[**Module 6 Group Project Final Report**](https://northeastern.instructure.com/courses/140385/assignments/1854124)

Logo

Description automatically generated with low confidence

College of Professional Studies, Northeastern University

ALY6150: Healthcare/Pharma Data and Applications

**Topic: Mortality Data Analysis - Assessing the Influence of Pollution on Circulatory and Respiratory Disease**

**By**

**Group 1**

Dinesh Chenaramchowdary

Shital Lathiya

Paritosh Singh

Yash Yadav

Prof. Dr. Roy Wada

June 27th, 2023

Contents

[Introduction 3](#_Toc139156077)

[Dataset 4](#_Toc139156078)

[Statistical Summary 6](#_Toc139156079)

[Analysis 11](#_Toc139156080)

[Modeling: 22](#_Toc139156081)

[Results: 31](#_Toc139156082)

[Conclusion 32](#_Toc139156083)

[References 34](#_Toc139156084)

# Introduction

This report presents a thorough examination of mortality data, with a specific focus on respiratory and circulatory diseases. The dataset utilized in this analysis provides valuable insights into the recorded number of deaths among different demographic groups and for various ICD codes. Furthermore, this report delves into the integration of pollution data sourced from the United States Environmental Protection Agency (EPA) to assess its potential impact on mortality rates.

The mortality dataset encompasses gender-specific fatality records, enabling an exploration of disparities between males and females in terms of mortality. By analyzing this dataset at the state level, we can identify variations across geographical regions and investigate potential factors that contribute to divergent mortality rates. To enrich the analysis, this report merges the mortality data with pollution data obtained from the EPA. The EPA is an authoritative body responsible for monitoring and regulating environmental pollution. By incorporating this pollution dataset, our aim is to investigate potential connections between different pollutants and mortality rates.

Alongside descriptive analysis, this report employs machine learning techniques used to predict the number of deaths based on relevant variables. By developing machine learning models, we can gain insights into the relationships between mortality and key factors such as pollution levels, demographic characteristics, and other variables of interest.

Through this comprehensive analysis, our objective is to enhance our understanding of mortality patterns, identify potential risk factors, identify effect of pollution on respiratory and circulatory diseases and to explore the predictive capabilities of various models in estimating the number of deaths. The findings presented in this report have the potential to inform public health interventions, guide policy formulation, and inspire further research on the intricate interactions among mortality, pollution, and other factors influencing population health.

# Dataset

The Compressed Mortality database contains mortality and population counts for all U.S. counties. The dataset we have chosen contains values from the year 1999-2016. A summary of the dataset is listed below:

|  |  |
| --- | --- |
| Number of rows | 419357 |
| Number of columns | 17 |
| Year range | 1999-2016 |
| Mortality ICD code format | ICD-10 |
| Number of states | 51 (50 + District of Columbia) |

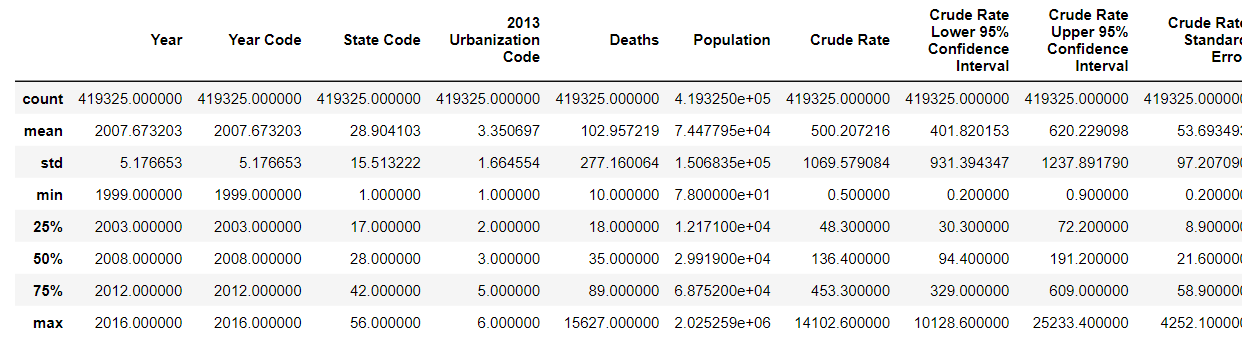
The data dictionary is as follows:

|  |  |
| --- | --- |
| Column Name | Description |
| ICD Chapter | The cause of death is indicated by the International Classification of Disease (ICD) code in the underlying cause of death section on each death certificate. Column contains textual representation of the ICD code. |
| ICD Chapter Code | Contains ICD-10 codes for the cause of death. Eg. A00-B99 |
| Year | Contains Year information. Ranges from 1999-2016 |
| Year Code | Same as Year |
| State | Contains name of State as text |
| State Code | Contains a number corresponding to each State |
| 2013 Urbanization | Contains text description of the region or locality type. Eg. Large Central Metro.  The Metro/Nonmetro list allows you to easily compare metropolitan and rural areas. The Metro group includes counties in these Urbanization categories: Large Central Metro, Large Fringe Metro, Medium Metro, and Small Metro. The Nonmetro group includes counties in these Urbanization categories: Micropolitan (non-metro) and NonCore (non-metro). |
| 2013 Urbanization Code | Contains code corresponding to the Urbanization zone. Ranges from 1-6 |
| Gender | Gender of cohorts as text. Contains Male or Female |
| Age Group | Age group of cohorts as text. Eg. 65-74 years |
| Age Group Code | Contains a code corresponding to the age group without any alphabets. Eg. 65-74 |
| Deaths | Count of deaths |
| Population | Size of population |
| Crude Rate | Rate of death. Crude Rates are expressed as the number of deaths reported each calendar year per factor. The default factor is per 100,000 population, reporting the death rate per 100,000 persons. Rates are marked as "unreliable" when the death count is less than 20. |
| Crude Rate Lower 95% Confidence Interval | Gives the lower 95% confidence threshold for the observed crude rate. |
| Crude Rate Upper 95% Confidence Interval | Gives the upper 95% confidence threshold for the observed crude rate. |
| Crude Rate Standard Error | Gives the standard error for the crude rate calculation. |

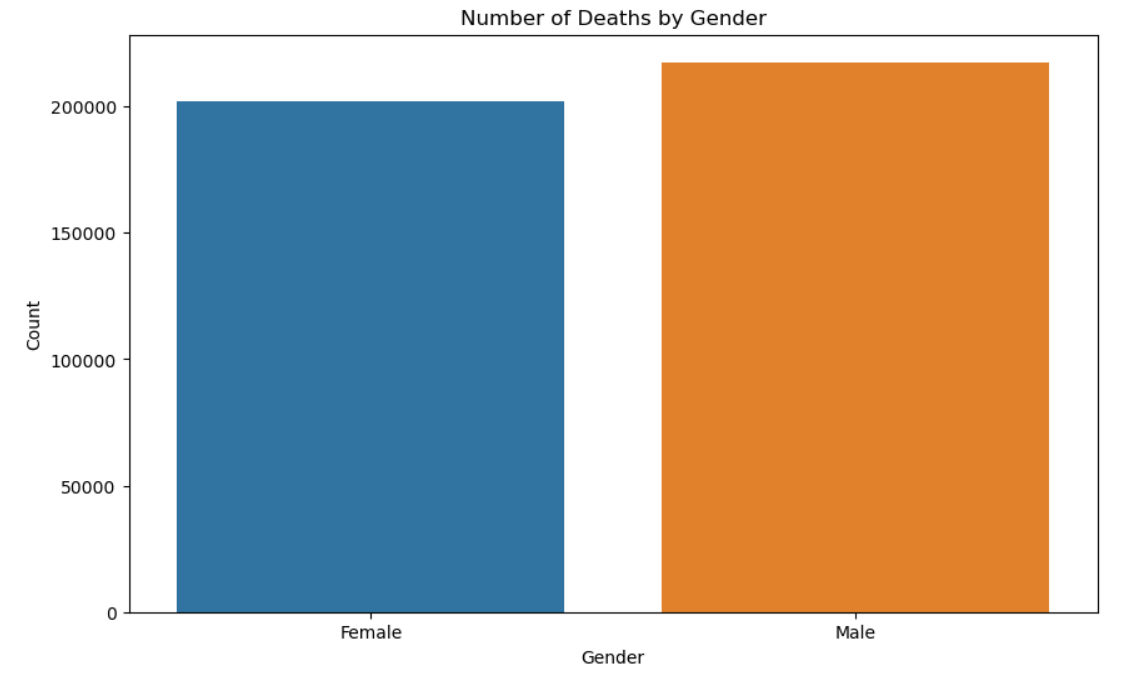
# 

# Statistical Summary

* By all columns, individually.



