Analysis of Respiratory Death Rates Across Regions: Impact of Socioeconomic and Environmental Factors

Regression Analysis SI 422

Anamika Basu Thakur (24N0084)
Jenithrika S. (24N0052)
Leena Patil (24N0045)
Preksha Agarwal (24N0048)

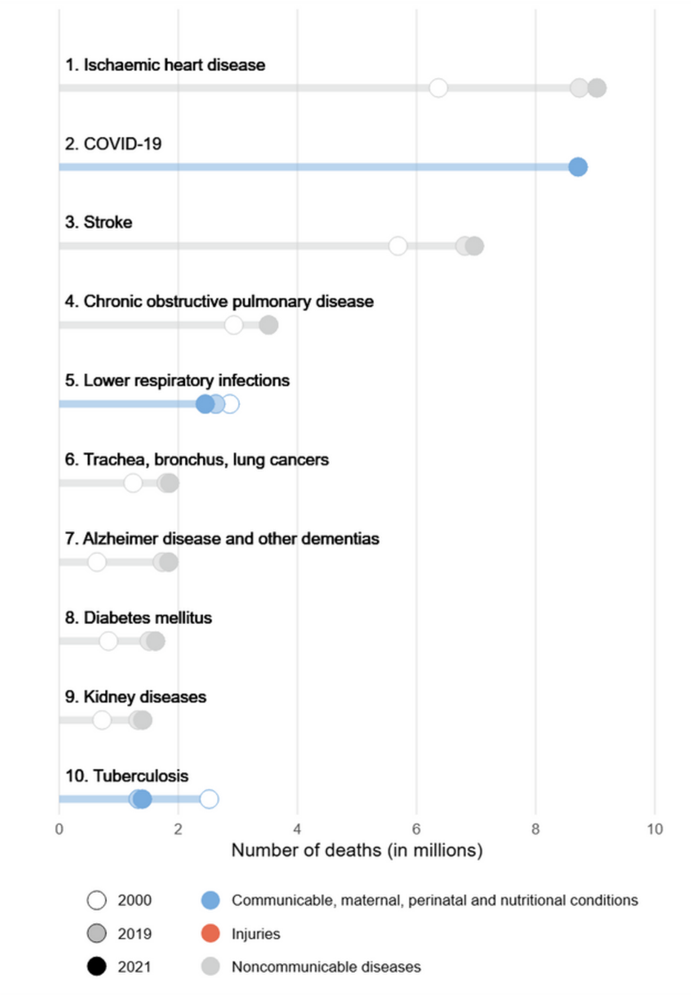
Acknowledgement

We would like to express our sincere gratitude to **Dr. Monika Bhattacharjee** for providing us with the opportunity, support, and guidance in completing this analysis project on the **"Analysis of Respiratory Death Rates Across Regions: Impact of Socioeconomic and Environmental Factors"** This project has been an invaluable learning experience. We would also like to take this opportunity to thank all those who dedicated their time to teaching us and assisting in the successful completion of this project. Furthermore, the success of this project would not have been possible without the collaboration and contributions of all our team members.

