Term Project Final

Urvi Ravindra Dhomne, Karthick Raj Manickam

Engineering Management

622 - Predictive Analytical & Statistical Learning

Prof Michael Prokle

Abstract:

As the COVID-19 pandemic has disproportionately impacted different demographic groups, it is crucial to examine death statistics broken down by age and sex. This project examines a comprehensive dataset from the Centers for Disease Control and Prevention on provisional COVID-19 deaths categorized by these factors. Through exploratory data analysis, visualizations, and predictive modeling techniques like k-nearest neighbors, time series analysis, and machine learning algorithms, the project aims to uncover trends, patterns, and correlations between COVID-19 mortality rates and variables such as pneumonia deaths, influenza deaths, and total deaths across different age groups and genders. The analysis also explores geographic variations and temporal trends to inform targeted public health interventions and policymaking. This study makes use of the complexity of the dataset to offer useful data about how the pandemic affects different groups, which will ultimately lead to more effective mitigation techniques and fair healthcare outcomes.

Introduction:

This study utilizes a dataset comprising approximately 100,000 records, providing a robust foundation for comprehensive analysis. This project capitalizes on the dataset's statistical reliability and extensive scope to conduct a thorough examination. It uses various analytical techniques, including Linear Regression, Neural Network, K-Nearest Neighbor (KNN), Scatter Plot, Bar Plot, Heatmap, Time Series Analysis, and Random Forest Classifier. These methodologies are used to address key research questions.

1. Are there any notable trends or patterns in COVID-19 mortality rates over time within different age and sex groups?
2. How do Pneumonia Deaths correlate with COVID-19 Deaths across different age and sex groups?
3. How do COVID-19 mortality rates vary geographically within different age and sex demographics?
4. Are there any unexpected correlations or relationships between Total Deaths, Influenza Deaths Rate, pneumonia deaths, and COVID-19 mortality that warrant further investigation or public health action?
5. How do the rates of COVID-19 deaths vary from year to year and month to month. Are there any significant deviations or anomalies that require further examination?
6. How does the incidence of Influenza Deaths vary between different sex groups, and what is the nature of this impact?

By employing these diverse analytical approaches, this study aims to offer nuanced insights into the complex dynamics of the Covid-19 pandemic, facilitating informed decision-making in public health strategies and policy formulation.

Methodologies:

Data Cleaning:

To ensure the quality of the dataset, data cleaning procedures are done. These processes involve removing duplicate entries, deleting unnecessary columns like "Month," "Data as Of," and "Footnote," and inserting zeros to any data that has a missing value which could affect the accuracy. Additionally, converting date columns to datetime format for consistency.

Linear Regression:

The Linear Regression model shows a relationship between a) Covid-19 deaths and gender (males and females) and b) Pneumonia death, Influenza death and Covid-19 deaths. The data is split into training and testing sets and standardizes features for model consistency. These models predict Covid-19 deaths based on the mentioned factors and finally, it evaluates the model's performance using mean squared error (MSE), providing insights into how well the models predict Covid-19.

Neural Network:

The models use hidden layers with 100 and 50 neurons, ReLU activation function, and the Adam solver. After training, the models make predictions on the test data, and their performance is evaluated using mean squared error (MSE).

KNN Classifier:

In the KNN Classifier the categorical variables like age group and sex is converted into numerical values using Label Encoder. Next, it trains and evaluates its performance on the test set, giving accuracy, precision, and recall scores.

Time Series Analysis:

Two time series analyses are conducted: yearly and monthly. For the yearly forecast, data is filtered by year, and a Prophet model predicts Covid-19 deaths for the next year. Similarly, for the monthly forecast, data is filtered by month, and predictions are made for the next year. Both forecasts are plotted against actual data for comparison.

Random Forest Classifier:

Random Forest Classifier predicts Influenza deaths based on gender (male and female). The data is split, trained and predictions are made to get the accuracy, precision and recall results for male and female datasets.

Plots:

To visualize the data, three scatter plots are generated. The first scatter plot shows Covid-19 deaths against age, while the second one displays Pneumonia deaths against Covid-19 deaths. Also, a bar plot showcases state-wise Covid-19 deaths, and a heatmap shows all numerical features in the dataset.

Result:

1. Covid-19 Mortality trends by Age and Sex
    a) Model Performance:
    Linear Regression Male MSE: 2,614,443
    Neural Network Male MSE: 2,614,510
    Linear Regression Female MSE: 1,287,045
    Neural Network Female MSE: 1,286,843

These metrics indicate the mean squared error (MSE) for different models predicting COVID-19 deaths based on gender. Lower MSE values suggest better model performance.

b) Scatter Plot Analysis:
The scatter plot reveals a concentration of COVID-19 deaths between ages 40 to 85 and over, indicating a notable mortality trend across both genders within these age ranges.

c) Time Series Analysis:
The time series plot spans from November 2019 to November 2023. Black dots represent actual COVID-19 death counts, while the blue line depicts predicted deaths. The model generally aligns with the actual trend of COVID-19 deaths, capturing peaks and troughs over time. However, there are noticeable deviations where actual and predicted values differ.

2. Pneumonia Death Correlation with Covid-19 deaths by Age and Sex

a) Model Performance:
The classifier achieved an accuracy of 63.63%, indicating that roughly two-thirds of the predictions were correct. Precision, at 54.66%, shows the proportion of true positive predictions among all positive predictions, while recall, also at 63.63%, represents the ratio of correctly predicted positive observations to all actual positives.

b) