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

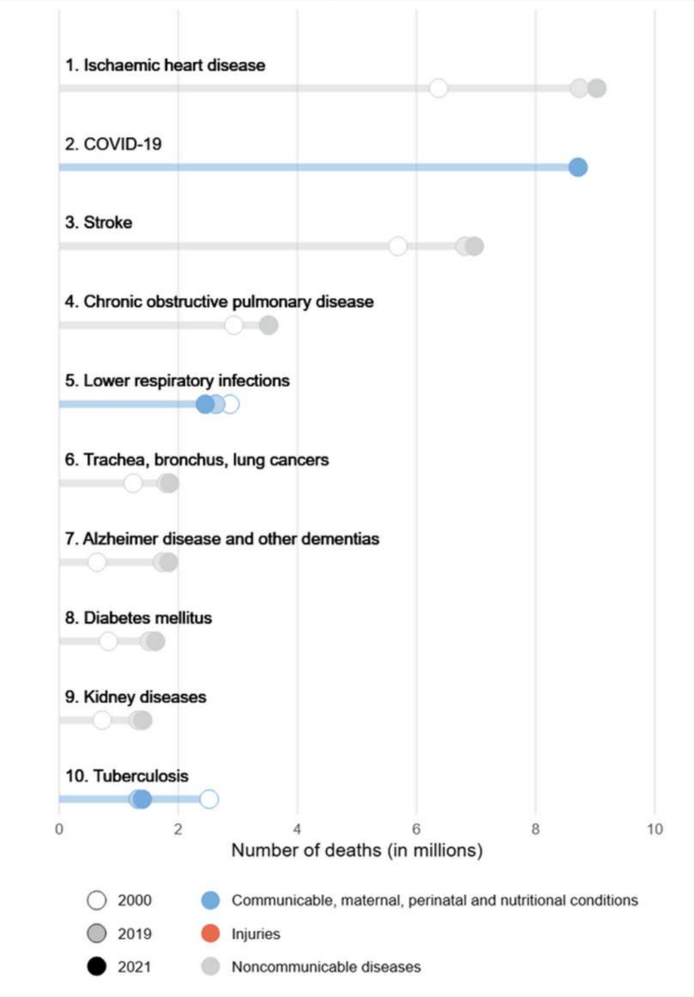
We would like to express our sincere gratitude to **Dr. Monika Bhattacharjee** for providing us with the opportunity, support, and guidance in completing this analysis project on the **"Analysis of Respiratory Death Rates Across Regions: Impact of Socioeconomic and Environmental Factors"** This project 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

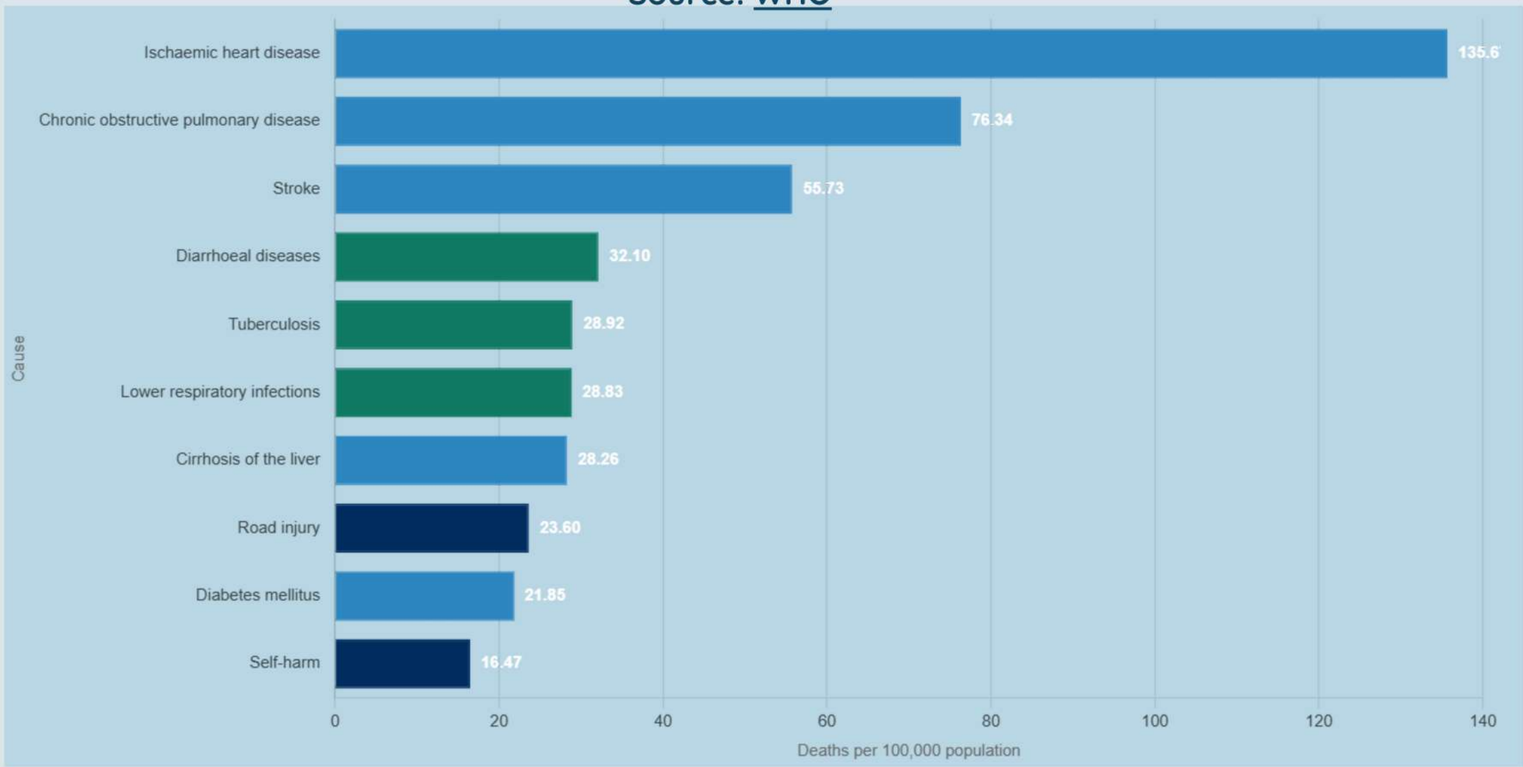
1. **Problem Statement**
2. **Data Sources**
3. **Data Description**
4. **Data Preprocessing**
5. **Exploratory Data Analysis (EDA)**
6. **Model Building: Multiple Linear Regression**
7. **Outlier Removal**
8. **Validation of Assumptions of Linear Regression**
9. **Model Building:**
    a. **Results of Multiple Linear Regression**
    b. **Stepwise Linear Regression**
    c. **Principal Component Regression**
    d. **Ridge Regression**
    e. **Least Absolute Shrinkage and Selection Operator Regression**
    f. **Partial Least Squares Regression**
10. **Model Evaluation**
11. **Conclusion & Future Scope**
12. **References**
13. **Appendix**

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

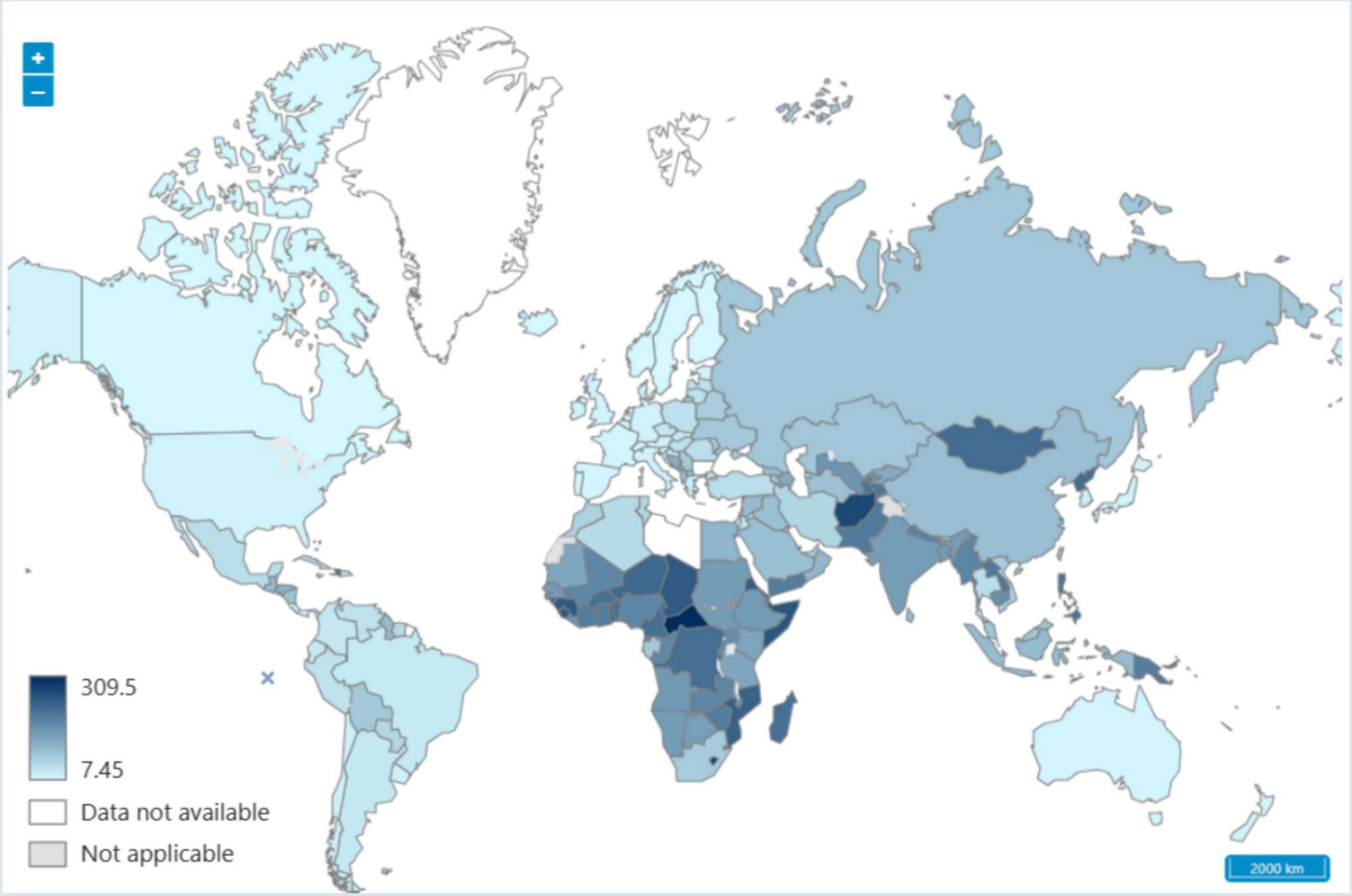


Image 1.3
Source: WHO

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

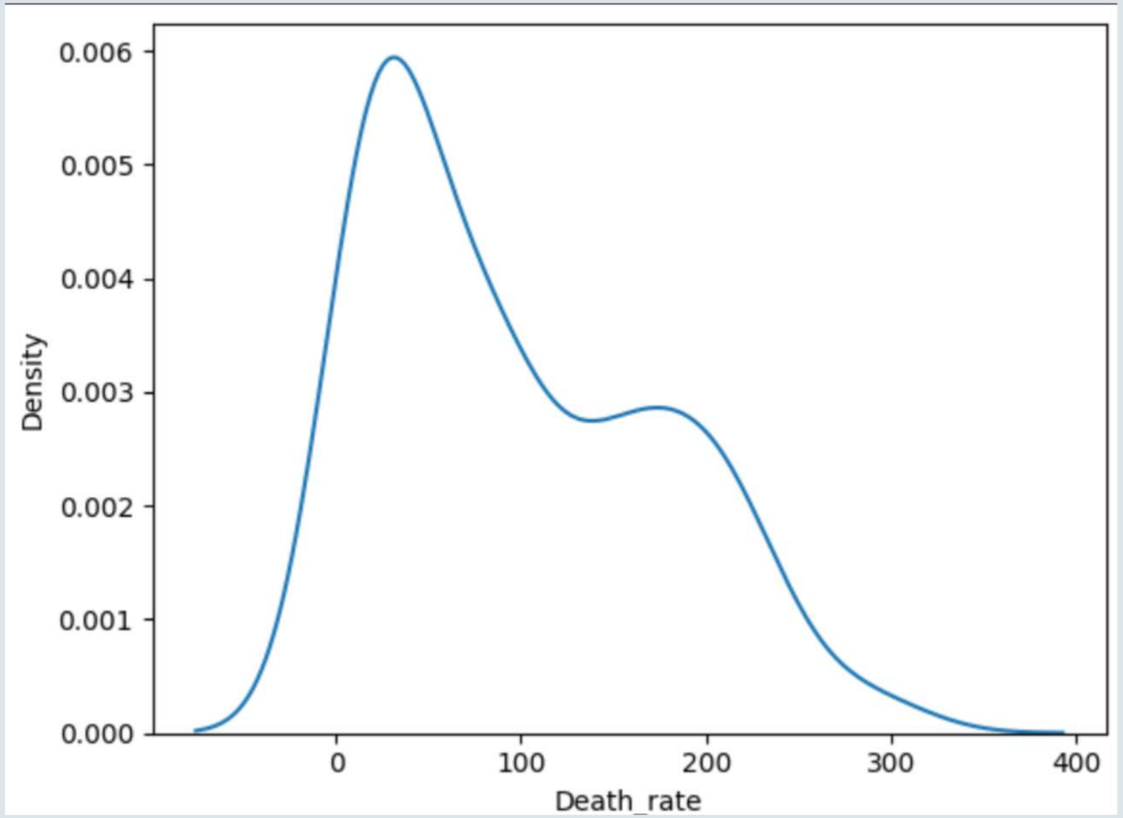
- **Bivariate Analysis:** The relationship between each predictor and Death Rate was visualized using:
  - **Scatter plots** to detect linear or non-linear associations.
  - **Box plots** for categorical variables to compare the distribution of Death Rate across different categories.

Exploratory Data Analysis

Distribution of Death Rate(Y)

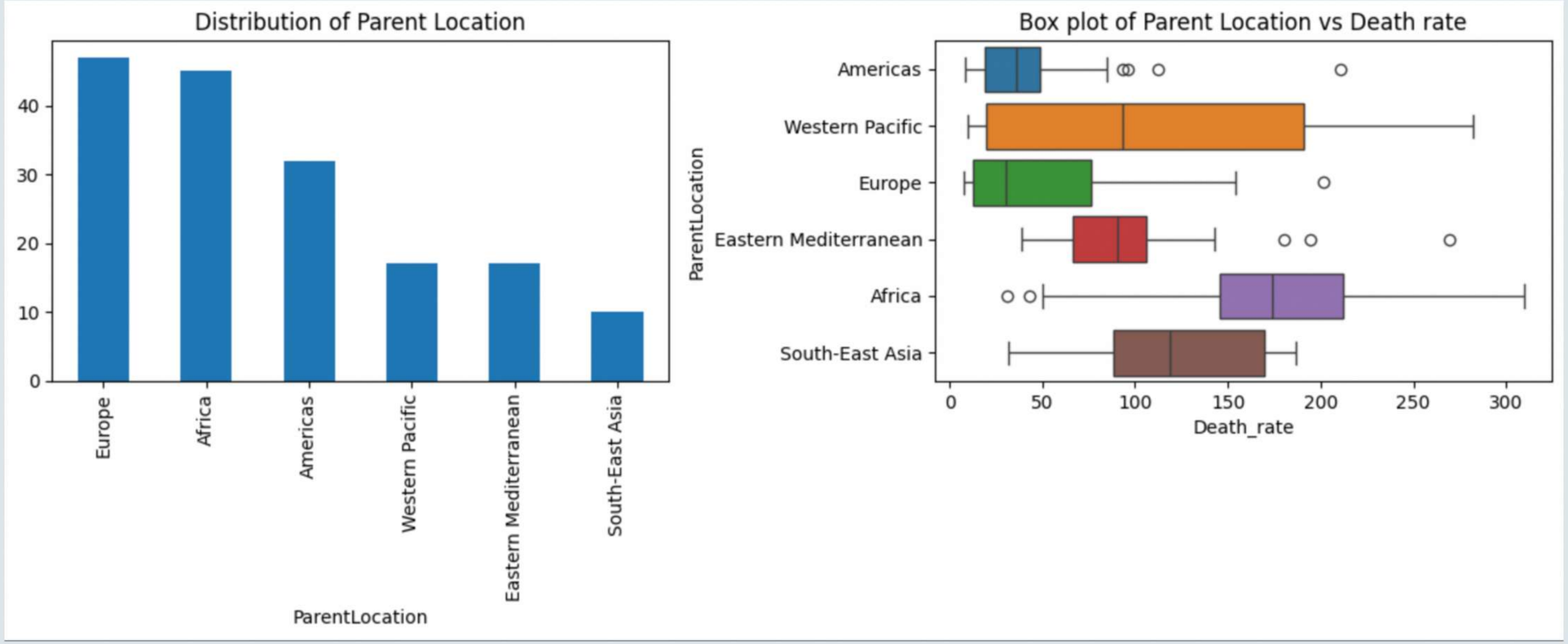
Analysis:

- **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

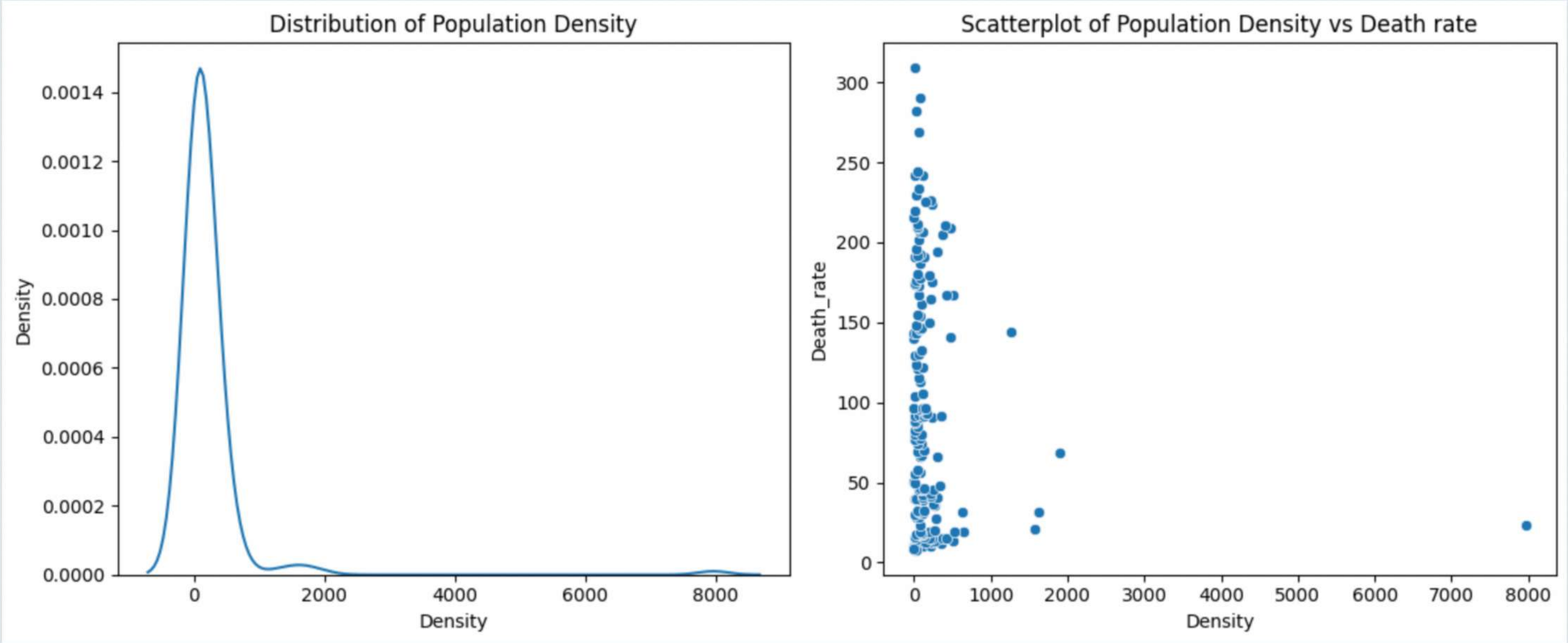
Analysis of Parent Location: Distribution and Relationship with Death Rate



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

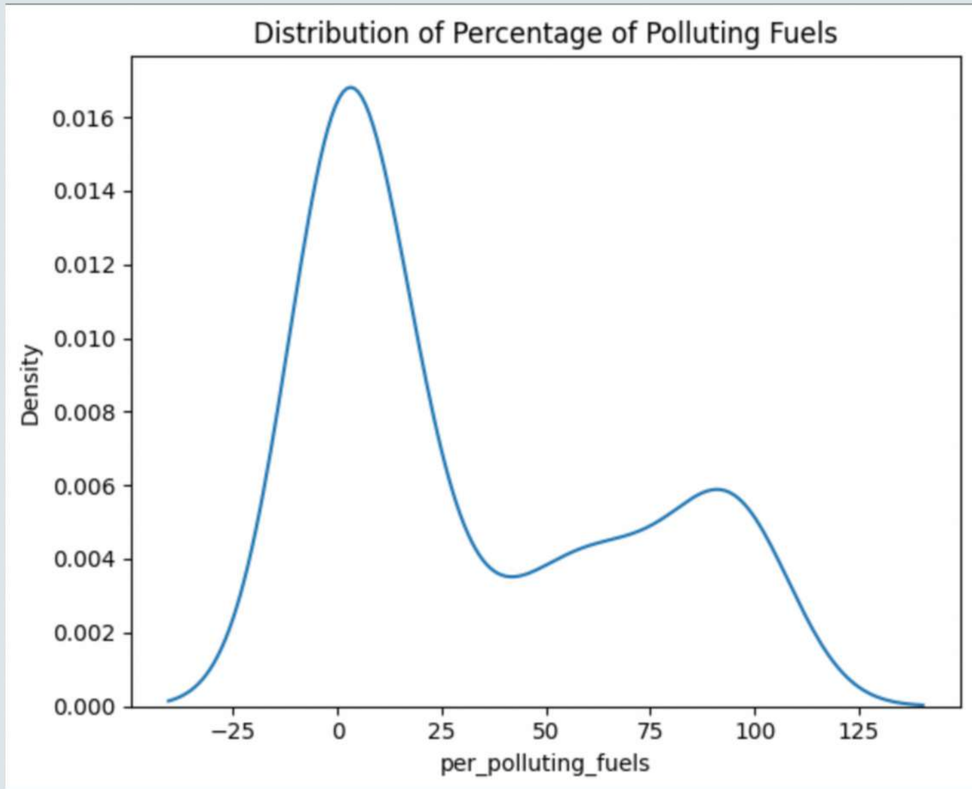
Analysis of Life Expectancy: Distribution and Relationship with Death Rate

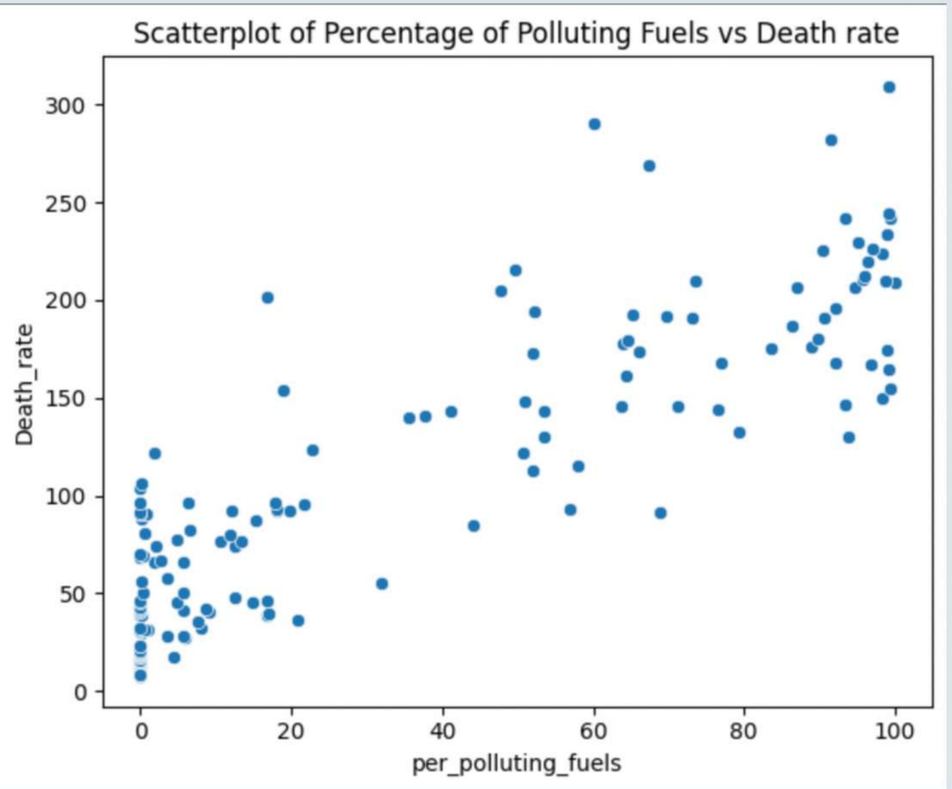
- **Mean:** 72.86 years
- **Median:** 73.92 years
- **Standard Deviation:** 7.26 years

- **Skewness:** -0.51
- **Kurtosis:** -0.45

Exploratory Data Analysis

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

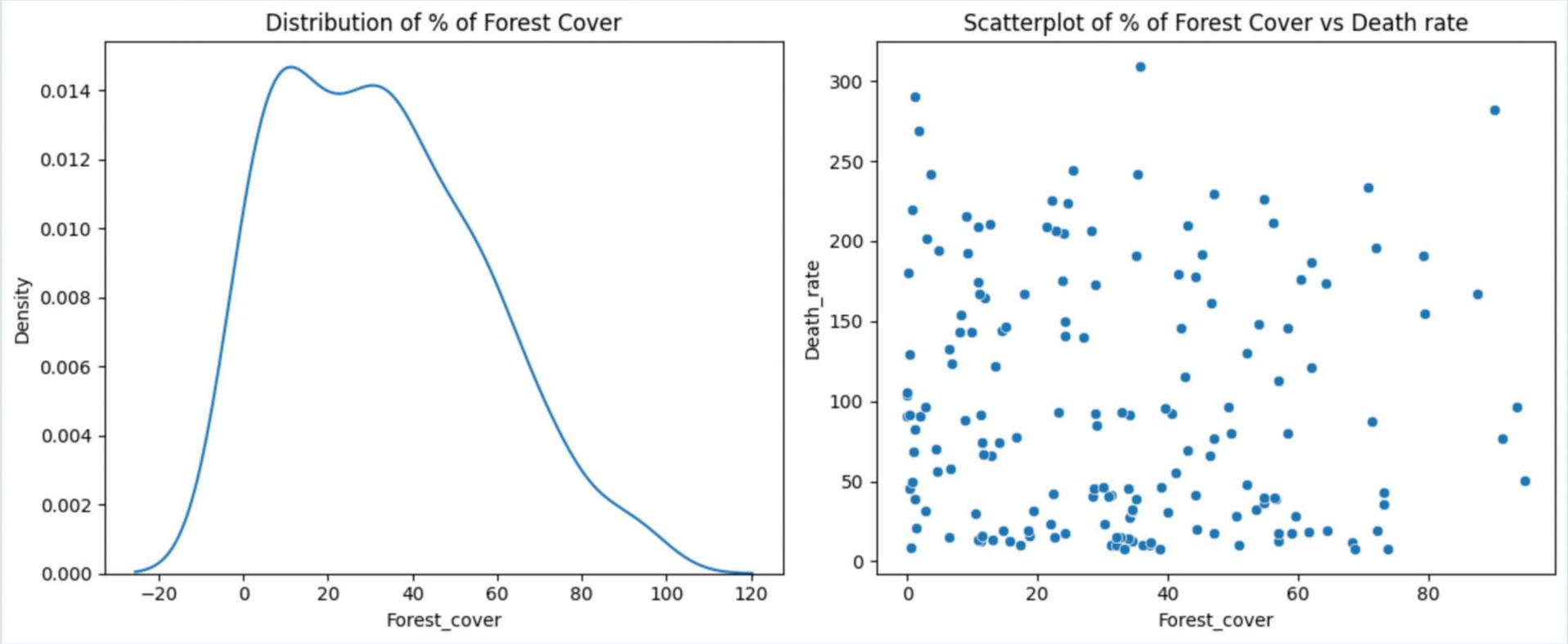




- **Mean:** 31.21 %
- **Median:** 8.90 %
- **Standard Deviation:** 37.53 %
- **Skewness:** 0.76
- **Kurtosis:** -1.09