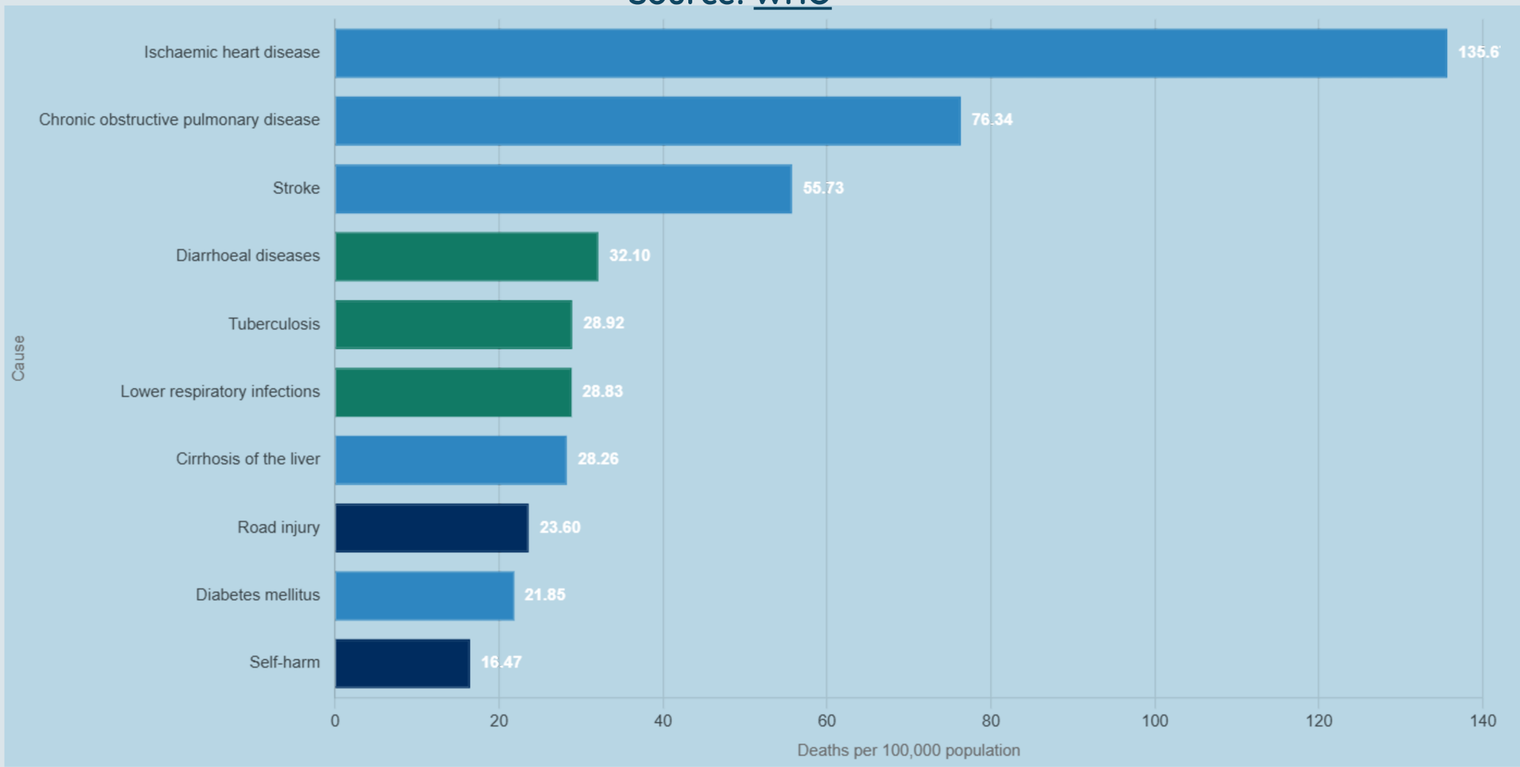
INDEX

Leading Causes of Death

1.1 Global leading causes of death
Source: WHO



1.2 Leading Causes of Death in India (2019)
Source: WHO



As seen in the **WHO** data, conditions like **chronic obstructive pulmonary disease, lower respiratory infections, and trachea/bronchus/lung cancers** are among the **top 10** causes of death worldwide. Notably, lower respiratory infections and tuberculosis show consistent mortality across decades, and during 2021, COVID-19 became the second leading cause of death globally, underlining the acute vulnerability of the respiratory system.