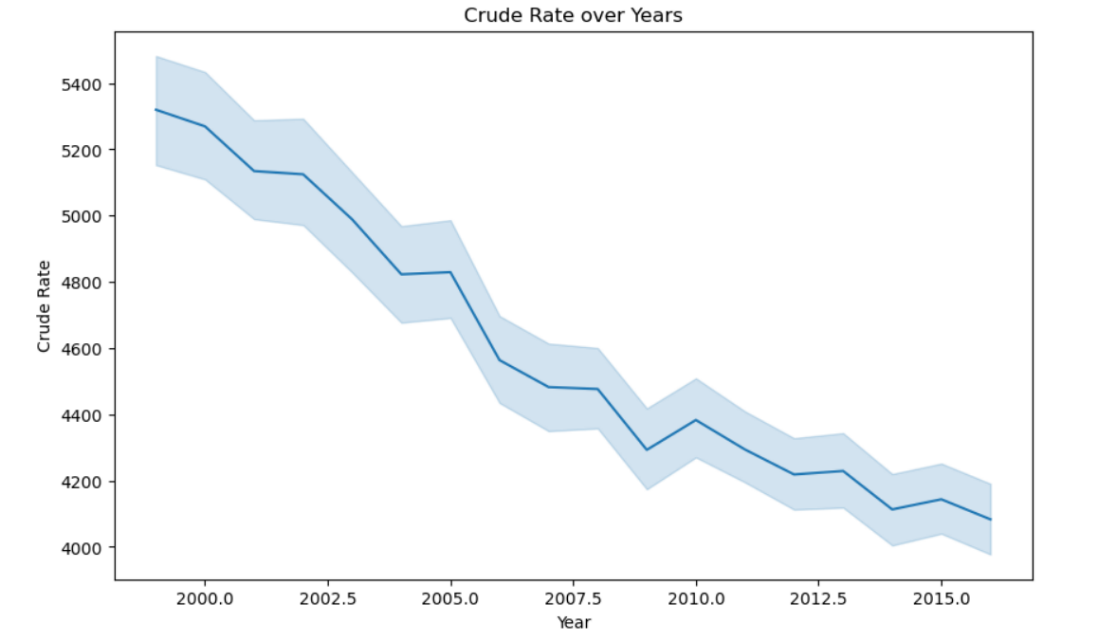
Bar plot using **seaborn's countplot** function. It specifies the variable to be plotted on the x-axis as **'Gender'** and the data source as the DataFrame **df**. The **countplot** function counts the occurrences of each unique value in the **'Gender'** column and displays them as bars.



Line plot (shown below) using **seaborn's lineplot** function. It specifies the variable to be plotted on the x-axis as **'Year'**, the variable to be plotted on the y-axis as **'Crude Rate**', and the data source as the DataFrame **df**. The line plot function connects the data points in a continuous line, showing the trend of the '**Crude Rate**' over the different '**Year**' values.

The Crude Rate in mortality data for the United States from 1999 to 2016 refers to a measure of mortality that is adjusted for the size and age distribution of the population. It represents the number of deaths per 1,000 or 100,000 population during a specific time period. In this context, the Crude Rate is calculated by dividing the total number of deaths in a given year by the corresponding population size and then multiplying by a scaling factor (e.g., 1,000 or 100,000) to express the rate per 1,000 or 100,000 population.

The Crude Rate provides a standardized measure to compare mortality across different populations and time periods. It allows for the assessment of overall mortality trends and patterns over a specific time span, in this case, from 1999 to 2016. By analyzing the Crude Rate over these years, it is possible to identify changes in mortality rates, understand the impact of various factors on population health, track the effectiveness of public health interventions, and identify potential areas of concern or improvement in healthcare systems. The Crude Rate can also serve as a basis for making international or regional comparisons of mortality rates.

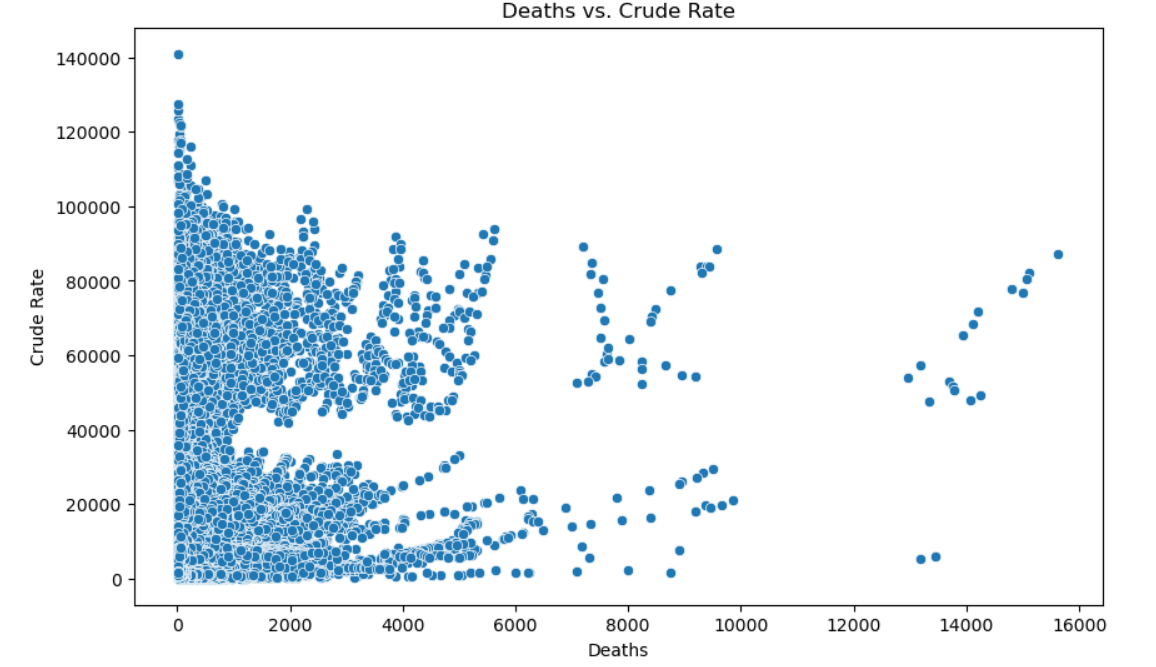


A scatter graph depicting "Deaths vs. Crude Rate" represents the relationship between the number of deaths and the corresponding Crude Rate for a specific population or time period. Unlike a line graph, a scatter graph shows individual data points rather than connecting them with a line. It helps visualize the distribution and clustering of data points and provides insights into the relationship between the variables.

In a scatter graph of "Deaths vs. Crude Rate," each data point represents a specific observation or data instance. The y-axis represents the Crude Rate, which indicates the mortality rate or deaths per 1,000 or 100,000 population. The x-axis represents the number of deaths, showing the actual count of deaths that occurred within a particular population or period.

The scatter graph can signify various patterns and relationships:

* **Positive correlation**: If the data points are clustered in an upward trend or show a general increasing pattern from left to right, it suggests a positive correlation between the Crude Rate and the number of deaths. This indicates that as the Crude Rate increases, the number of deaths also tends to increase.
* **Negative correlation**: If the data points are clustered in a downward trend or show a general decreasing pattern from left to right, it signifies a negative correlation between the Crude Rate and the number of deaths. In this case, as the Crude Rate decreases, the number of deaths also tends to decrease.
* **No correlation**: If the data points are scattered randomly without any discernible pattern, it suggests no significant correlation between the Crude Rate and the number of deaths. This implies that the mortality rate and the number of deaths is not strongly associated or influenced by each other.



The dataset will be analyzed in the report using a structured methodology, starting with exploratory data analysis to learn more about the properties and distributions of the variables. Additionally, we will investigate the significance and impact of various features on a customer's choice to apply for a credit card. Credit card issuers will be able to adapt their marketing strategies because of this analysis, which will aid in understanding the crucial factors that influence customer interest.

The report's goal is to offer a thorough analysis of the dataset, highlighting important insights into the variables affecting the prediction of credit card leads. We aim to empower credit card issuers to make informed decisions, improve customer targeting, and increase the conversion of potential leads into happy credit card customers by utilizing the power of data mining and classification techniques.

Our current dataset is about mortality which contains information about deaths and population for various ICD codes grouped by state, year, age group and urbanization. We decided to add pollution data here based on state and year to find correlation between pollution and mortality rate.

* Pollution Dataset:

**A screenshot of a computer

Description automatically generated with low confidence**

Figure 1: Emission Dataset

The pollution dataset consists of PM10-PRI (Particulate Matter 10 micrometers or less) emission values obtained from the United States Environmental Protection Agency (EPA). The dataset provides information on pollution levels measured across different years and states within the United States. These values are crucial in understanding the air quality and potential health impacts associated with particulate matter pollution. On the other hand, the mortality dataset contains information about deaths, specifically focusing on state and year. This dataset allows us to analyze mortality rates and trends across different states over a period of 1999 to 2016.

* Merged Dataset:

**A picture containing text, font, line, screenshot

Description automatically generated**

Figure 2: Mortality + Pollution data

To gain a comprehensive understanding of the relationship between pollution and mortality, we merged these two datasets based on the common fields of state and year. By merging the datasets, we can explore potential correlations and analyze the impact of particulate matter pollution on mortality rates at the state level.

This merged dataset provides a valuable resource for conducting in-depth analyses and investigations into the association between pollution and mortality. It allows researchers, policymakers, and public health professionals to assess the health risks posed by particulate matter pollution and develop targeted interventions and policies to mitigate its adverse effects. This dataset contributes to our understanding of the complex relationship between pollution and public health, aiding in the formulation of evidence-based strategies to protect and improve the well-being of individuals and communities.

# Analysis

* Summary Statistics for full dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **​** | **Year​** | **Year Code​** | **State Code​** | **2013 Urbanization Code​** | **Deaths​** | **Emissions\_CO​** | **Emissions\_PM10​** |
| **count**​ | 419356​ | 419356​ | 419356​ | 419356​ | 419356​ | 419356​ | 419356​ |
| **mean**​ | 2007.67​ | 2007.67​ | 28.90​ | 3.35​ | 102.95​ | 1797.75​ | 370.55​ |
| **std**​ | 5.18​ | 5.18​ | 15.51​ | 1.66​ | 277.15​ | 1434.20​ | 333.49​ |
| **min**​ | 1999​ | 1999​ | 1​ | 1​ | 10​ | 29.79​ | 3.46​ |
| **25%**​ | 2003​ | 2003​ | 17​ | 2​ | 18​ | 851.75​ | 197.29​ |
| **50%**​ | 2008​ | 2008​ | 28​ | 3​ | 35​ | 1427.23​ | 267.73​ |
| **75%**​ | 2012​ | 2012​ | 42​ | 5​ | 89​ | 2281.44​ | 424.83​ |
| **max**​ | 2016​ | 2016​ | 56​ | 6​ | 15627​ | 9618.34​ | 2697.85​ |

Figure 3: Summary statistics

These statistics provide summary information about the variables 'Deaths' and 'Population' based on the given dataset.