Exploratory Data Analysis

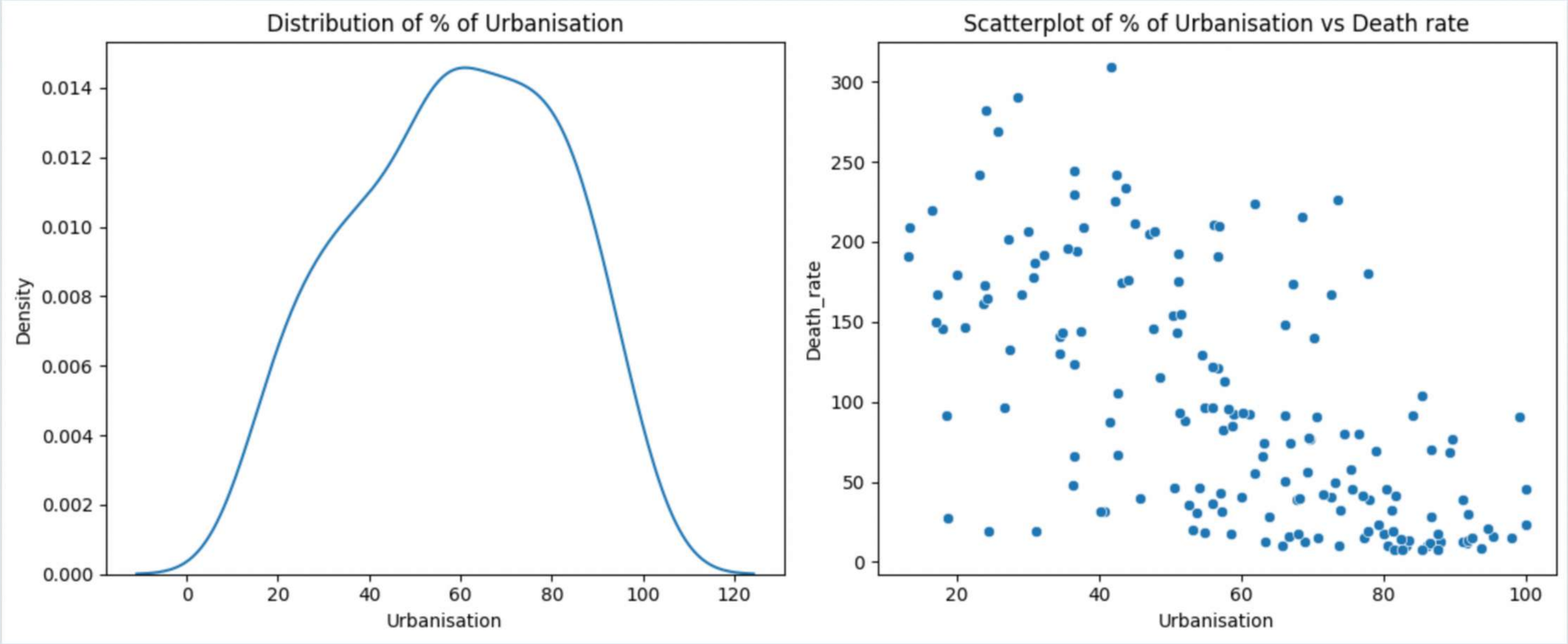
Analysis of Forest Cover: Distribution and Relationship with Death Rate



- **Mean:** 32.53 %
- **Median:** 31.09 %
- **Standard Deviation:** 23.70 %
- **Skewness:** 0.53
- **Kurtosis:** -0.45

Exploratory Data Analysis

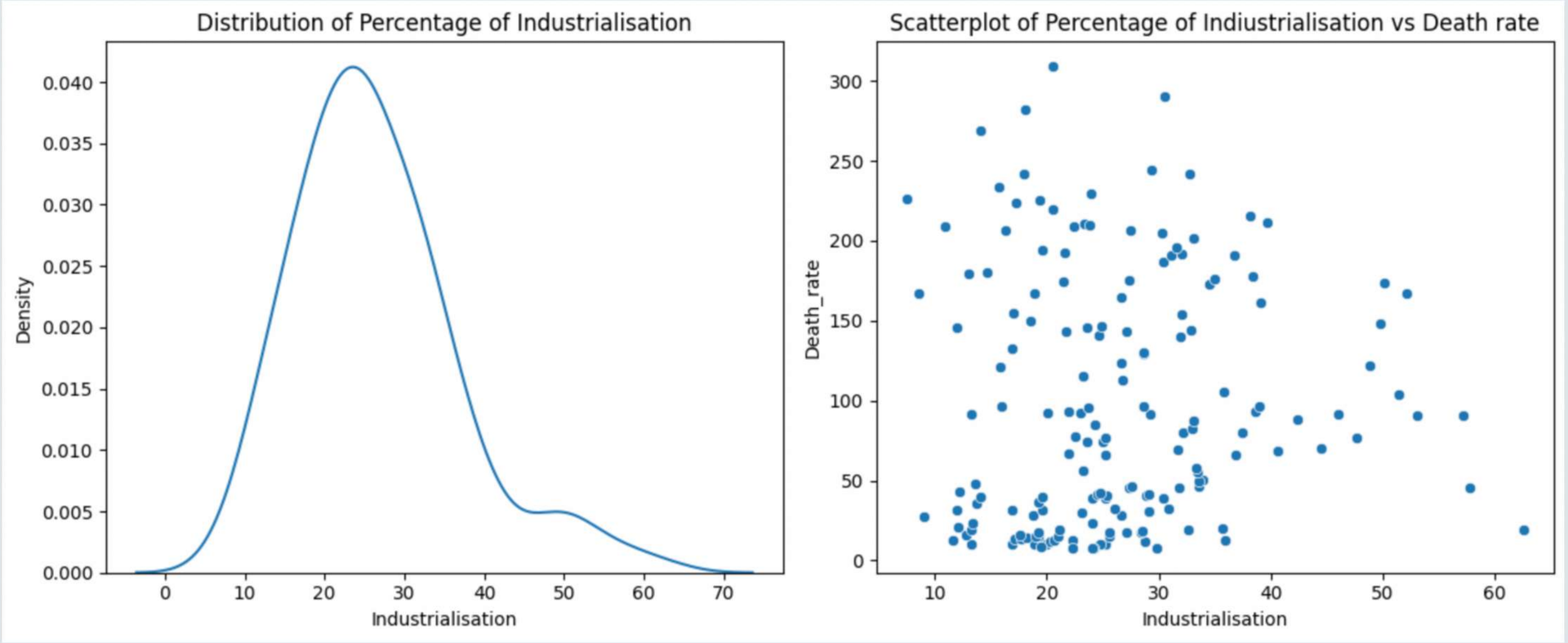
Analysis of Urbanisation: Distribution and Relationship with Death Rate



- **Mean:** 58.74 %
- **Median:** 58.64%
- **Standard Deviation:** 22.50%
- **Skewness:** -0.16
- **Kurtosis:** -0.94

Exploratory Data Analysis

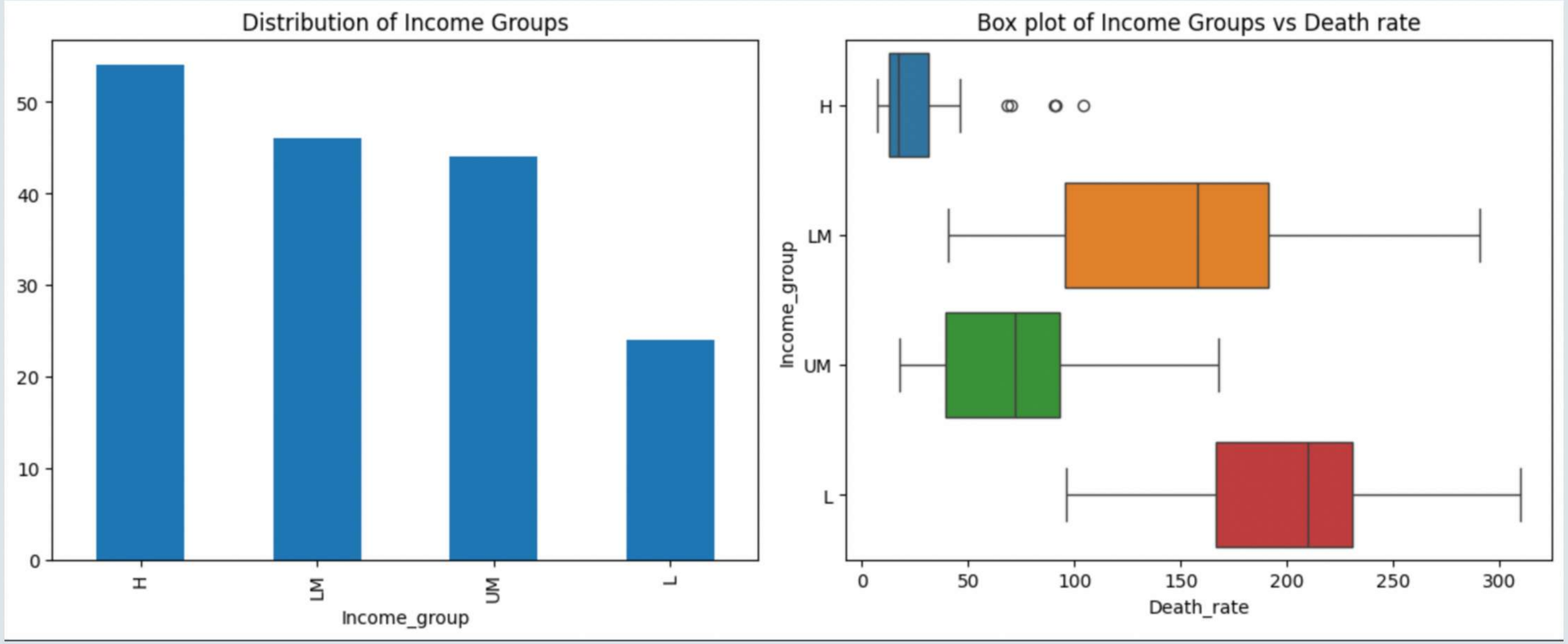
Analysis of Industrialisation: Distribution and Relationship with Death Rate



- **Mean:** 26.39 %
- **Median:** 24.77 %
- **Standard Deviation:** 10.30 %
- **Skewness:** 0.95
- **Kurtosis:** 1.16

Exploratory Data Analysis

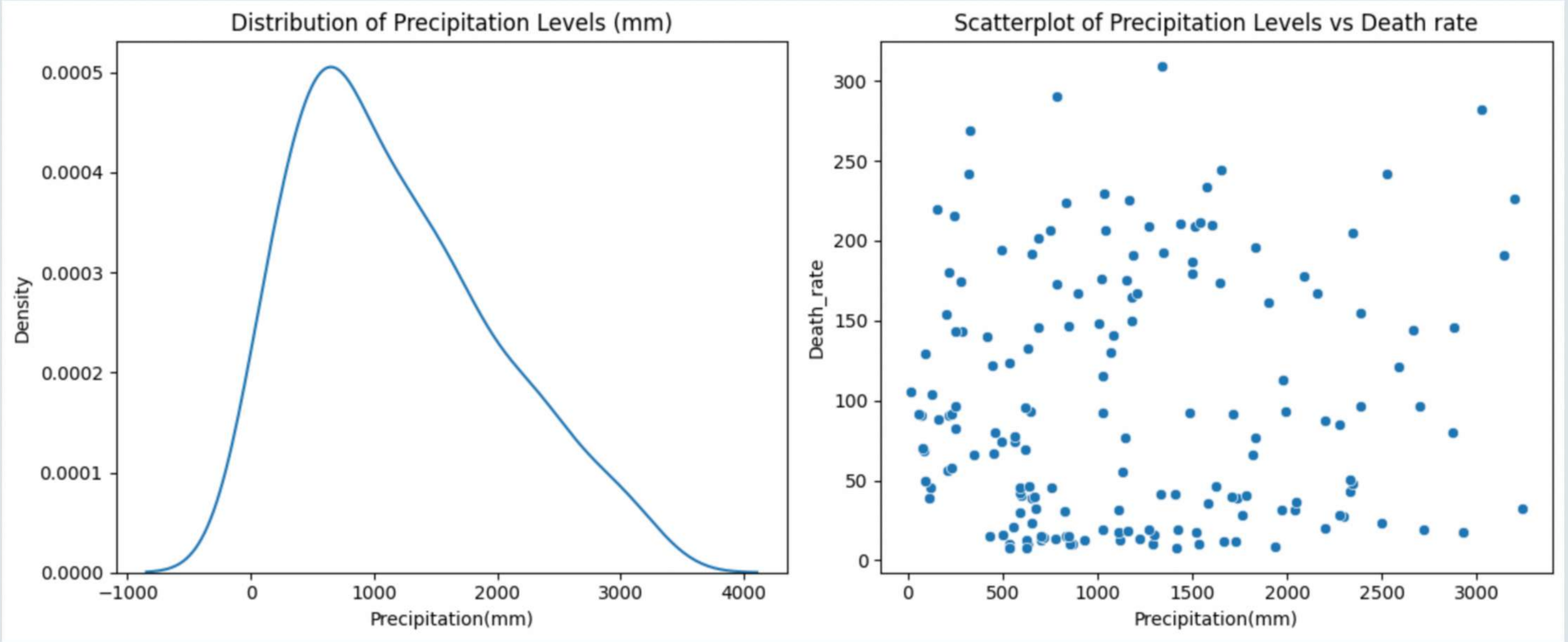
Analysis of Income Group: Distribution and Relationship with Death Rate



Comments: High-income countries have the lowest death rates, while **upper-middle income** countries show slightly higher but still moderate rates. **Lower-middle income** countries have a wider range and higher median death rates. **Low-income** countries exhibit the highest median death rates and more variability, indicating greater burden.

Exploratory Data Analysis

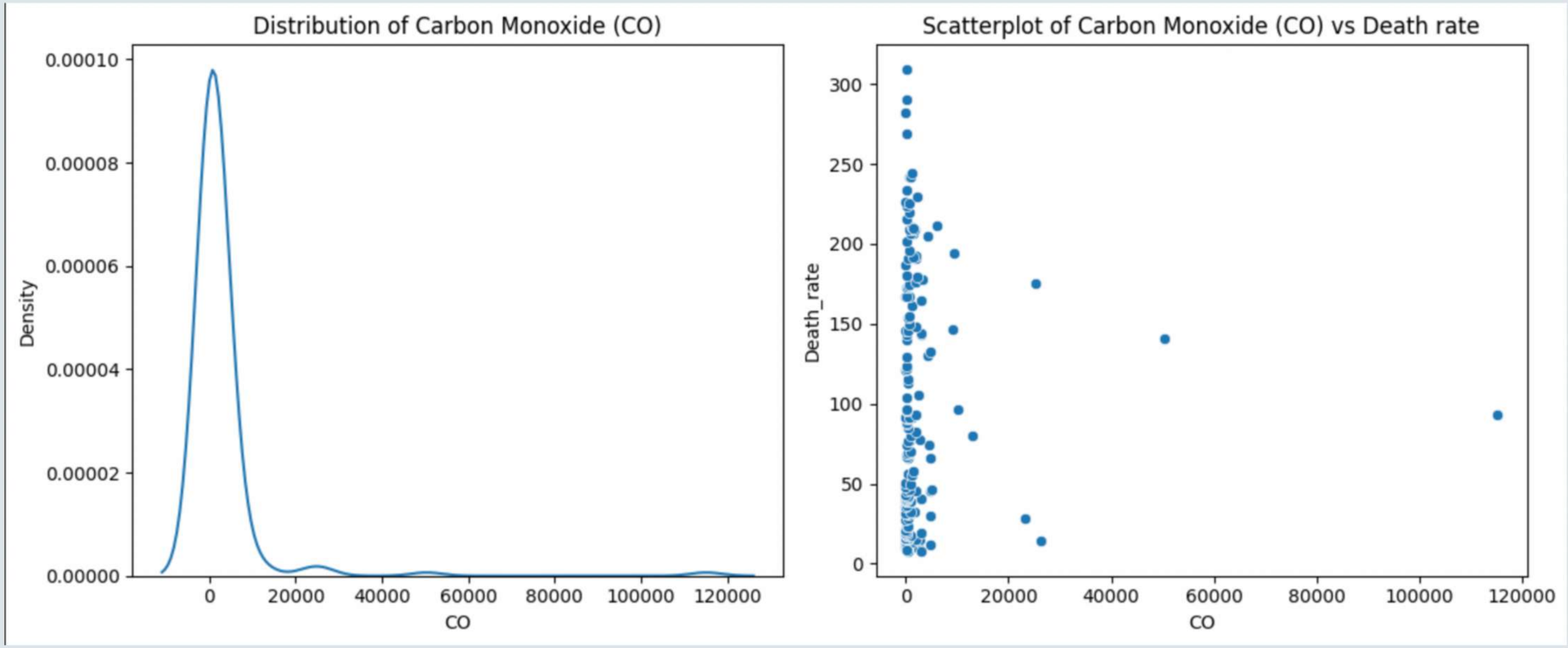
Analysis of Precipitation: Distribution and Relationship with Death Rate



- **Mean:** 1184.16 mm
- **Median:** 1035.50 mm
- **Standard Deviation:** 804.25 mm
- **Skewness:** 0.65
- **Kurtosis:** -0.40

Exploratory Data Analysis

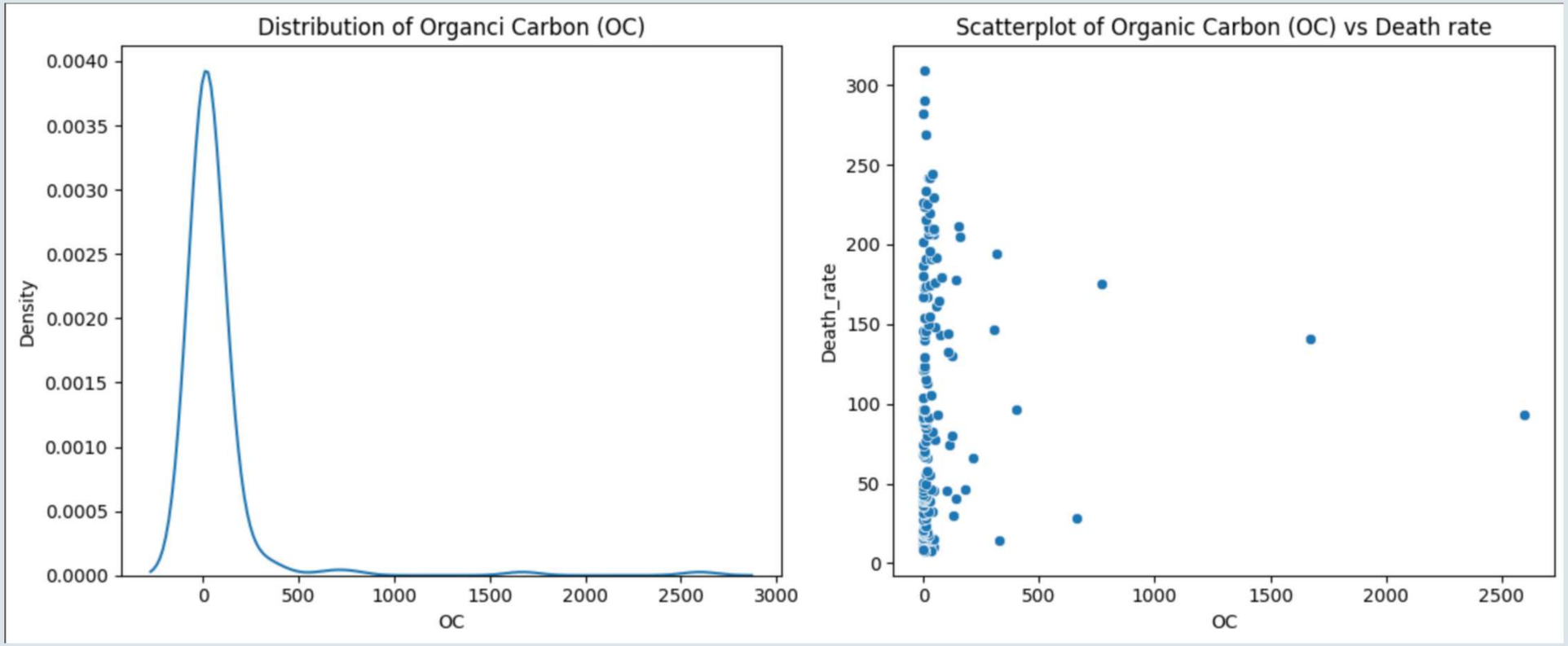
Analysis of CO: Distribution and Relationship with Death Rate



- **Mean:** 2583.72 gigagrams/year
- **Median:** 444.50 gigagrams/year
- **Standard Deviation:** 10199.94 gigagrams/year
- **Skewness:** 8.96
- **Kurtosis:** 92.41

Exploratory Data Analysis

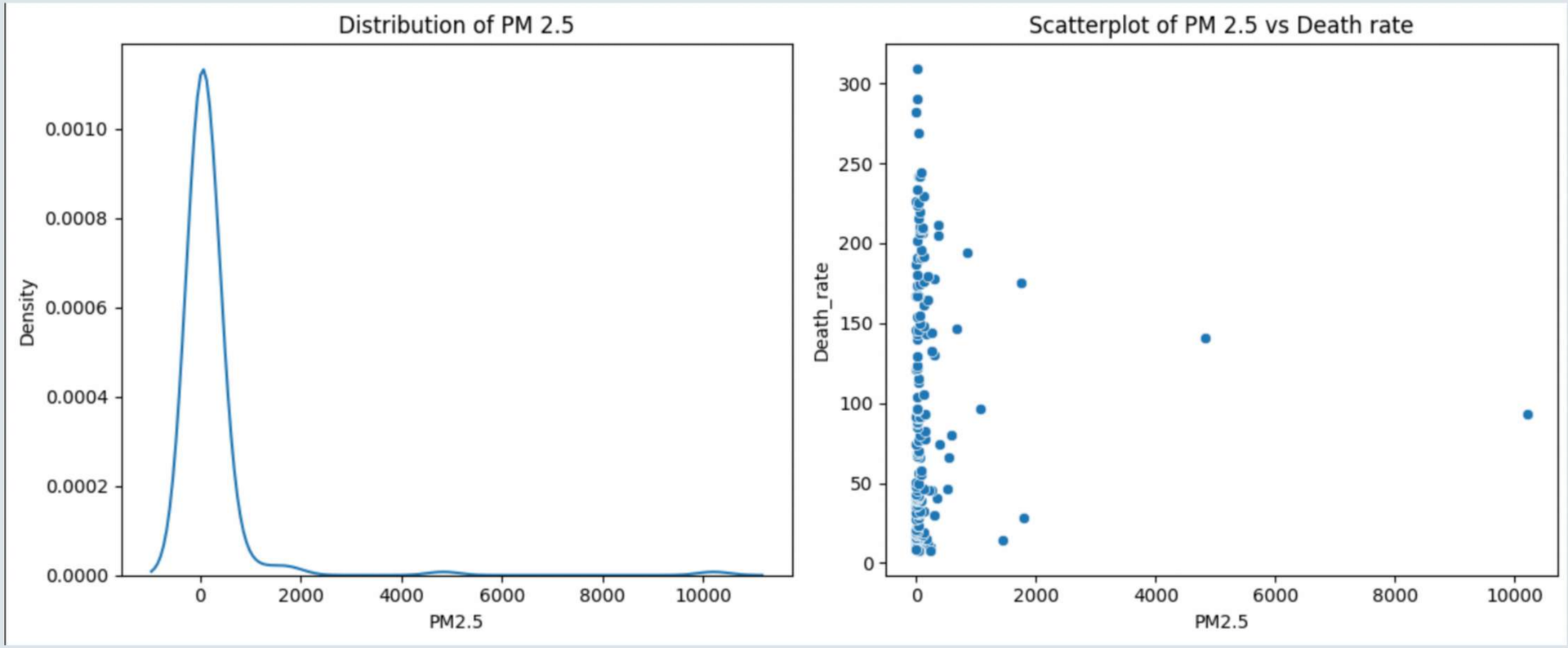
Analysis of OC: Distribution and Relationship with Death Rate



- **Mean:** 65.62 gigagrams/year
- **Median:** 9.48 gigagrams/year
- **Standard Deviation:** 252.57 gigagrams/year
- **Skewness:** 7.91
- **Kurtosis:** 70.46

Exploratory Data Analysis

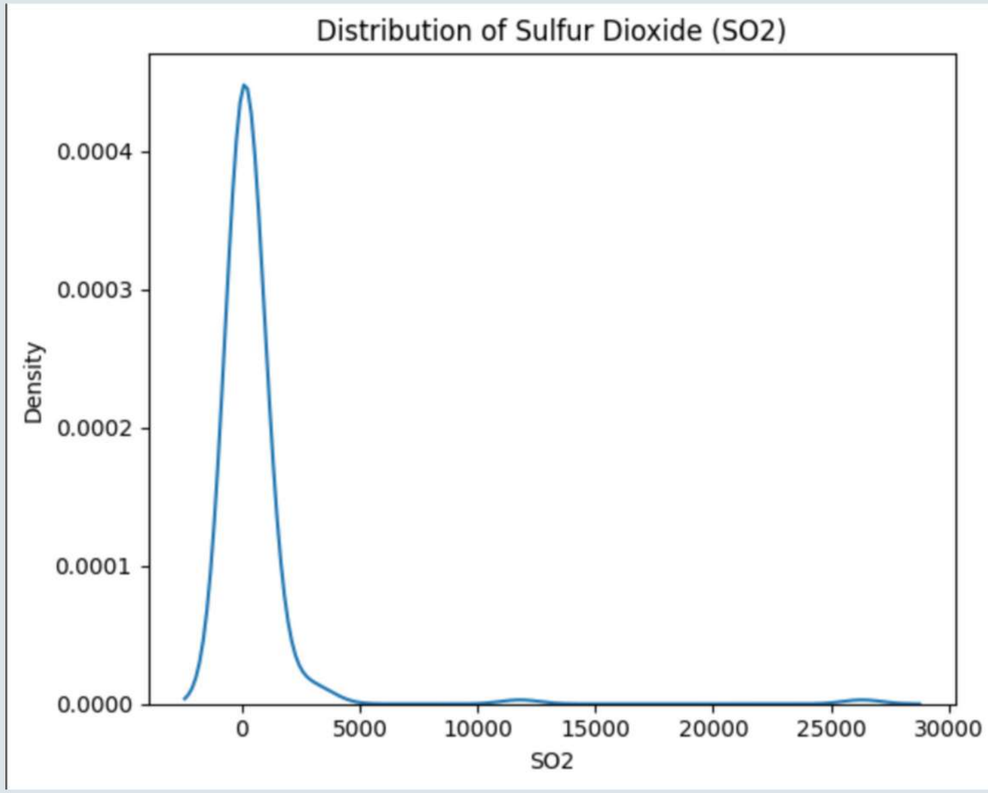
Analysis of PM2.5: Distribution and Relationship with Death Rate