Geographic Distribution of Deaths Due to Respiratory Conditions

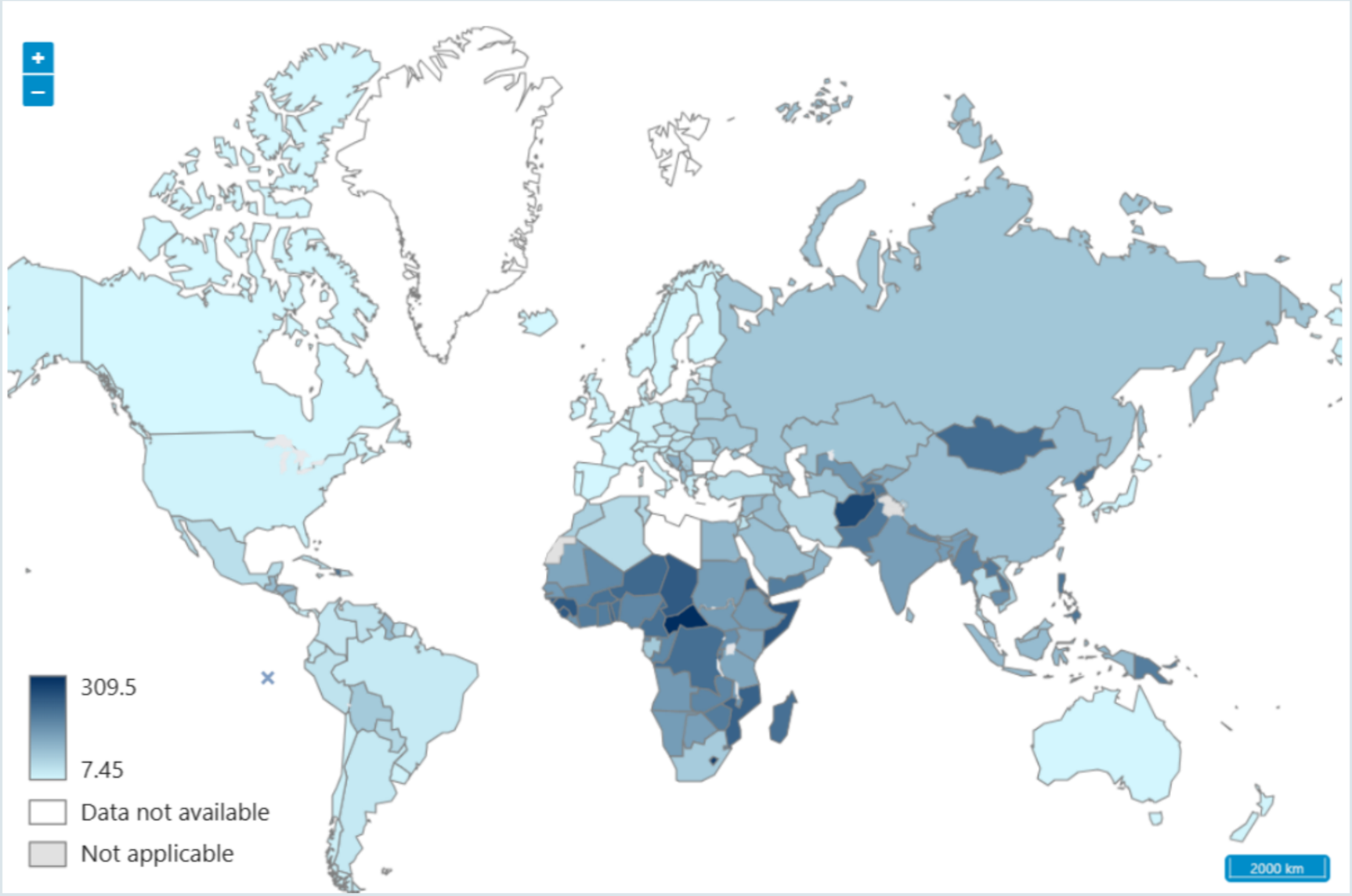


Image 1.3
Source: WHO

Problem Statement and Motivation

Respiratory diseases are a leading cause of mortality worldwide, with death rates varying significantly across regions. These differences are likely influenced by a combination of environmental pollutants and socioeconomic factors such as life expectancy, income group, and urbanization. The complexity of these interactions, coupled with challenges like multicollinearity and outliers in real-world data, makes it difficult to isolate the true drivers of respiratory mortality.

We chose this topic to explore how various environmental and demographic indicators jointly affect respiratory death rates. The motivation lies in generating data-driven insights that can inform policy decisions aimed at improving public health, especially in vulnerable regions. Through this analysis, we aim to build a statistically sound model that highlights the most impactful factors contributing to respiratory health outcomes.

Data Sources

SOURCE NAME	DATA USED	LINK
WHO Global Health Observatory	Estimated Death Rate, Life Expectancy, Population with primary reliance on polluting fuels and technologies for cooking(%)	who.int
World Bank Open Data	Forest Cover, Industrialisation, Urbanisation, Population Density, Precipitation levels	data.worldbank.org/
EDGAR Emissions Database	CO, OC, PM2.5, SO2, BC, NH3, NMVOC, NOx, PM10	edgar.jrc.ec.europa.eu/
World Bank	Income Groups	datahelpdesk.worldbank.org

Dataset Description

Dataset:

- **Total Records** : 180 observations
- **Total Variables** : 19 predictors, 1 target

Feature Breakdown:

Categorical features

- **ParentLocation** – Region of the country
- **Location** – Country
- **Income_group** - Income group assigned based on gross national income (GNI) per capita, in U.S. dollars

Numerical features

- **Density –** Population density (people per square km)
- **Life expectancy** – Average life expectancy at birth (in years)
- **per_polluting_fuels** – Proportion of population with primary reliance on polluting fuels and technologies for cooking (%)
- **Urbanisation** – % of population living in urban areas
- **Industrialisation** – Industry (including construction), value added (% of GDP)

Numerical features(cont.)