The mean (average) gives the average value of each variable. For 'Deaths', the mean is approximately 102.95, indicating the average number of deaths. Similarly, for 'Emissions', the mean is approximately 372.09, representing the average emissions value. The standard deviation (std) measures the dispersion or variability of the data around the mean. A higher standard deviation suggests greater variability in the values. In this case, 'Deaths' has a standard deviation of approximately 277.15, indicating a relatively high variability in the number of deaths. 'Emissions' has a standard deviation of approximately 333.35, indicating significant variability in emission values as well. The minimum (min) represents the smallest value observed for each variable. In this case, the minimum value for 'Deaths' is 10, indicating the lowest recorded number of deaths. For 'Emissions', the minimum value is 7.45185, representing the lowest emission value. The maximum (max) represents the largest value observed for each variable. In this case, the maximum value for 'Deaths' is 15,627, indicating the highest recorded number of deaths. For 'Emissions', the maximum value is 2,697.854607, representing the highest emission value in the dataset.

These summary statistics offer insights into the distribution, central tendency, and variability of the 'Deaths' and 'Emissions' variables in the dataset.

* Summary statistics by Gender

**A screenshot of a computer

Description automatically generated with medium confidence**

Figure 4: Summary statistics by Gender

The summary statistics for the 'Deaths' variable, categorized by gender, reveal interesting patterns. On average, females have a slightly higher number of recorded deaths (mean: 107.50) compared to males (mean: 98.73). The median values indicate that 50% of recorded deaths for both females (median: 34.0) and males (median: 35.0) fall at or below these values, suggesting a relatively similar central tendency. However, there is greater variability in the number of deaths for females, as indicated by a higher standard deviation (approximately 311.45), compared to males (approximately 240.93). This suggests that the range of deaths among females is more spread out. The minimum values for both genders are the same at 10 deaths, while the maximum values differ, with females having a higher maximum of 15,627 deaths compared to males with a maximum of 9,494 deaths. Overall, these statistics highlight the variations in recorded deaths by gender, indicating higher averages and greater variability among females.

* Summary Statistics by State

**A screenshot of a graph

Description automatically generated with low confidence**

Figure 5: Summary statistics by State

The summary statistics for deaths and emissions, categorized by state, provide insights into important trends and features within the dataset.

For deaths:

- The average number of deaths per state ranges from a minimum of 24.38 in Alaska to a maximum of 280.42 in California. The overall mean is approximately 459.91 deaths.

- The median values range from a minimum of 18 deaths in Alaska to a maximum of 56 deaths in Florida.

- The standard deviation indicates the variability in the number of deaths within each state. States like California and Illinois have higher standard deviations, suggesting greater variation in the number of deaths, while states like Connecticut and Rhode Island have relatively lower standard deviations, indicating less variability.

- The minimum and maximum values represent the range of deaths observed within each state.

For emissions:

- The average emissions per state range from a minimum of 29.00 in Hawaii to a maximum of 1468.25 in Texas. The overall mean is approximately 459.91 emissions.

- The median values range from a minimum of 19.06 emissions in Delaware to a maximum of 455.24 emissions in Missouri.

- The standard deviation reflects the variability in emissions levels across states. States like Texas and California have higher standard deviations, indicating a wider range of emissions values, while states like Rhode Island and Vermont have lower standard deviations, suggesting less variability.

Overall, these statistics highlight significant variations in deaths and emissions across different states. Some states exhibit higher averages and greater variability, indicating potentially higher mortality rates and varying levels of pollution.

* Summary Statistics by year

A picture containing text, screenshot, document, number

Description automatically generated

Figure 6: Summary statistics by Year

The summary statistics reveal important patterns in deaths and emissions over the years. The data shows that the average number of deaths ranged from around 100 to 107 per year, with some years experiencing significantly higher mortality rates. However, the median number of deaths remained relatively stable at 32 to 38 per year. In terms of emissions, the mean values decreased from 591.61 to 323.06, indicating a gradual reduction over time. The median emissions remained consistent around 504.15 to 239.63. These statistics demonstrate both the overall trends of declining deaths and emissions and the significant variations within the data.

In addition to the main trends, there are some other interesting observations within the summary statistics. The standard deviation of deaths ranged from 271.20 to 292.72, indicating considerable variability in mortality rates across different years. Similarly, the standard deviation of emissions ranged from 259.74 to 291.27, highlighting the fluctuations in emission levels over time. Looking at the extreme values, the minimum number of deaths and emissions remained constant at 10 throughout the years. On the other hand, the maximum number of deaths varied from 13,428 to 14,291, showcasing the years with higher mortality rates. Similarly, the maximum emissions ranged from 1,164.56 to 2,697.85, representing the years with the highest levels of pollution.

* Adding New geographical variables

To enhance the mortality dataset with additional geographical information, we added three new columns: state short name (two-letter abbreviation), latitude, and longitude. The coordinates for each state of the USA can be obtained from the dataset available on developers.google.com, which provides comprehensive information on geographic locations. By incorporating the state short name column, the dataset will include a standardized two-letter abbreviation for each state, allowing for easier data manipulation and analysis. This column can serve as a unique identifier for each state within the dataset.

A picture containing text, font, line, screenshot

Description automatically generated

Figure 7: Added new variables Latitude and Longitude

The latitude and longitude columns will provide precise geographic coordinates for each state. These coordinates are essential for spatial analysis and mapping applications. By including latitude and longitude values, we can explore geographical patterns in mortality data, identify regional variations, and assess the impact of location on mortality rates. This expanded dataset enables a wide range of investigations, such as examining regional disparities in mortality, exploring spatial clusters of high or low mortality rates, and identifying areas with specific health challenges.

The combination of mortality data with geographic information enhances our understanding of the spatial aspects of public health and supports evidence-based decision-making. It facilitates the identification of geographical hotspots and aids in the development of targeted interventions to address health disparities and improve health outcomes at the state level.

* Visualizations

1. Map based visualizations.

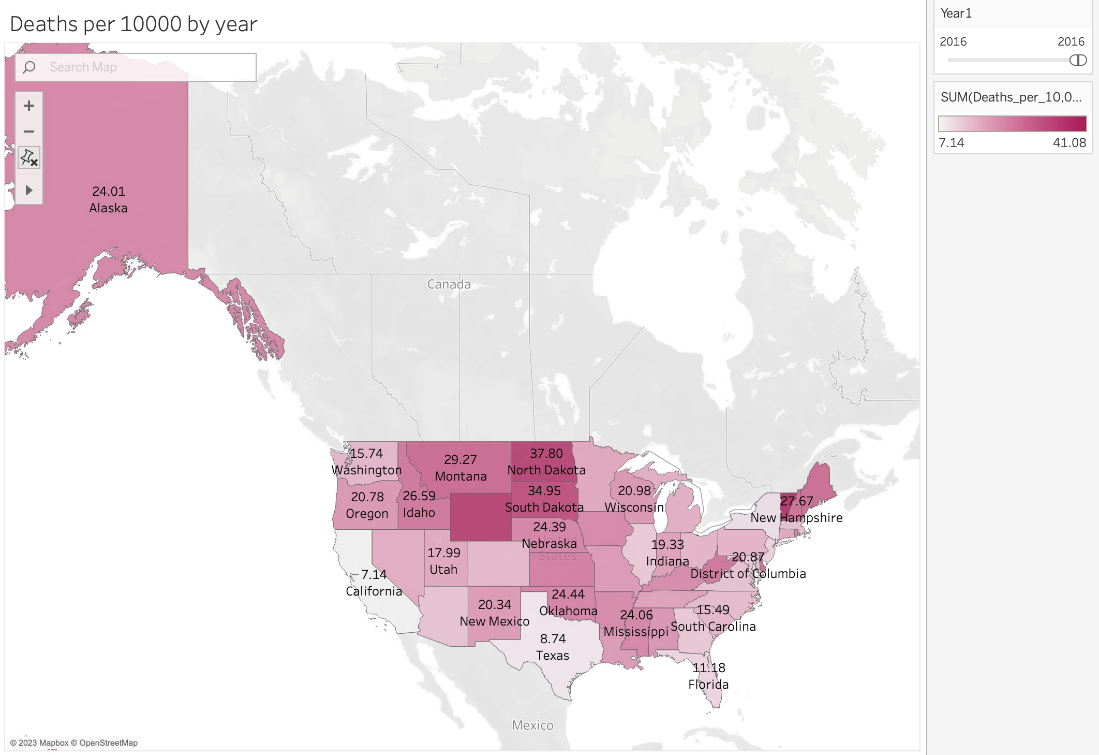


Figure 8: Deaths per 1000 for year 2016

* North Dakota shows the highest death rate at 37.8 deaths per 10,000 in 2016.
* California has high number of total deaths but it’s quite low deaths per capita due to its large population and size.