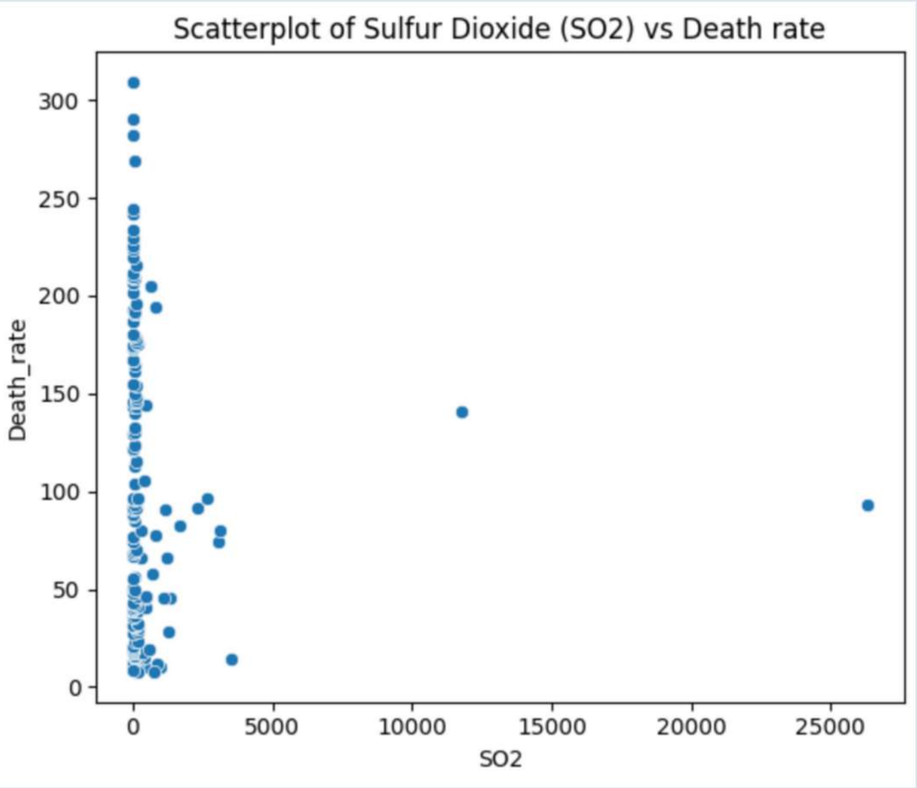


- **Mean:** 204.02 gigagrams/year
- **Median:** 34.99 gigagrams/year
- **Standard Deviation:** 895.38 gigagrams/year
- **Skewness:** 9.37
- **Kurtosis:** 98.18

Exploratory Data Analysis

Analysis of SO2: Distribution and Relationship with Death Rate

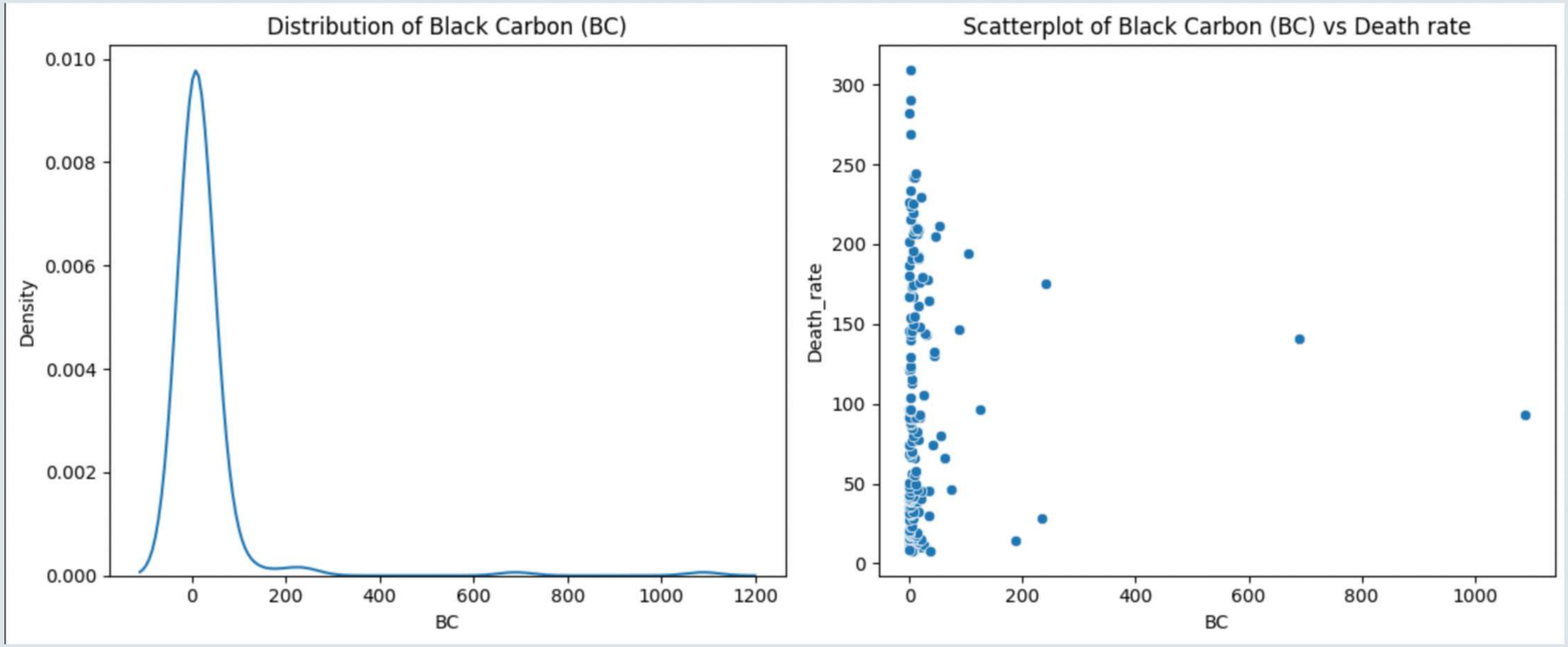




- **Mean:** 465.92 gigagrams/year
- **Median:** 48.03 gigagrams/year
- **Standard Deviation:** 2265.58 gigagrams/year
- **Skewness:** 9.81
- **Kurtosis:** 106.18

Exploratory Data Analysis

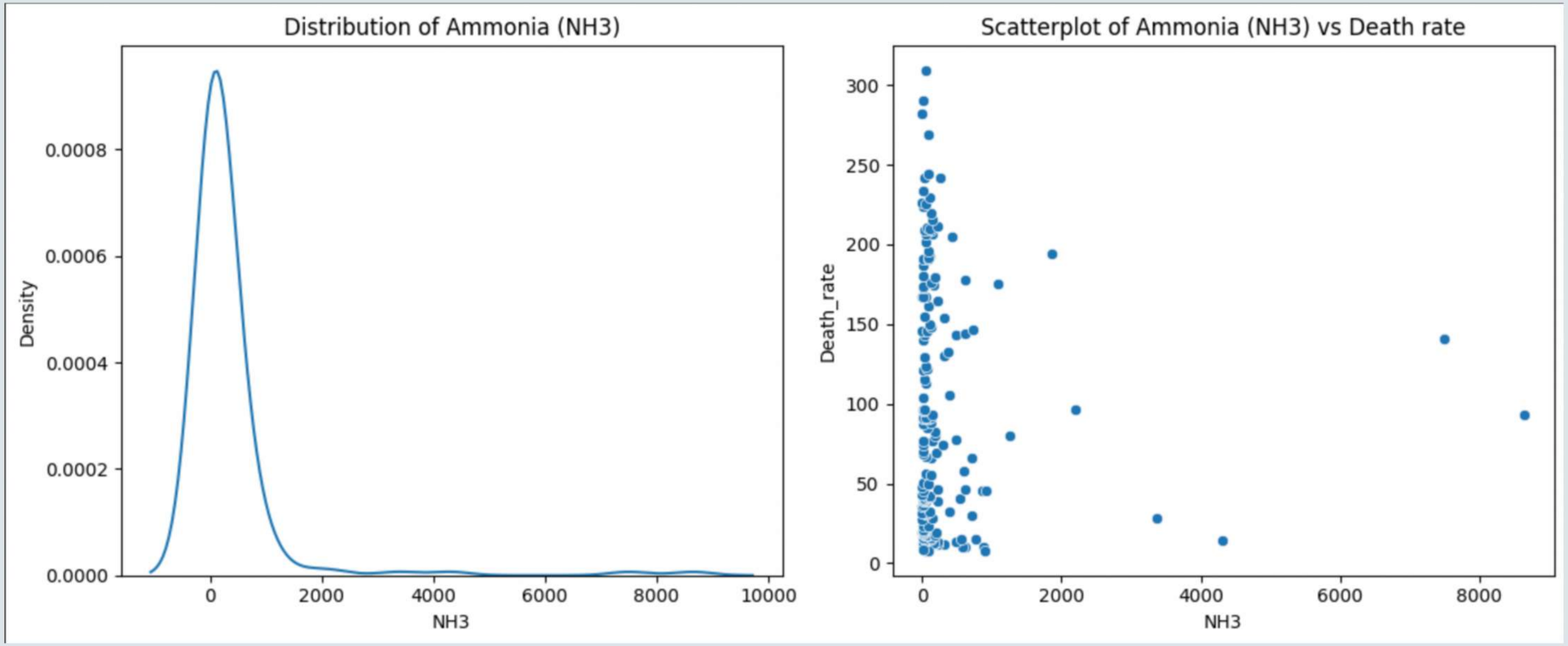
Analysis of BC: Distribution and Relationship with Death Rate



- **Mean:** 25.61 gigagrams/year
- **Median:** 4.61 gigagrams/year
- **Standard Deviation:** 103.04 gigagrams/year
- **Skewness:** 8.39
- **Kurtosis:** 78.00

Exploratory Data Analysis

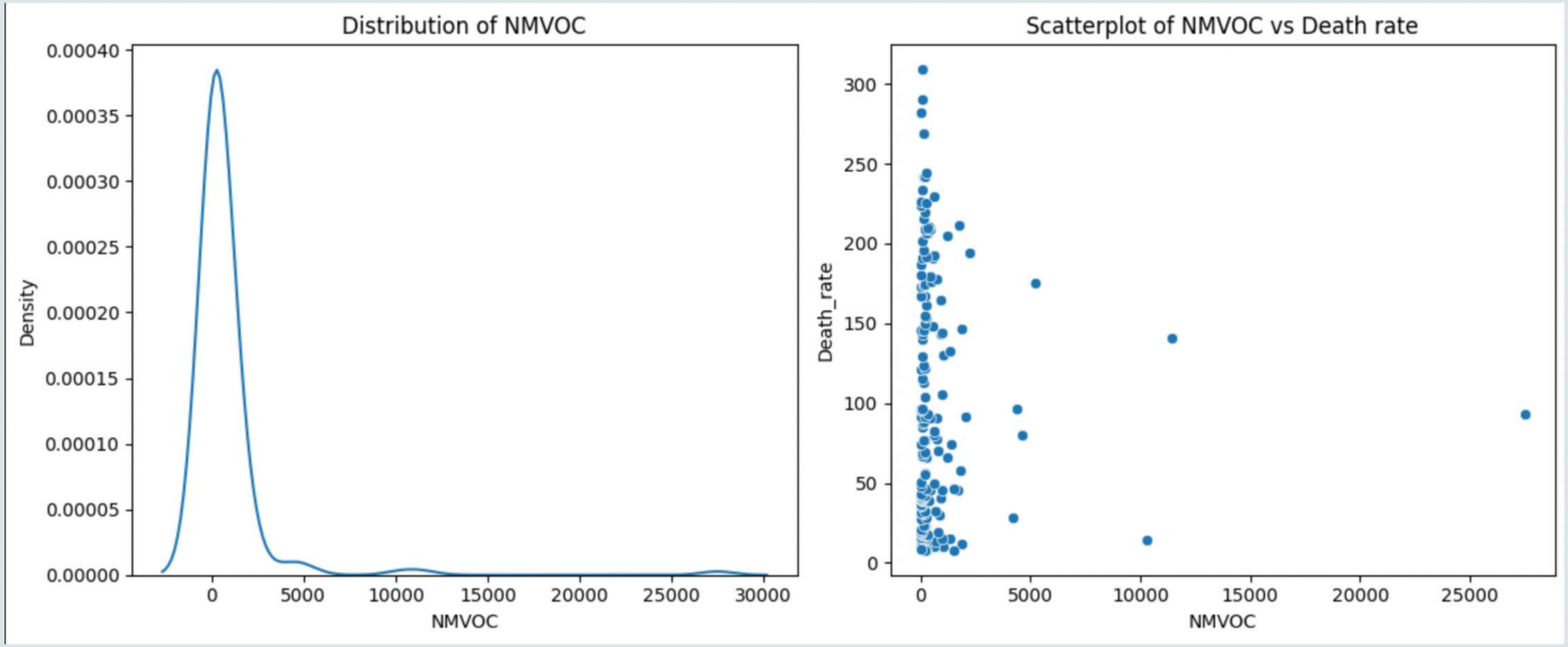
Analysis of NH3: Distribution and Relationship with Death Rate



- **Mean:** 323.89 gigagrams/year
- **Median:** 76.93 gigagrams/year
- **Standard Deviation:** 994.45 gigagrams/year
- **Skewness:** 6.45
- **Kurtosis:** 46.52

Exploratory Data Analysis

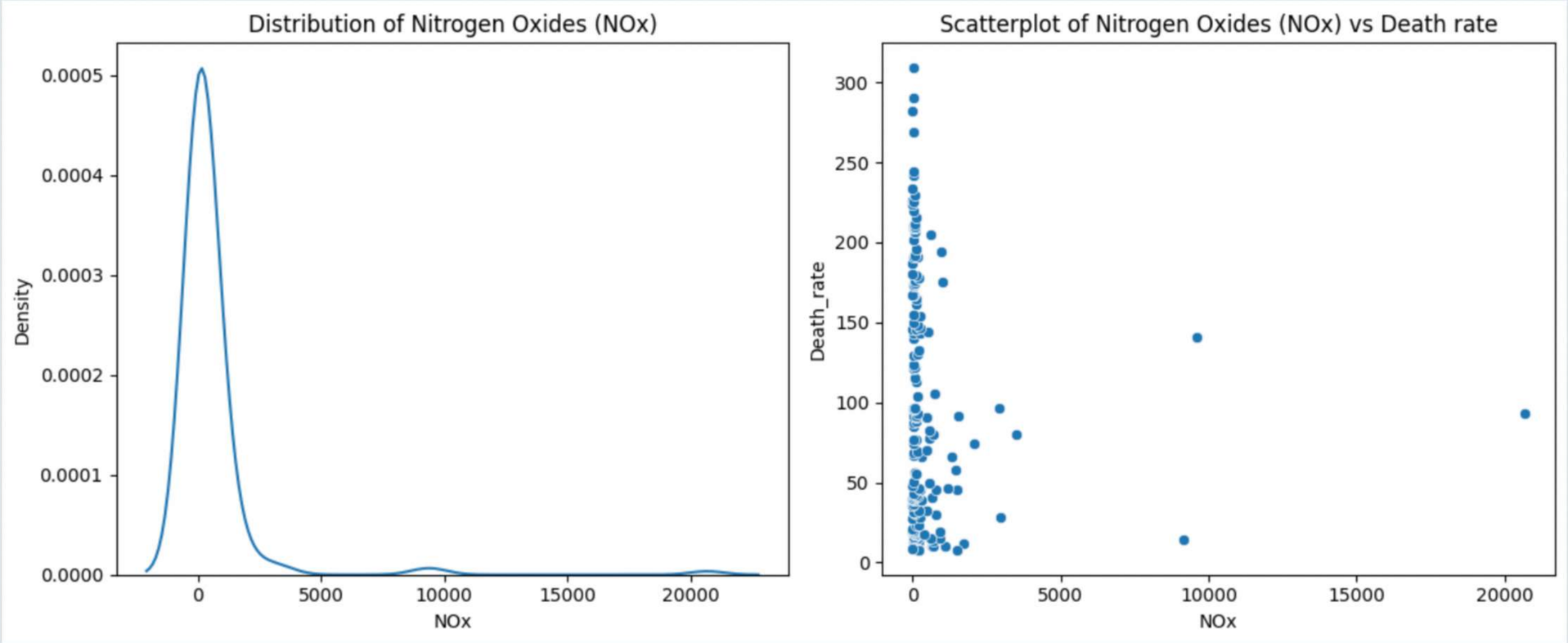
Analysis of NMVOC: Distribution and Relationship with Death Rate



- **Mean:** 751.29 gigagrams/year
- **Median:** 172.29 gigagrams/year
- **Standard Deviation:** 2497.64 gigagrams/year
- **Skewness:** 8.34
- **Kurtosis:** 82.20

Exploratory Data Analysis

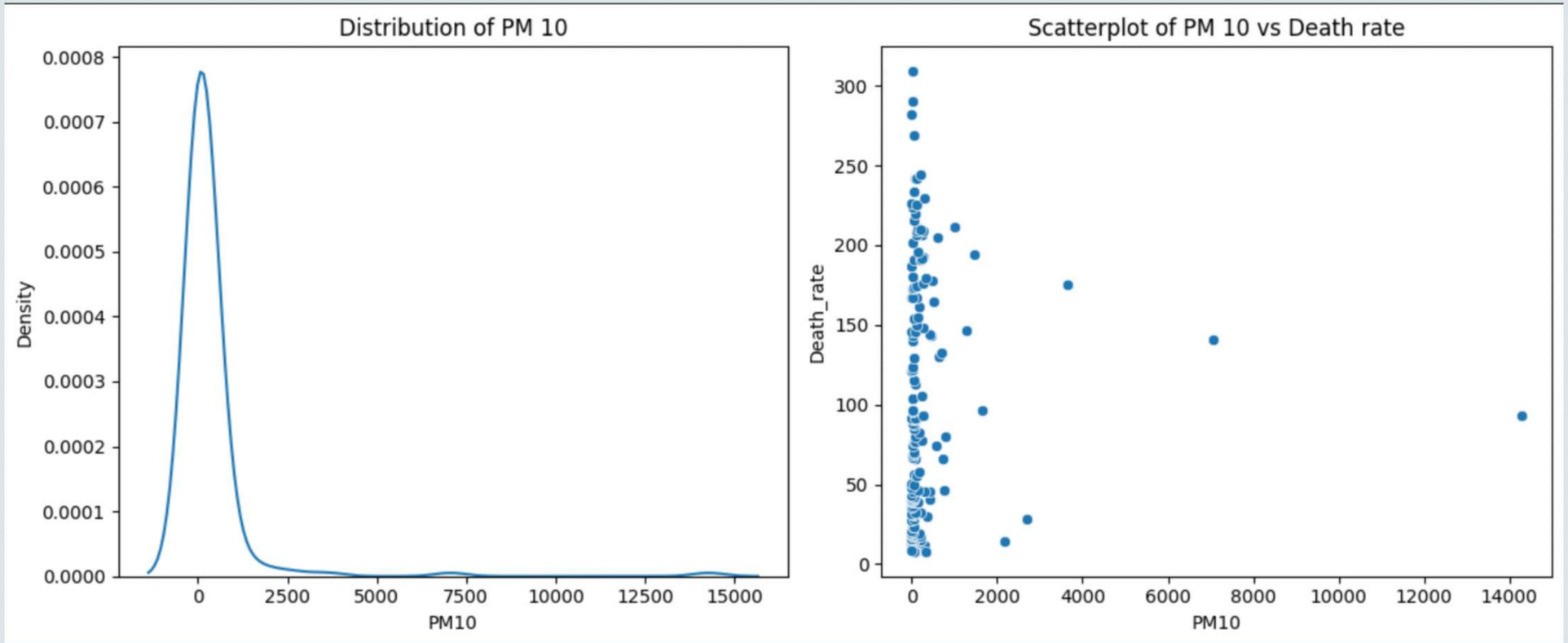
Analysis of NOx: Distribution and Relationship with Death Rate



- **Mean:** 519.93 gigagrams/year
- **Median:** 86.45 gigagrams/year
- **Standard Deviation:** 1928.22 gigagrams/year
- **Skewness:** 8.15
- **Kurtosis:** 76.76

Exploratory Data Analysis

Analysis of PM10: Distribution and Relationship with Death Rate



- **Mean:** 323.10 gigagrams/year
- **Median:** 55.03 gigagrams/year
- **Standard Deviation:** 1284.61 gigagrams/year
- **Skewness:** 8.81
- **Kurtosis:** 88.47

Exploratory Data Analysis

Summary of the Analysis:

- Death rates show **noticeable variation across different regions**, highlighting geographic influence on health outcomes.
- There is **no clear relationship** between **population density** and death rate.
- **Higher life expectancy** is generally associated with **lower** death rates.
- Death rates **tend to rise** with **higher usage of polluting fuels for cooking**.
- The graph suggests **no visible link** between **forest cover** and death rate, nor between **industrialisation** and death rate
- **Higher urbanisation** seems to be linked to **lower** death rates
- **Lower-income** groups tend to have **higher** death rates, while **higher-income** groups show **lower** and more stable death rates.
- **Precipitation levels** appear to have **minimal impact** on the death rate.
- **All the polluting particles** (e.g. CO, NH3, PM2.5, PM10) **do not seem to affect** the death rate and nearly all of them have a very **high variance**.

Multiple Linear Regression Model

Multiple Linear Regression models the relationship between one dependent variable and several independent variables, helping to predict outcomes and understand the combined effect of multiple factors.

Equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Where:

- Y = response variable
- $X_1, X_2, \ldots, X_p$ = predictor (independent) variables
- β_0 = intercept term
- $\beta_1, \beta_2, \ldots, \beta_p$ = regression coefficients
- ε = random error term
- n = total number of observations in the dataset
- p = total number of predictors (independent variables)

Outlier & Influence Diagnostics

In regression analysis, certain data points can have a disproportionate impact on the model's results. These observations — known as **outliers or influential points** — can **distort coefficients, inflate errors, and lead to misleading conclusions** if not properly identified and addressed.

This section explores key diagnostic tools — including **Leverage, Cook's Distance, DFBETAs, DFFITs and COVRATIO** — used to:

- Detect **outliers and influential observations**
- Assess their **impact on model stability**
- Guide decisions on **data cleaning** or **model refinement**

Outlier & Influence Diagnostics

1. Leverage

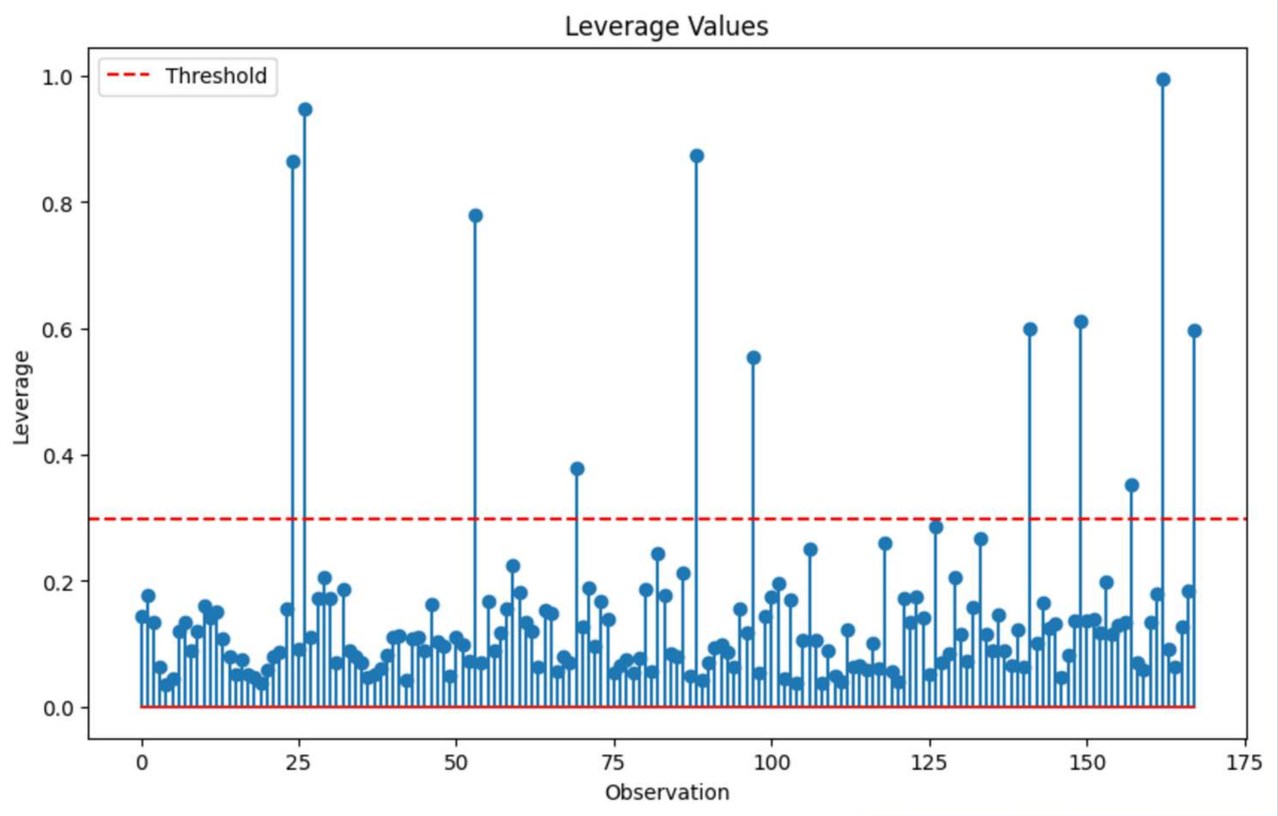
Purpose: Leverage identifies observations with extreme predictor values that can strongly influence model predictions.

Threshold: i^{th} point is a leverage point if

Leverage> 2(p+1)/n

Interpretation from the plot:

- Most points have leverage values well below the threshold, indicating normal predictor influence.
- 13 observations exceeded the leverage threshold



Outlier & Influence Diagnostics

2. Cook's Distance

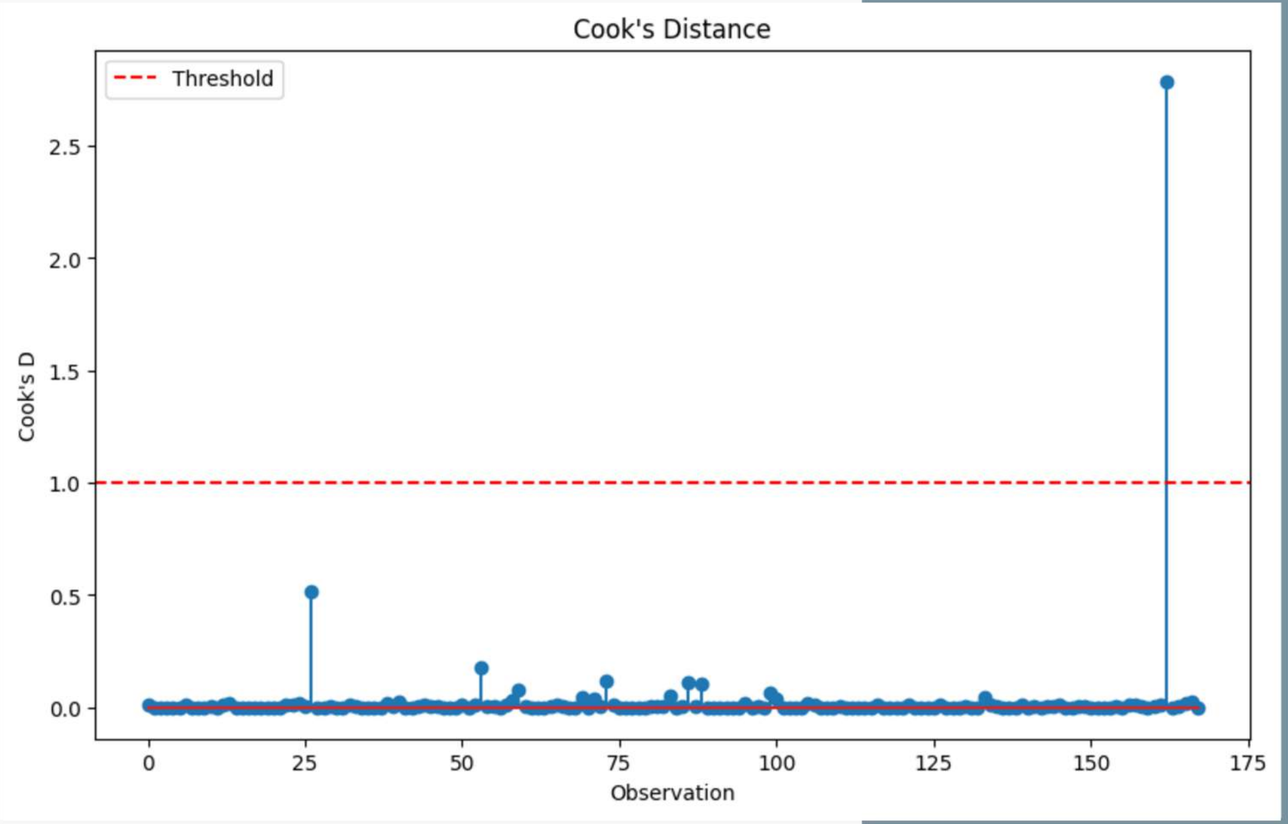
Purpose: Cook's Distance detects points that both have high leverage and large residuals, making them especially impactful on the model's fit. High Cook's Distance means the point affects both coefficients and predictions

Threshold: i^{th} point is an influential point if

Cook's Distance(i) > 1

Interpretation from the plot:

- Almost all points are under the threshold; a single observation show mild global influence.