- **Precipitation(mm)** – Average annual precipitation (in mm)
- **Forest_cover** – % of land area covered by forests in a country

Air Pollutant emmisons (Gigagrams/year):

- **CO** - Carbon Monoxide
- **OC** - Organic Carbon
- **PM 2.5** - Fine particulate matter (≤ 2.5 microns)
- **SO2** - Sulfur Dioxide
- **BC** - Black Carbon
- **NH3** - Ammonia
- **NMVOC** - Non-methane volatile organic compounds
- **NOx** - Nitrogen Oxides
- **PM10** - Coarse particulate matter (≤ 10 microns)

Target Variable:

- **Death rate** - Estimated deaths due to respiratory problems per 100,000 individuals

Data Preprocessing

Data Cleaning & Merging

- Selected relevant features and standardized variable names for clarity and consistency.
- Applied filters to retain meaningful and comparable records across datasets.
- Combined multiple datasets into a single structured format using common keys.

Handling Missing Values

- Missing values caused by inconsistent country names were identified and resolved.
- Name discrepancies for identical countries were corrected to ensure accurate merging.
- Remaining incomplete entries were dropped, yielding **171** complete observations for analysis.

Feature Scaling & Encoding

- Numerical variables were scaled appropriately to ensure uniform influence across the model.
- Categorical variables were encoded using suitable encoding techniques to make them machine-readable.
- These steps ensured that all variables were prepared for effective modeling and interpretation.

Exploratory Data Analysis

In order to gain deeper insights into the dataset and the relationship between predictors and the target variable (Death Rate), a comprehensive exploratory data analysis (EDA) was conducted. This step was crucial for understanding the data's structure, distributions, and potential patterns before modeling.

- **Univariate Analysis:** Each predictor's distribution was analyzed using histograms, box plots, and density plots to examine:
  - The shape of distributions (e.g., skewness, normality)
  - The presence of outliers and extreme values

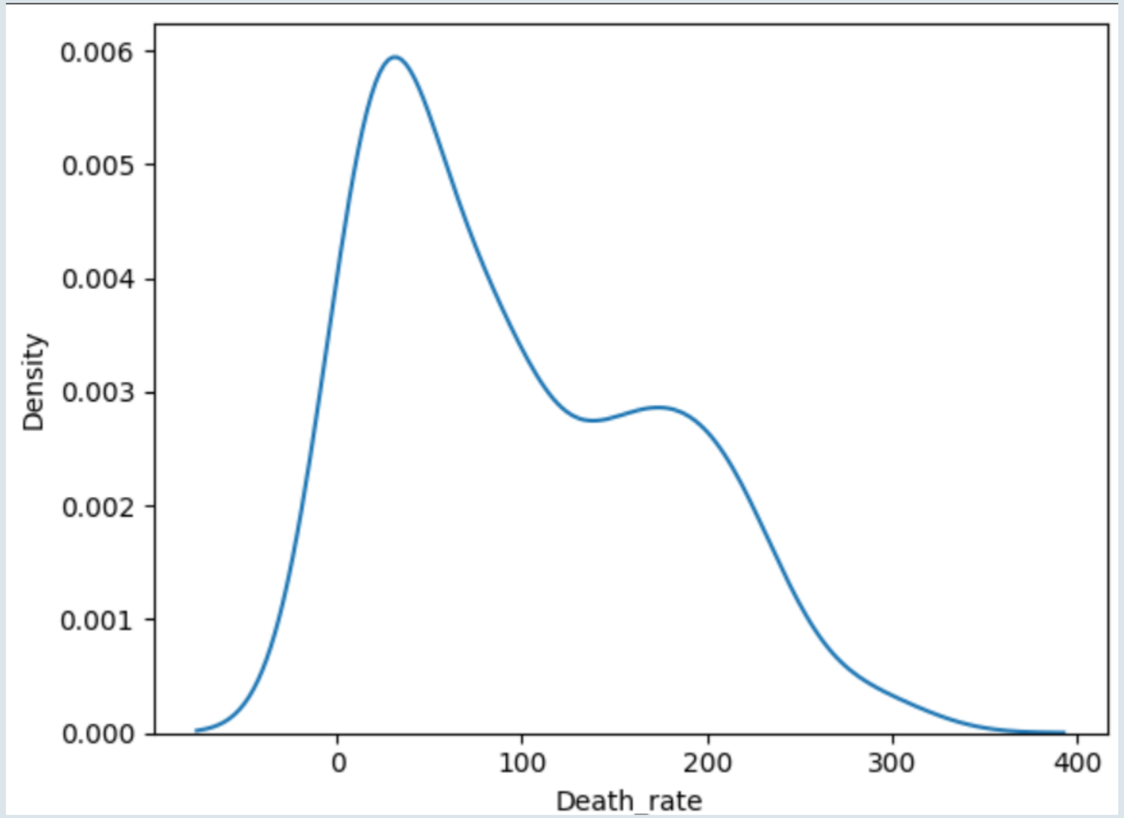
- **Bivariate Analysis:** The relationship between each predictor and Death Rate was visualized using:
  - **Scatter plots** to detect linear or non-linear associations.
  - **Box plots** for categorical variables to compare the distribution of Death Rate across different categories.

