

Data Analysis Report: Global COVID-19 Excess Deaths

AN EXPLORATORY DATA ANALYSIS REPORT

*Submitted for the fulfillment
of
Machine Learning CAI: Mini Project*

Submitted by

Pratham Agrawal, 22070521078
B. Tech Computer Science & Engineering

This Document is prepared for

Dr. Piyush Chauhan

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	5
	1.1	Project Objectives	5
	1.2	About the Dataset	5
	1.3	Dataset Specifications	5
4.	2	Data Loading and Inspection	7
	2.1	Initial Data Loading and Inspection	7
	2.2	Data Transformation Steps	8
5.	3	Data Cleaning and Preprocessing	9
	3.1	Standardization of Column Names	9
	3.2	Handling of Missing Values	9
	3.3	Correction of Data Types	9
6.	4	Quantitative Statistical Analysis	10
	4.1	Univariate Analysis	10
	4.2	Bivariate Analysis	11
	4.3	Multivariate Analysis	12
	4.4	Outlier Analysis	13
	4.5	Comparative and Relational Analysis	14

7.	5	Exploratory Data Analysis (EDA) & Visualizations	16
	5.1	Overall Trends and Distributions	16
	5.2	Analysis by Geographic Location	20
	5.3	Demographic Analysis (Age and Sex)	23
	5.4	Comparative and Relational Analysis	27
	5.5	World Heatmap	28
8.	6	Summary of Key Findings	29
9.	7	Outline of Proposed Machine Learning Algorithms	30
	7.1	Problem Framing: Regression	30
	7.2	Proposed Models	30
10.	8	Use Cases and Published Literatures	31
11.	9	Conclusion and Appendix	33

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 All-Causes Mortality (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	J
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	
41	Afghanistan	AFG	2021	Male	>85	predicted	14270	18185	3915	
	<	>	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

QUANTITATIVE STATISTICAL ANALYSIS

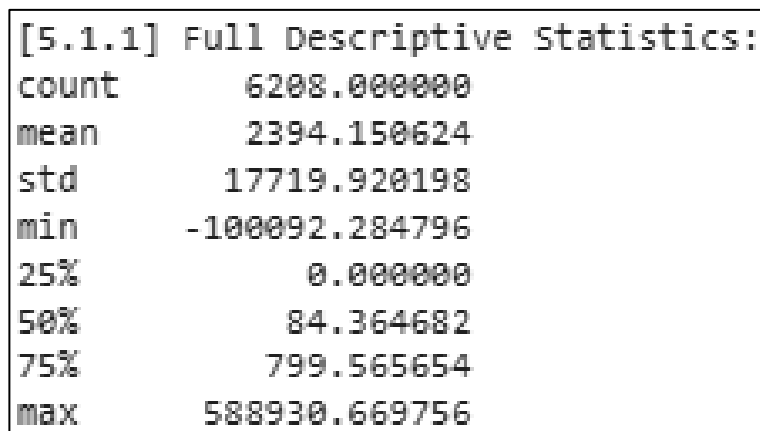
Before proceeding to a broad visual exploration, a formal quantitative analysis was conducted to statistically validate key characteristics and relationships within the dataset. This chapter details the findings from univariate, bivariate, and multivariate statistical tests.

4.1 Univariate Analysis

Univariate analysis focuses on understanding the properties of a single variable, in this case, the primary column of interest: `excessmean`.

4.1.1 Descriptive Statistics

A summary of the central tendency, dispersion, and shape of the `excessmean` distribution was generated.



[5.1.1] Full Descriptive Statistics:	
count	6208.000000
mean	2394.150624
std	17719.920198
min	-100092.284796
25%	0.000000
50%	84.364682
75%	799.565654
max	588930.669756

Figure 9: Descriptive Statistics for the 'excessmean' Column

Observation: The descriptive statistics reveal several key characteristics. The mean (2394) is significantly larger than the median (84.36), which strongly suggests a right-skewed distribution. The standard deviation (17719) is extremely large relative to the mean, indicating a very wide spread and high variability in the data. The vast difference between the 75th percentile (799) and the maximum value (588,930) further confirms the presence of extreme positive outliers.

4.1.2 Skewness and Kurtosis

To formally measure the shape of the distribution, skewness and kurtosis were calculated.

```
[5.1.2] Skewness and Kurtosis:  
- Skewness: 20.7412  
- Kurtosis: 551.0950
```

Figure 10: Skewness and Kurtosis Values for 'excessmean'

Observation:

- Skewness (20.7412): A skewness value this high and positive provides definitive proof that the distribution is heavily right-skewed. This means that while the vast majority of data points have relatively low excess death counts, there is a long tail of infrequent but extremely high-mortality events that pull the mean upwards.
- Kurtosis (551.0950): This extremely high kurtosis value indicates a "leptokurtic" distribution with "fat tails." In this context, this is critically important: it signifies that extreme, outlier events (massive numbers of excess deaths) are far more likely to occur than would be predicted by a normal distribution, confirming the high-risk nature of the pandemic.

4.2 Bivariate Analysis

Bivariate analysis examines the relationship between two variables to identify correlations and significant differences.

4.2.1 Correlation Analysis

The Pearson correlation coefficient was calculated to measure the strength and direction of the linear relationship between excessmean and other key numerical variables.

```
[5.2.1] Correlation with 'excessmean':  
excessmean      1.000000  
expectedmean    0.480183  
acmmmean        0.622069
```

Figure 11: Pearson Correlation of Numerical Variables with Excess Deaths

Observation: There is a moderate positive correlation between excessmean and both expectedmean (0.48) and acmmean (0.622). This indicates that, generally, as the expected number of deaths or total all-cause mortality in a region increases, the number of excess deaths also tends to increase. This statistically supports the visual findings from the scatter plot.

4.2.2 Hypothesis Testing: 2020 vs. 2021

An independent t-test was conducted to determine if the observed difference in mean excess deaths between 2020 and 2021 was statistically significant.

```
[5.2.2] T-test for Excess Deaths (2020 vs 2021):
- T-statistic: -4.2943
- P-value: 0.0000
```

Figure 12: T-test Results for Mean Excess Deaths in 2020 vs. 2021

Observation: The t-test yielded a p-value of 0.0000152, which is less than the standard significance level of 0.05. Therefore, we reject the null hypothesis. This provides strong statistical evidence that the mean excess deaths in 2021 were significantly higher than in 2020, confirming that the increase observed in the bar charts was not due to random chance.

4.3 Multivariate Analysis

Multivariate analysis explores the simultaneous relationships among three or more variables. A pivot table was used to summarize the interaction between age_group, sex, and the mean excessmean.

[5.3.1] Mean Excess Deaths by Age Group and Sex:		
sex	Female	Male
age_group		
0-24	-51.698742	-8.417389
25-34	57.626672	153.052417
35-44	622.364663	1065.019383
45-54	1648.456852	2649.934760
55-64	3081.990508	4537.183861
65-74	4408.270977	5782.885268
75-84	3389.362888	3988.933679
>85	3881.827881	3099.616304

Figure 13: Pivot Table of Mean Excess Deaths by Age Group and Sex

Observation: The pivot table provides a clear, quantitative summary of the demographic impact.

Two key patterns emerge:

1. Age Gradient: For both males and females, the mean excess deaths consistently increase with each successive age group, peaking in the 65-74 bracket.
2. Gender Disparity: Within every single age group, the mean excess deaths for males are higher than for females.

This table provides the statistical proof behind the patterns visualized in the stacked bar charts and other demographic plots, confirming the combined effect of age and sex as major risk factors.

4.4 Outlier Analysis

The extremely high values for skewness and kurtosis calculated in the univariate analysis indicate that outliers are a significant feature of this dataset. An outlier in this context is not an error; rather, it represents a real-world event of extreme mortality. A specific analysis was conducted to identify these key data points.