Outlier & Influence Diagnostics

3. DFBETA

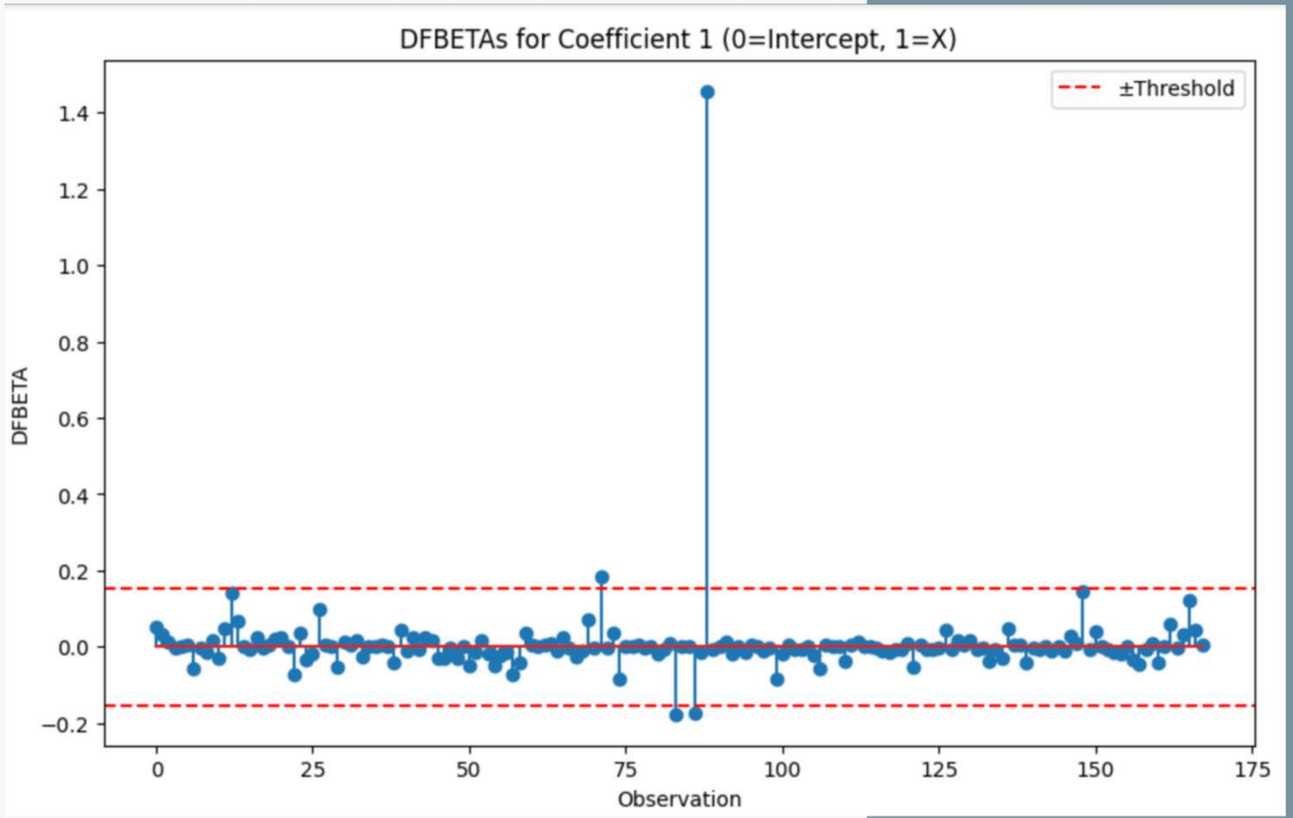
Purpose: It measures how much a specific regression coefficient (β_j) changes when observation (i) is removed, helping identify observations that strongly influence individual predictors.

Threshold: i^{th} point is influential if

$$| DFEBTAj,(i) | < 2 \times \sqrt{(1 / n)}$$

Interpretation from the plot:

- This plot shows the influence of each observation on the Population Density.
- Most points lie within the safe ± threshold.
- A clear outlier at index ~85 strongly influences the coefficient (DFBETA > 1.4).
- Similar checks were done for all other coefficients.



Outlier & Influence Diagnostics

4. DFFIT

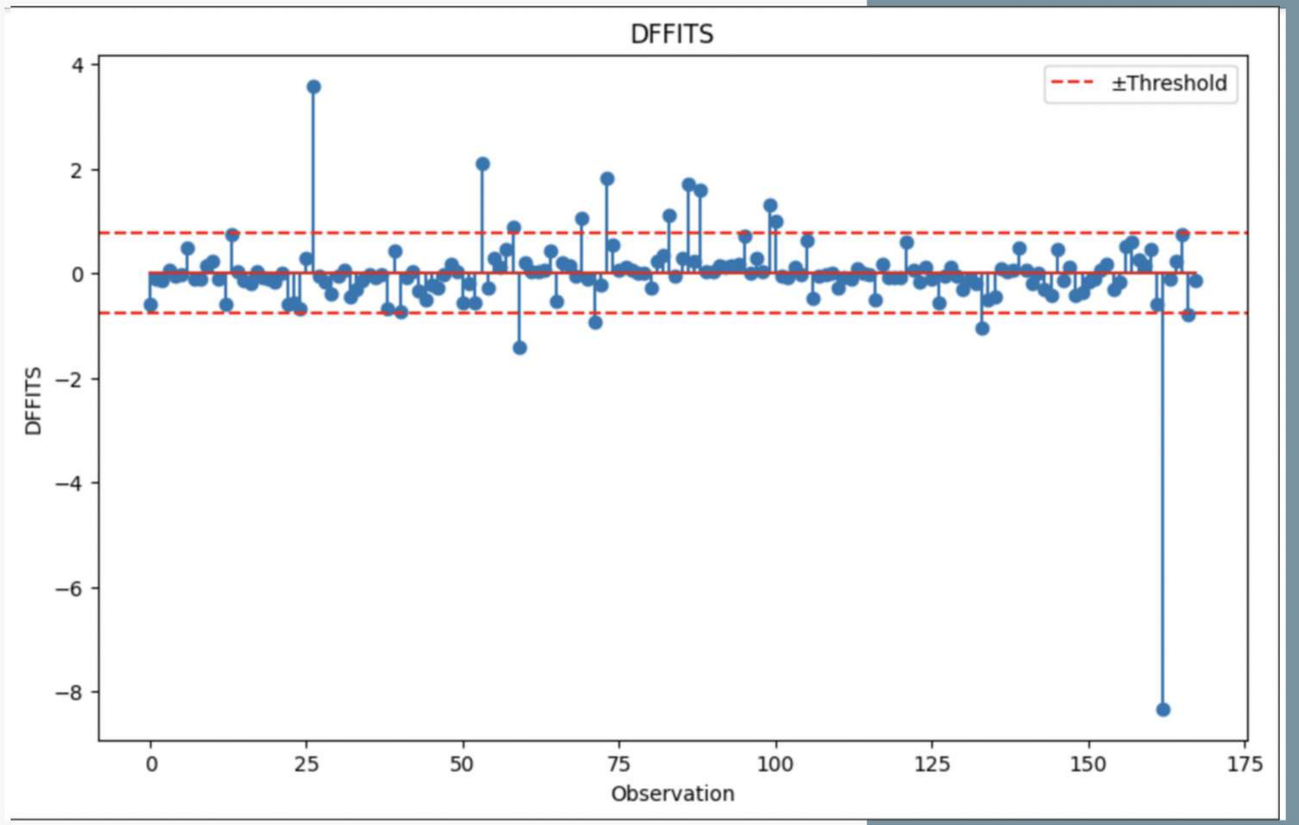
Purpose: Measures how much the predicted value for an observation changes when that observation is excluded from the model. It identifies points that have a strong influence on the fitted values.

Threshold: i^{th} point is influential if

$$|DFFITS_i| > 2 \times \sqrt{((p+1) / n)}$$

Interpretation from the plot:

- Most points lie within the ± threshold bounds, indicating limited influence.
- A major outlier is present near index ~170 with DFFITS < -8, indicating very strong influence.
- Other smaller spikes between indexes 25–100 also cross the threshold slightly.



Outlier & Influence Diagnostics

5. COVRATIO

Purpose: It measures how much the covariance matrix of the regression coefficients changes when a particular observation is removed from the model, thereby helping to detect influential outliers.

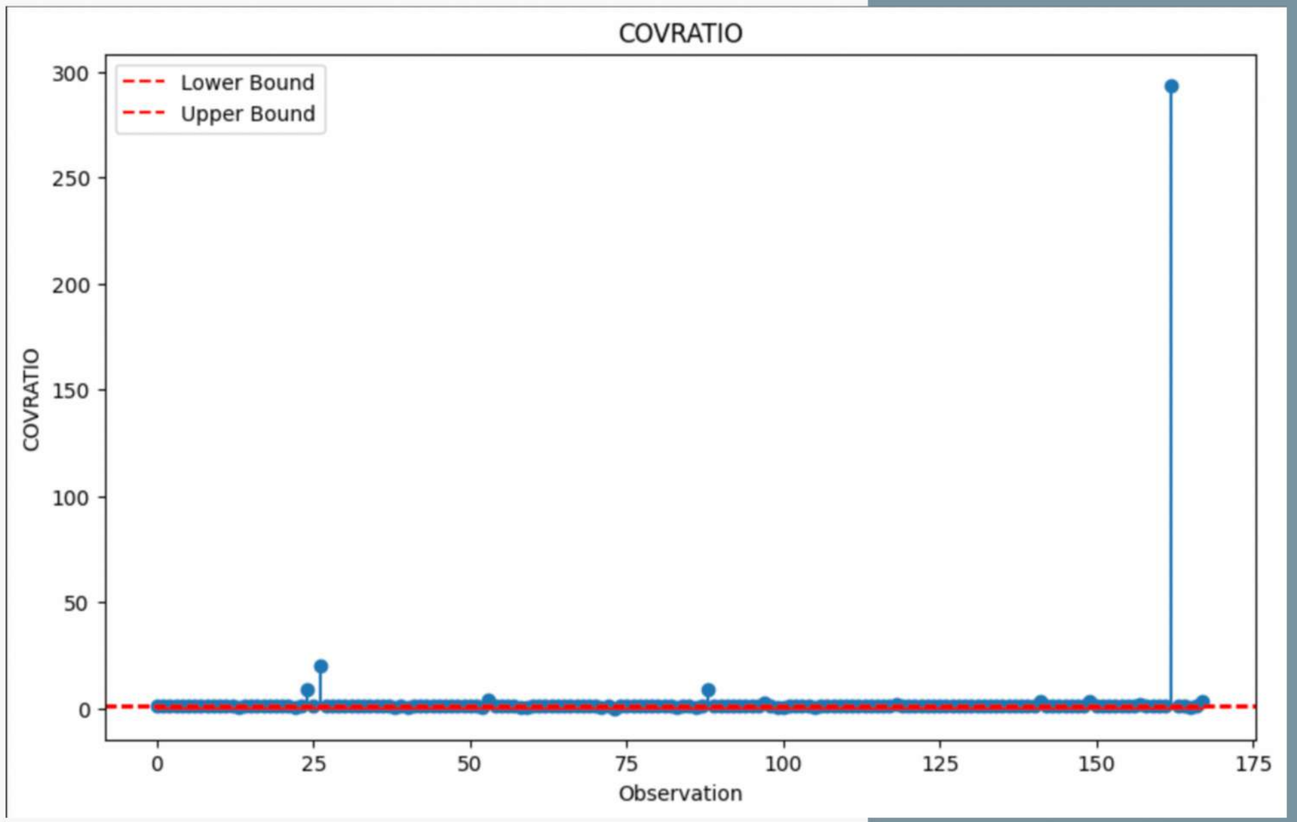
Threshold: i^{th} point is influential if

$$COVRATIOi < 1- 3(p+1)/n$$

$$COVRATIOi > 1+ 3(p+1)/n$$

Interpretation from the plot:

- Most points lie close to the red bounds — this is normal.
- A few moderately influential points (e.g., around 25 and 90).
- One extreme outlier near index 170 with COVRATIO > 300.



Outlier & Influence Diagnostics

Consensus Diagnostic Framework

Approach:

- Applied multiple diagnostic checks – Leverage, Cook's Distance, DFFITS, COVRATIO, and DFBETAs (per coefficient)
- Flagged each observation that exceeded any standard threshold.
- Computed a total flag count for each observation.

Filtering rule:

- Observations with 8 or fewer flags were retained.
- This ensures robust results by excluding highly influential or extreme points.

GAUSS-MARKOV ASSUMPTIONS OF LINEAR REGRESSION

Assumptions of Multiple Linear Regression Model

- **Linearity:** The relationship between independent and dependent variables is linear.
- **Normality of Errors**: Errors (residuals) are normally distributed.
- **Homoscedasticity**: Constant variance of errors across all levels of independent variables.
- **No Multicollinearity**: Independent variables are not highly correlated with each other.
- **Independence:** Observations are independent of each other.

The model's assumptions were thoroughly tested to ensure that the analysis meets the required conditions for accurate interpretation.

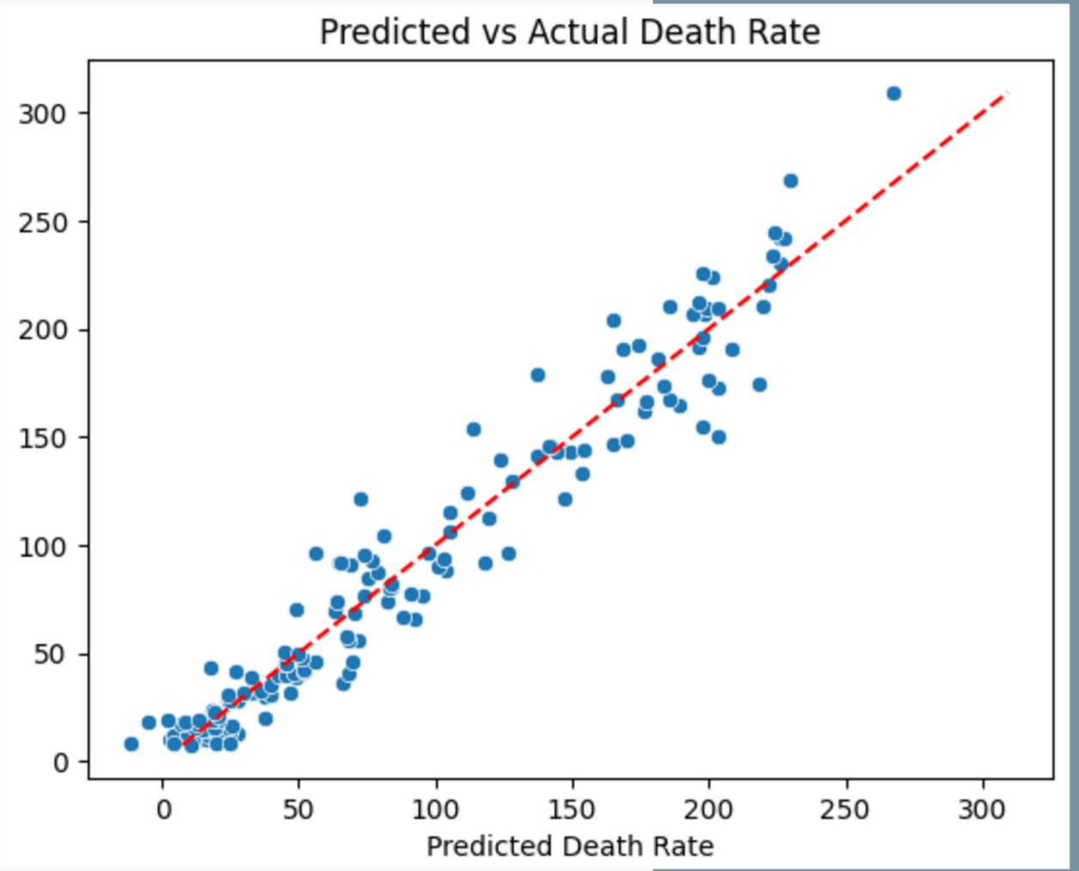
NOTE: We fit the model on the data after removing outliers. Here n=**154**

Linearity of the Relationship

- **Meaning**: The expected value of the dependent variable is a linear function of the independent variables.

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- As predictions match the actual values closely and form a straight-line pattern, it means the model understands the relationship well — and that relationship is likely linear.

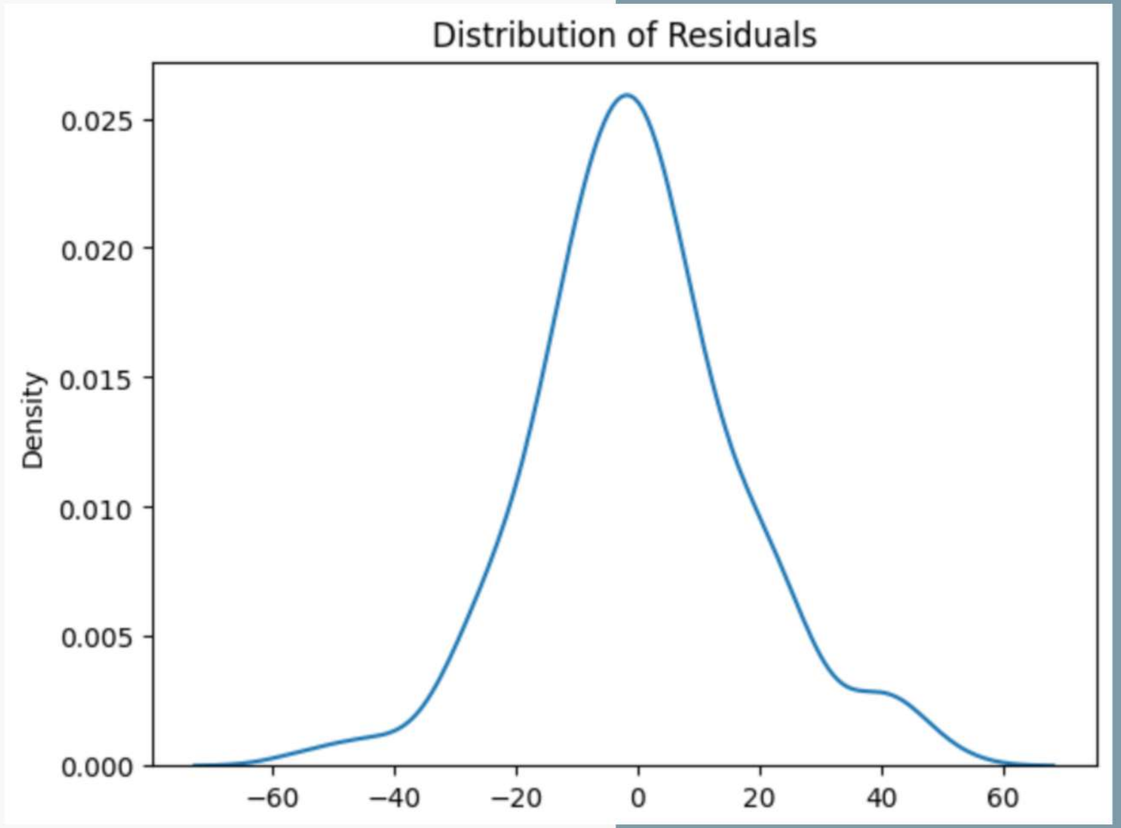


Normality of Errors

Residuals are normally distributed with mean 0 and constant variance.

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

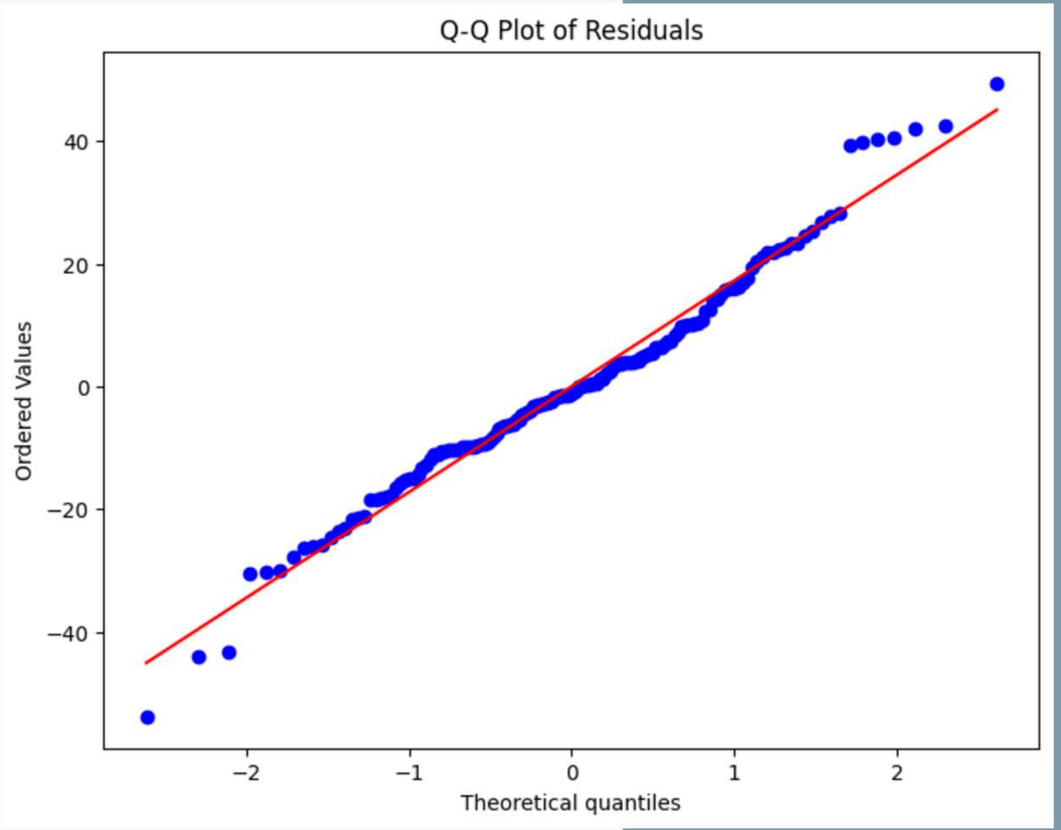
The residuals demonstrate a **unimodal**, **approximately symmetric** distribution **centered at zero**, supporting the assumption of normality



Normality of Errors

Interpretation of the Q-Q plot

The blue points closely follow the red diagonal line, especially in the middle range, indicating that the residuals are approximately normally distributed. so, the assumption of normality is reasonably satisfied .



Homoscedasticity

Objective: We check the homogeneity of error variance using Breusch Pagan Test

H_0: Residuals have constant variance (homoscedastic)

vs

H_1: Residual variance is not constant (heteroscedastic)

For this test, we first fit the linear regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

We then calculate the residuals from the OLS regression and regress the squared residuals on the independent variables

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_k x_{ik} + u_i.$$

Homoscedasticity

Test Statistic: The test statistic is based on the explained sum of squares from the auxiliary regression relative to the total number of observations n.
The statistic LM is given by

$$LM = \frac{n \cdot R^2}{2}$$

R^2 : coefficient of determination from the auxiliary regression

Under the null hypothesis,

$$\mathrm{LM} \sim \chi^2_{P-1}$$

P : number of predictors

Observation:

p-value = 0.1677307318980863 > 0.05

Conclusion:

We fail to reject null hypothesis at 5% level of significance
So this is an evidence of homoscedasticity

Multicollinearity

Variance Inflation Factor (VIF)

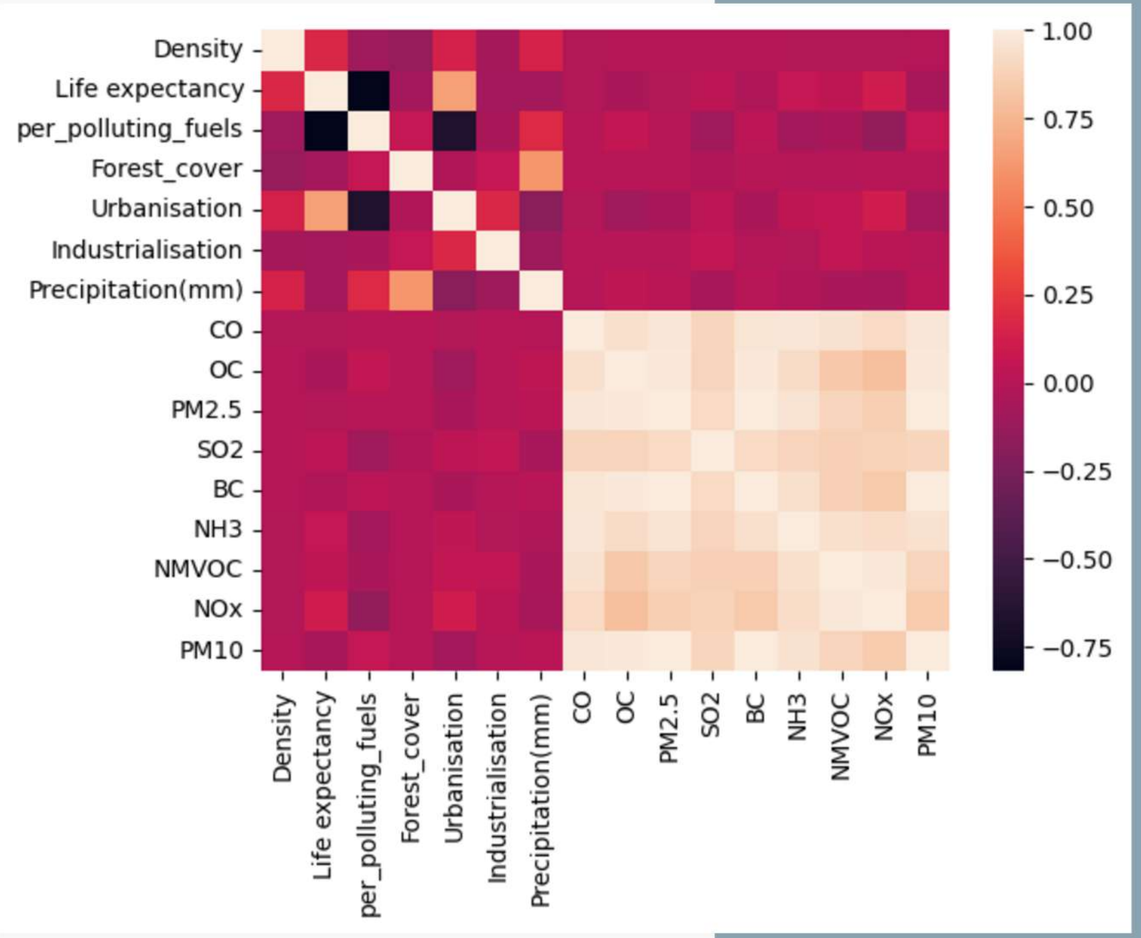
- Multicollinearity occurs predictor variables in a regression model are highly correlated, meaning they carry similar information.
- VIF greater than 10 suggests that the predictor is highly correlated with other variables in the model
- Air pollutants like PM2.5, PM10, NOx, SO2, NH3, CO, OC exhibit high VIF values, indicating strong multicollinearity.

feature	VIF
Density	1.460010
Life expectancy	53.761611
per_polluting_fuels	9.385249
Forest_cover	6.615466
Urbanisation	21.431650
Industrialisation	12.064442
Precipitation(mm)	10.635934
CO	136.280045
OC	617.595806
PM2.5	1785.597210
SO2	21.154253
BC	402.267936
NH3	62.589229
NMVOC	187.479350
NOx	156.580316
PM10	652.199287
Income_group_L	5.762940
Income_group_LM	4.056679
Income_group_UM	2.680349
ParentLocation_Americas	4.358206
ParentLocation_Eastern Mediterranean	2.733015
ParentLocation_Europe	6.750319
ParentLocation_South-East Asia	2.363674
ParentLocation_Western Pacific	2.220969