Exploratory Data Analysis

Distribution of Death Rate(Y)

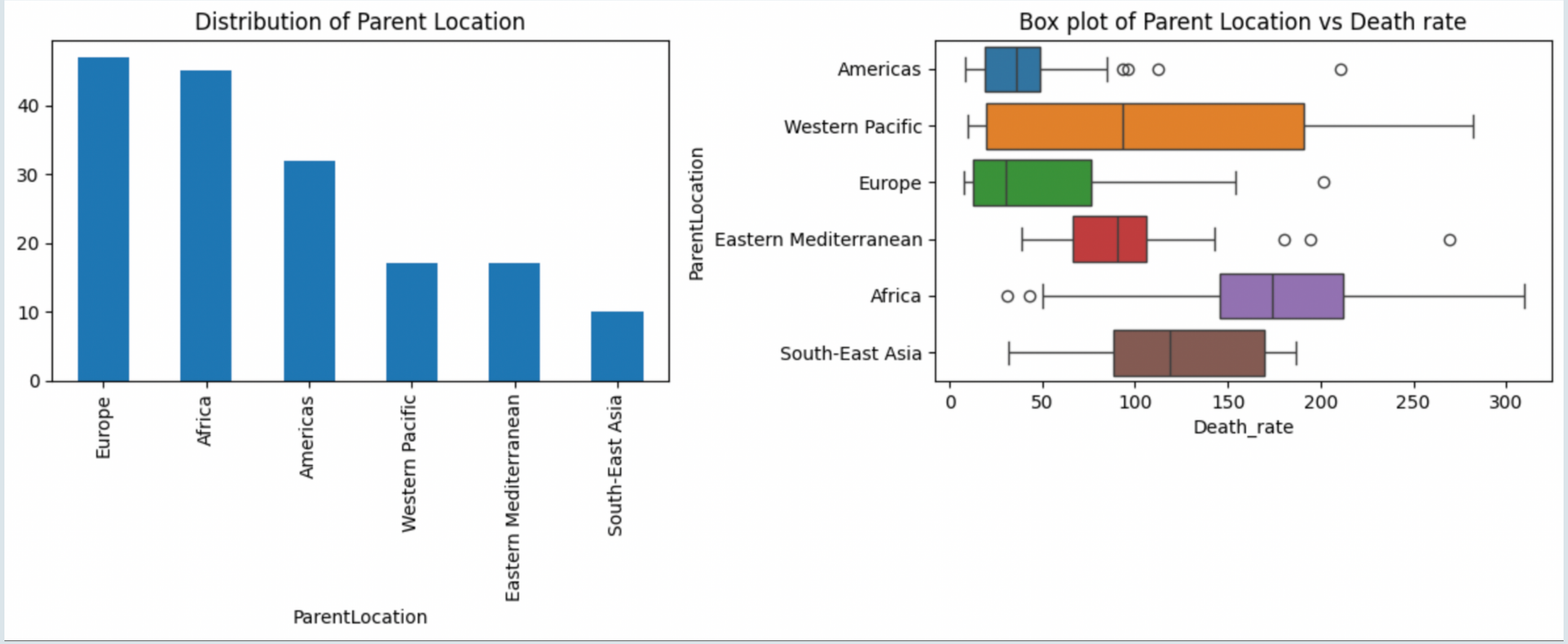
Analysis:

- **Skewness of the Distribution:**
  - The target variable is right-skewed, with most values concentrated at lower levels
- **Bimodal Tendency:**
  - A bimodal pattern suggests the presence of two distinct subpopulations.
- **Outliers:**
  - There are a few extreme outliers (Death_rate > 300), which could impact model performance.