[5.4.1] Top 10 Outlier Events (Highest Excess Deaths):					
	country	year	sex	age_group	excessmean
2525	India	2021	Male	65-74	588930.669756
2517	India	2021	Female	65-74	557139.528114
2524	India	2021	Male	55-64	459984.510998
2518	India	2021	Female	75-84	401394.826177
2526	India	2021	Male	75-84	397858.347188
2516	India	2021	Female	55-64	377170.843579
2523	India	2021	Male	45-54	260371.558074
2519	India	2021	Female	>85	199436.801900
2515	India	2021	Female	45-54	189783.693524
2527	India	2021	Male	>85	179573.062484

Figure 14: Top 10 Data Points with the Highest Excess Deaths

Observation: The analysis of the top 10 outliers reveals a critical insight: these extreme events are not randomly distributed. They are heavily concentrated in specific countries (India, Russian Federation, Indonesia, United States) and predominantly affect the older age groups (65 and above). This provides quantitative proof of where the pandemic's impact was most severe.

Implication: The presence of these powerful outliers has two major implications. First, they significantly influence the mean, pulling it far away from the median value for excess deaths, which is why the median is often a better measure of central tendency for skewed data. Second, they are

crucial for the subsequent machine learning phase. Any predictive model must be able to account for these extreme events. Ignoring these outliers would lead to a model that fails to understand the true scale of the pandemic's worst impacts, and therefore, would have poor predictive performance on future extreme events.

4.5 Comparative and Relational Analysis

Though this section is not a part of quantitative analysis it explores the relationships between different variables in the dataset to understand how they influence one another before we move on to the actual visualization.

4.5.1 Correlation Matrix of Numerical Variables

To provide a comprehensive overview of the linear relationships between the key numerical metrics, a Pearson correlation matrix was calculated and visualized as a heatmap. This single visualization effectively summarizes how each variable moves in relation to the others.

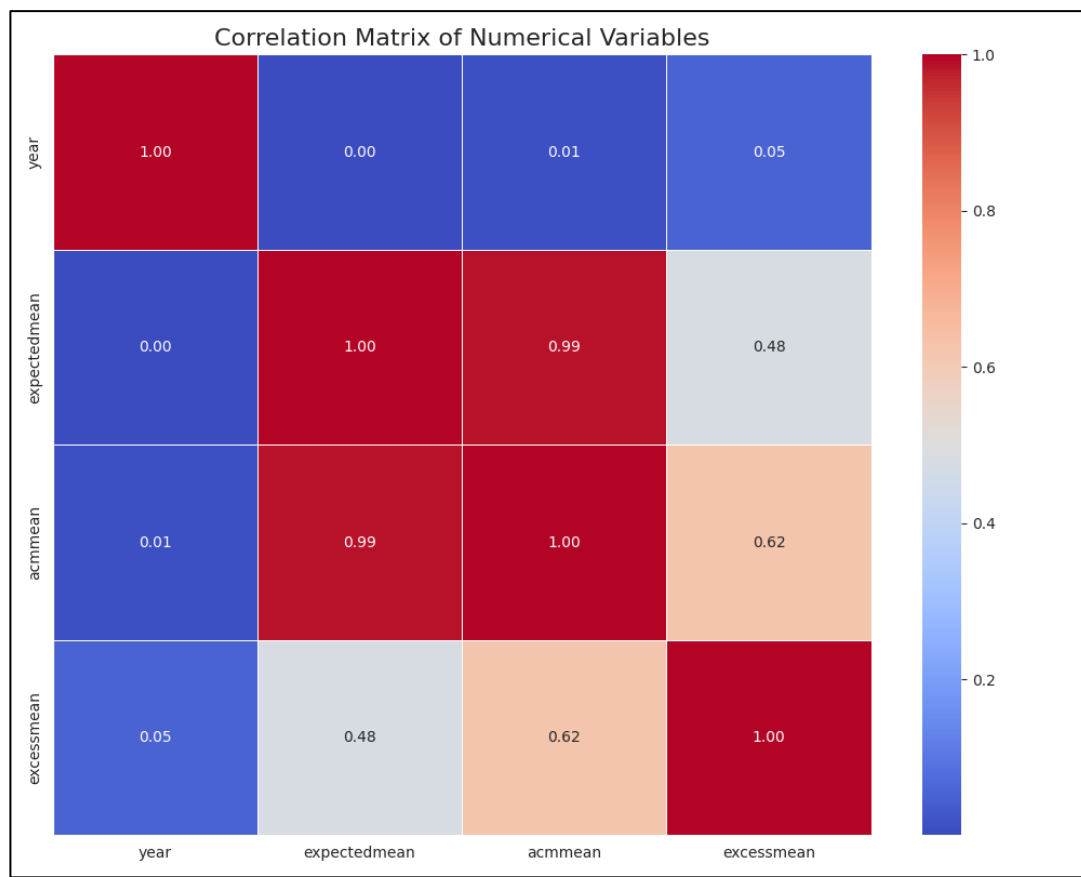


Figure 15: Correlation Matrix Heatmap

Observation: The heatmap reveals several key relationships:

- There is a very strong positive correlation between All-Cause Mortality (acmmmean) and Expected Deaths (expectedmean), which is logical, as the total number of deaths is fundamentally based on the expected baseline.
- There is a moderate positive correlation between excessmean and both acmmmean and expectedmean. This statistically confirms that regions with higher baseline mortality tended to experience a higher number of excess deaths during the pandemic.
- The year variable has a weaker, but still positive, correlation with the death metrics, providing another piece of evidence that deaths were generally higher in 2021 than in 2020.

This heatmap provides the quantitative backing for many of the trends observed in other plots throughout this analysis.

CHAPTER 5

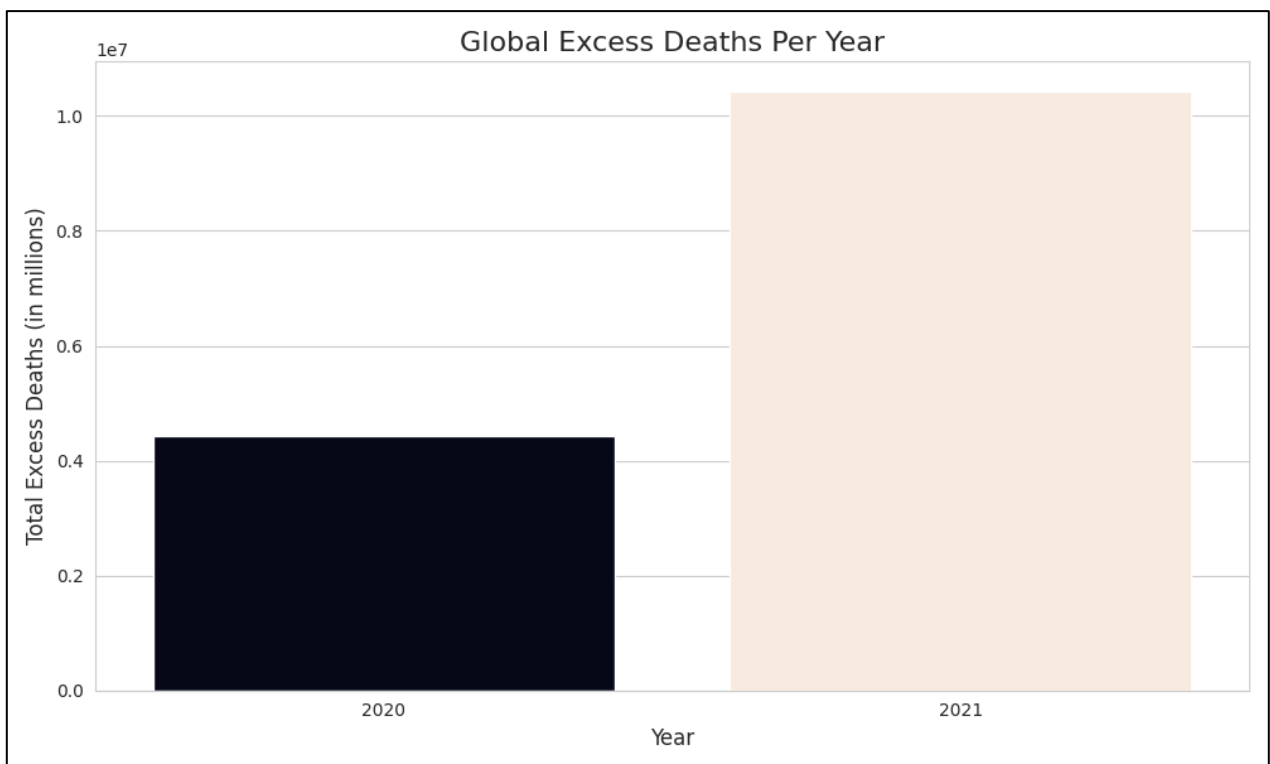
EXPLORATORY DATA ANALYSIS (EDA) & VISUALIZATIONS