Multicollinearity

Key Points from the Heatmap:

- **Life expectancy** and **Forest_cover** is **strongly negatively correlated** with **per_polluting_fuels.**
- **Air pollutants** like PM2.5, PM10, NOx, SO2, NH3, etc. show **strong positive correlations** with each other
- **Moderate positive correlation** between **Urbanisation, Industrialisation**, and **per_polluting_fuels**



Independence of Errors

Objective: We check the independence of errors using Run's Test

H_0: The residuals are independent and randomly distributed

vs

H_1: The residuals are not random (dependent)

Test Statistic:

$$z = \frac{r - \mu_r}{\sigma^r}$$

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1 \qquad \sigma_r = \sqrt{\frac{(2n_1n_2)\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$$

n1: number of positive residuals
n2: number of negative residuals

Observation:

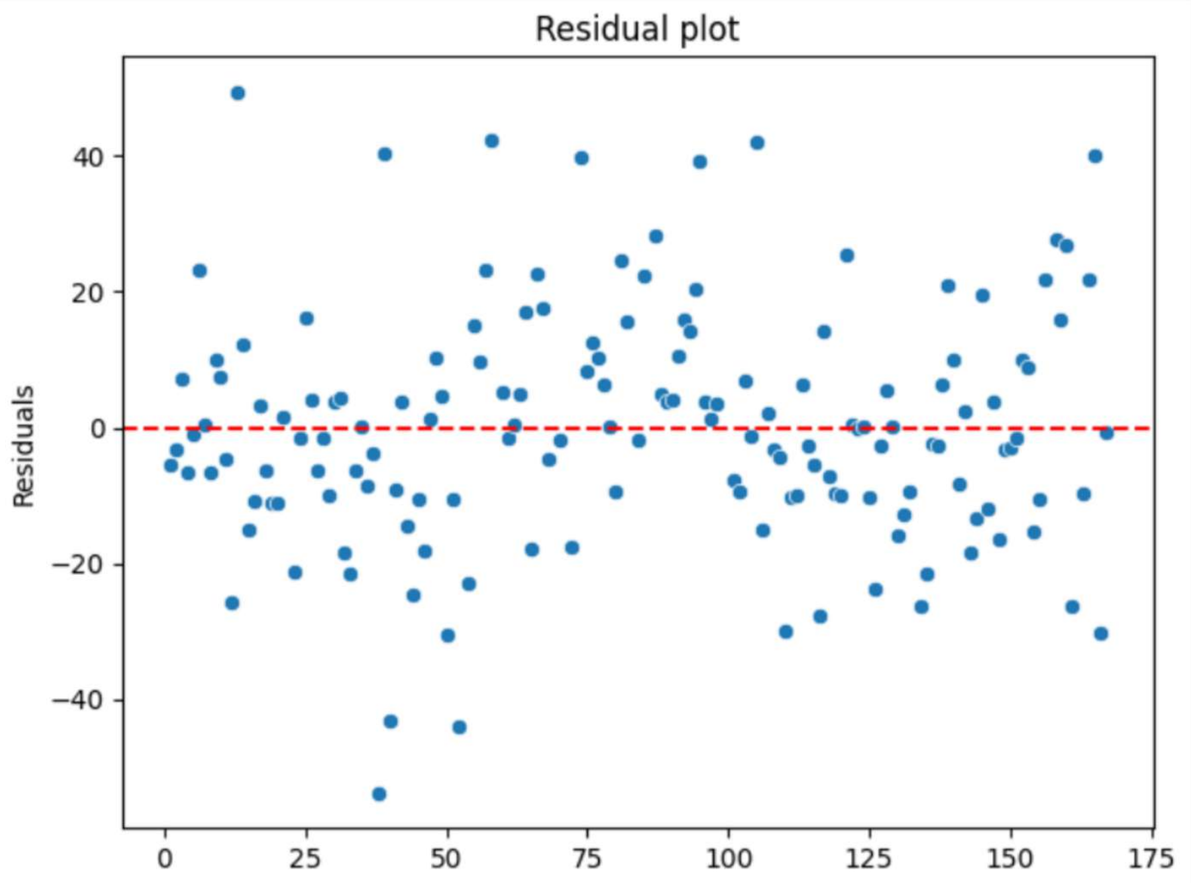
p-value = 0.036 > 0.01

Conclusion:

We fail to reject null hypothesis at 1% level of significance

So this is an evidence of independence of errors

Independence of Errors



Interpretation:

Although our data might exhibit spatial dependence, however from the residual plot, we can observe that the residuals shows no discernible pattern or clustering, and the residuals appear randomly scattered around zero. Combined with the results of Run's test, which also indicate randomness, we can reasonably conclude that there is no significant dependency in the residuals.

MODEL BUILDING

Multiple Linear Regression Model

Applying Multiple Linear Regression model to our dataset, we get the following results:

R2 score: 0.885306

Adjusted R2 score: 0.863968

MSE: 625.459071

RMSE: 25.009180

MAE: 17.993338

The following factors came out to be significant:

	coef	std err	t	P>|t|	[0.025	0.975]
const	437.8630	41.530	10.543	0.000	355.695	520.031
Density	1.158e-05	0.003	0.005	0.996	-0.005	0.005
Life expectancy	-5.7786	0.559	-10.331	0.000	-6.885	-4.672
per_polluting_fuels	1.0269	0.103	9.950	0.000	0.823	1.231
Forest_cover	-0.2310	0.099	-2.336	0.021	-0.427	-0.035
Urbanisation	0.1607	0.113	1.418	0.159	-0.064	0.385
Industrialisation	0.3990	0.200	1.996	0.048	0.004	0.795
Precipitation(mm)	0.0023	0.003	0.652	0.516	-0.005	0.009
CO	-0.0016	0.003	-0.485	0.629	-0.008	0.005
OC	0.0666	0.237	0.281	0.779	-0.403	0.536
PM2.5	0.1169	0.139	0.838	0.403	-0.159	0.393
SO2	-0.0052	0.006	-0.842	0.401	-0.018	0.007
BC	0.3444	0.481	0.715	0.476	-0.608	1.297
NH3	-0.0160	0.015	-1.091	0.277	-0.045	0.013
NMVOC	0.0191	0.014	1.366	0.174	-0.009	0.047
NOx	-0.0082	0.015	-0.529	0.598	-0.039	0.022
PM10	-0.1105	0.056	-1.970	0.051	-0.222	0.000
Income_group_L	24.7618	10.287	2.407	0.017	4.408	45.116
Income_group_LM	21.1671	6.825	3.101	0.002	7.663	34.671
Income_group_UM	6.0406	5.187	1.165	0.246	-4.223	16.304
ParentLocation_Americas	9.7970	7.433	1.318	0.190	-4.910	24.504
ParentLocation_Eastern Mediterranean	41.3957	8.430	4.911	0.000	24.717	58.075
ParentLocation_Europe	36.4740	7.588	4.807	0.000	21.462	51.486
ParentLocation_South-East Asia	23.3952	9.950	2.351	0.020	3.709	43.082
ParentLocation_Western Pacific	36.4576	8.717	4.182	0.000	19.211	53.704

Stepwise Variable Selection

Variable selection is method for choosing a subset of predictor variables for a regression model by iteratively adding or removing variables based on statistical significance. It can be implemented through forward selection, backward elimination, or a combination of both.

- We have chosen Stepwise selection method.
- Stepwise selection model retained **9 predictors** out of 24 original variables. These variables were selected based on statistical significance (p-value < 0.05)

The following factors came out to be significant:

- **per_polluting_fuels**
- **Life expectancy**
- **ParentLocation_Eastern Mediterranean**

Stepwise Variable Selection

- **Income_group_LM**
- **ParentLocation_Europe**
- **ParentLocation_Western Pacific**
- **ParentLocation_South-East Asia**
- **Forest_cover**
- **Industrialisation**

Applying Stepwise selection method, we get the following results:

R2 score: 0.929822

Adjusted R2 score: 0.916766

MSE: 382.698674

RMSE:19.562686

MAE: 15.033224

Principal Component Regression

- **Principal Component Analysis (PCA):**

Transform correlated predictors $X_1, X_2, \ldots, X_n$ into uncorrelated principal components $Z_1, Z_2, \ldots, Z_m$:

$$Z_j = \sum_{i=1}^{n} w_{ij} X_i$$

where w_{ij} are the weights (or loadings) for each component.

- **Regression on Principal Components:**

Fit a linear regression model:

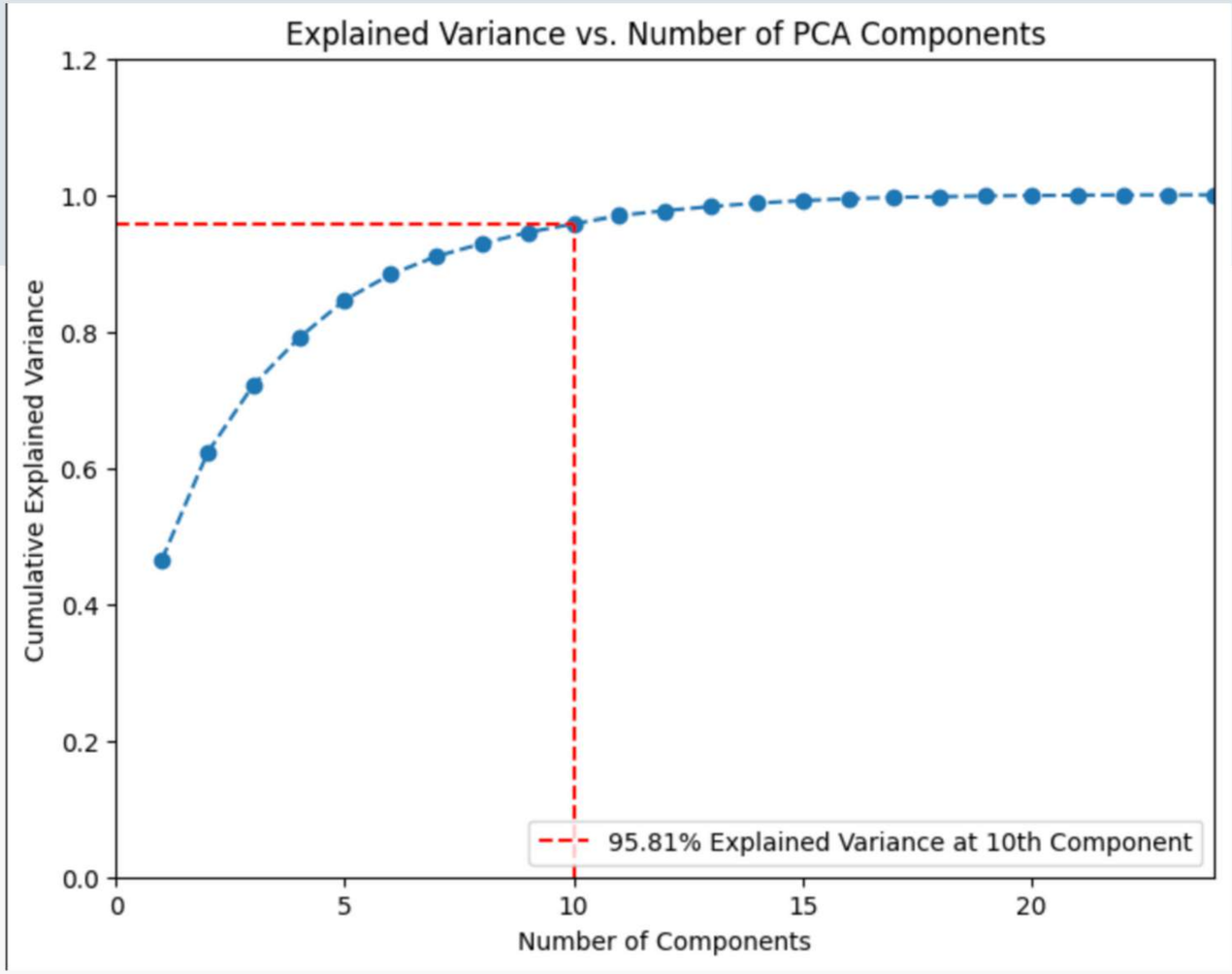
$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_m Z_m + \epsilon$$

where Y is the dependent variable, and $Z_1, Z_2, \ldots, Z_m$ are the selected principal components.

PCR helps handle **multicollinearity, reduces overfitting, and improves model stability.**

Principal Component Regression

Explained Variance: A total of **10** principal components were selected, which together explain approximately **95%** of the total variance in the data.



Applying Principal Component Regression model to our dataset, we get the following results:

R2 score: 0.874289

Adjusted R2 score: 0.850901

MSE: 685.536831

RMSE: 26.182758

MAE: 18.985216

Ridge Regression

Ridge Regression is a type of linear regression that includes L2 regularization. It adds a penalty term to the loss function that shrinks the regression coefficients towards zero.

In this, we try to solve the following optimization problem

$$\underset{\beta}{\text{Minimize}} \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{subject to } \boldsymbol{\beta}'\boldsymbol{\beta} \leq d$$

Applying Ridge Regression to our dataset, we get the following results

R2 score: 0.901385
Adjusted R2 score: 0.883038
MSE: 537.776270
RMSE: 23.190004
MAE: 17.009438

Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that uses L1 regularization. It adds a penalty equal to the absolute values of the coefficients to the loss function. Lasso can shrink some coefficients exactly to zero, removing unimportant features from the model.

In this, we try to solve the following optimization problem

$$\underset{\beta}{\text{Minimize}} \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \text{ subject to } \|\beta\|_1 \leq t$$

Applying Lasso Regression to our dataset, we get the following results

R2 score: 0.909932
Adjusted R2 score: 0.893175
MSE: 491.166203
RMSE: 22.162270
MAE: 16.159237

Lasso Regression

Predictor	Coefficient
per_polluting_fuels	37.226447
Life expectancy	-30.875212
ParentLocation_Eastern Mediterranean	15.508773
ParentLocation_Americas	-11.523299
Income_group_LM	6.249965
ParentLocation_Western Pacific	5.398163
Industrialisation	5.302142
Forest_cover	-4.382460
SO2	-0.463103
Precipitation(mm)	-0.000000
Urbanisation	-0.000000
Density	0.000000
CO	-0.000000
OC	-0.000000
PM2.5	-0.000000
BC	-0.000000
PM10	-0.000000
NOx	-0.000000
NMVOC	-0.000000
NH3	-0.000000
Income_group_UM	0.000000
Income_group_L	0.000000
ParentLocation_Europe	0.000000
ParentLocation_South-East Asia	0.000000

Lasso is performing feature selection: keeping only the most relevant variables.

Key predictors retained:
per_polluting_fuels, Life expectancy, Industrialisation, Forest cover, certain Income_groups, and ParentLocation regions — these have the most influence on the response variable.

Zero coefficients:
Variables like PM2.5, CO, NOx, and other pollutants were excluded, meaning they were either not significant or highly correlated with selected variables.

Partial Least Squares

PLS is a regression technique that models the relationship between **predictors and response by projecting both** into a new lower-dimensional space. It is useful especially when the predictors have high multicollinearity

PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via least squares using these M new features.

$$Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_m Z_m + \epsilon$$

PLS components are chosen to maximize the covariance between X and Y, not just the variance of X. So we often compute:

Explained variance in X: How much of the variance in the predictors is explained by the PLS components.

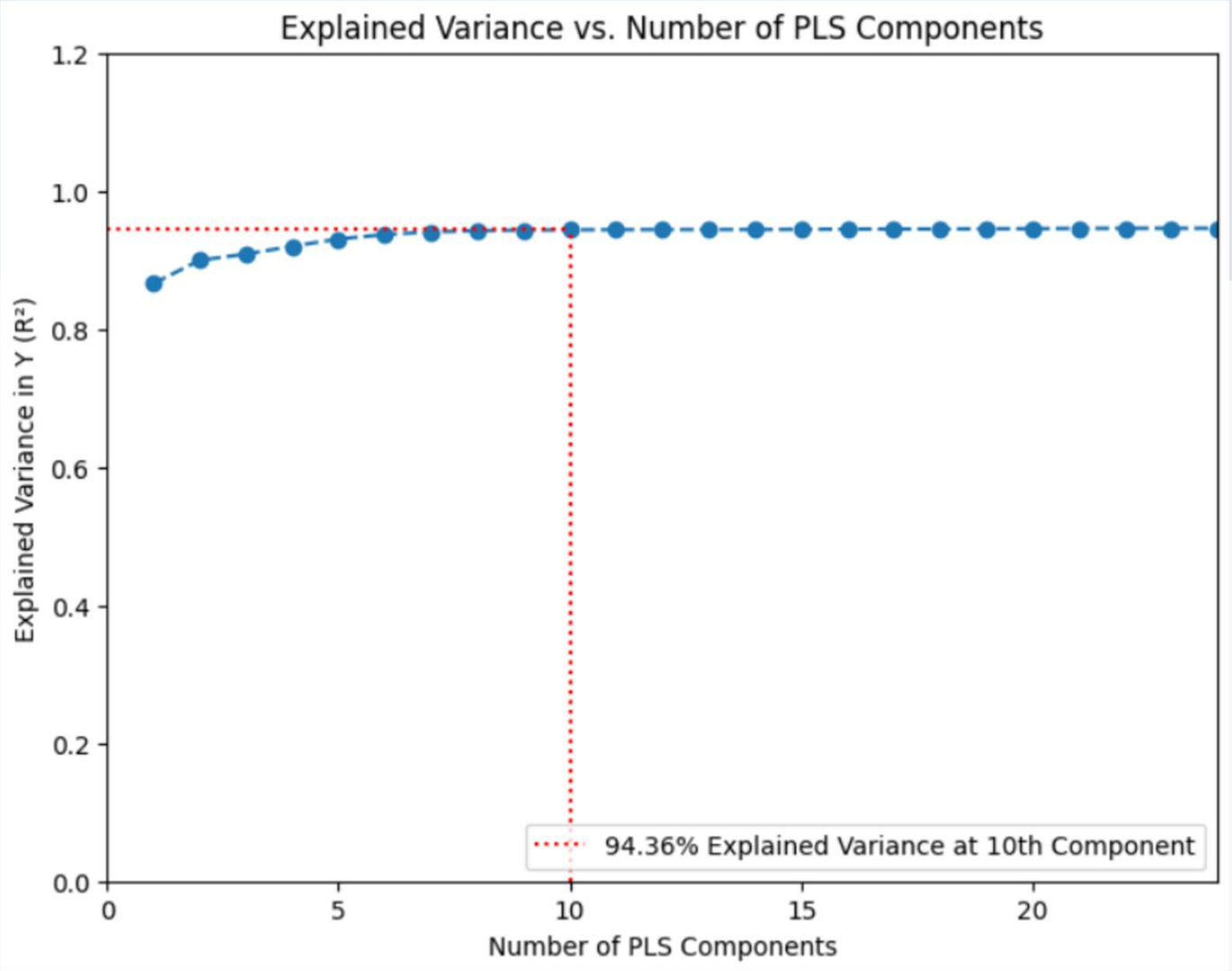
Explained variance in Y: How much of the variance in the response is explained by the PLS components.

Partial Least Squares

Explained Variance: A total of **10** PLS components were selected, which together explain **94.36%** of the total variance.

Applying Partial Least Squares Regression to our dataset, we get the following results

R2 score: 0.911691
Adjusted R2 score: 0.895261
MSE: 481.575376
RMSE: 21.944826
MAE: 16.350599



Model Evaluation

Model	R^2	Adjusted R^2	MSE	RMSE	MAE
Stepwise selection	0.929822	0.916766	382.70	19.56	15.03
Partial Least Squares	0.911691	0.895261	481.58	21.94	16.35
Lasso Regression	0.909932	0.893175	491.17	22.16	16.16
Ridge Regression	0.901385	0.883038	537.78	23.19	17.01
Multiple Linear Regression	0.885306	0.863968	625.46	25.01	17.99
Principal Component Regression	0.874289	0.850901	685.54	26.18	18.99

- Stepwise Regression showed the best performance (R^2 = 0.93, lowest errors), but it's sensitive to multicollinearity and may include redundant variables.
- To overcome this, we considered more robust models:
  - **Lasso Regression**: Performs variable selection and regularization, reducing overfitting.
  - **Partial Least Squares (PLS)**: Converts predictors into orthogonal components to handle multicollinearity.

Thus, Lasso and PLS are preferred for their balance of accuracy and reliability.

Conclusion

The model equation obtained from lasso regression is the following:

$\hat{Y}$ = *Respiratory Death Rate*

$$\hat{Y} = 37.226 \cdot \text{per\_polluting\_fuels} - 30.875 \cdot \text{Life\_expectancy} + 15.509 \cdot \text{ParentLocation\_Eastern\_Mediterranean} - 11.523 \cdot \text{ParentLocation\_Americas} + 6.250 \cdot \text{Income\_group\_LM} + 5.398 \cdot \text{ParentLocation\_Western\_Pacific} + 5.302 \cdot \text{Industrialisation} - 4.382 \cdot \text{Forest\_cover} - 0.463 \cdot SO_2$$

Conclusion

In conclusion, several factors significantly influence respiratory health outcomes. Increased forest cover improves air quality, reducing mortality and respiratory symptoms. Socioeconomic disparities, particularly income inequality, exacerbate respiratory issues, leading to worsened symptoms and higher disease prevalence. Geographical disparities also exist, with rural areas experiencing higher mortality from respiratory diseases compared to urban centers. The use of polluting fuels, such as biomass for cooking, is a significant health risk, contributing to a substantial number of pneumonia-related deaths. Improvements in life expectancy are closely tied to reduced deaths from respiratory infections, underscoring the critical role of respiratory health in overall longevity. Finally, industrialization contributes to respiratory mortality, particularly among vulnerable populations. These factors highlight the complex interplay of environmental, socioeconomic, and regional influences on respiratory health.

Future Scope

Predictive Risk Index: Developing a predictive index for respiratory mortality risk based on key features from the project could aid policymakers in identifying high-risk regions and prioritizing interventions.

Policy Simulation, Evaluation and Public Health Planning: The project can be adapted to simulate the potential impact of policy changes, it can help identify high-risk regions and population segments that require targeted interventions such as pollution control measures, improved healthcare access, or education initiatives, enabling data-driven decision-making.

Real-time and Global Expansion: Integrating real-time environmental data from satellite or sensor sources could enhance responsiveness. Additionally, expanding the study across countries or cities would provide comparative insights and account for varying socio-political contexts.

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Introduction to Statistical Learning with Applications in Python. Springer.
- Twohig-Bennett, C., & Jones, A. (2018). The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. Environmental Research, 166, 628-637. DOI:10.1016/j.envres.2018.06.030
- U.S. Forest Service. (2017). The role of urban forests in human health and well-being: Effects on air quality and public health. U.S. Forest Service.
- Gaffney AW, Himmelstein DU, Christiani DC, Woolhandler S. Socioeconomic Inequality in Respiratory Health in the US From 1959 to 2018. JAMA Intern Med. 2021;181(7):968–976. doi:10.1001/jamainternmed.2021.2441

References

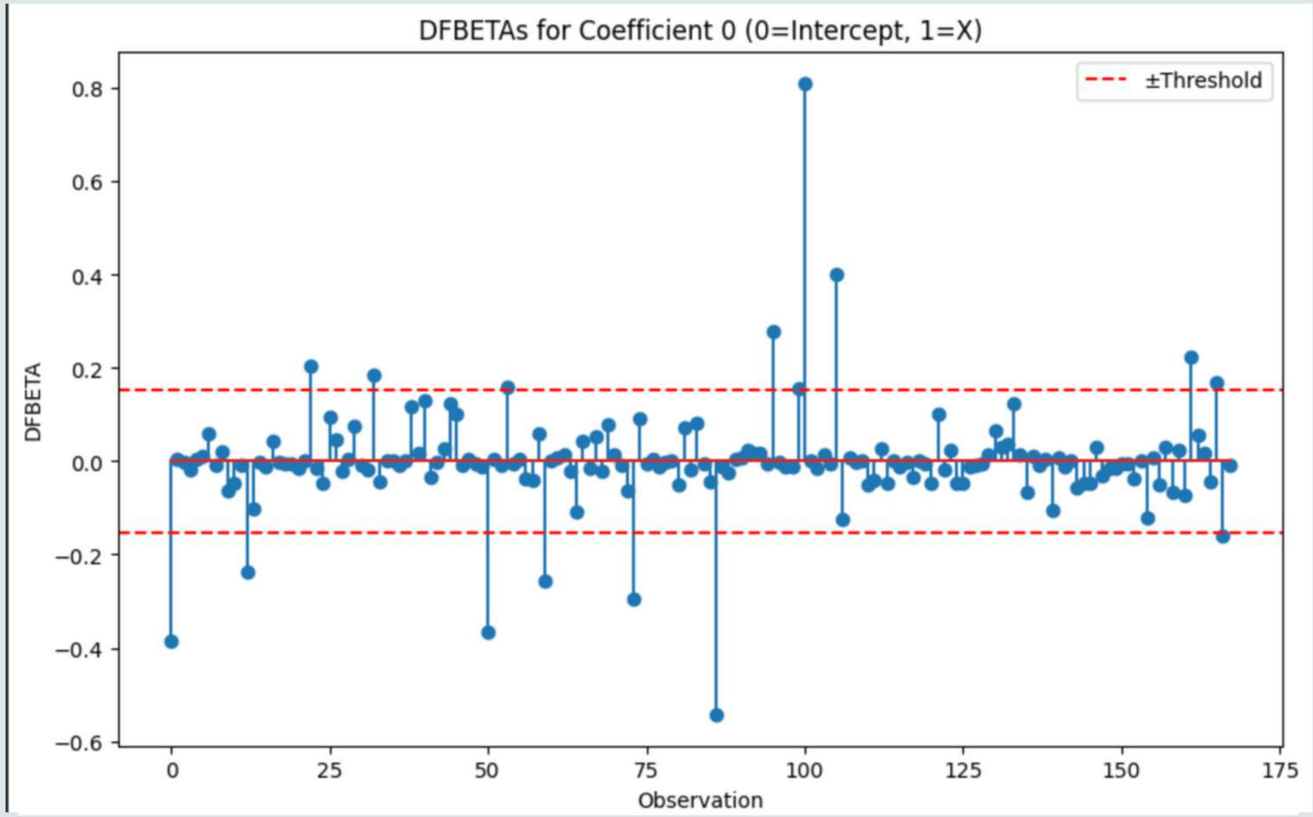
- Introduction to Linear Regression Analysis, by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining
- Jamil, A., Rehman, O., Shakoor, M., Kamdi, A., Fatima, E., Sohail, H., Nadeem, Z., Pirih, F. N. U., Khalid, Z., Inam, S., & Haq, M. I. U. (2025). Regional Disparities in Chronic Respiratory Diseases and Stroke Mortality Across the United States from 1999 to 2020 (S8.004). Neurology, 104(7 Supplement 1). DOI:10.1212/WNL.0000000000208680
- Household air pollution
- Life expectancy increased as world addressed major killers including diarrhea, lower respiratory infections, and stroke
- BMJ. (2009). Income inequality, mortality, and self-rated health: Meta-analysis of multilevel studies. BMJ, 339, b4471. DOI:10.1136/bmj.b4471