Exploratory Data Analysis

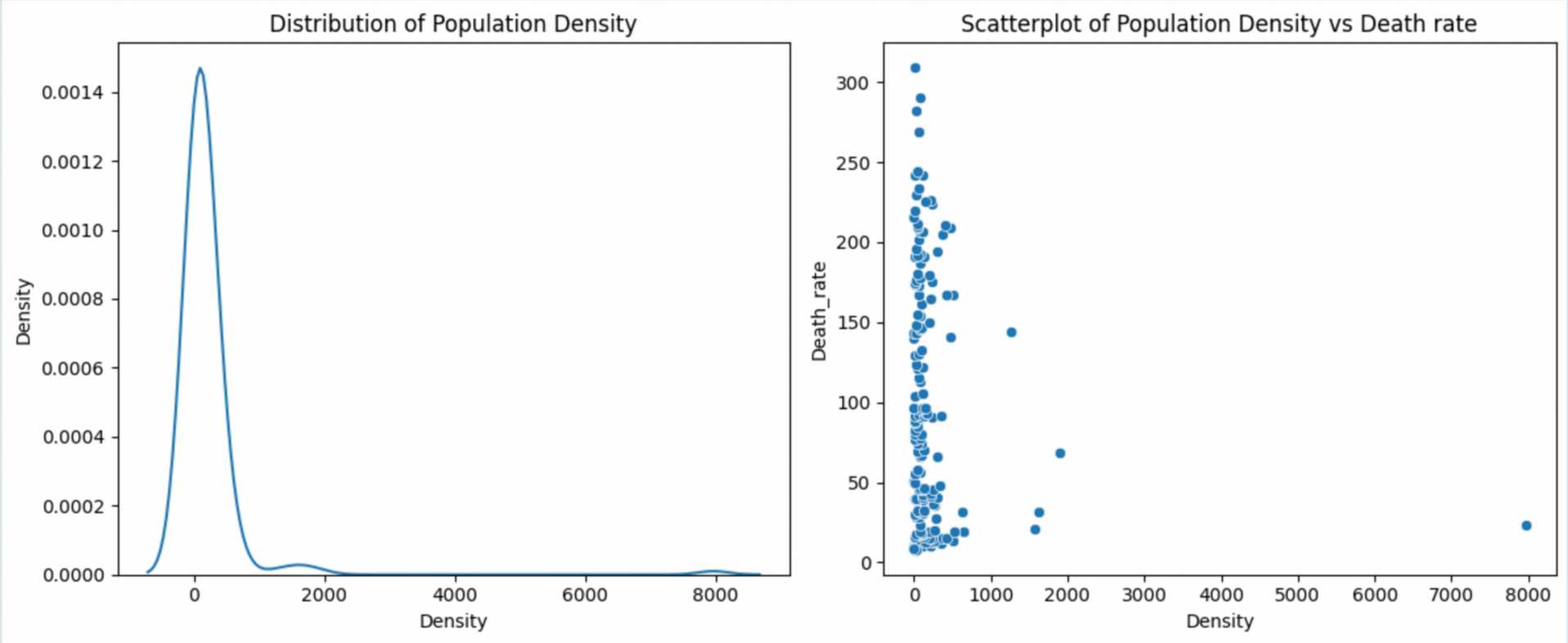
Analysis of Parent Location: Distribution and Relationship with Death Rate



Comments: Western Pacific shows the highest variability. **Africa** also has a wide range and some very high death rates. **Europe** and the **Americas** tend to have lower and more consistent death rates. **South-East Asia** has a narrower distribution, but still moderately high rates.