1. Line visualizations

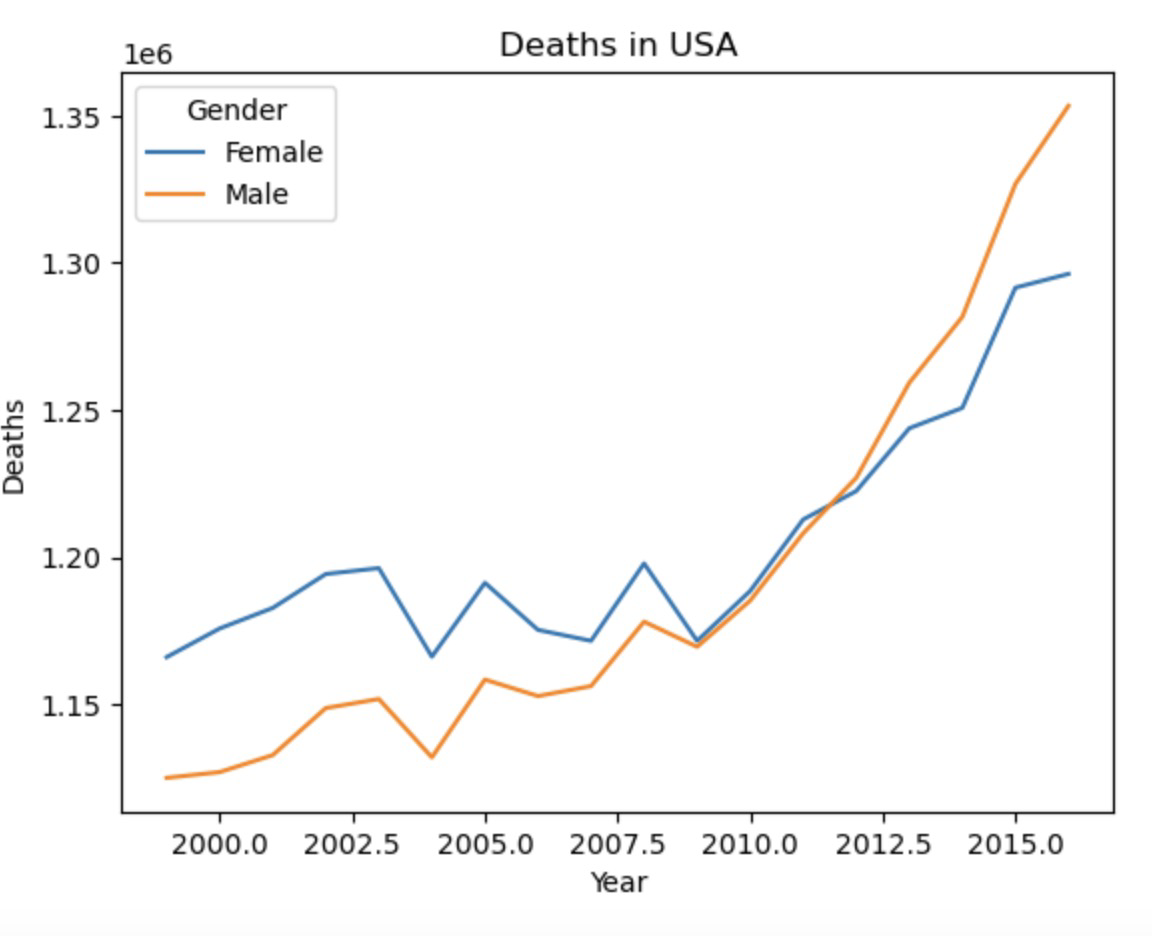


Figure 9: Deaths of Male vs Female from 2000 to 2015

We plotted total number of deaths across time. We see that there seems to be a stronger increase in male deaths compared to female deaths. The male deaths went from around 1,120,000 to 1,350,000. The female deaths rose from 1,170,000 to about 1,300,000. We should normalize the data against population before looking further.

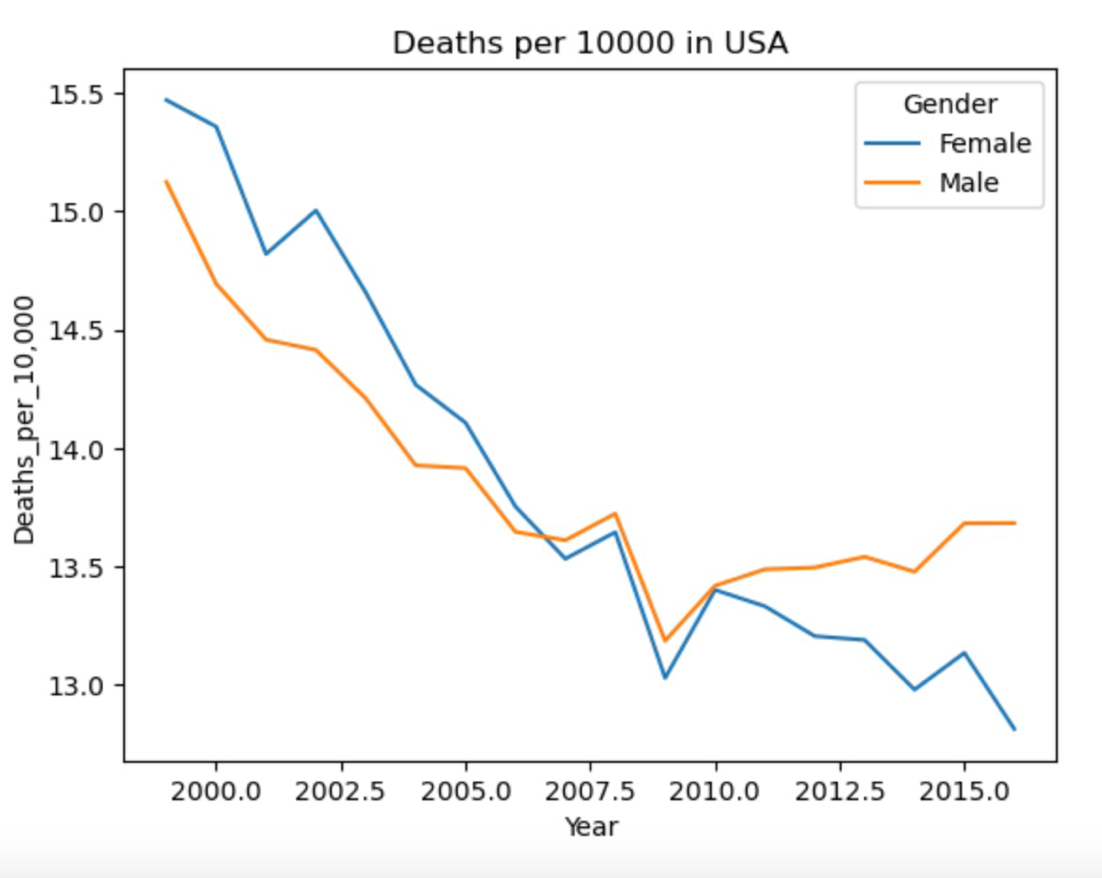


Figure 10: Deaths per 1000 in USA

Normalizing by population, we see that Female deaths seemed to show a roughly steady decline, whereas male deaths increased slightly from 2008 onwards, ending at 13.5 deaths per 10000.

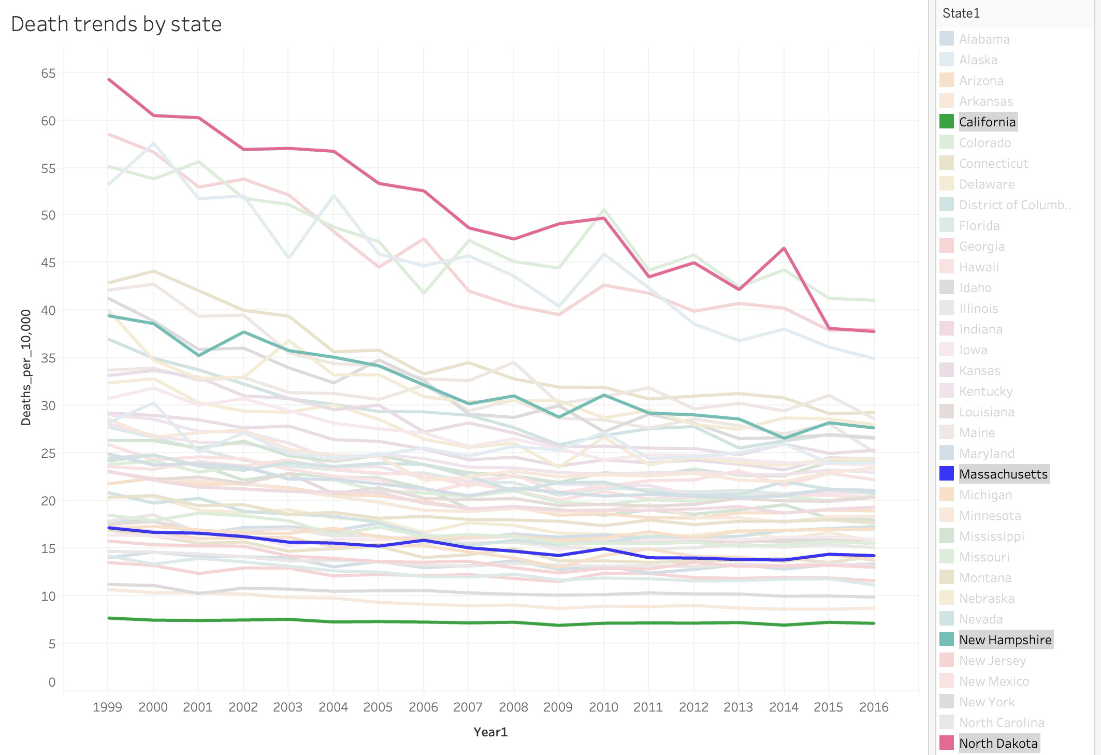


Figure 11: Death trends by state

Trend lines of all states. We highlighted a few states for comparison. California has one of the lowest death rates per capita at around 7 per 10,000. North Dakota has the highest but it’s showing a decreasing trend, starting at around 65 deaths per 10,000 in 1999 and ending at around 40 deaths per 10,000. Massachusetts sits somewhere in the middle with a steady 17 deaths per 10,000.