Thank you

APPENDIX

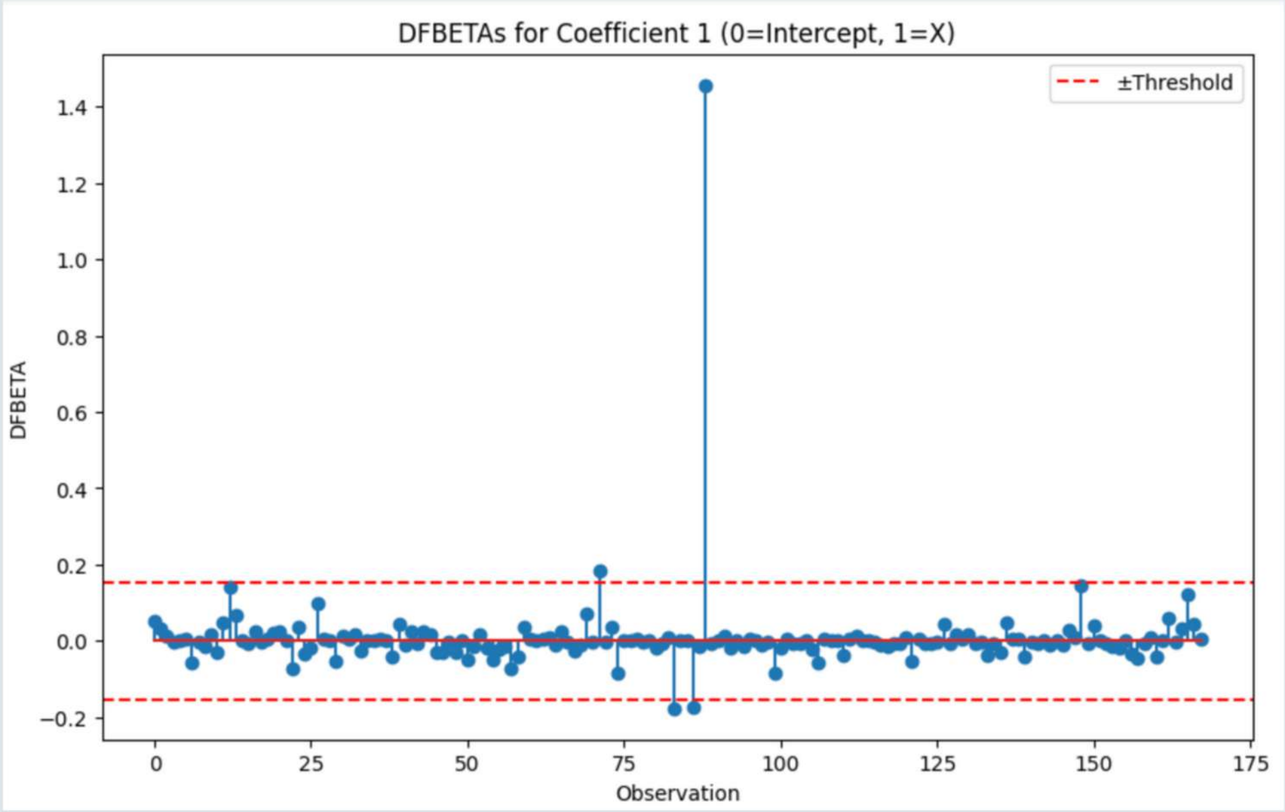
DFBETAs

Intercept(β_0)



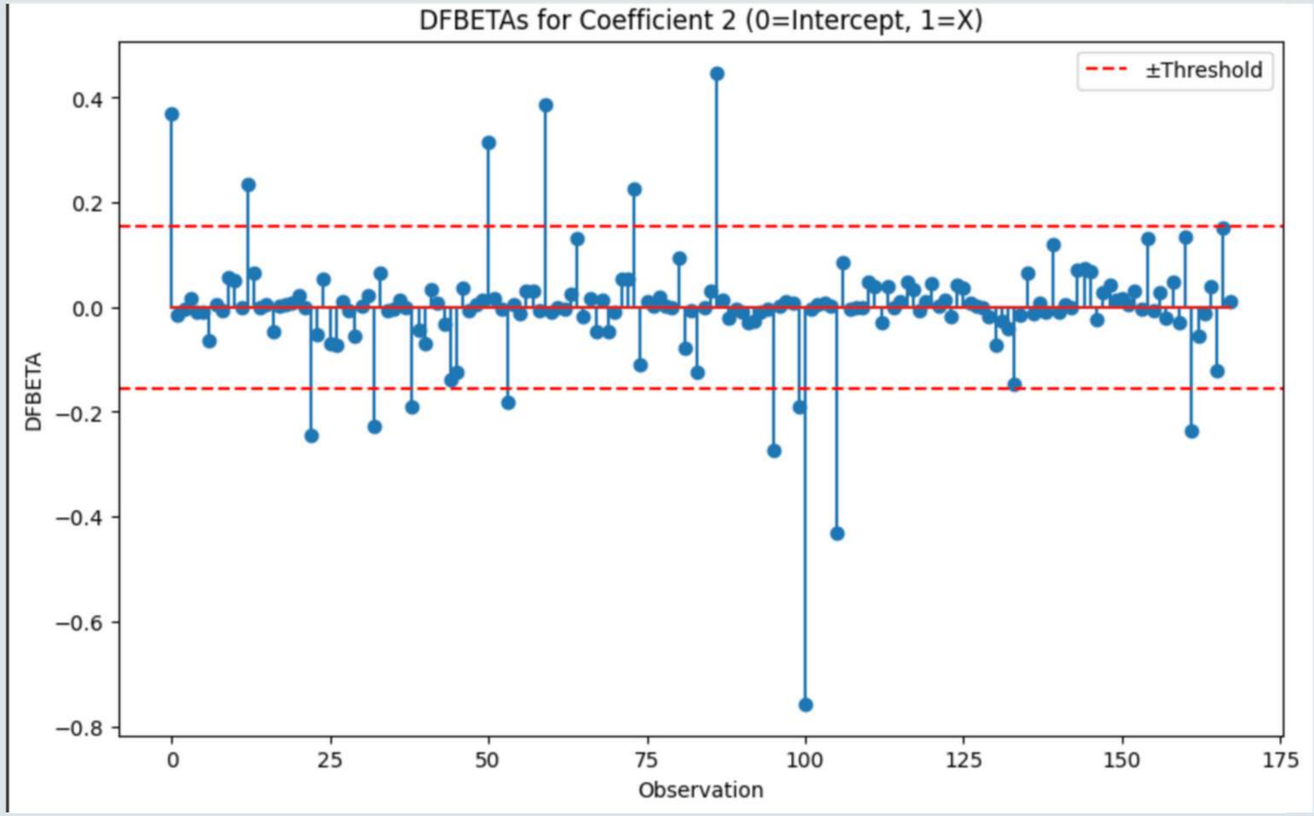
DFBETAs

Density(β_1)



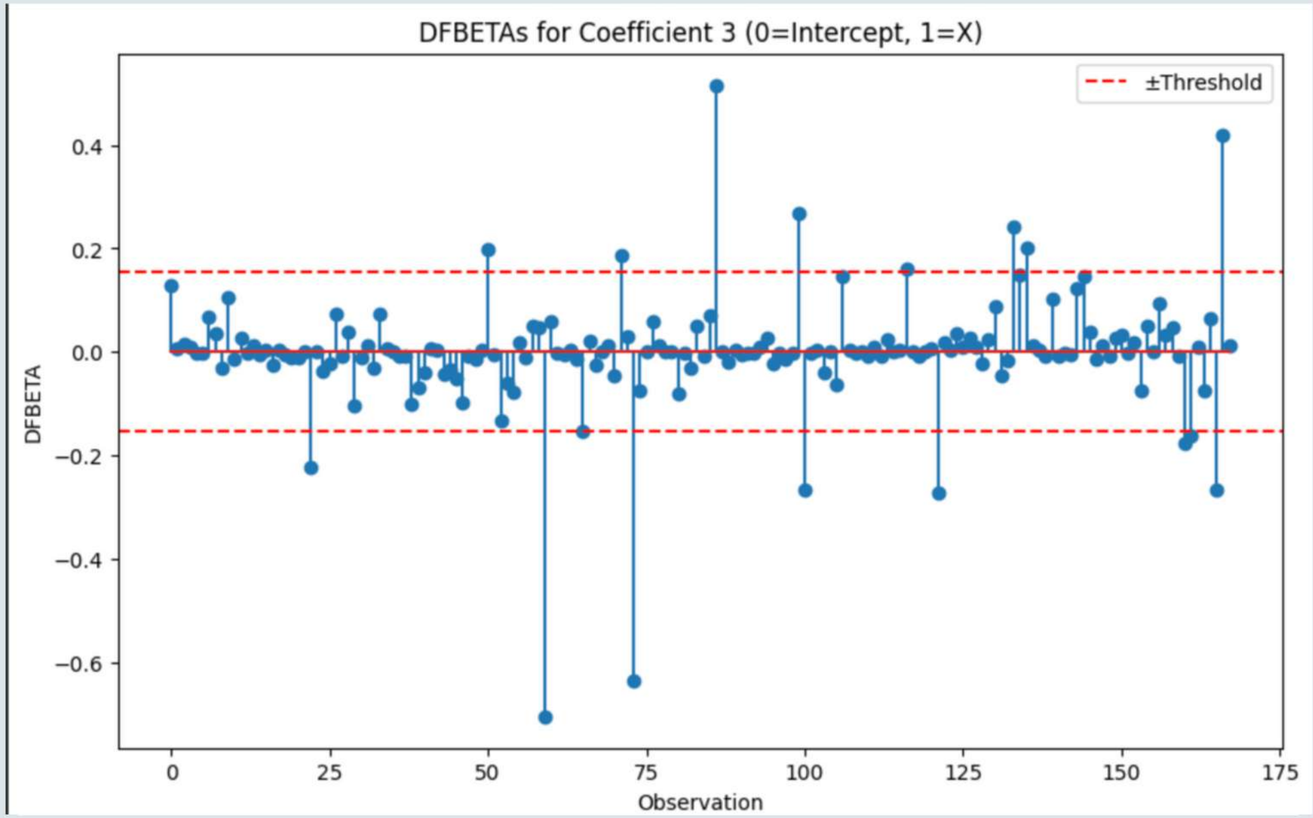
DFBETAs

Life Expectancy(β_2)



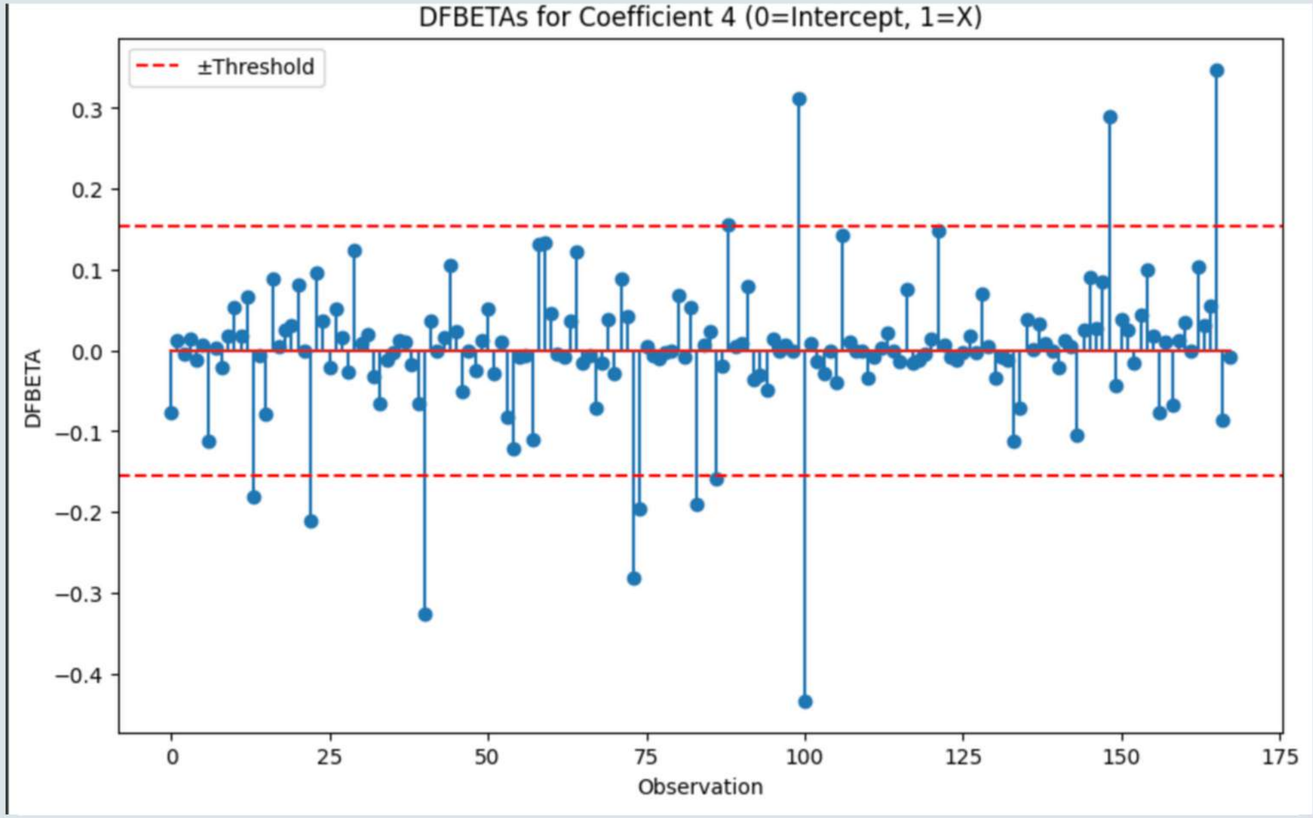
DFBETAs

per_polluting_fuels (β_3)



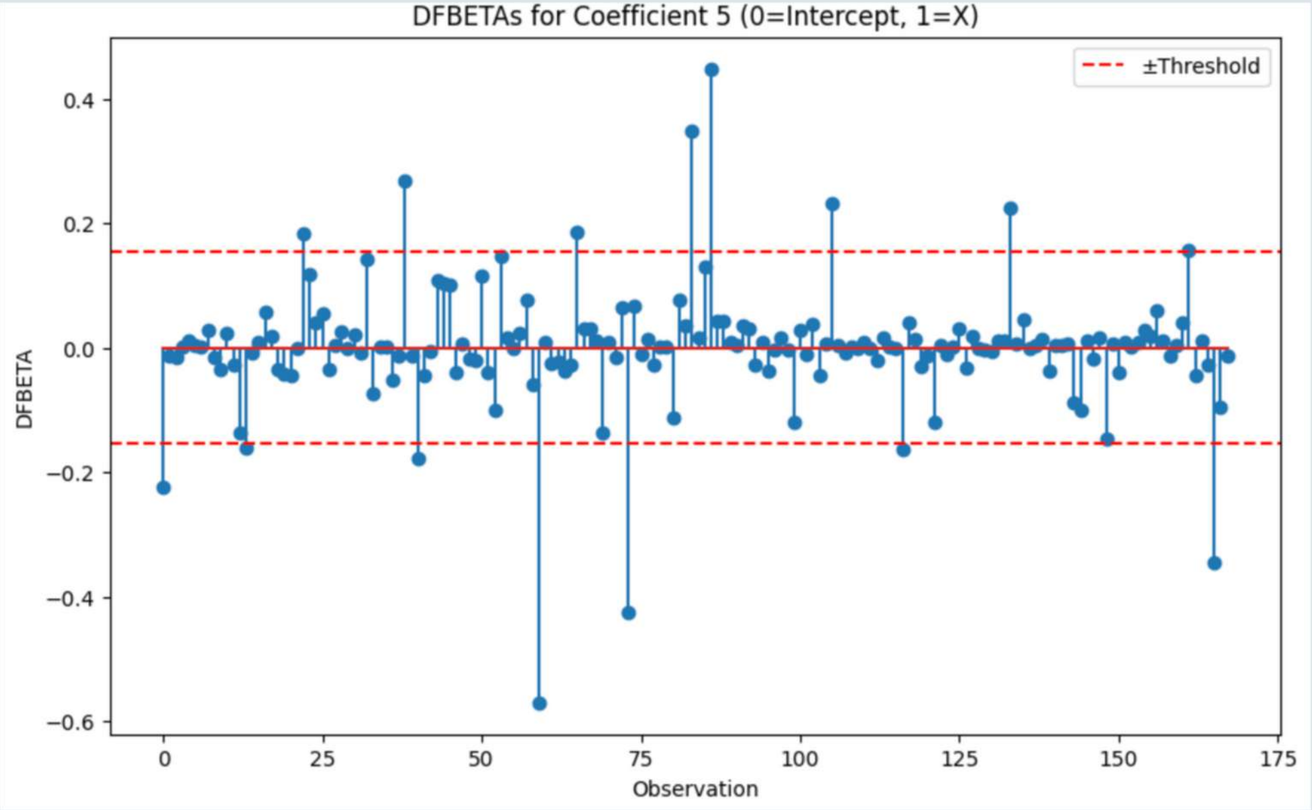
DFBETAs

Forest_cover (β_4)



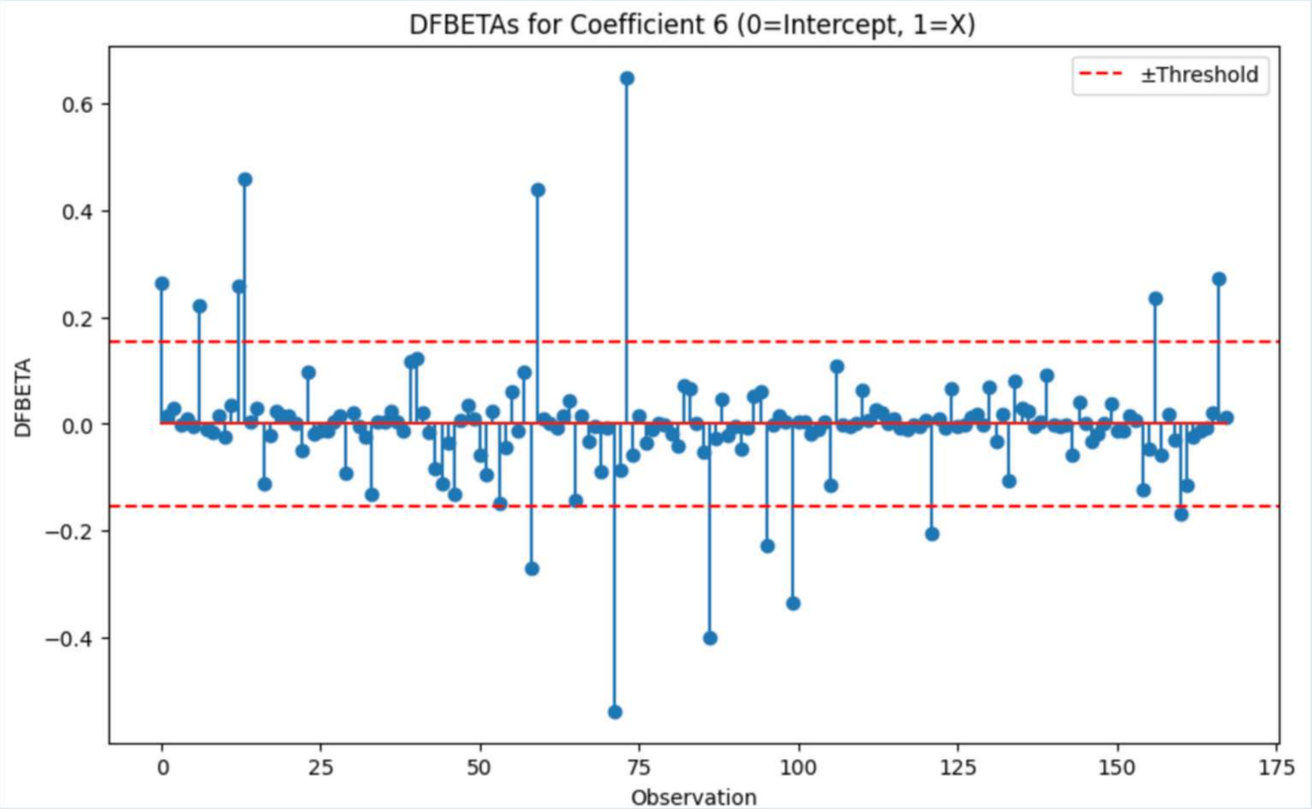
DFBETAs

Urbanisation (β_5)



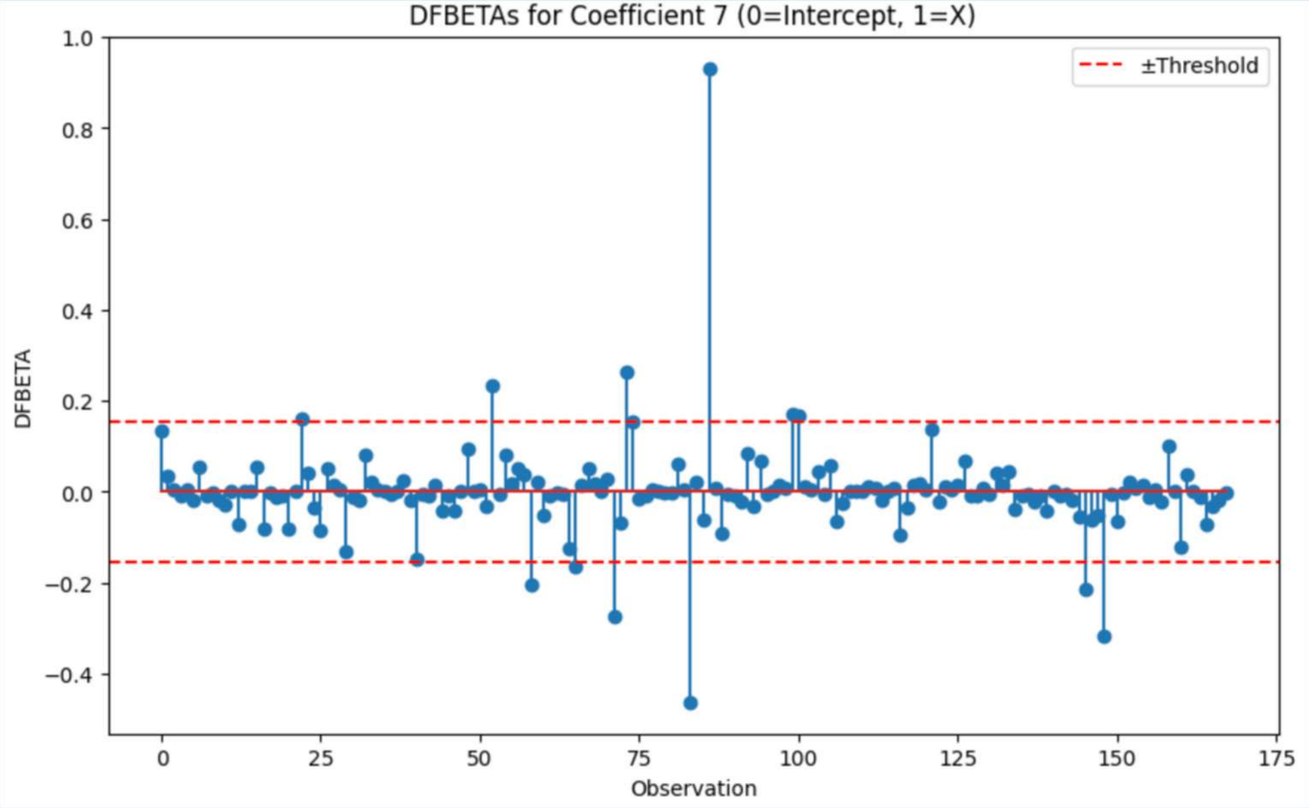
DFBETAs

Industrialisation (β_6)



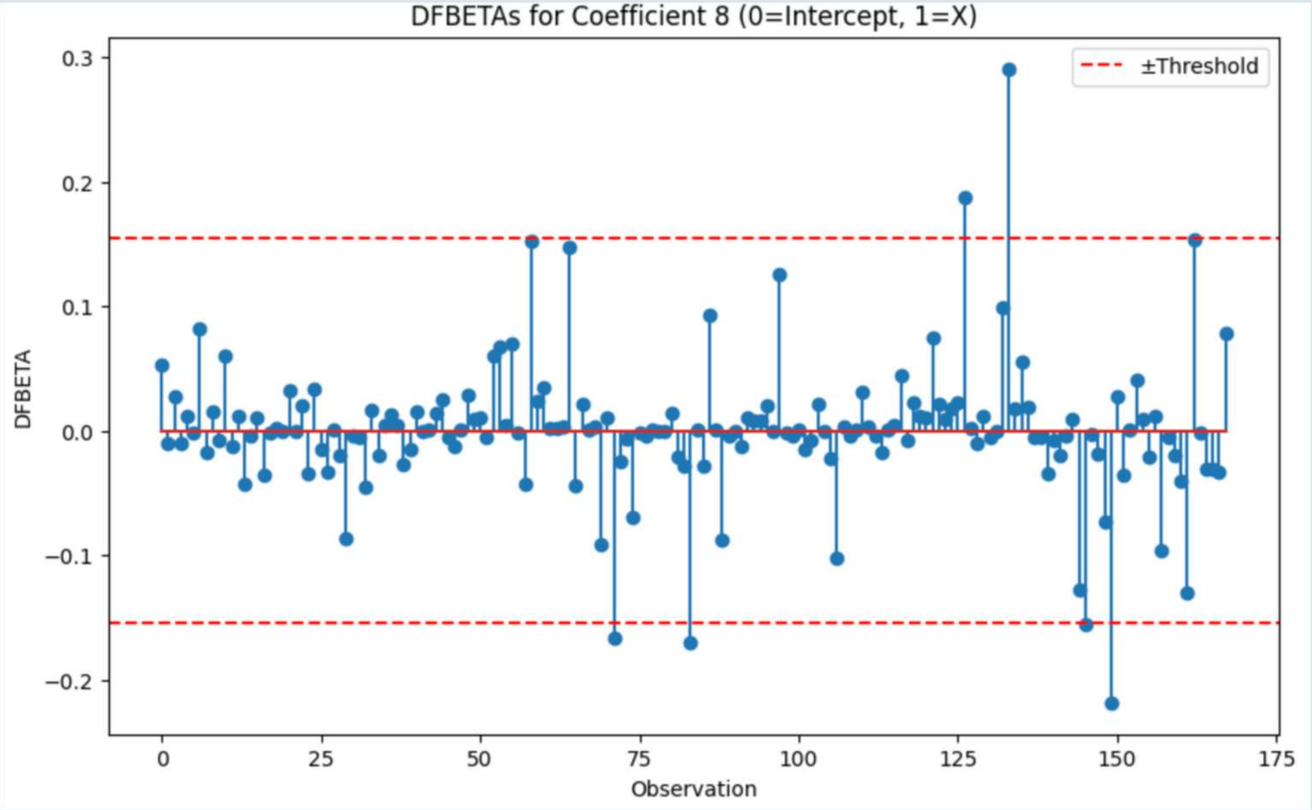
DFBETAs

Precipitation (mm) (β_7)



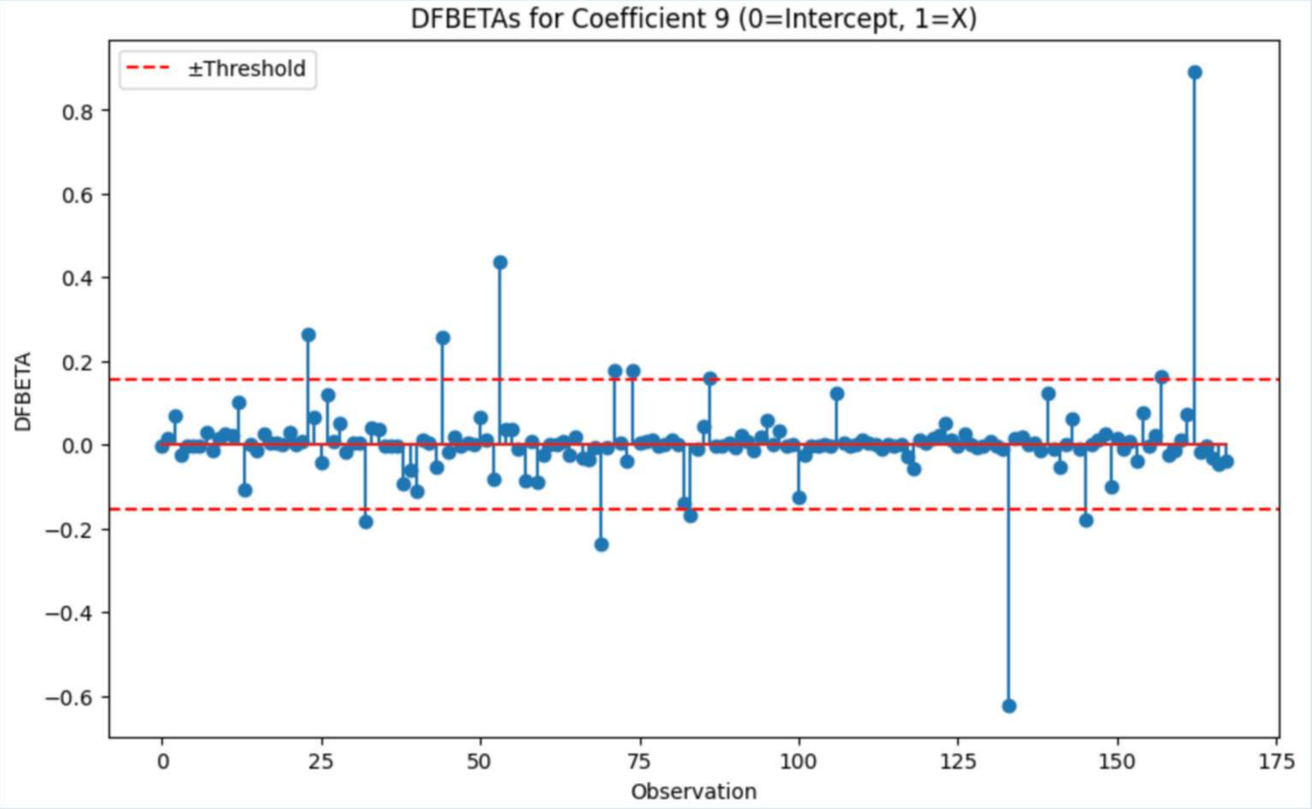
DFBETAs

CO (β_8)