Exploratory Data Analysis

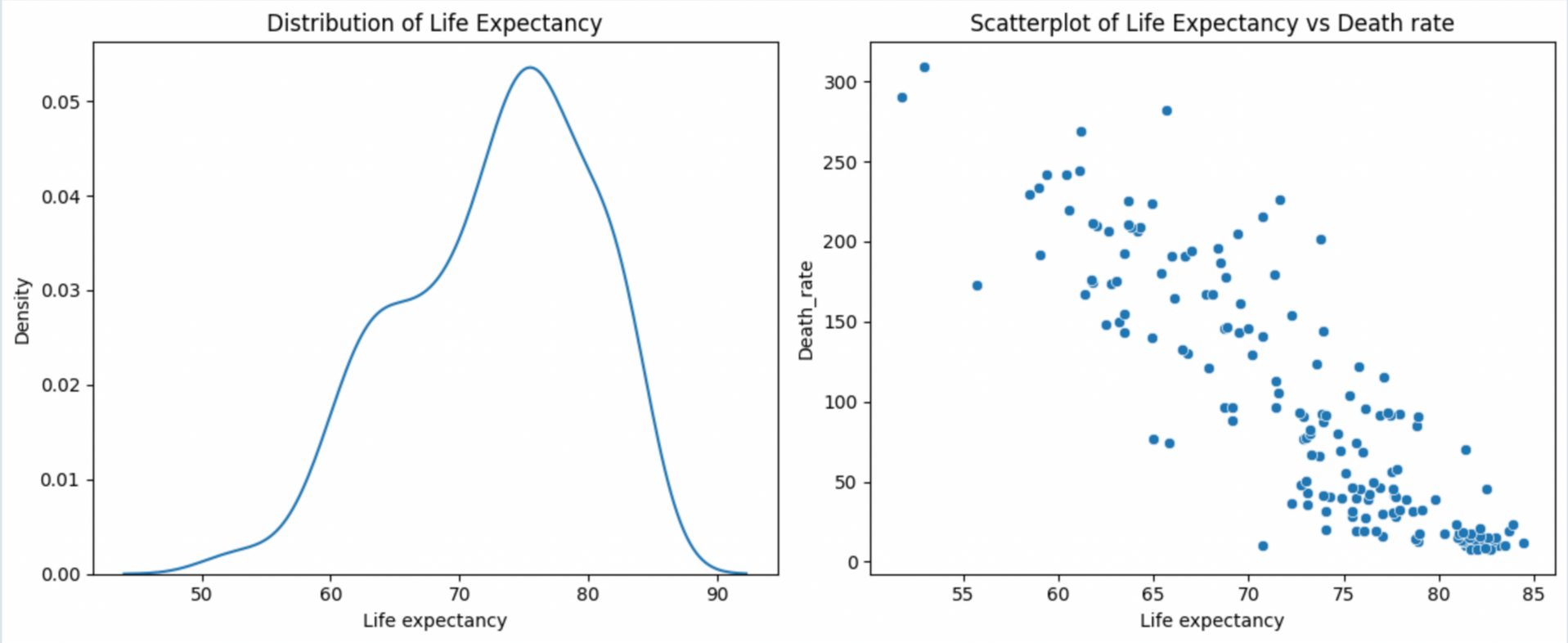
Analysis of Population Density: Distribution and Relationship with Death Rate



- **Mean:** 206.16 people per sq. km.
- **Median:** 84.10 people per sq. km.
- **Standard Deviation:** 656.73 people per sq. km.
- **Skewness:** 10.24
- **Kurtosis:** 118.51