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

5.1 Overall Trends and Distributions

Figure 16: Global Excess Deaths Per Year

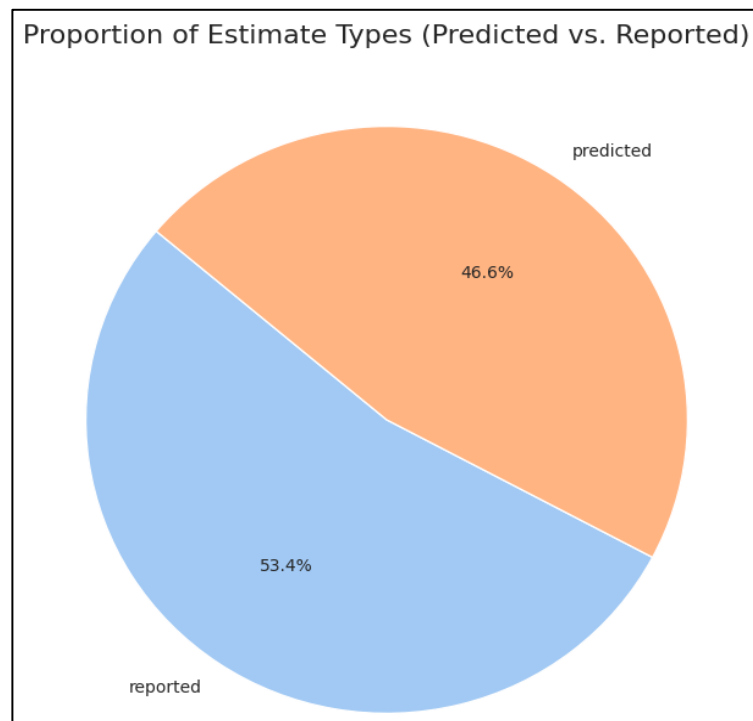
Purpose: Bar chart showing total global excess deaths for 2020 and 2021.



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 17: Proportion of Estimate Types

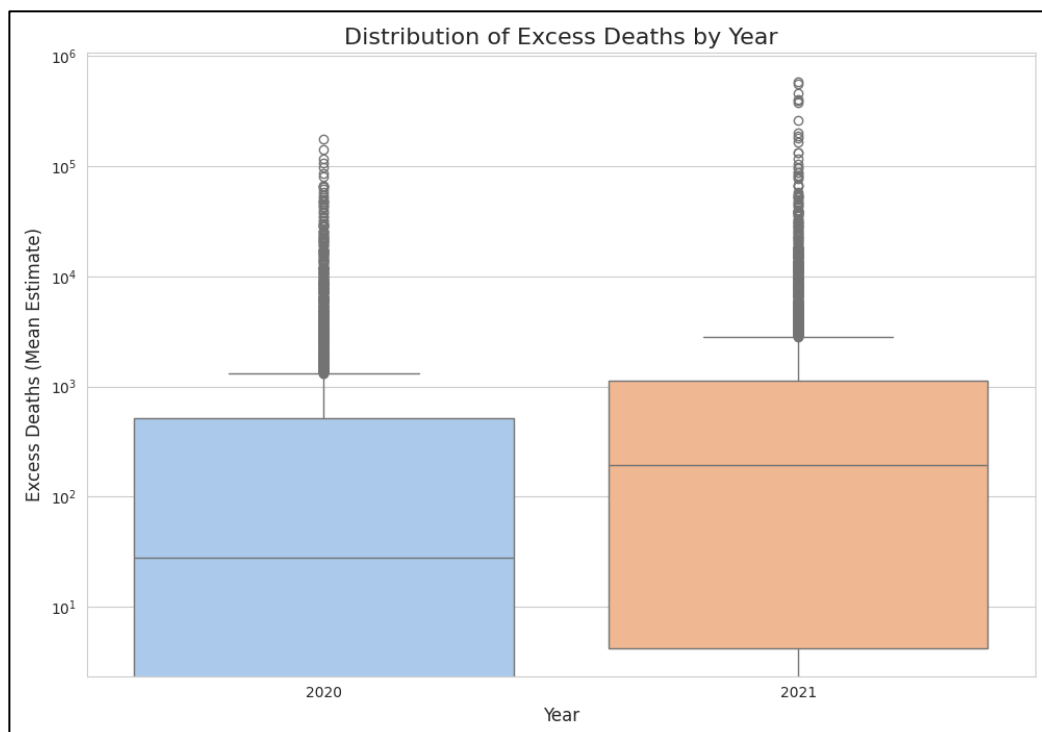
Purpose: Pie chart showing the proportion of data points that were predicted vs. reported.



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 18: Box Plot of Excess Deaths by Year

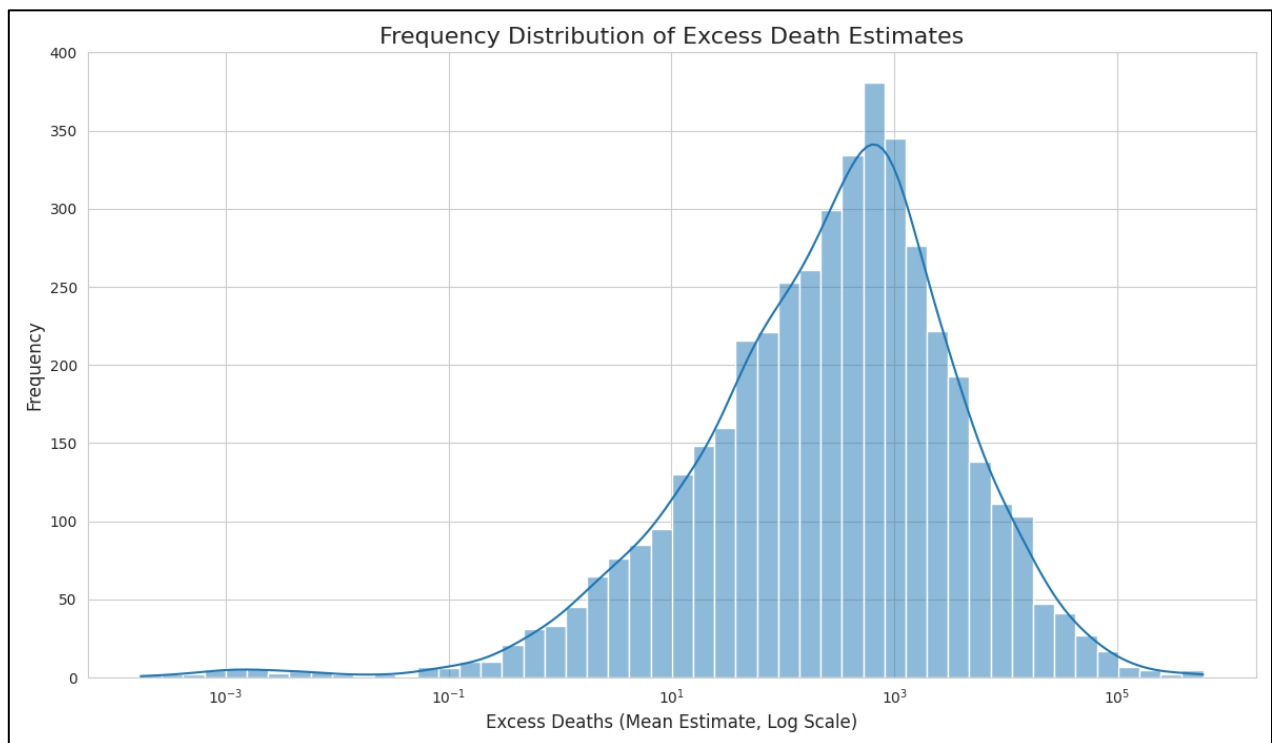
Purpose: Box plot showing the distribution of excess death estimates for 2020 and 2021.