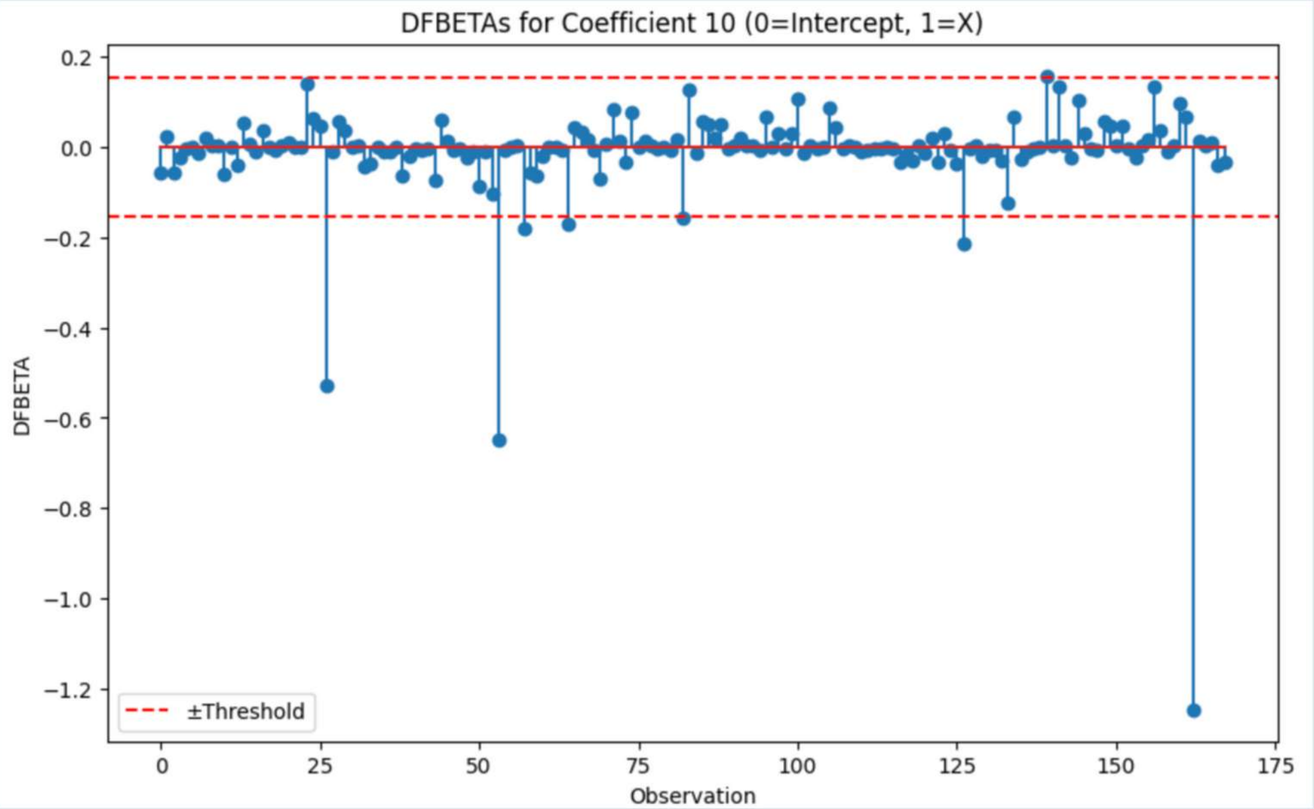
DFBETAs

OC (β_9)



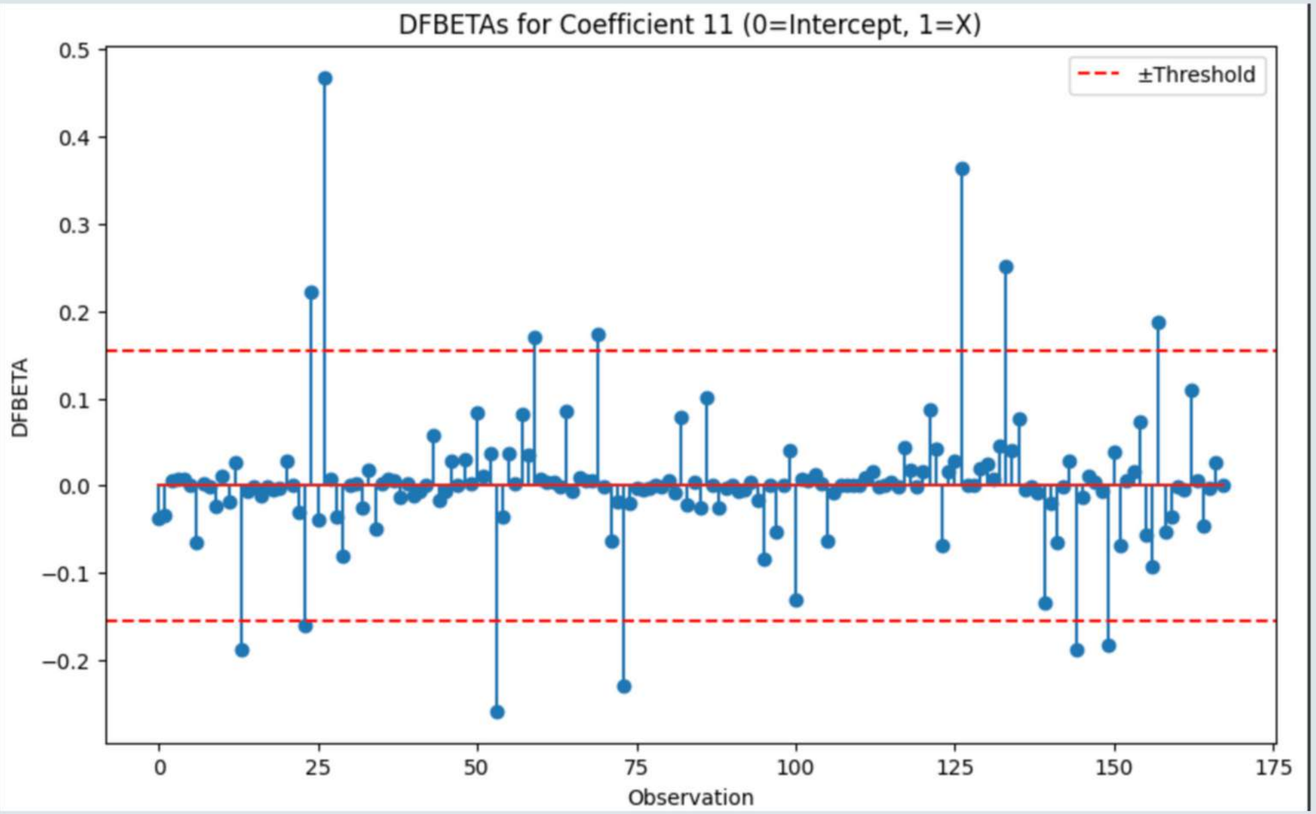
DFBETAs

PM2.5 (β_{10})



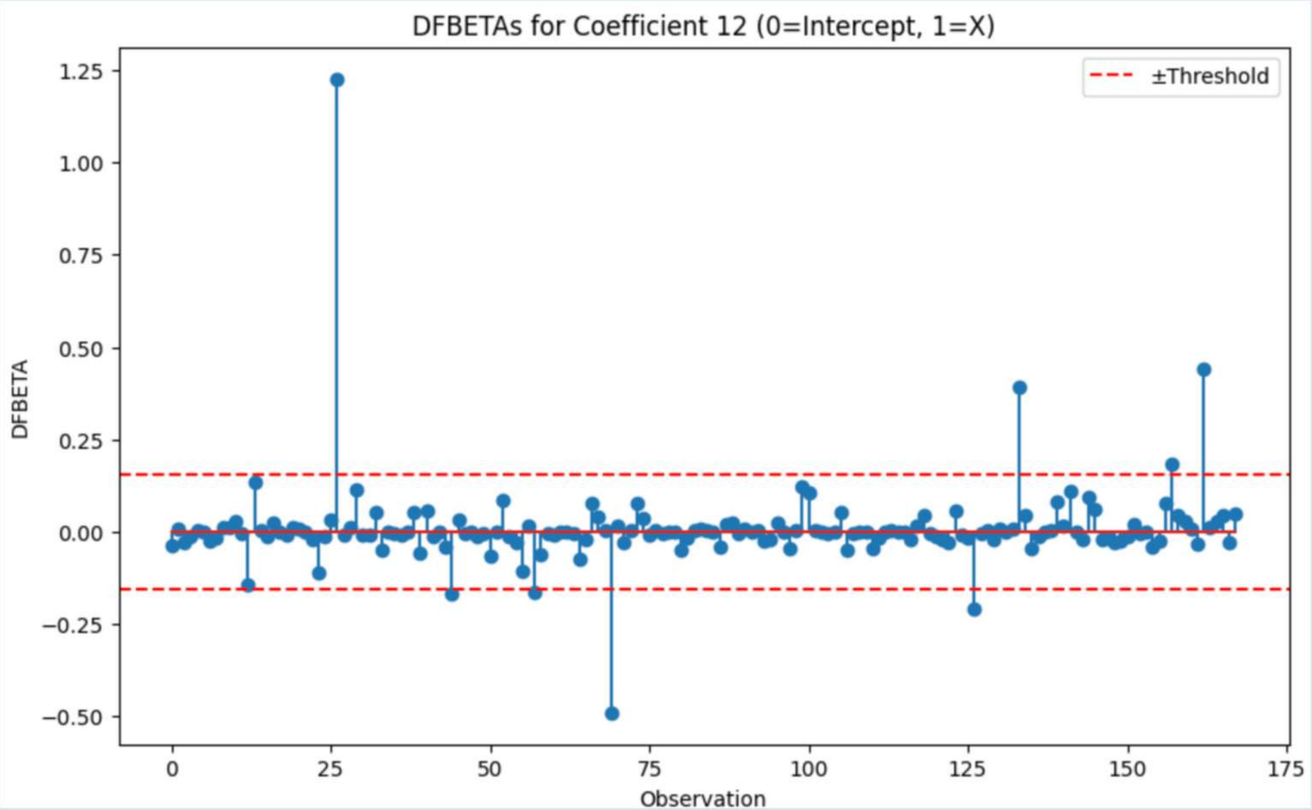
DFBETAs

SO_2 (β_{11})



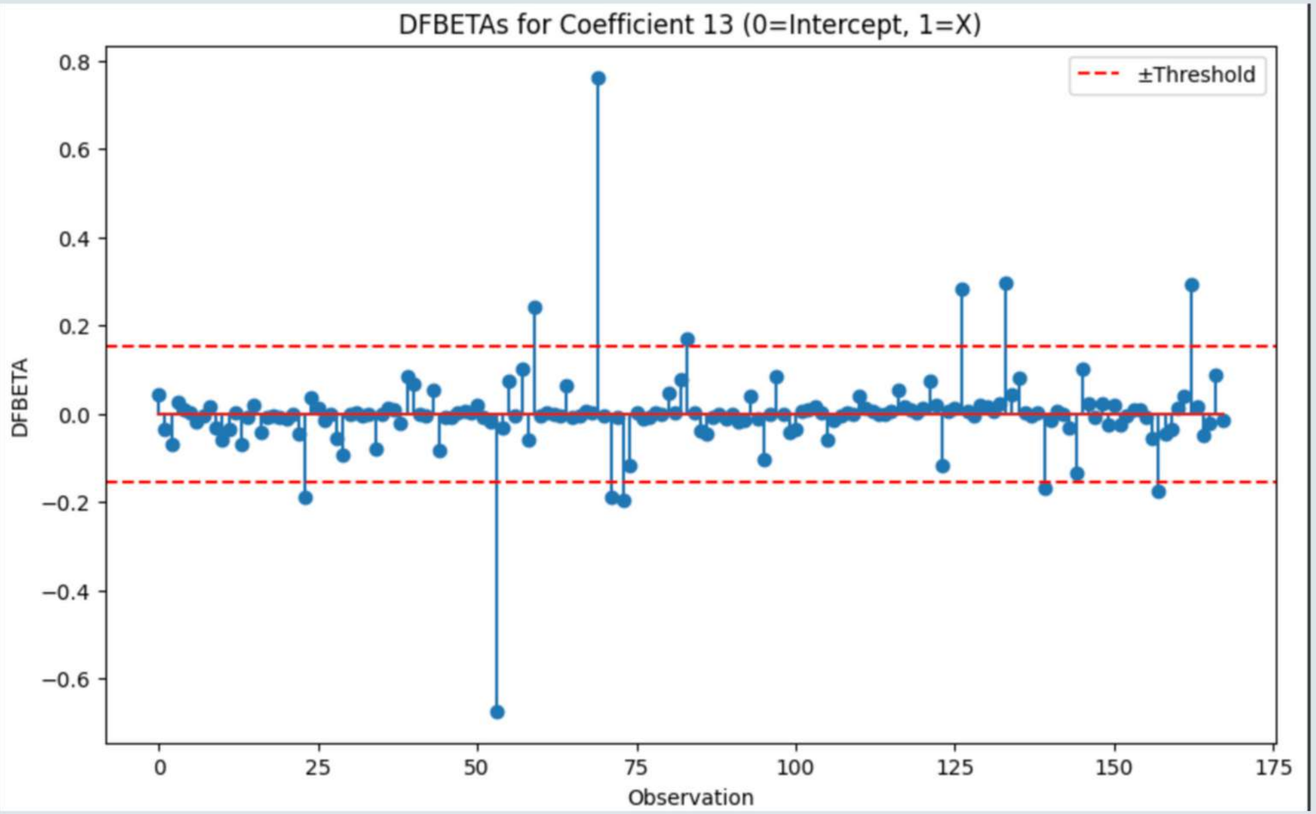
DFBETAs

BC (β_{12})



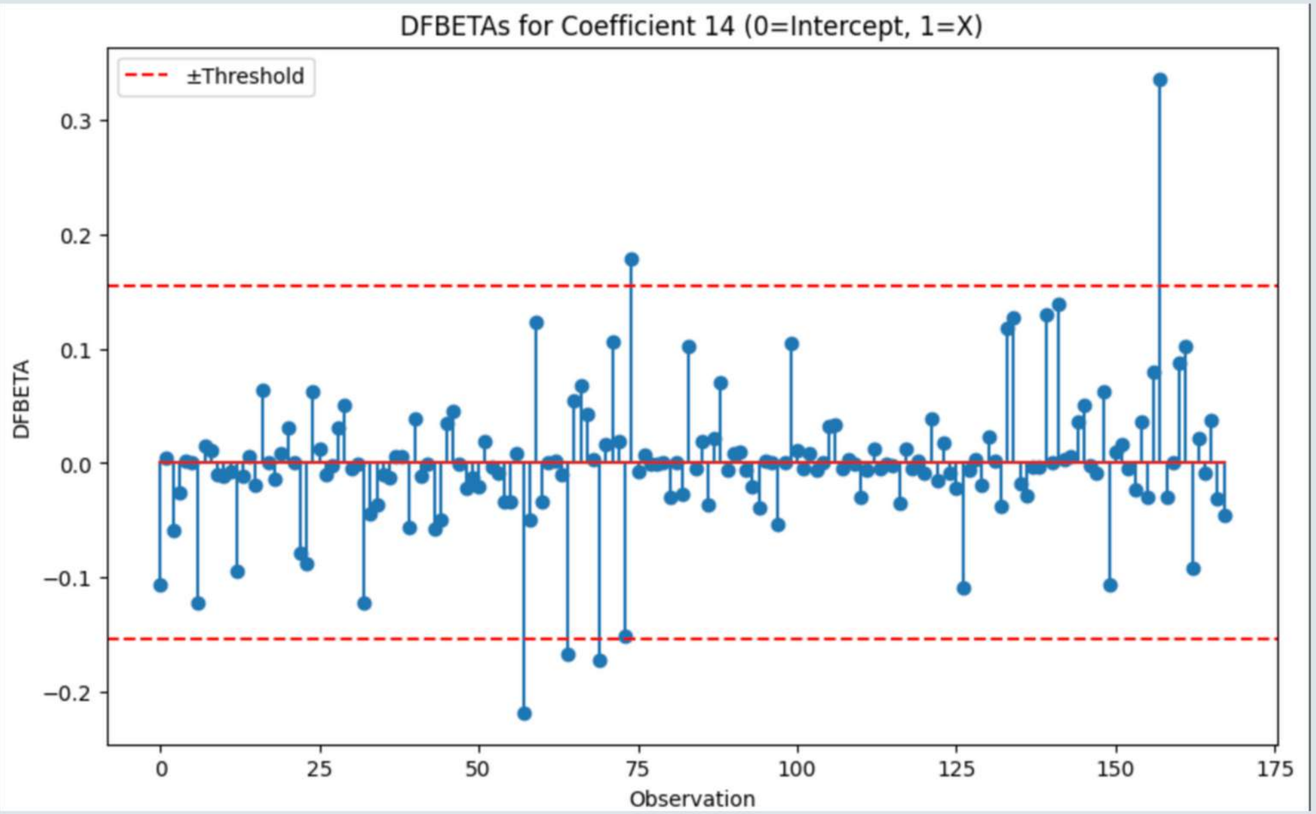
DFBETAs

NH_3 (β_{13})



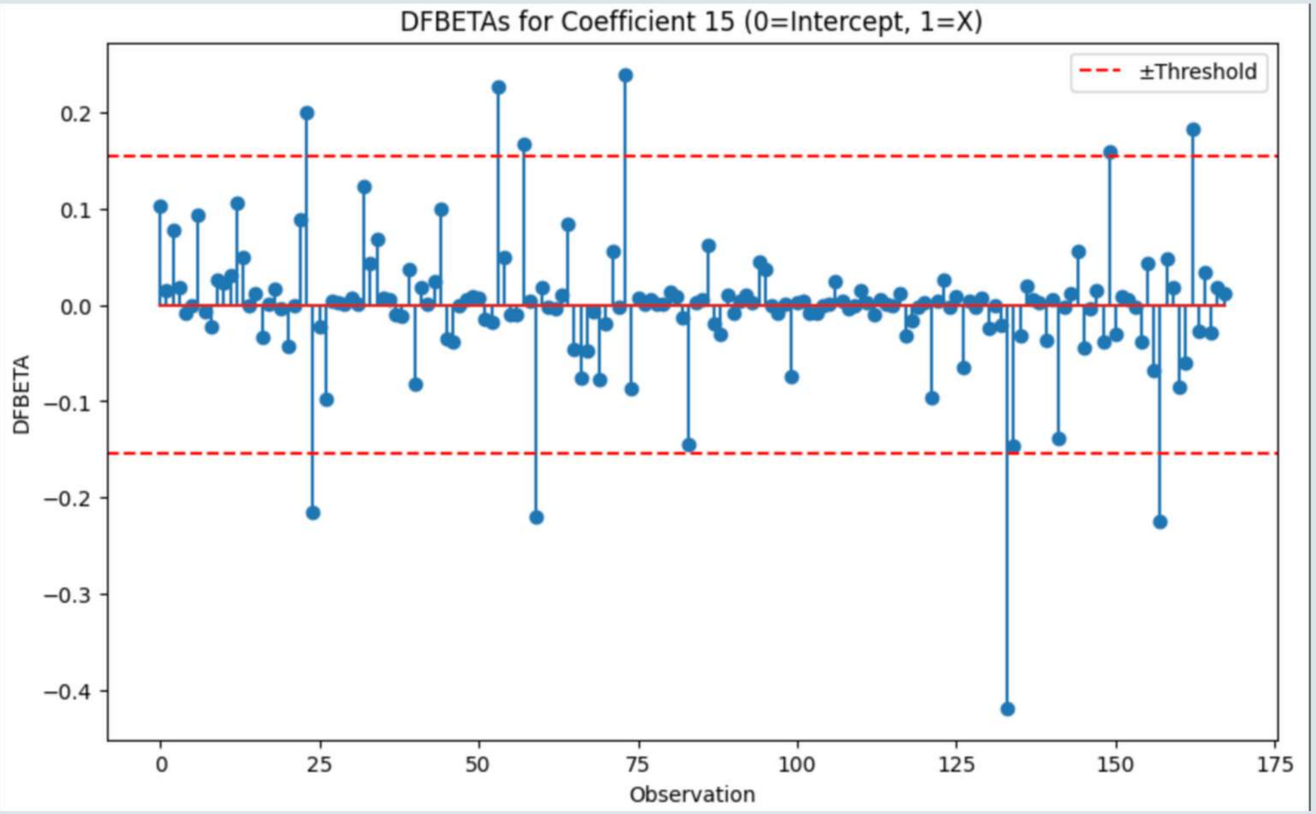
DFBETAs

NMVOC (β_{14})



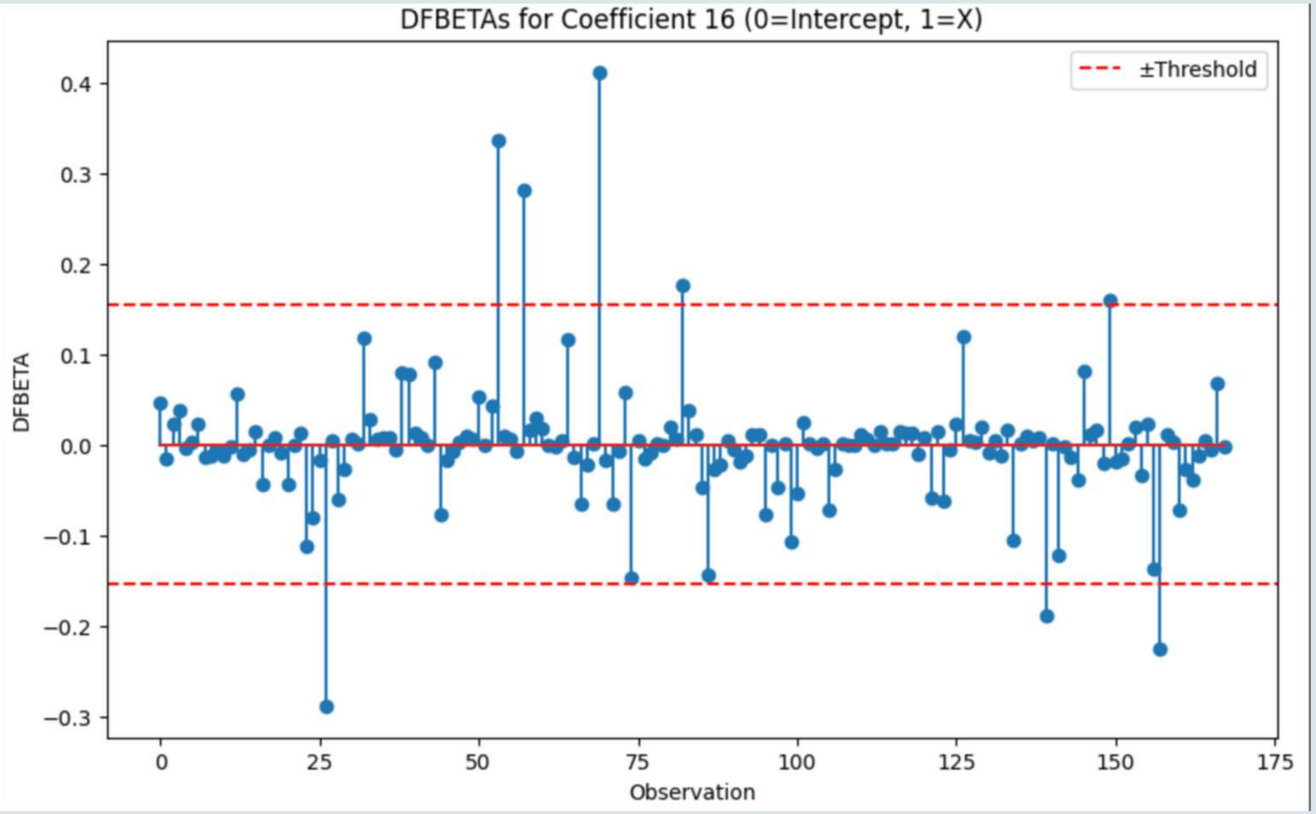
DFBETAs

NO_x (β_{15})



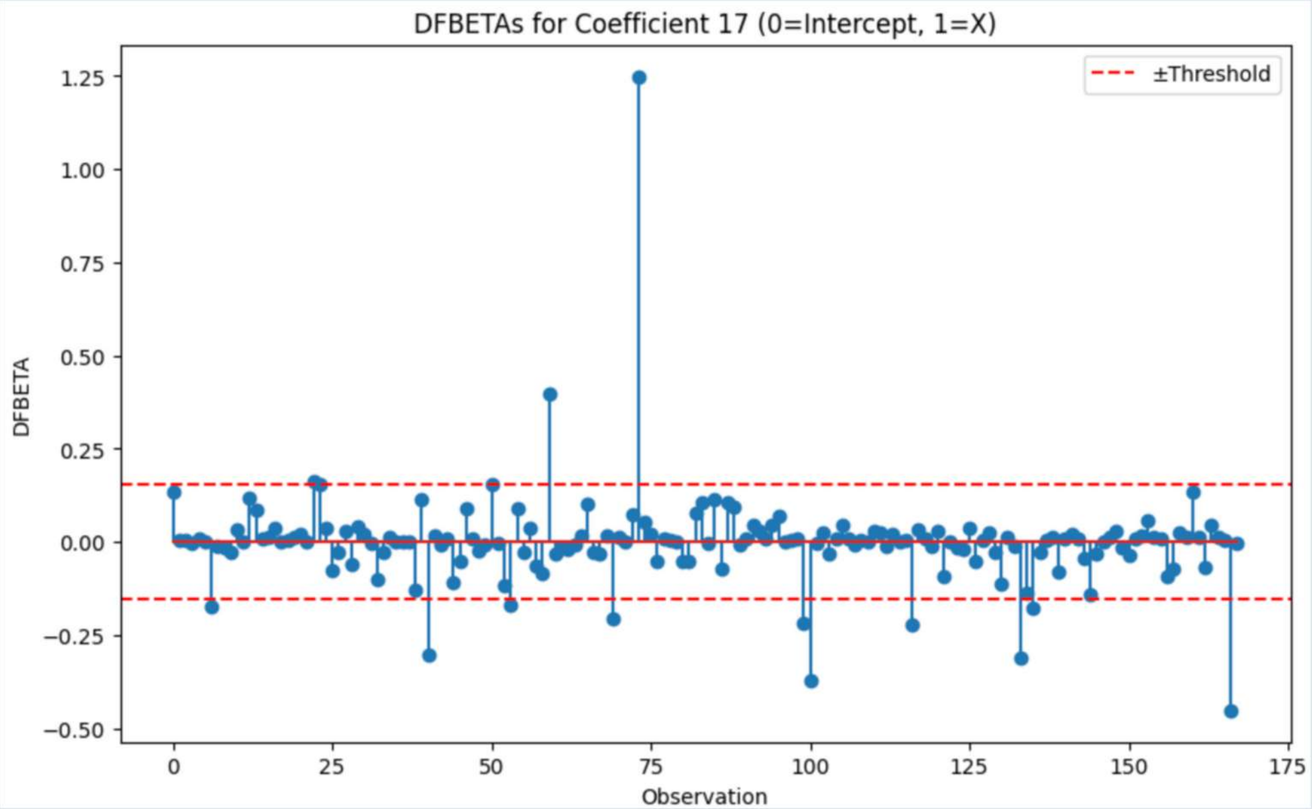
DFBETAs

PM10 (β_{16})