Exploratory Data Analysis

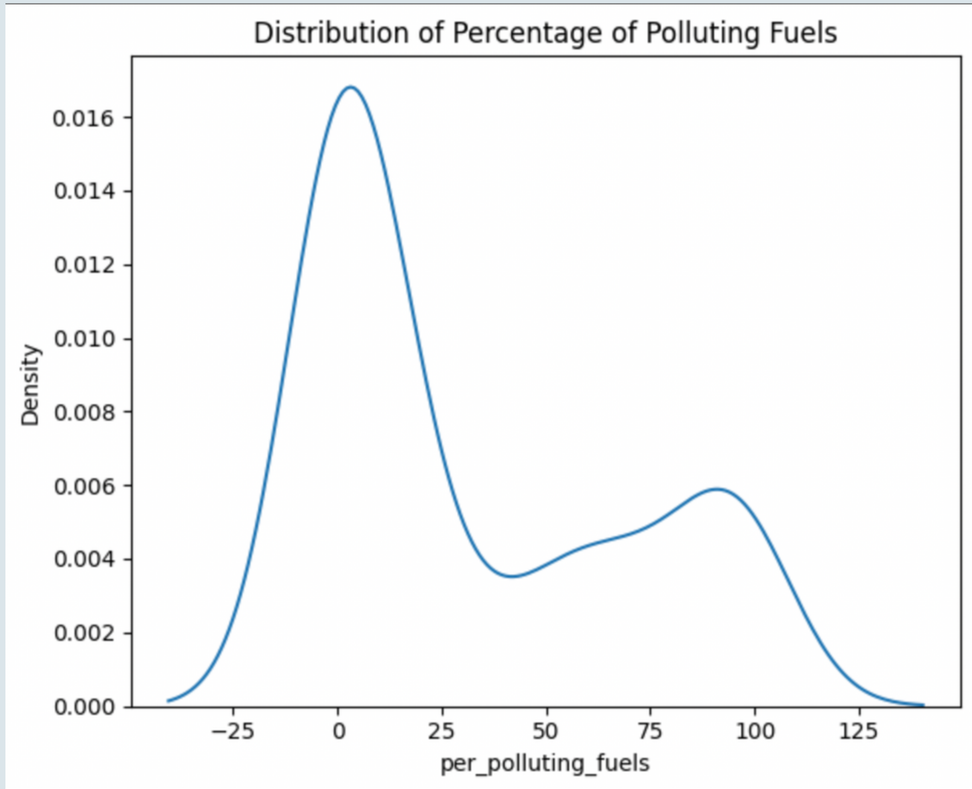
Analysis of Life Expectancy: Distribution and Relationship with Death Rate

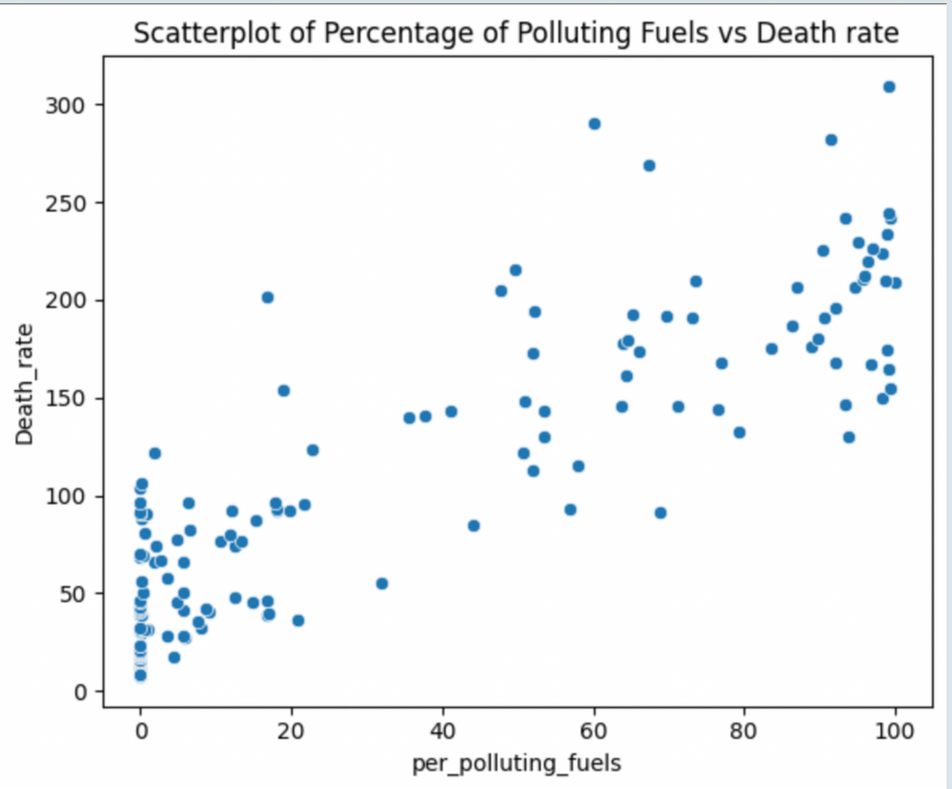


- **Mean:** 72.86 years
- **Median:** 73.92 years
- **Standard Deviation:** 7.26 years
- **Skewness:** -0.51
- **Kurtosis:** -0.45

Exploratory Data Analysis

Analysis of Usage of polluting fuels for cooking (%): Distribution and Relationship with Death Rate





- **Mean:** 31.21 %
- **Median:** 8.90 %
- **Standard Deviation:** 37.53 %
- **Skewness:** 0.76
- **Kurtosis:** -1.09