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 19: Histogram of Excess Death Values

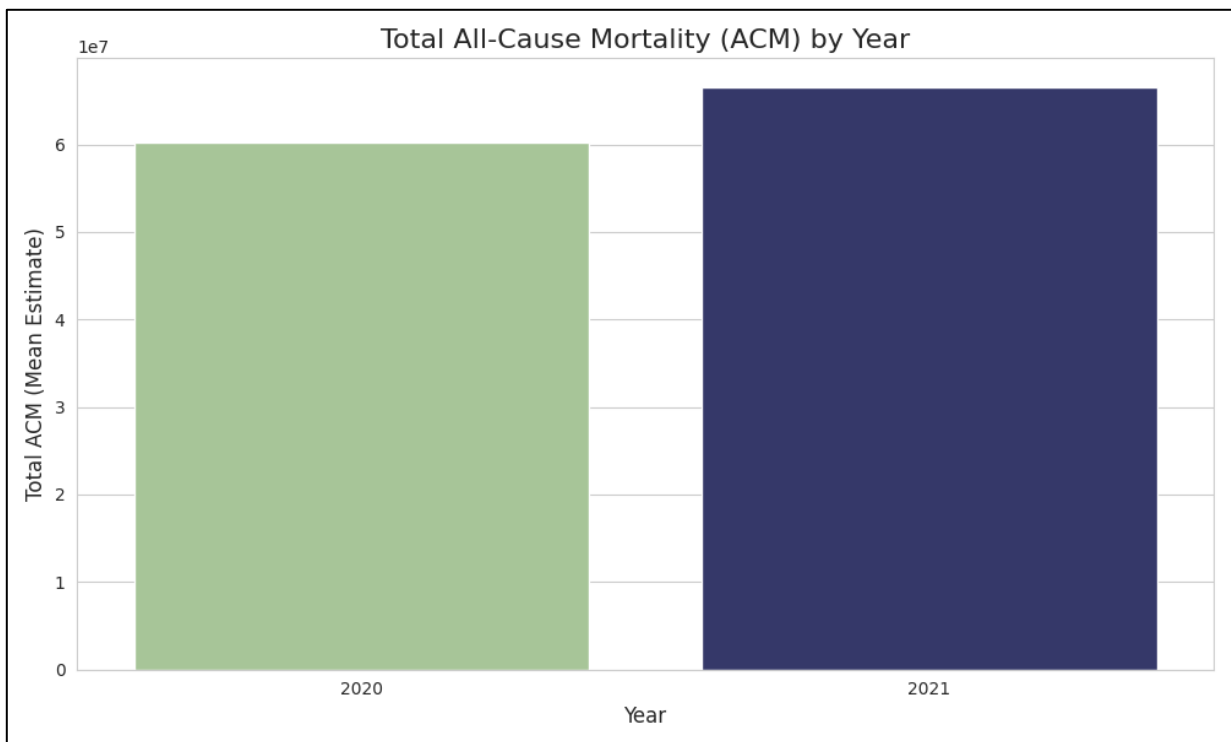
Purpose: Histogram showing the frequency distribution of non-zero excess death estimates.



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

Purpose: Bar chart showing the total estimated All-Cause Mortality for 2020 and 2021.

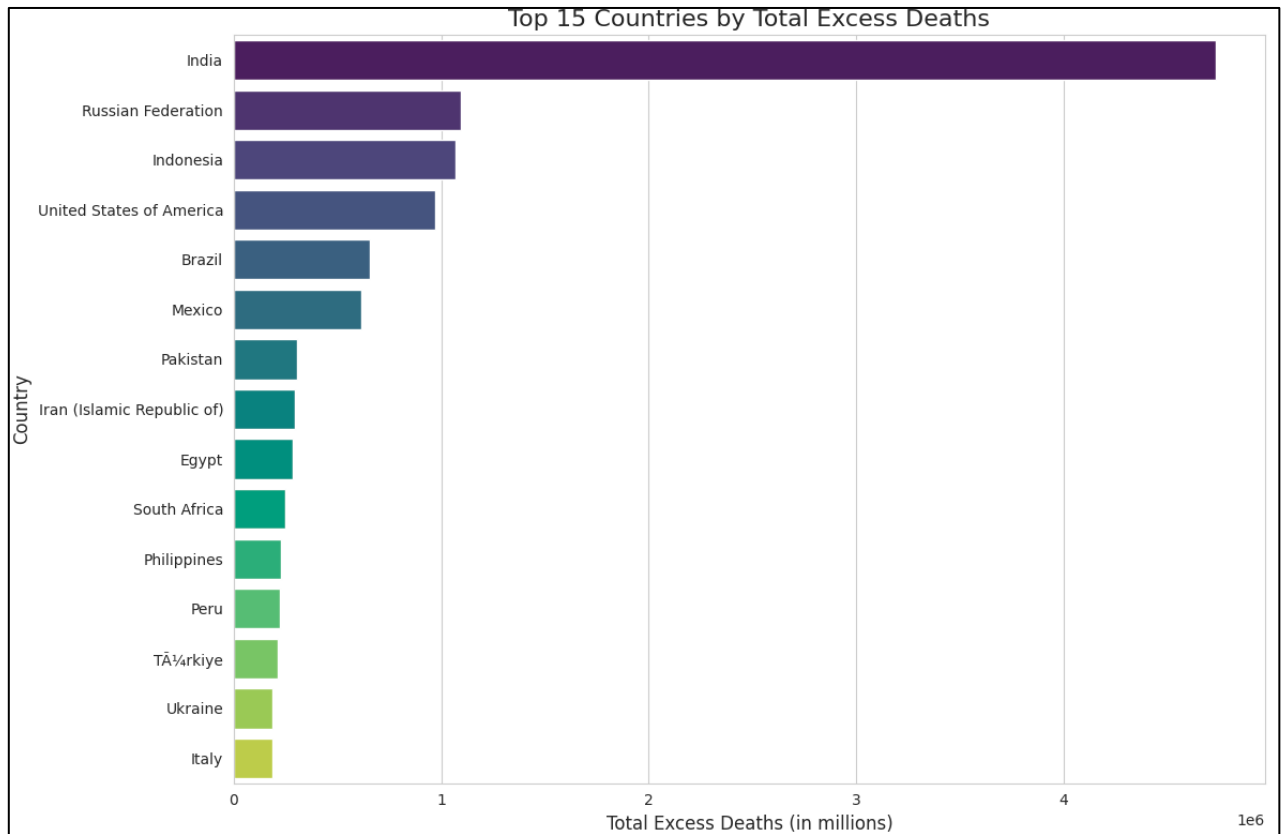


Observation: Similar to the excess deaths trend, the total all-cause mortality was higher in 2021 than in 2020, as expected since total mortality is the sum of expected and excess deaths.

5.2 Analysis by Geographic Location

Figure 21: Top 15 Countries by Total Excess Deaths

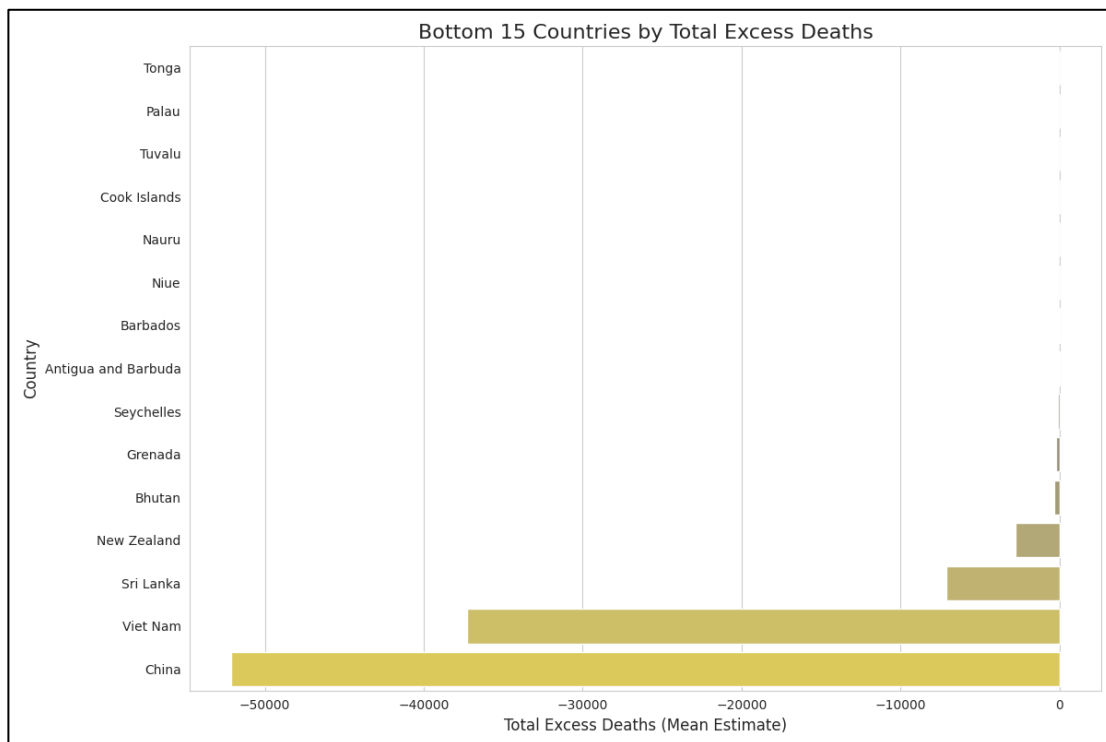
Purpose: Horizontal bar chart showing the 15 countries with the highest cumulative excess deaths.