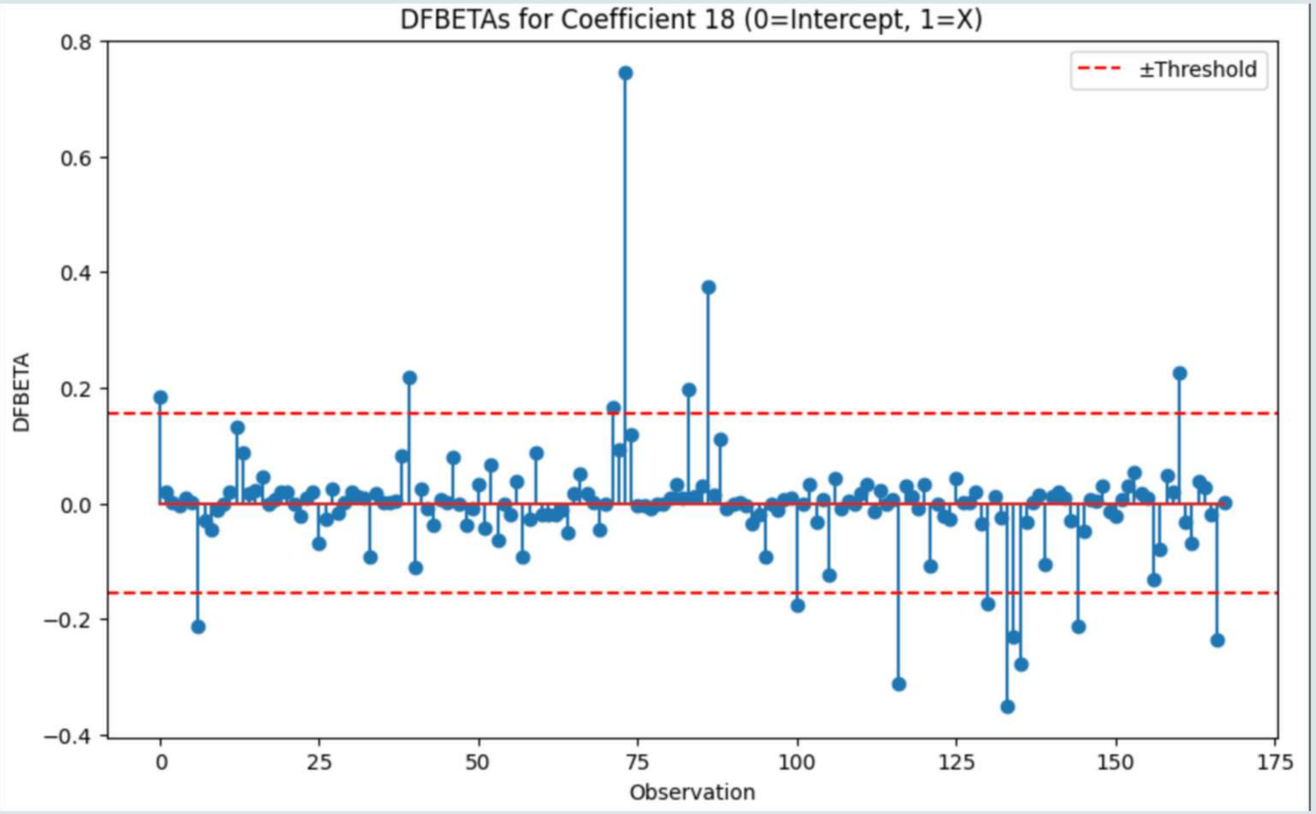
DFBETAs

Income_group_L (β_{17})



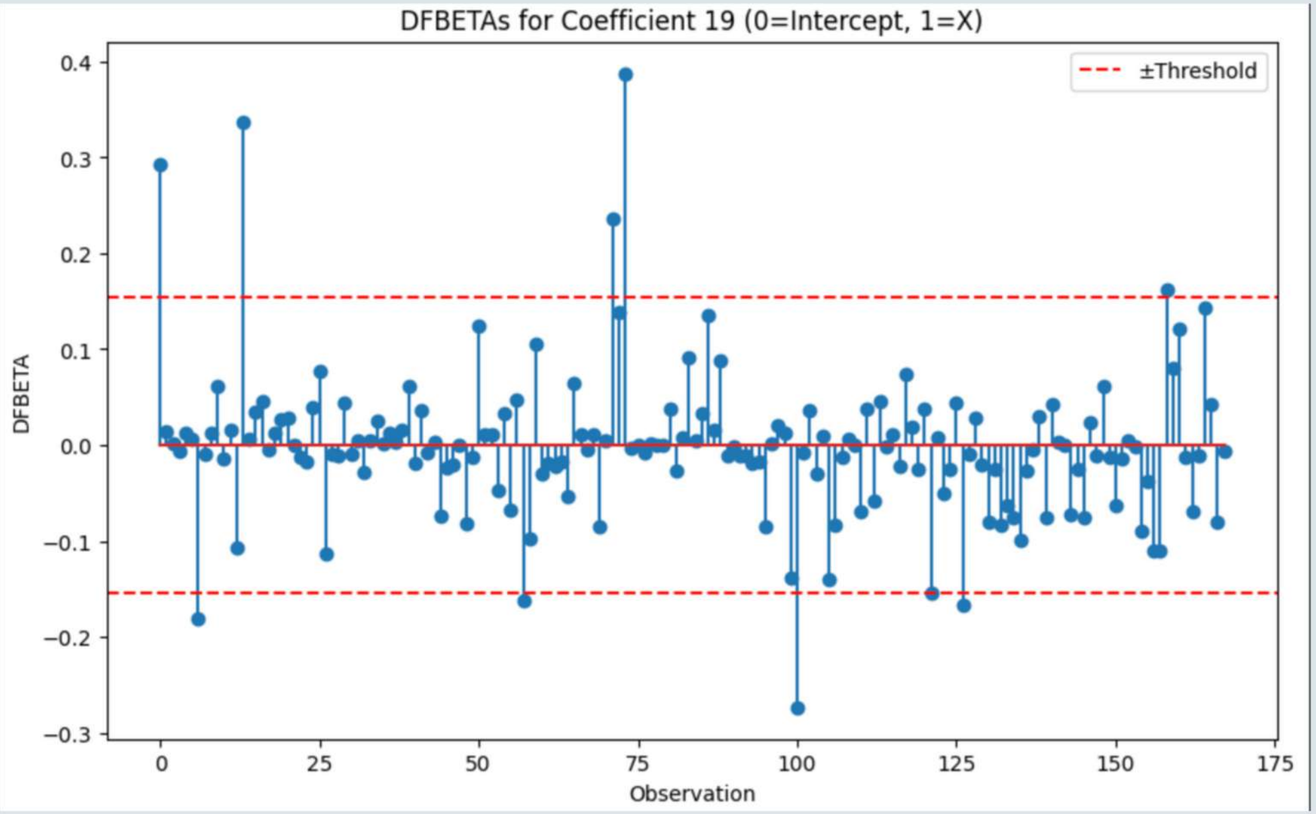
DFBETAs

Income_group_LM (β_{18})



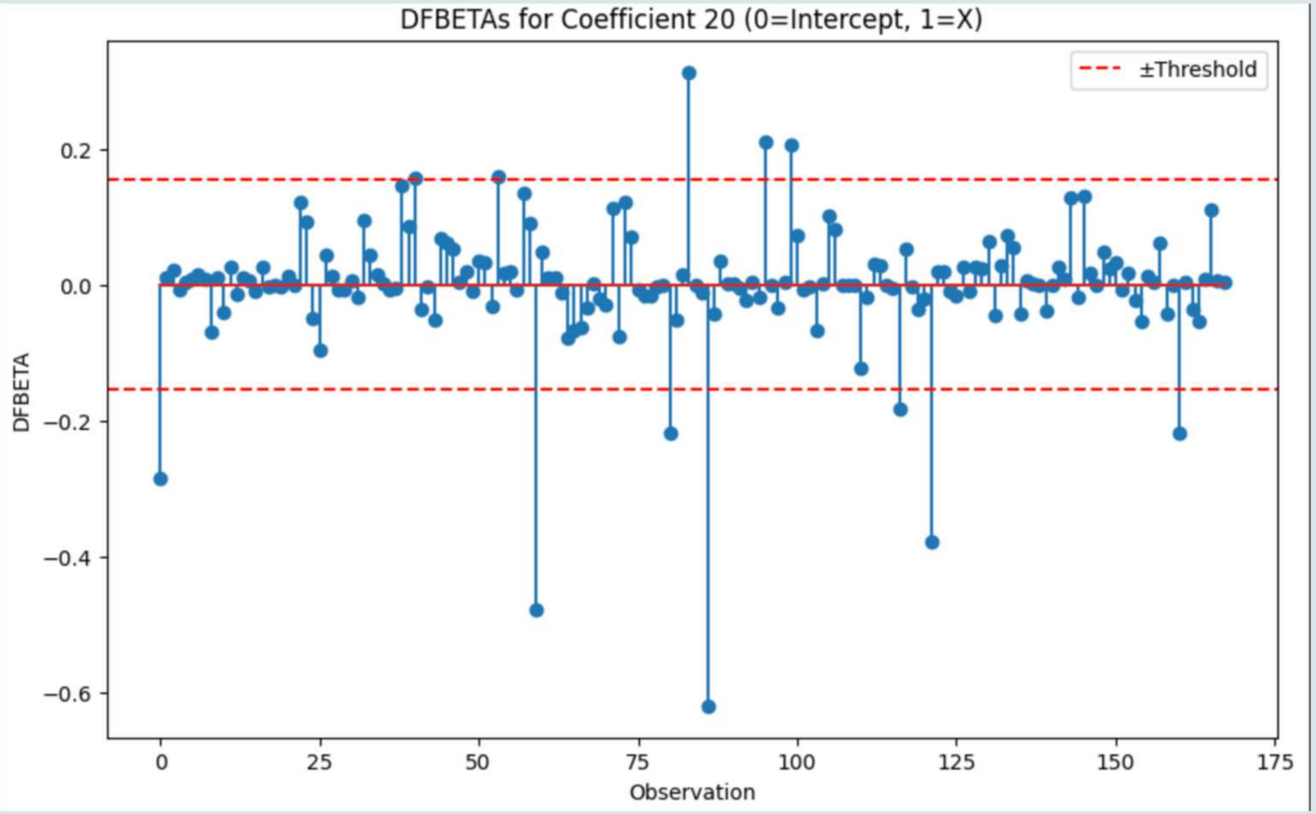
DFBETAs

Income_group_UM (β_{19})



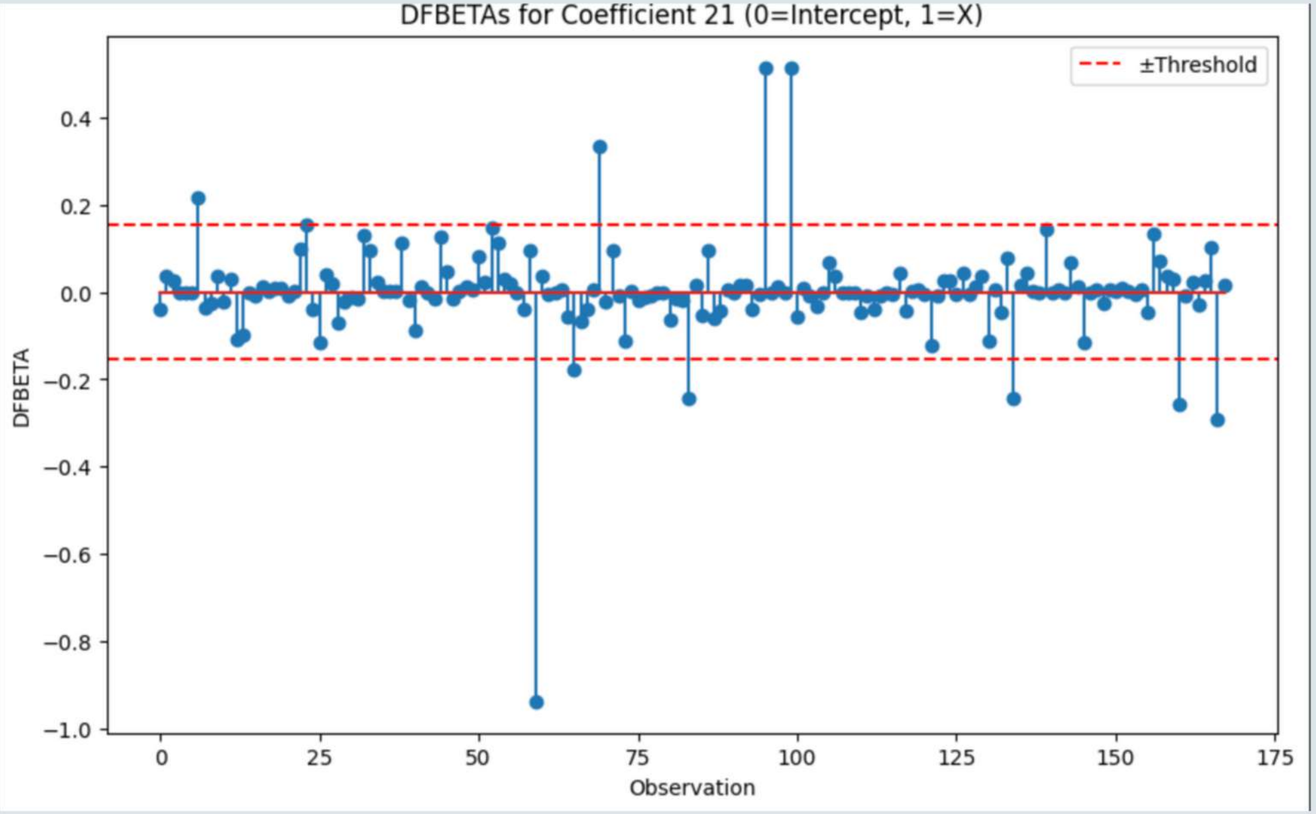
DFBETAs

ParentLocation_Americas (β_{20})



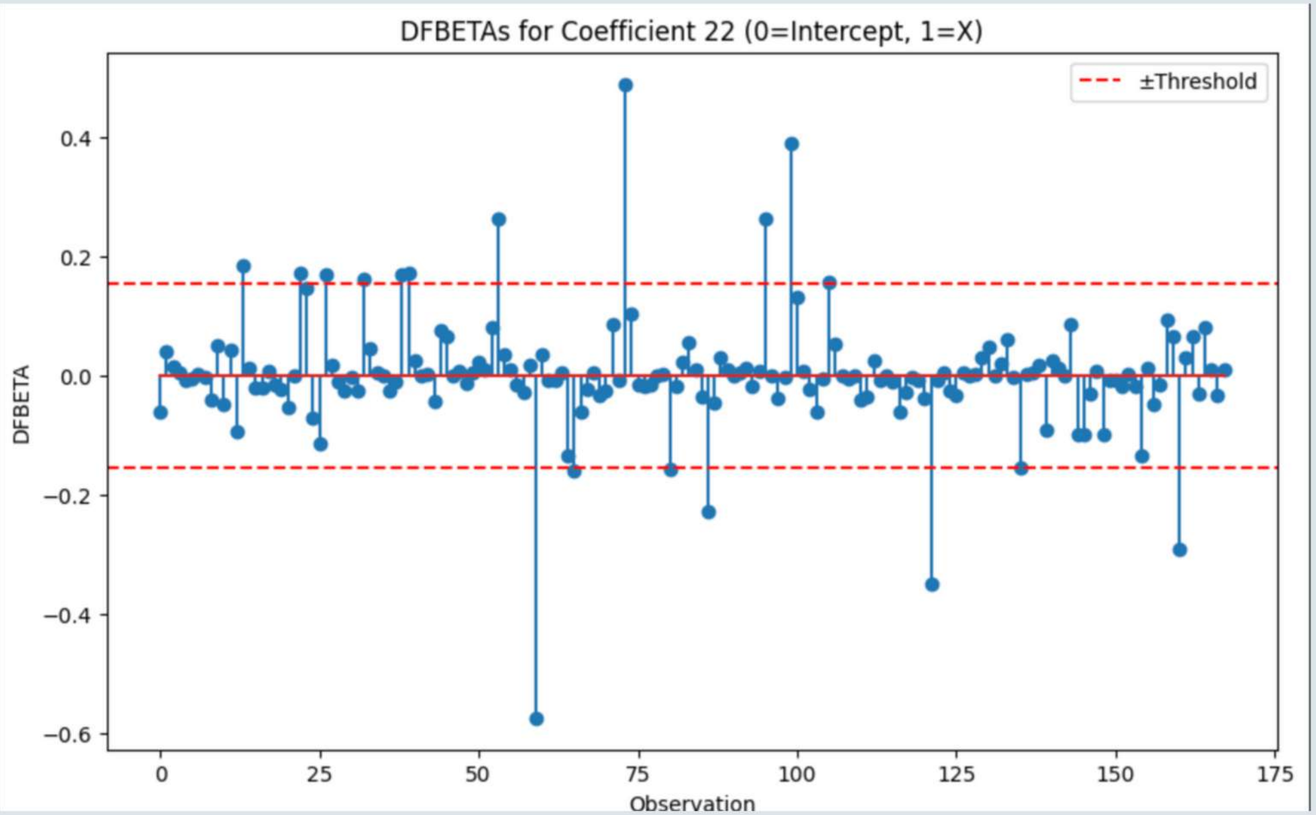
DFBETAs

ParentLocation_Eastern Mediterranean (β_{21})



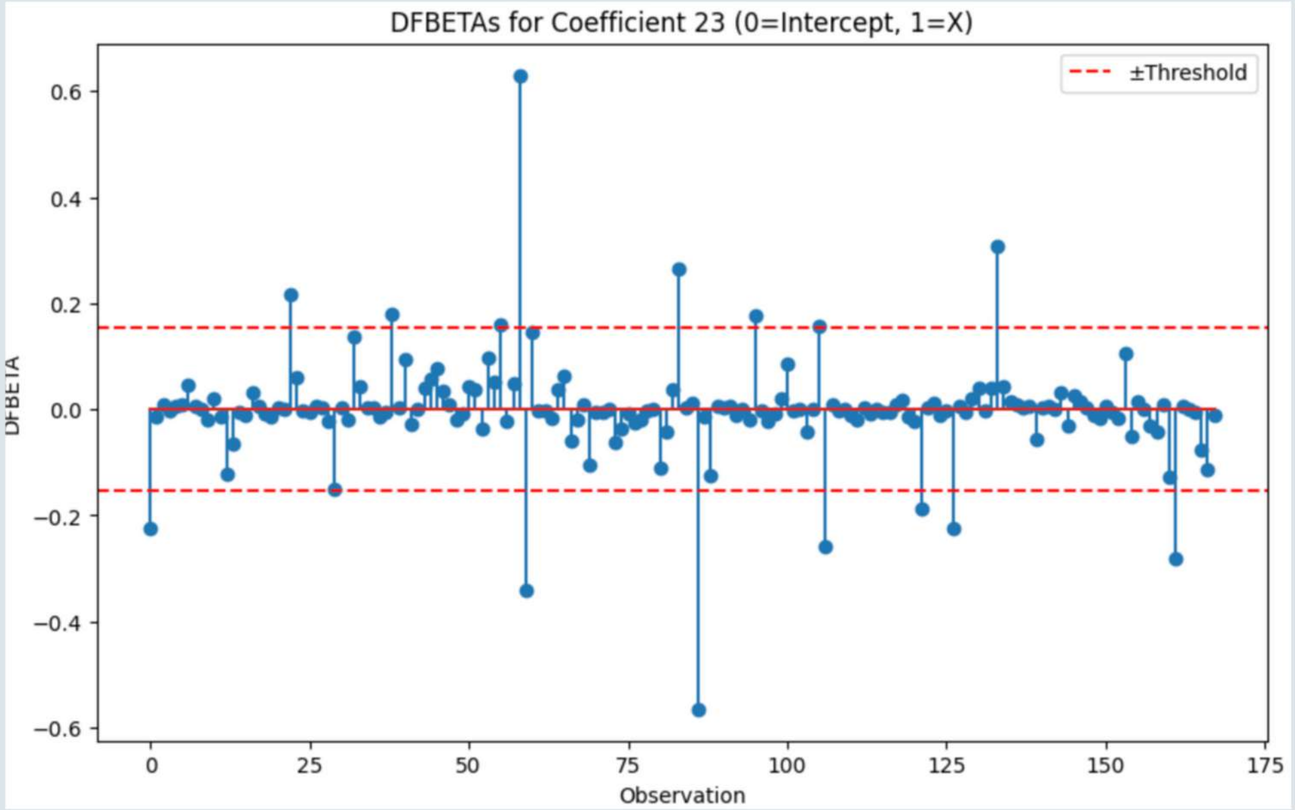
DFBETAs

ParentLocation_Europe (β_{22})



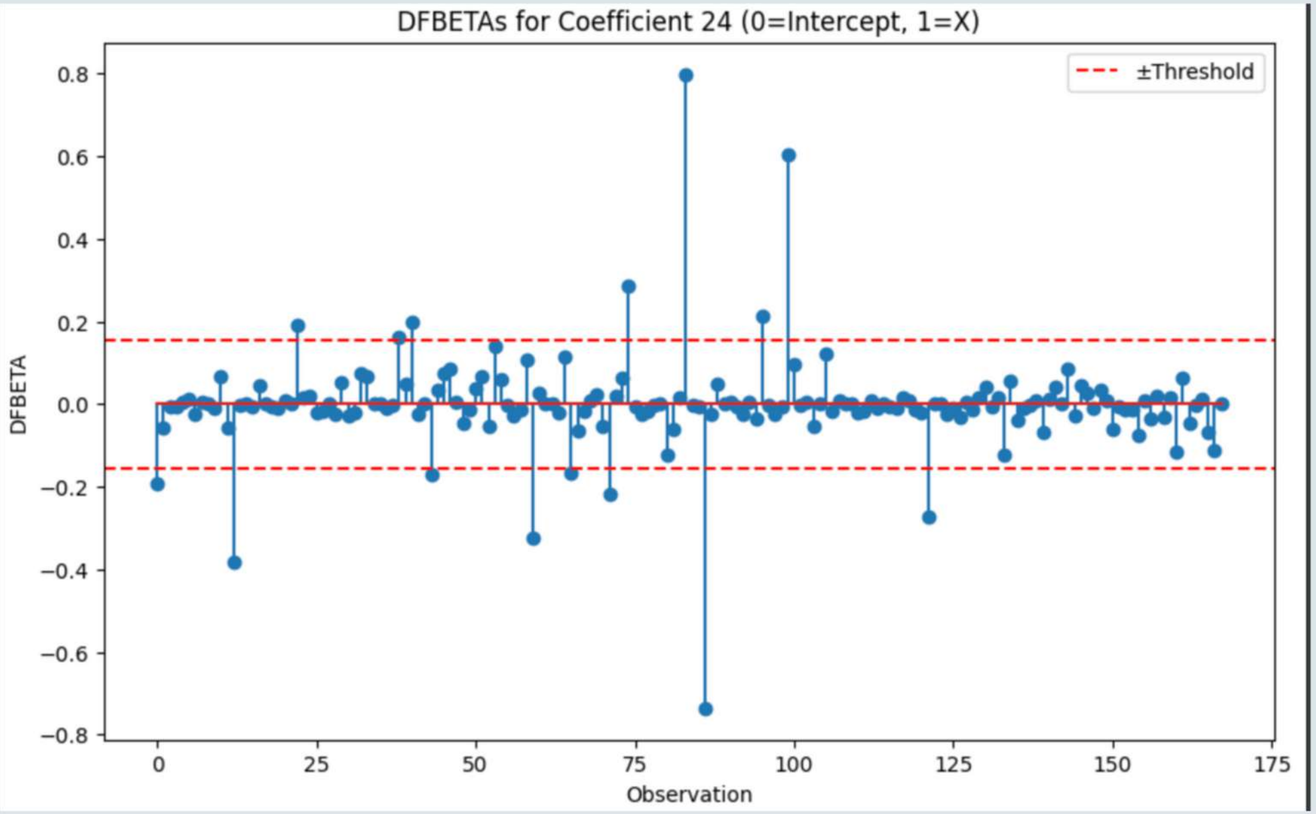
DFBETAs

ParentLocation_South-East Asia (β_{23})



DFBETAs

ParentLocation_Western Pacific (β_{24})



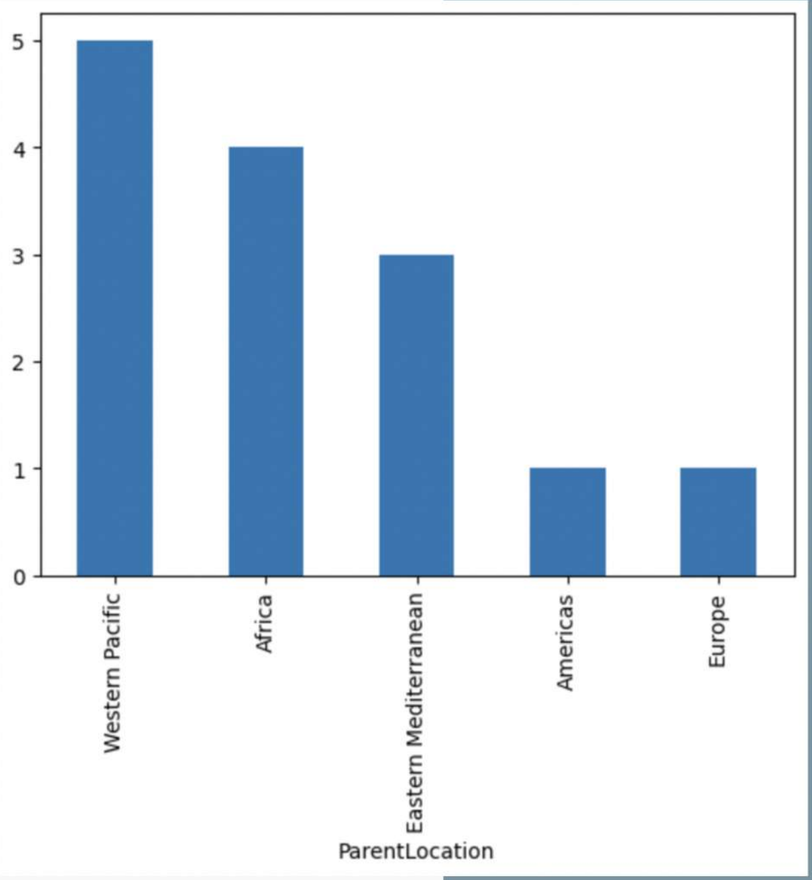
Outlier & Influence Diagnostics

Analysis of Parent Locations that correspond to high influential points

Interpretation of the plot:

- **Western Pacific , Africa ,** and **Eastern Mediterranean** have the highest number of countries with serious diagnostic concerns
- Such regions appear to struggle with structural concerns—like **weak data systems, under-resourced health services, or misaligned policies**
- **Europe and the Americas** show minimal concerns, likely due to stronger systems, reliable data, and better policy alignment.

Number of countries marked as high influence in each Parent Location



Formula Sheet for Outliers

MEASURE	FORMULA
Leverage	$h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$
Cook's Distance	$D_i = \frac{r_i^2}{p \cdot MSE} \cdot \frac{h_i}{(1 - h_i)^2}$
DFBETA	$\text{DFBETA}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{s_{\hat{\beta}_j}}$
DFFIT	$\text{DFFITS}_i = \frac{r_i}{s_{\hat{y}}} \cdot \sqrt{\frac{h_i}{1 - h_i}}$
COVRATIO	$\text{COVRATIO}_i = \frac{\det(\mathbf{X}^T \mathbf{X})}{\det(\mathbf{X}_i^T \mathbf{X}_i)}$

VIF of predictors after Stepwise Selection

feature	VIF
per_polluting_fuels	2.128108
Life expectancy	12.039377
ParentLocation_Eastern Mediterranean	1.841125
Income_group_LM	1.606118
ParentLocation_Europe	2.433549
ParentLocation_Western Pacific	1.234463
ParentLocation_South-East Asia	1.237896
Forest_cover	4.021953
Industrialisation	9.125508