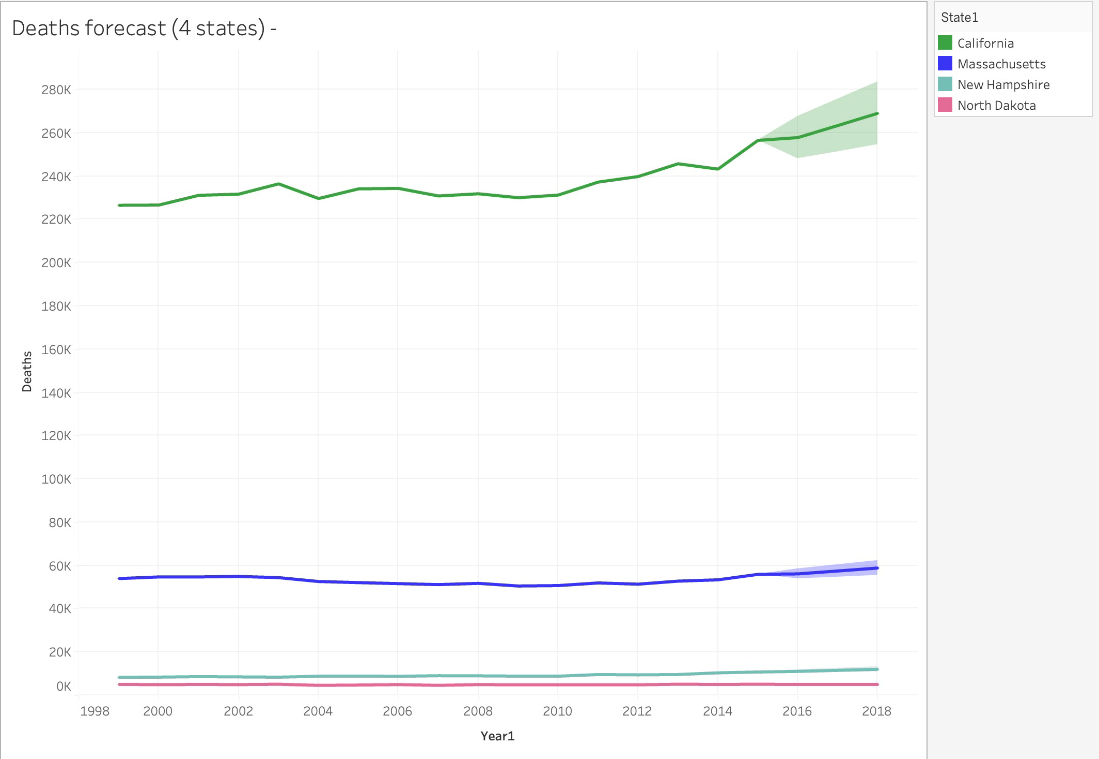


Figure 12: Deaths forecast

Based on the 4 states, looking at deaths directly we see that North Dakota has a very small population, which is why even with fewer deaths, we see a higher death per capita. Death projections for California have the most uncertainty, whereas other states are relatively stable.

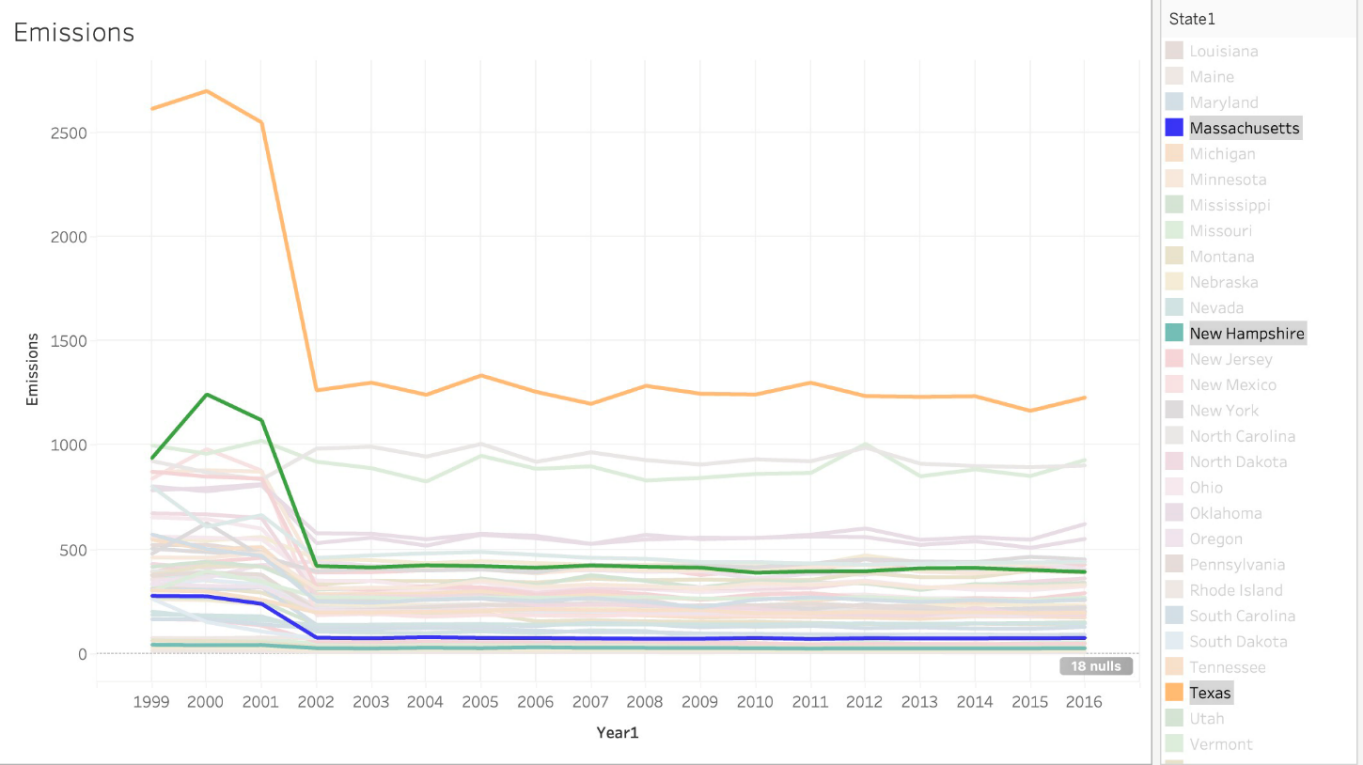


Figure 13: Emissions over years

We also looked at emission trends of PM10 in various states over the years. Most states appear to have a fixed emission rate with minimal decreases over the last few years. This does indicate that perhaps these emissions may not have a lot of predictive power for this task. Texas has the highest emissions at around 1250.

# Modeling:

The models we used were Decision Tree, Random Forest and SVR(Support Vector Regressor). These models have been utilized for both Circulatory and Respiratory diseases. And we tried to predict mortality rate per 10,000 deaths for year 2014, 2015 and 2016. Below are the usage of models and reasons to use them.

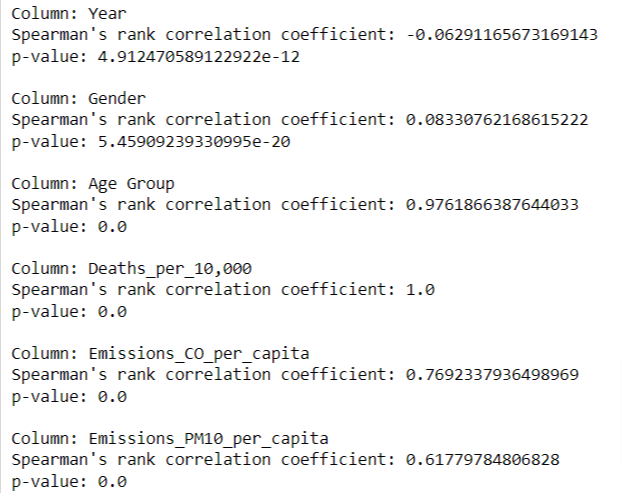
The dataset used in this modeling was ‘Year’, ‘Gender’, ‘Age Group’, ‘Deaths\_per\_10,000’, ‘Emissions\_CO\_per\_capita’, ‘Emissions\_PM10\_per\_capita’, ‘State’. "Carbon monoxide emissions amount per capita" refers to the measurement of carbon monoxide (CO) emissions produced by a particular state divided by its population. It is a way to quantify the average carbon monoxide emissions generated by each individual within a given population.

Carbon monoxide (CO) can have several adverse effects on the circulatory system. When inhaled, CO binds to hemoglobin in red blood cells, forming carboxyhemoglobin (COHb), which reduces the oxygen-carrying capacity of the blood. This leads to a decrease in the amount of oxygen available to vital organs and tissues, including the heart and brain. The primary adverse effects of CO on the circulatory system include:

* Reduced oxygen delivery: CO competes with oxygen for binding sites on hemoglobin. As CO binds to hemoglobin more readily than oxygen, it can displace oxygen from hemoglobin, resulting in reduced oxygen delivery to tissues.
* Impaired cardiac function: Reduced oxygen supply due to CO exposure can impair the function of the heart. CO exposure can also cause irregular heart rhythms (arrhythmias) and, in severe cases, may contribute to the development of heart disease.
* Increased risk of cardiovascular events: Prolonged or high levels of CO exposure have been associated with an increased risk of cardiovascular events, such as heart attacks and strokes.
* Vasoconstriction: CO exposure can cause vasoconstriction, narrowing the blood vessels and reducing blood flow to various organs.
* Increased blood clotting: CO exposure has been shown to increase blood clotting, leading to an increased risk of thrombotic events such as deep vein thrombosis (DVT) and pulmonary embolism (PE).