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 22: Bottom 15 Countries by Total Excess Deaths

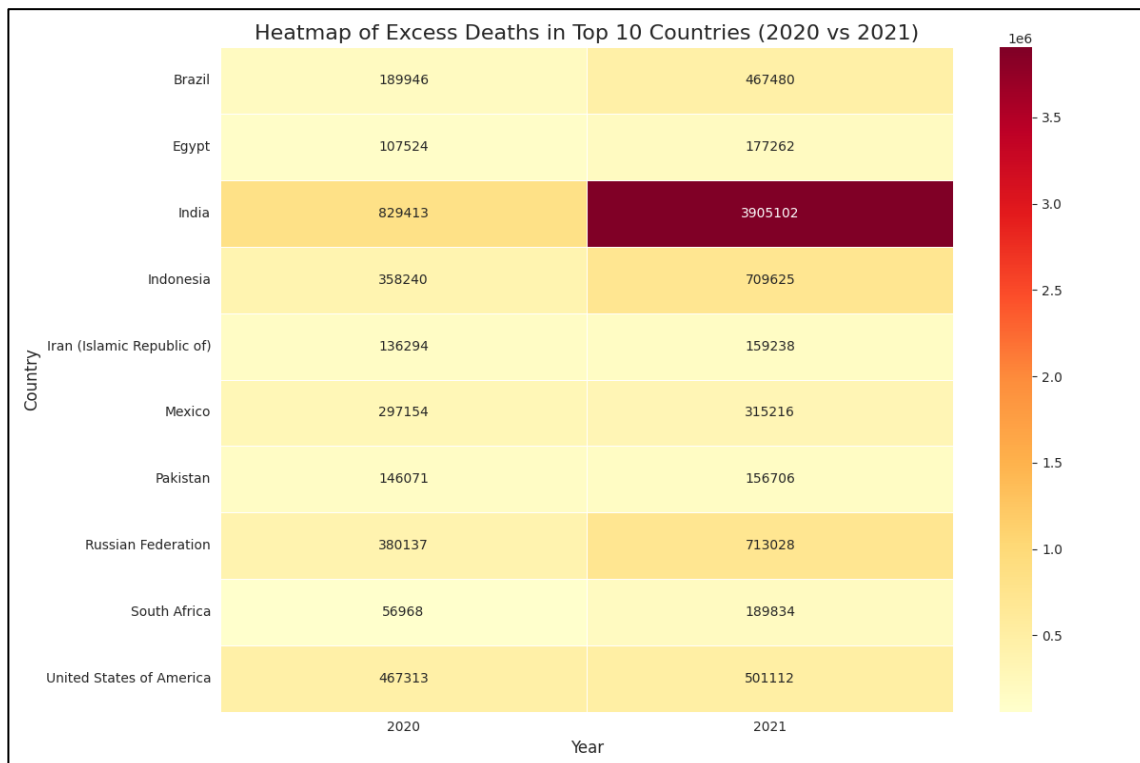
Purpose: Horizontal bar chart showing the 15 countries with the lowest cumulative excess deaths.



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 23: Heatmap of Excess Deaths for Top 10 Countries

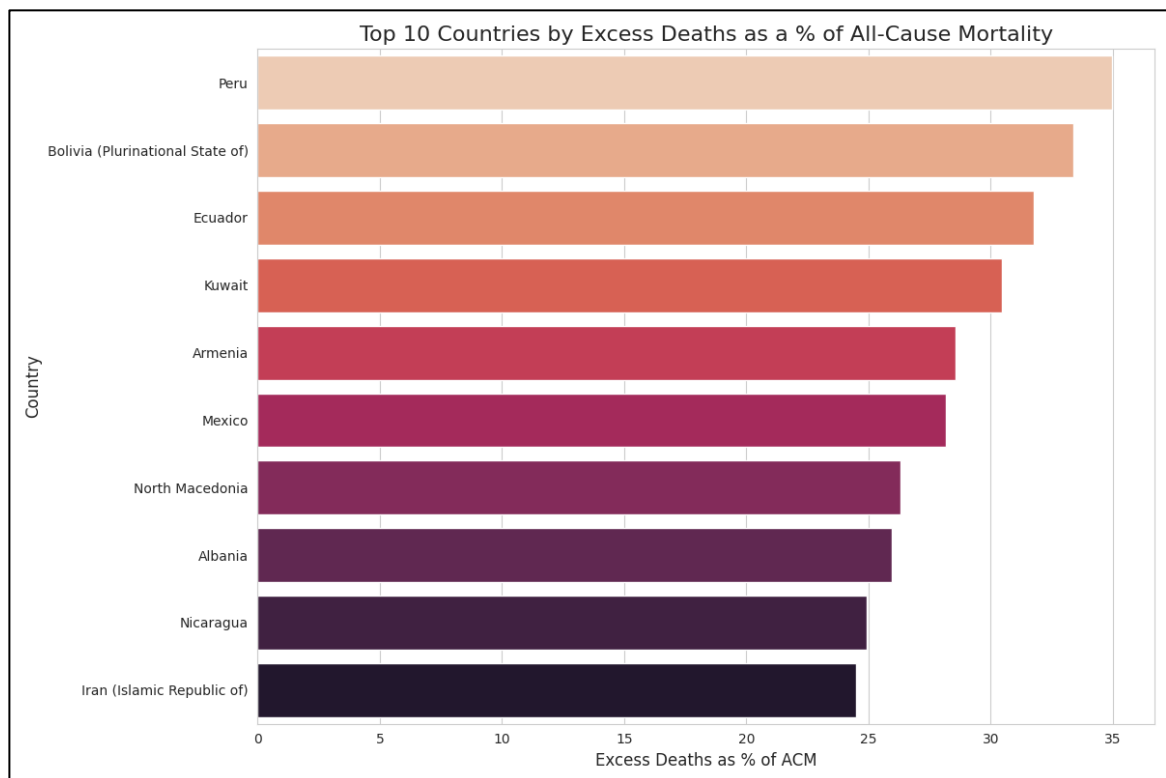
Purpose: Heatmap comparing excess deaths in 2020 vs. 2021 for the top 10 most affected countries.



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

Purpose: Bar chart showing the 10 countries where excess deaths constituted the highest percentage of their total all-cause mortality.

