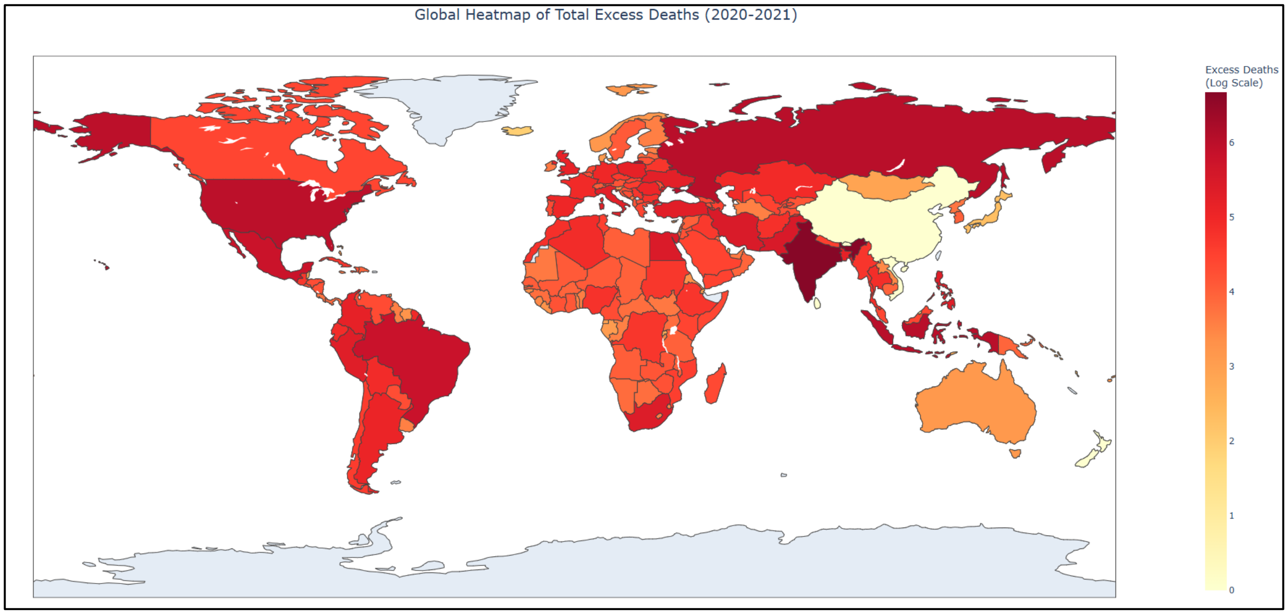
Figure 23: Excess Deaths vs. Expected Deaths



Observation: There appears to be a positive correlation between expected deaths and excess deaths. This suggests that regions with higher baseline mortality also tended to experience a higher number of excess deaths during the pandemic.

4.5 World Heatmap

Figure 25: World Heatmap of Total Cumulative Excess Deaths (2020-2021)



Observation: This world map provides a definitive global overview of the pandemic's cumulative toll. The color intensity, which is on a logarithmic scale to better visualize variations, clearly shows the epicentres of the crisis. North and South America, Europe, and South Asia (particularly India) are shaded in the darkest reds, indicating the highest concentration of excess deaths. In contrast, regions in Africa and Oceania show significantly lighter shading, reflecting lower estimated mortality. This single visualization encapsulates the geographic disparity of the pandemic's impact. An HTML file is added in the Visualization folder in the repository which can be opened to view an interactive visualization of the above world heatmap.

CHAPTER 5

SUMMARY OF KEY FINDINGS

The exploratory data analysis has yielded several critical insights into the nature of the COVID-19 pandemic's impact on global mortality:

- **Temporal Escalation:** The impact of the pandemic was not uniform over time. Mortality was substantially greater in **2021** than in 2020, indicating a significant worsening of the crisis globally in its second year.
- **Geographic Disparity:** The burden of excess deaths was not evenly distributed. A handful of countries, particularly those with large populations like **India, Russia, and the USA**, accounted for a disproportionately large share of the absolute excess deaths.
- **Demographic Vulnerability:** The analysis identified clear high-risk demographics. The risk of excess death was consistently higher for **males** than for females across all age groups. Furthermore, risk increased dramatically with **age**, establishing the elderly as the most vulnerable population.
- **Data Characteristics:** The dataset relies heavily on **statistical predictions** rather than direct reporting. This is a crucial context, implying that while the trends are robust, the exact numbers are estimates and should be treated as such.

CHAPTER 6

OUTLINE OF PROPOSED MACHINE LEARNING ALGORITHMS

6.1 Problem Framing: Regression

Based on the EDA, the dataset is perfectly suited for a supervised machine learning regression task. The primary goal will be to predict the continuous numerical value of excessmean. The features for this model will be the categorical variables (country, sex, age_group) and the numerical variable (year). Categorical features will be transformed using one-hot encoding to be compatible with machine learning algorithms.

6.2 Correction of Data Types

A multi-tiered modelling strategy is proposed to benchmark performance and build towards a highly accurate model:

1. **Linear Regression (Baseline):** This model will be implemented first to establish a baseline performance. While it is likely too simple to capture the complex, non-linear relationships in the data, its performance will serve as a crucial benchmark against which more sophisticated models can be compared.
2. **Random Forest Regressor:** As a powerful ensemble model, the Random Forest can effectively capture non-linear relationships and complex feature interactions. It is expected to provide a significant improvement in accuracy over the baseline. A key advantage is its ability to calculate feature importance, which can provide insights into which factors (e.g., age, country) are most predictive of excess deaths.
3. **Gradient Boosting Regressor (e.g., XGBoost, LightGBM):** This is expected to be the highest-performing model. Gradient Boosting algorithms build decision trees sequentially, with each new tree correcting the errors of the previous ones. They are renowned for their state-of-the-art performance in structured data competitions and are capable of capturing the most intricate patterns in the data to deliver highly precise predictions.

CHAPTER 7

CONCLUSION AND APPENDIX

This Exploratory Data Analysis has successfully processed and analysed the WHO dataset on Global Excess Deaths, transforming raw data into a series of actionable insights. Through a methodical process of data cleaning, preprocessing, and extensive visualization, this report has illuminated the profound and varied impact of the COVID-19 pandemic across the globe.

The analysis conclusively demonstrates that the pandemic's toll on mortality was not uniform; it escalated significantly in **2021**, disproportionately affected **males** and the **elderly**, and was heavily concentrated in specific geographic regions, including the **Americas, Europe, and South Asia**. Visualizations such as the world heatmap and demographic breakdowns have effectively quantified these disparities.

Furthermore, this EDA has successfully prepared the dataset for the next phase of the project. The patterns and correlations identified here provide a solid foundation for building predictive models. The proposed machine learning approach, aiming to forecast excess deaths, is a logical next step that builds directly upon the findings of this report. In essence, this analysis has not only provided a clear picture of the past but has also paved the way for developing tools to anticipate future public health challenges.

APPENDIX

Dataset Name: WHO_COVID_Excess_Deaths_Estimates_By_Countries.xlsx

Dataset Link: https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modelled-estimates

GitHub Link:
https://github.com/PrathamAgrawal51/Pratham_Agrawal_22070521078_ML_CA1