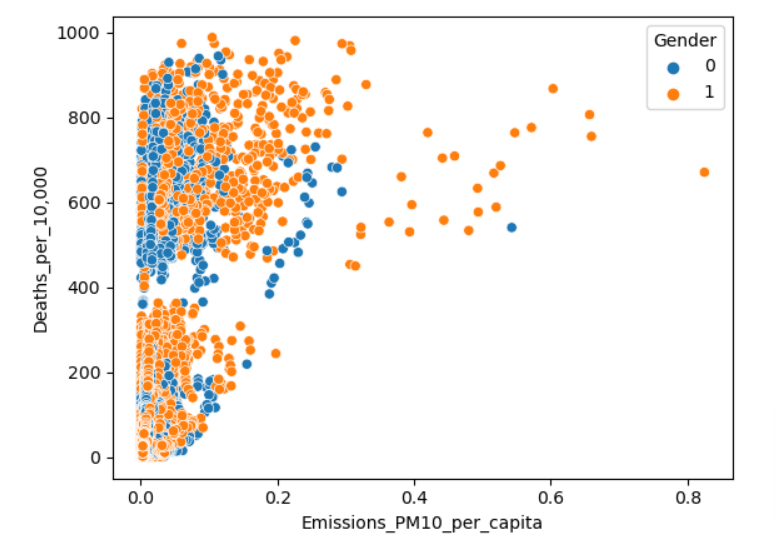
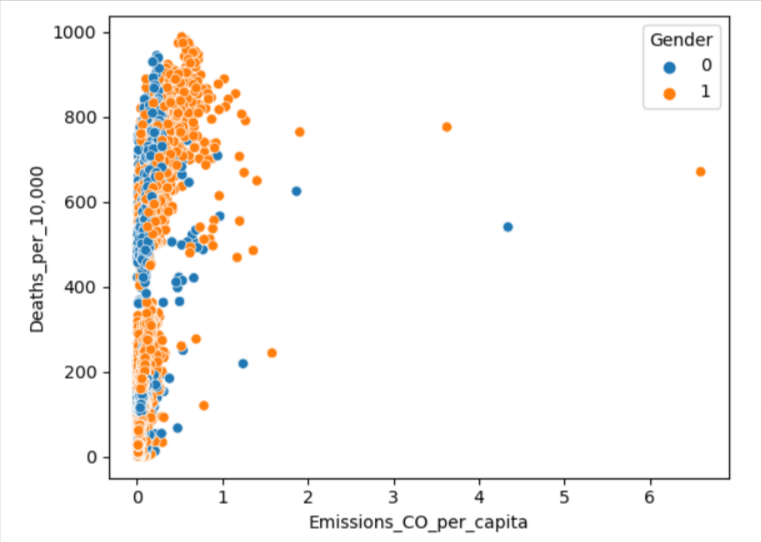
Emissions of PM10 has somewhat same effects on human body, that is why we used these emissions and used specific ICD chapters that are related to deaths by circulatory diseases, especially due to pollutants.

We started off with null hypothesis(that there is no correlation between independent and dependent variables) and from the below Spearman correlations, we can safely say that there is no correlation, so we dropped alternative hypothesis(that there is heavy correlation).



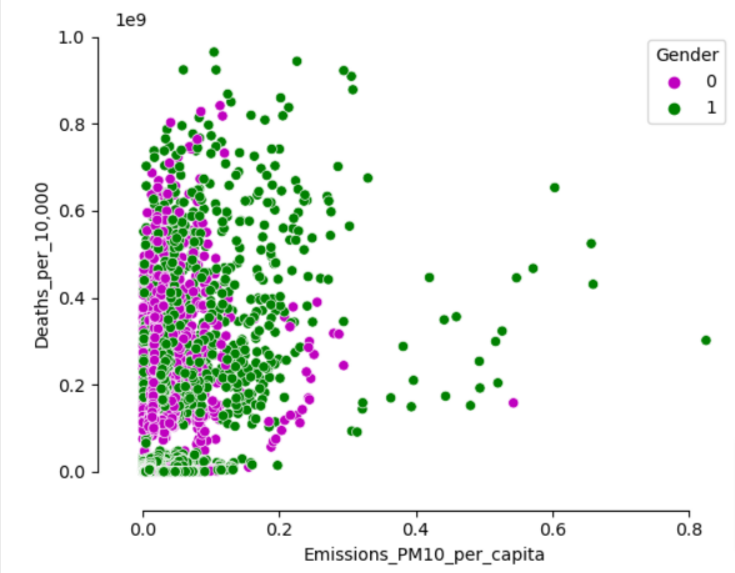
*Figure: 14: Spearman’s correlations of important variables and their p-values used to eliminate alternative hypothesis*

Therefore, the next step was to visualize the relationships of deaths by emissions, below are the charts for it.



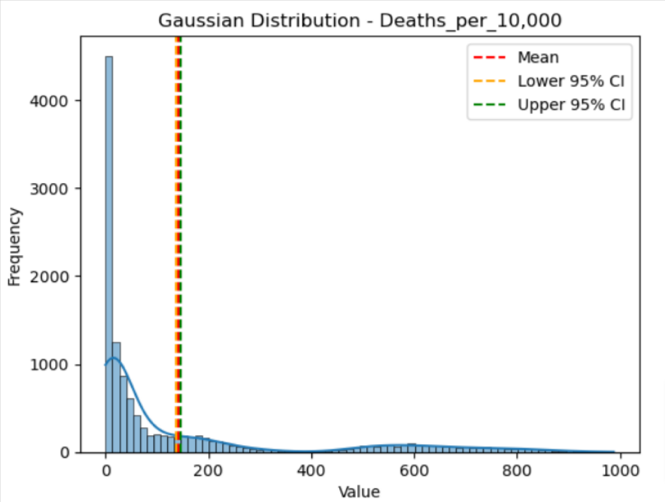
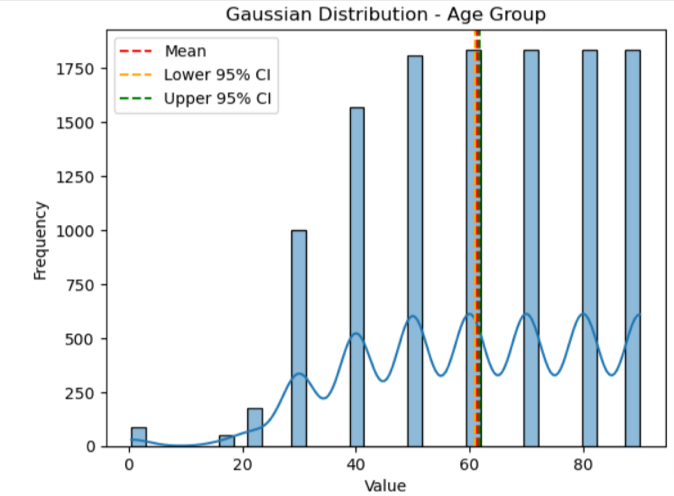
*Figure 15: (on left or top) Deaths per 10,000 vs emissions of CO per capita and (on right or bottom) Deaths per 10,000 vs emissions of PM10 per capita*

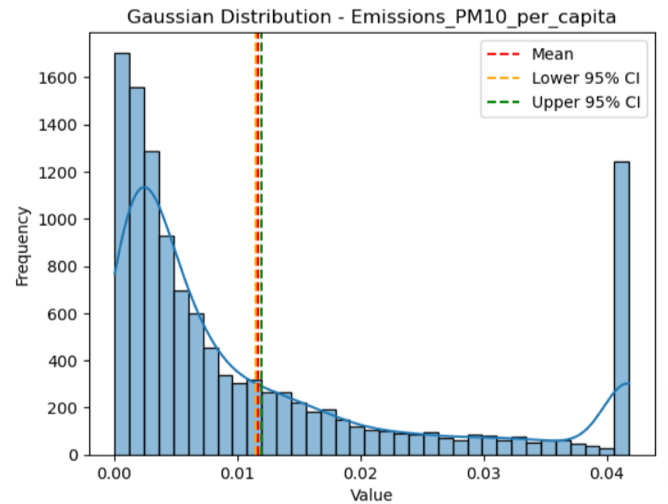
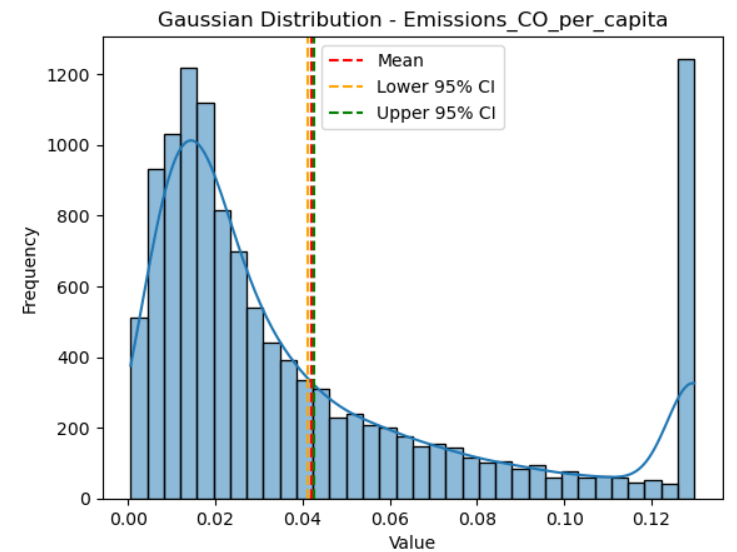
From the above scatterplot we observed that there was little to no relationship between Deaths and Emissions. Insights from the above plots are, more deaths can be observed due to low emissions, even high emissions means high deaths per 10,000. If we hue this plot with gender, below is the graph.



*Figure 16: Scatterplot of deaths per 10,000 vs emissions of PM10(all particles less than 10 picometer) along with gender differentiation*

The above plot does not exhibit a discernible relationship, indicating that the data may lack normalization or, in simpler terms, the sample dataset we obtained may not have been appropriately selected. This suspicion can be verified by examining another plot known as a Gaussian distribution plot. By representing the 'Age Group' column as a Gaussian distribution, this plot offers valuable insights into the dataset's central tendency, variability, and distribution shape. Here is what the Gaussian distribution plot can unveil:





*Figure 17: Gaussian distributions of all variables individually, along with mean, lower 95% confidence intervals and upper 95% confidence intervals.*

* Spread or variability: The width or spread of the Gaussian curve reflects the variability or dispersion of ages within the dataset. A wider curve indicates a larger spread of ages, while a narrower curve suggests less variability. The standard deviation(20) can be used to quantify the spread of the distribution. Moreover, all the graphs shown above strengthens the fact that these data points are skewed or does not follow normal distribution curve. Which further states that the data is not correctly sampled. However, this is obvious because, as per the definition of age group provided in data description, we had to take a central limit of every age group by taking mean of higher and lower limit of ages.
* Skewness: The Gaussian distribution plot can reveal any skewness in the emissions and deaths distribution. A symmetric bell-shaped curve with no skewness indicates that the age groups are evenly distributed on both sides of the mean. However, if the curve is skewed to the left (negative skew) or right (positive skew), it suggests an uneven distribution with more data points towards one side. Therefore, from the above plots we can surely say that deaths and emissions are positive skewed and may have non-linear relationships.
* Outliers: The presence of outliers, which are extreme values or data points that deviate significantly from the rest of the distribution, can be observed on the plot. Outliers can appear as individual points or clusters outside the main body of the Gaussian curve. As per emission’s gaussian plots, we observed high amount of outliers, and to deal with those we imputed outliers with 95 percentile values which helped us to remove skewness and bring the datapoints into 95% confidence interval range. More box plots to confirm this method are in python code provided in Final project code submissions.