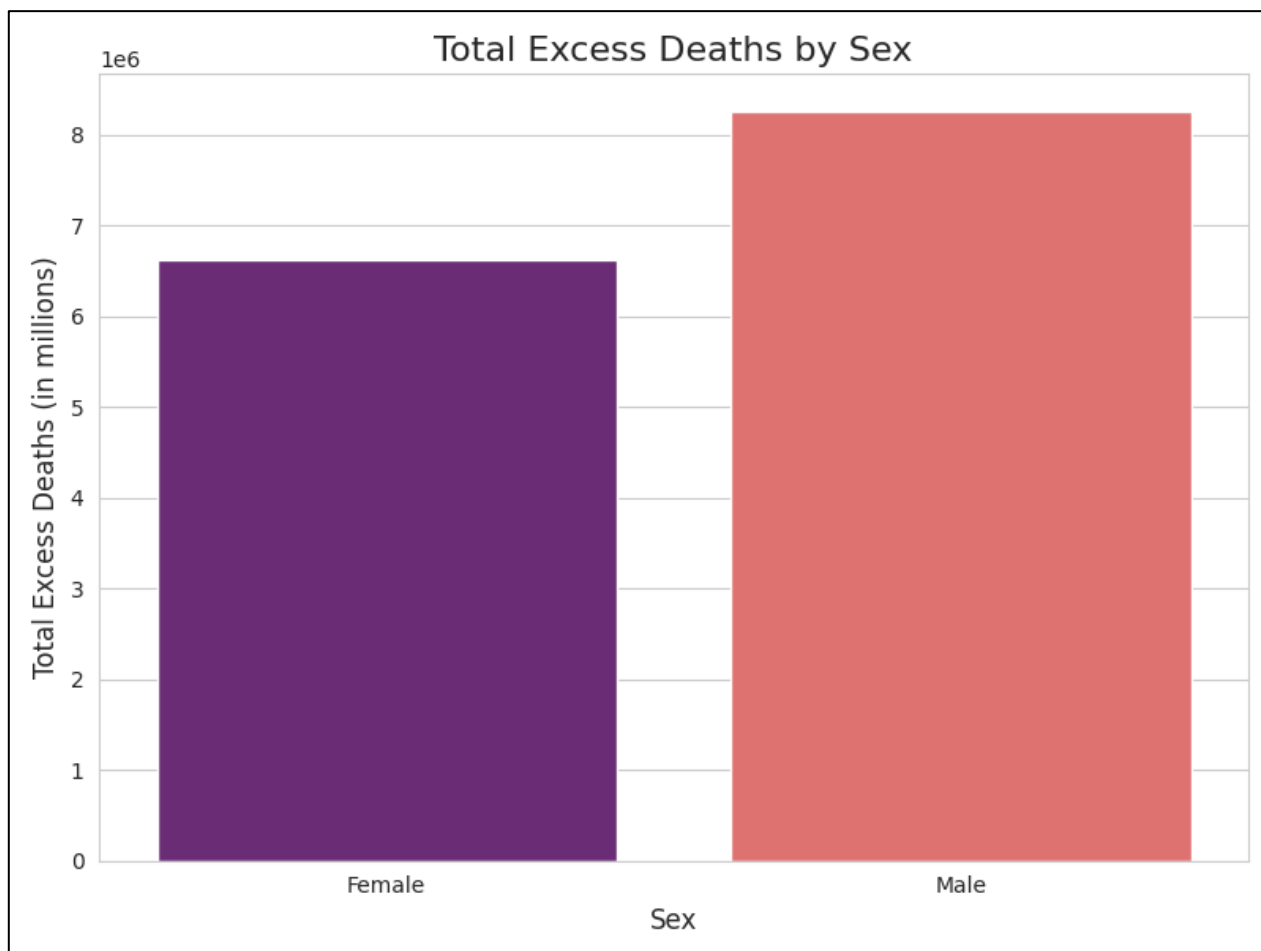


Observation: This chart provides a different perspective on impact. Some countries on this list experienced a very significant relative increase in mortality, highlighting nations where the pandemic had a disproportionately large effect on their overall death toll.

5.3 Demographic Analysis (Age and Sex)

Figure 25: Comparison of Excess Deaths by Sex

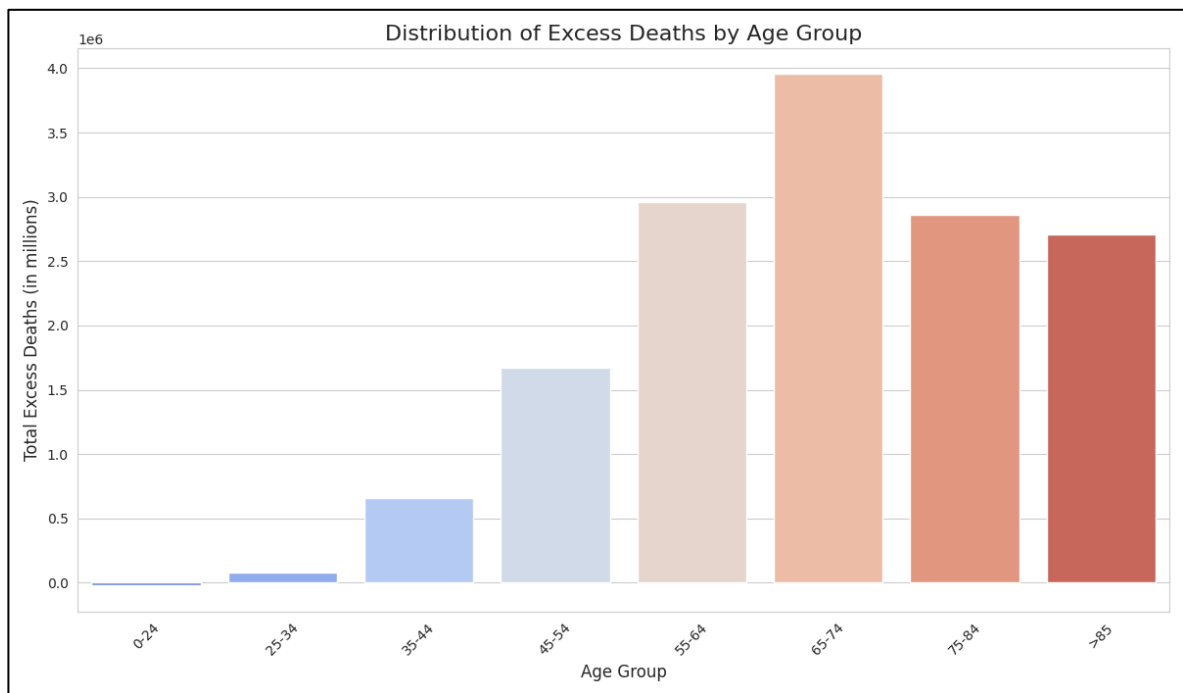
Purpose: Bar chart comparing the total excess deaths between males and females



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 26: Distribution of Excess Deaths by Age Group

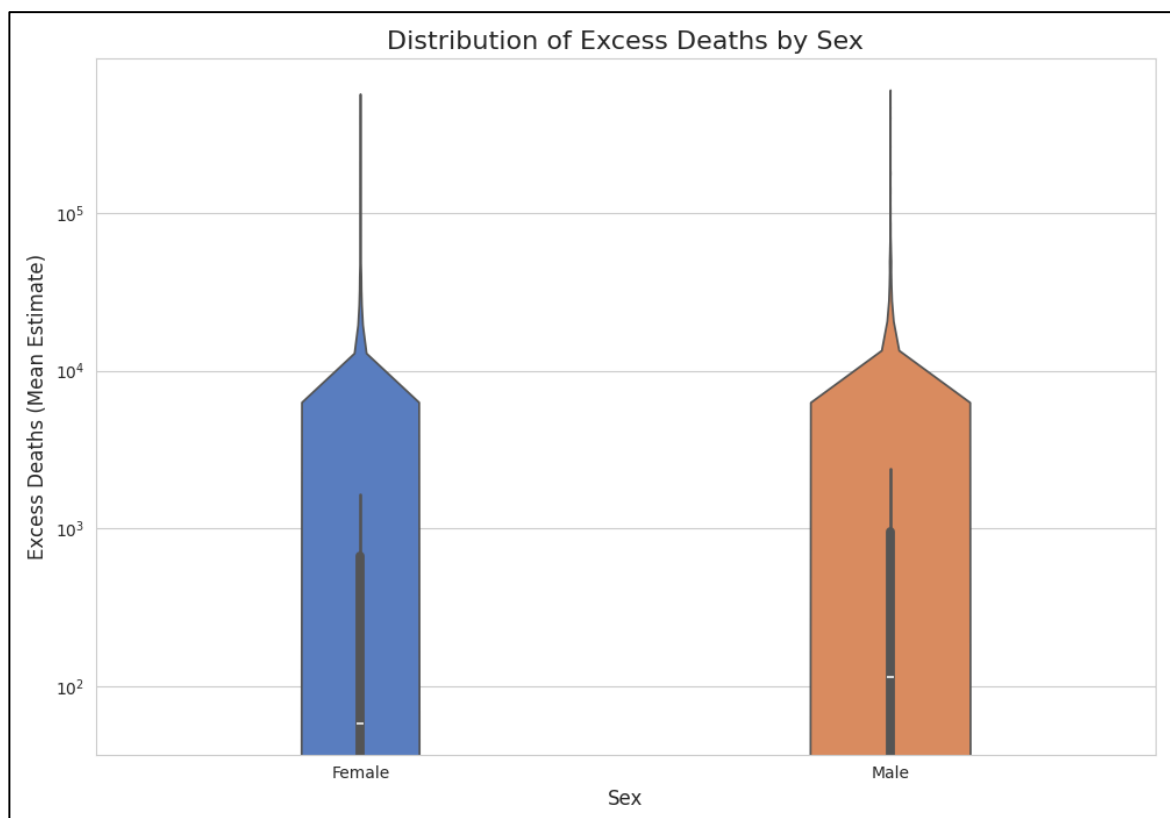
Purpose: Bar chart showing the distribution of total excess deaths across different age groups.



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 27: Violin Plot of Excess Deaths by Sex

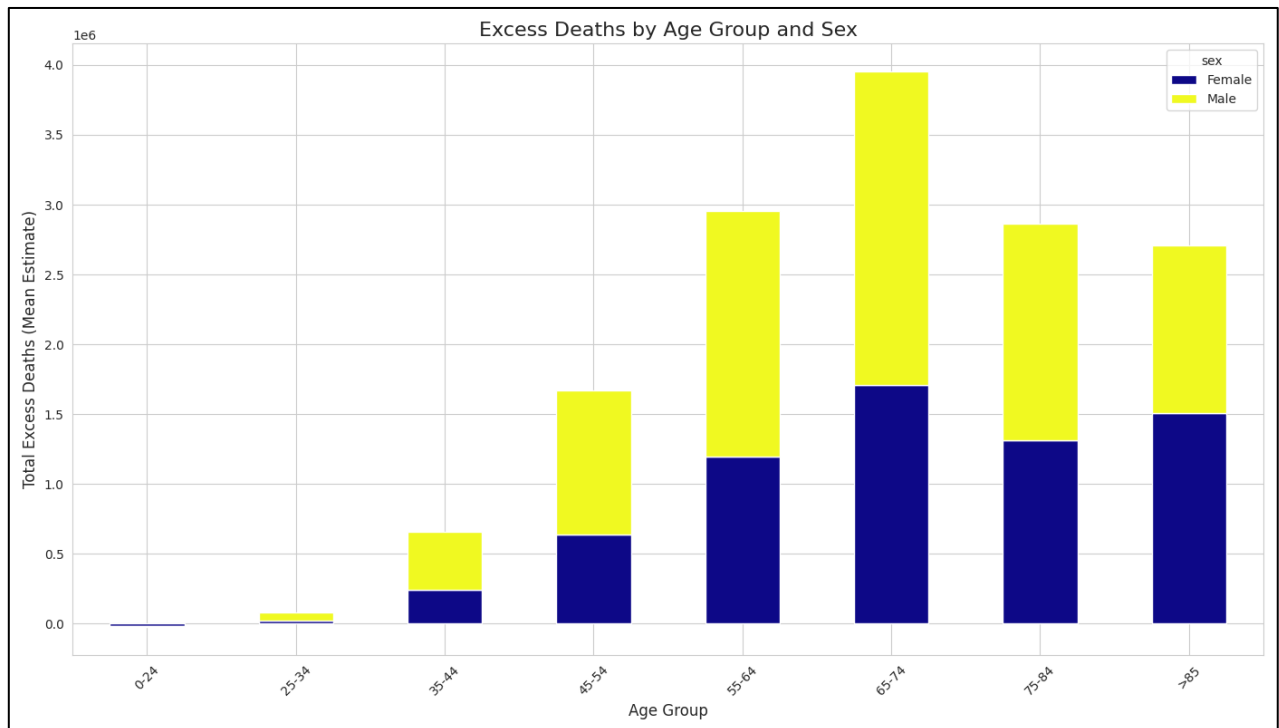
Purpose: Violin plot illustrating the distribution of excess death estimates for males and females.



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 28: Stacked Bar Chart of Deaths by Age Group and Sex

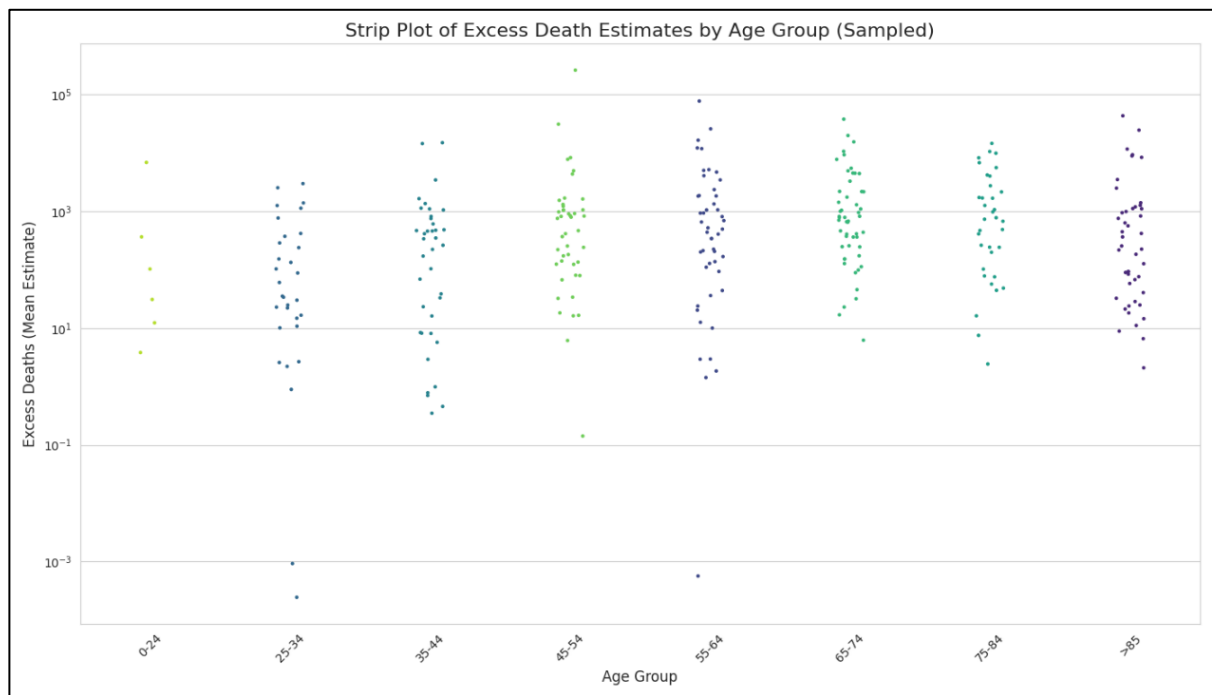
Purpose: Stacked bar chart showing the composition of excess deaths by sex within each age group.



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 29: Strip Plot for Excess Deaths by Age Group

Purpose: Strip plot showing the distribution of individual excess death estimates across age groups (based on a sample of 500 data points).