VIF, which stands for Variance Inflation Factor, is a measure used in statistics to assess multicollinearity between predictor variables in a regression analysis.

* Age Group: The VIF score of 2.205711 suggests that there is relatively low multicollinearity between the "Age Group" variable and the other predictor variables. A VIF score around 2 indicates that the variance of the estimated regression coefficient for "Age Group" is only slightly inflated due to multicollinearity.
* Gender: The VIF score of 1.825026 indicates that there is also low multicollinearity between the "Gender" variable and the other predictors. Similar to the "Age Group" variable, the variance of the estimated regression coefficient for "Gender" is only slightly inflated.
* Emissions\_CO\_per\_capita: The VIF score of 2.345250 suggests a low-to-moderate level of multicollinearity between the "Emissions\_CO\_per\_capita" variable and the other predictors. The variance of the estimated regression coefficient for "Emissions\_CO\_per\_capita" is somewhat inflated due to multicollinearity.
* Emissions\_PM10\_per\_capita: The VIF score of 2.372605 indicates a similar level of multicollinearity as the "Emissions\_CO\_per\_capita" variable. There is a low-to-moderate level of multicollinearity between the "Emissions\_PM10\_per\_capita" variable and the other predictors.

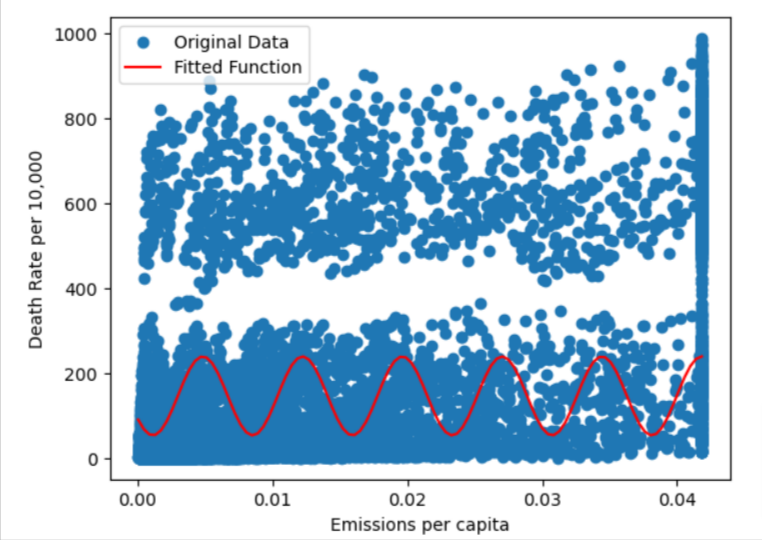
If multicollinearity is detected among predictor variables in a regression task, there are several strategies you can employ to address the issue. Here are some common approaches:

* **Remove one of the correlated variables**: We could not employ this strategy because we need these variables to predict mortality as these are important for our study.
* **Combine correlated variables**: Instead of including multiple correlated variables separately, you can create composite variables by combining them. This could not be done because if we combine Emissions columns this will increase its collinearity with deaths.
* **Regularization techniques**: Regularization methods, such as Ridge regression or Lasso regression, can be effective in handling multicollinearity. These techniques add a penalty term to the regression model, which reduces the impact of correlated variables and helps stabilize the coefficients. This strategy was tested but did not produce accurate results.
* **Non-parametric regression models**: These can be advantageous in situations where the relationship between the predictors and the response is complex and not easily captured by a linear or parametric model. These models have the flexibility to capture non-linear relationships, interactions, and other complex patterns in the data.

We experimented with various non-linear functions in an attempt to extract meaningful information from our dataset for model training, but unfortunately, we did not achieve satisfactory results. However, we did identify a function that yielded the highest level of accuracy:

**𝑓(𝑥) = (𝑎+10) ⋅ cos( 𝑏⋅𝑥+1000 ) + 𝑐**

The equation we utilized to discover patterns in our dataset involved additional coefficients, denoted as 'a' and 'b'. These coefficients were incorporated to enhance the accuracy of the function. 'x' represents our independent features, while 'c' represents the intercept term. The graph below visually illustrates the relationship depicted by this equation.



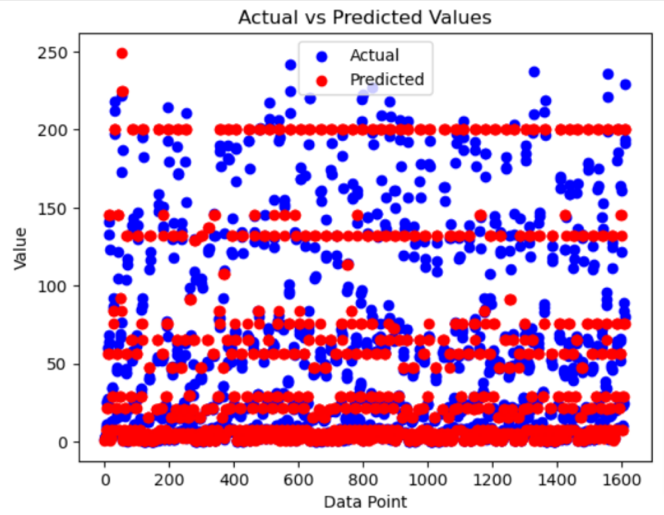
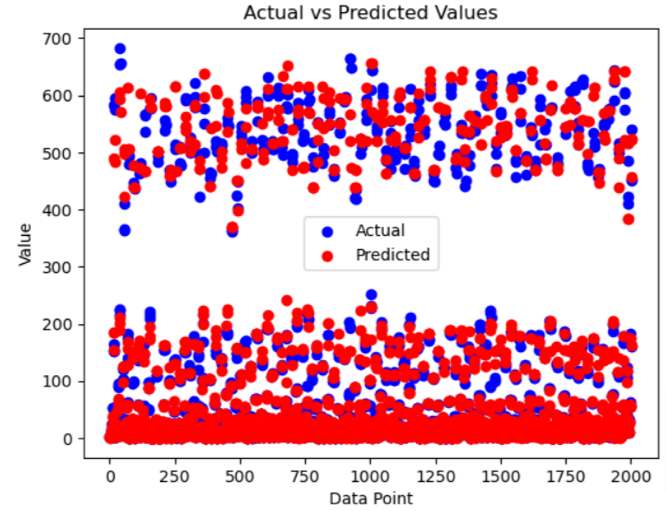
*Figure 18: Scatterplot of Deaths and Emission data with fitted non-linear function to find some relationship between independent variables and dependent variable(deaths per 10,000)*

This is the reason why we chose to work with non-parametric models. Non-parametric models are better in cases where there is no linear relationship among the data for several reasons:

* Flexibility: Non-parametric models do not assume a specific functional form or distribution of the data. They have the flexibility to capture complex and non-linear relationships that may exist in the data. This makes them suitable for situations where the relationship between the predictors and the response variable is not well-defined or cannot be accurately represented by a linear model.
* Robustness: Non-parametric models are robust to outliers and violations of assumptions. They do not rely on assumptions about the distribution of the data or the relationship between variables. As a result, they can handle data with nonlinear patterns, irregularities, or anomalies without compromising the model's performance.
* Adaptability: Non-parametric models can adapt to different types of data and variable interactions. They can handle both continuous and categorical predictors, as well as interactions between variables. This adaptability makes them suitable for a wide range of data types and provides more flexibility in modeling complex relationships.
* Interpretability: Non-parametric models often provide interpretable results, as they can directly estimate relationships between predictors and the response variable without relying on specific assumptions. This can be particularly useful when the relationship is complex and difficult to express using a parametric model.

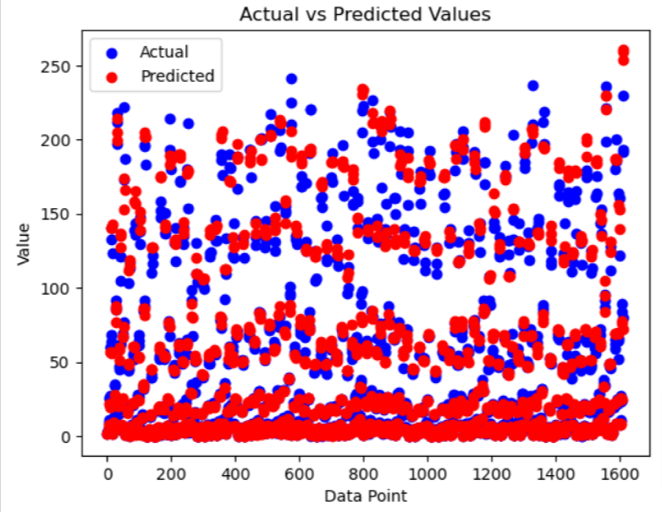
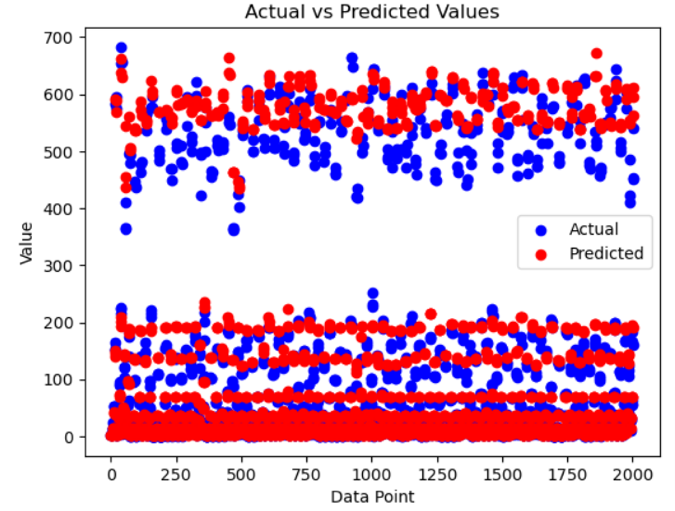
Below are some non Parametric models that we employed:

**Decision Trees**: Decision trees recursively split the data based on the predictor variables to make predictions. They can handle non-linear relationships and interactions effectively. Below is the actual vs predicted scatterplot to see how accurately our model fits to the data without overfitting it.