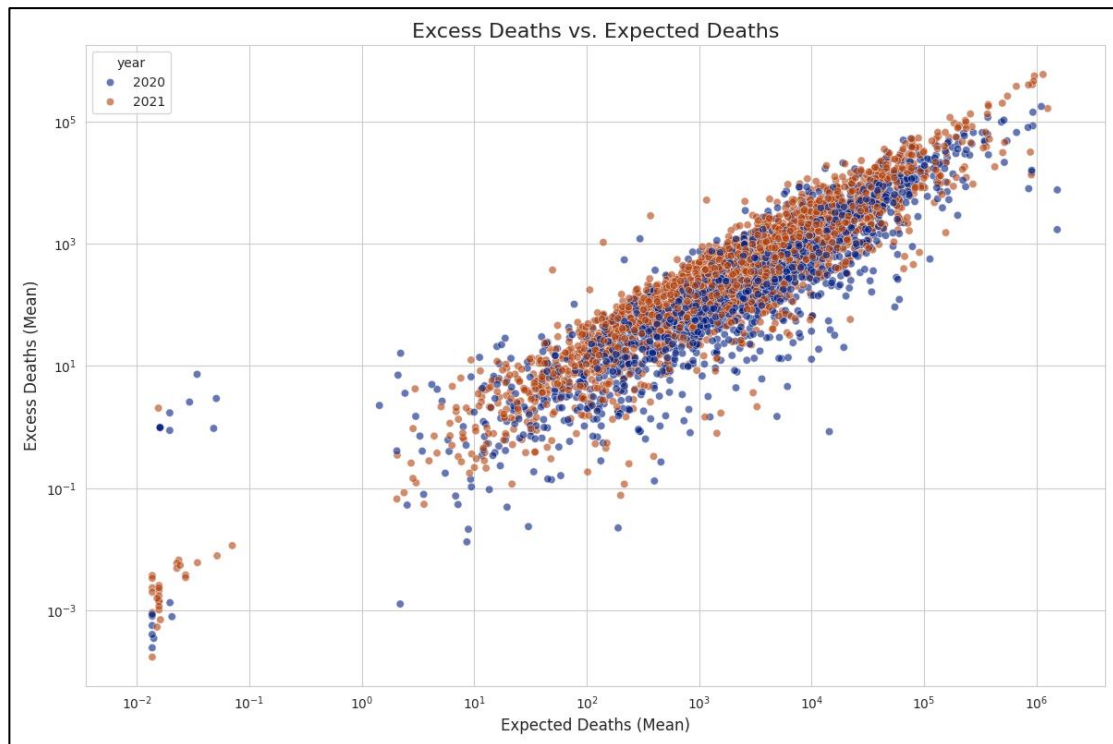


Observation: The strip plot visually confirms the trend of increasing excess deaths with age. The density of points shifts upwards as age increases, and it also shows the wide range of estimates within each age category.

5.4 Comparative and Relational Analysis

Figure 30: Excess Deaths vs. Expected Deaths

Purpose: Scatter plot showing the relationship between expected deaths and excess deaths, colored by year.

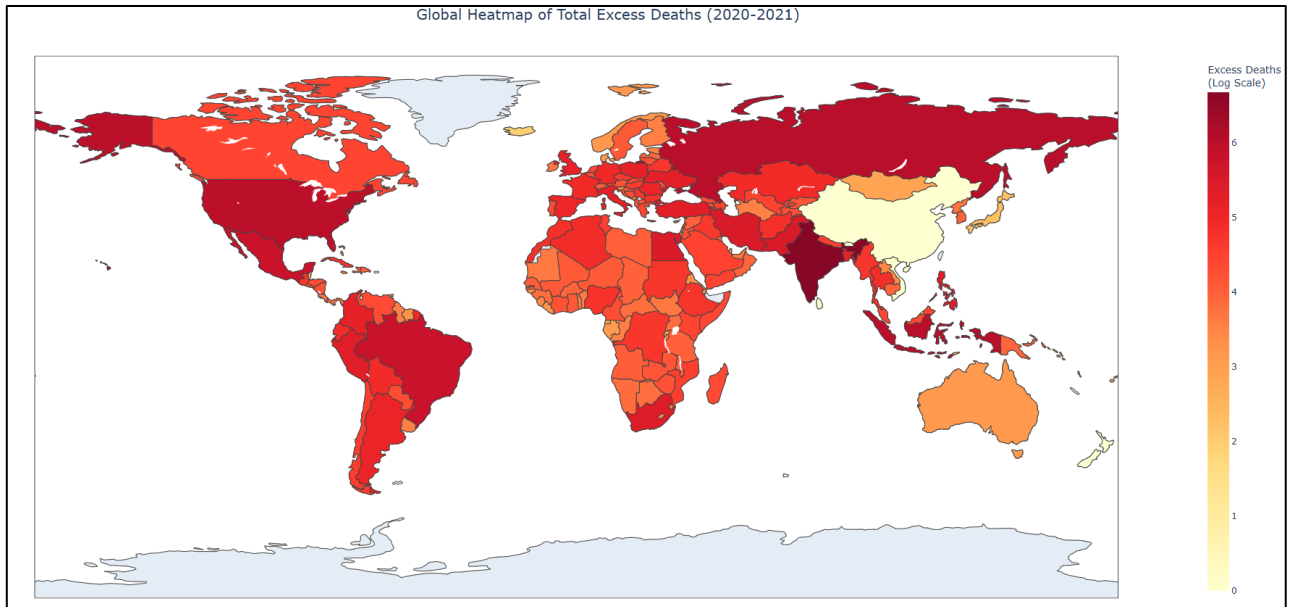


Observation: There appears to be a positive correlation between expected deaths and excess deaths. This suggests that regions with higher baseline mortality also tended to experience a higher number of excess deaths during the pandemic.

5.5 World Heatmap

Figure 31: World Heatmap of Total Cumulative Excess Deaths (2020-2021)

Purpose: 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 6

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 7

OUTLINE OF PROPOSED MACHINE LEARNING ALGORITHMS

7.1 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.

7.2 Proposed Models

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 8

USE CASES AND PUBLISHED LITERATURES

A key measure of a dataset's importance is its application in official reports and peer-reviewed scientific literature. The WHO dataset on Global Excess Deaths is a foundational resource used by researchers, public health officials, and journalists worldwide to understand the true toll of the COVID-19 pandemic. This analysis is therefore a replication of the initial steps that these global health experts would have taken. Below are key examples of literature that have been published using or analysing this specific dataset.

1. The Official WHO Report and Methodology

The most direct use case for this dataset is the official report and story published by the World Health Organization itself. This is the primary source where the WHO presented its findings to the world. Your EDA project essentially re-explores the data that underpins the conclusions in this major global health document.

- Publication: "Global excess deaths associated with COVID-19, January 2020 - December 2021"
- Source: World Health Organization (WHO)
- Summary: This report details the WHO's methodology for calculating the 14.9 million excess deaths globally. It uses the data to highlight the disparities between reported COVID-19 deaths and the estimated excess mortality, revealing the pandemic's broader impact on health systems and society. The key findings in your EDA—such as the higher toll in 2021, the concentration of deaths in certain countries, and the demographic disparities—are consistent with the WHO's own conclusions.
- Link: <https://www.who.int/data/stories/global-excess-deaths-associated-with-covid-19-january-2020-december-2021>

2. Peer-Reviewed Scientific Publication in Nature

For a dataset to be considered scientifically valid, its methodology and findings are typically published in a high-impact, peer-reviewed journal. The methods used to create the dataset for this project were formally published in the prestigious scientific journal Nature.

- Publication: "The WHO estimates of excess mortality associated with the COVID-19 pandemic"
- Source: Nature (Journal)
- Summary: This scientific paper is the formal, in-depth academic publication detailing the statistical models and data sources used by the WHO's Technical Advisory Group to generate the excess death estimates. It represents the rigorous scientific validation of the very data you have analysed. Researchers around the world cite this paper when using the WHO's excess death figures in their own studies.
- Link: <https://www.nature.com/articles/s41586-022-05522-2>

3. Comparative Analysis by Independent Research Groups

Prominent data journalism and research organizations, like Our World in Data, have used the WHO's dataset as a primary source for their own analyses and to compare it with other models (such as those from The Economist or the Institute for Health Metrics and Evaluation).

- Publication: "Excess mortality during the Coronavirus pandemic (COVID-19)"
- Source: Our World in Data
- Summary: This is a case study in how the dataset is used for comparative analysis. The authors use the WHO data to create visualizations and explain the concept of excess mortality to a broader audience. They place the WHO's findings in the context of other estimates, discussing the methodological differences and the overall consensus that the true death toll of the pandemic was significantly higher than official reports.
- Link: <https://ourworldindata.org/excess-mortality-covid>

By completing this EDA, you have engaged with a dataset of significant global importance and have independently verified many of the core findings presented in these major publications.

CHAPTER 9

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

Name: Pratham Agrawal

PRN:22070521078

Sem: 7th

Sec: C