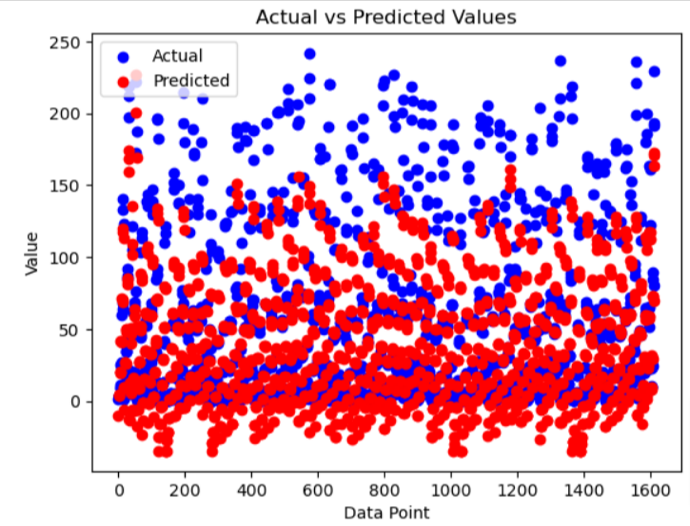
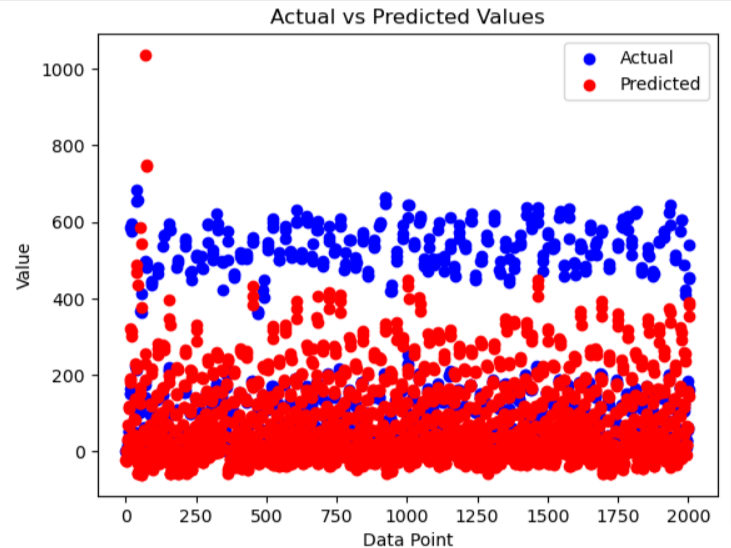
*Figure 19: (left/top) Decision tree model for CIRCULATORY ICD Chapters applied with friedman mean squared error, (right/bottom) Decision tree model for RESPIRATORY ICD Chapters applied with friedman mean squared error, min\_samples\_split = 150, min\_samples\_leaf = 2, max\_depth = 5, min\_impurity\_decrease = 0.2*

**Random Forests**: Random forests combine multiple decision trees to improve prediction accuracy. They are robust to overfitting and can handle high-dimensional data. Below is the actual vs predicted scatterplot to see how accurately our model fits to the data without overfitting it.



*Figure 20: (left/top) Random Forest model for CIRCULATORY ICD Chapters applied with max\_depth = 7, n\_estimators = 500, n\_jobs = -1, (right/bottom) Random Forest model for RESPIRATORY ICD Chapters applied with max\_depth = 8, n\_estimators = 600, n\_jobs = -1.*

**Support Vector Regression**: Support vector regression (SVR) uses support vector machines to perform regression. It can capture non-linear relationships by mapping the data to a higher-dimensional feature space. Below is the actual vs predicted scatterplot to see how accurately our model fits to the data without overfitting it.



*Figure 21: (left/top) Support Vector Regressor model for CIRCULATORY ICD Chapters applied with polynomial kernel with degree 2 and C = 2, (right/bottom) Support Vector Regressor model for RESPIRATORY ICD Chapters applied with polynomial kernel with degree 2 and C = 2.*

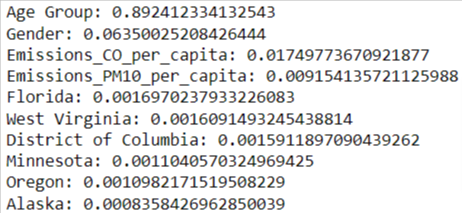
# Results:

Before interpreting which model performed well, we need to understand the metrics that we will use to identify that. MSE (Mean Squared Error), R-squared score, and adjusted R-squared score are commonly used metrics for evaluating the performance of regression models. Here's what each of these scores tells us:

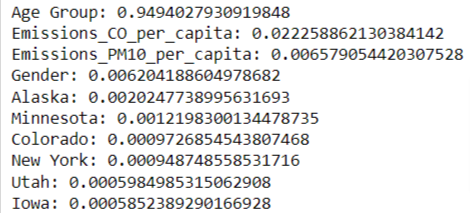
* Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values in the regression model. It quantifies the overall model fit by assessing how close the predicted values are to the actual values. A lower MSE indicates better model performance, where a value of zero indicates a perfect fit.
* R-squared score (Coefficient of Determination): R-squared score measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the regression model. It ranges between 0 and 1, with 1 indicating a perfect fit where all the variation in the target variable is explained by the model.
* Adjusted R-squared score: Adjusted R-squared score adjusts the R-squared score by penalizing the addition of irrelevant variables in the model. It takes into account the number of predictors and the sample size to provide a more reliable estimate of the model's explanatory power. Adjusted R-squared generally decreases as more variables are added to the model. It is useful for model comparison and selection, especially when dealing with models with a different number of predictors.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Respiratory** | | | **Circulatory** | | |
| R2 | MSE | Adj. R2 | R2 | MSE | Adj. R2 |
| **Decision Tree** | 0.962870 | 127.8105149 | 0.961536 | 0.986460 | 464.6271162 | 0.986070 |
| **Random Forest** | 0.982744 | 59.4056327 | 0.982076 | 0.978558 | 735.7844684 | 0.977962 |
| **SVR** | 0.798890 | 692.3304030 | 0.791648 | 0.659507 | 11684.1639919 | 0.649719 |

Hence, based on the scores obtained, we can conclude that the Random Forest model performed most effectively in predicting the mortality rate per 10,000 attributed to Respiratory diseases listed under the Respiratory ICD Chapters. On the other hand, the Decision Tree model demonstrated superior performance in predicting the mortality rate per 10,000 associated with Circulatory diseases listed under the Circulatory ICD Chapters. The following are the feature importance rankings obtained from the best models for both respiratory and circulatory predictions.



*Figure 22: Feature importance from Random Forest applied on Respiratory data for mortality rate prediction*



*Figure 23: Feature importance from Decision Tree applied on Circulatory data for mortality rate prediction*

# Conclusion

In conclusion, this project aimed to analyze mortality data and assess the influence of pollution on circulatory and respiratory diseases, focusing on the pollutants Emission\_co and emissions\_pm10. Through the analysis, we identified the key features that contribute to the prediction of these diseases.

For respiratory diseases, the age group emerged as the most significant factor, with a high importance percentage of 89.24%. This highlights the strong association between age and respiratory health, indicating that older individuals are more susceptible to respiratory diseases. Gender also played a notable role, although its contribution (6.35%) was relatively smaller compared to age. Additionally, the emissions of CO and PM10 were identified as minor factors, with importance percentages of 1.74% and 0.91%, respectively.

Similarly, in the analysis of circulatory diseases, the age group stood out as the dominant feature, accounting for an overwhelming 96.52% of the predictive power. This underscores the substantial impact of age on circulatory health, suggesting that older age groups are at higher risk. The emissions of CO followed with a contribution of 1.94%, while gender and PM10 emissions had lesser importance percentages of 0.57% and 0.3%, respectively.

These findings emphasize the significance of age as a major risk factor in both respiratory and circulatory diseases. It also highlights the potential role of pollutants, such as CO and PM10, although their impact appears to be relatively minor compared to age. Gender was found to have a modest influence on both disease types.

The results of this project contribute to our understanding of the relationship between pollution and health outcomes, specifically in the context of circulatory and respiratory diseases. They underscore the need for targeted interventions and policies to address the impact of pollution, particularly for vulnerable age groups. Future research could delve deeper into the specific mechanisms through which pollution affects these diseases and explore additional factors that may contribute to their occurrence and severity.

# References

* Centers for Disease Control and Prevention. (n.d.). Centers for Disease Control and Prevention <https://wonder.cdc.gov/controller/datarequest/D140>
* *Air Data: Air Quality Data Collected at Outdoor Monitors Across the US | US EPA*. (2022, October 13). US EPA. <https://www.epa.gov/outdoor-air-quality-data>
* Google. (n.d.). *States.csv | dataset publishing language | google for developers*. Google. <https://developers.google.com/public-data/docs/canonical/states_csv>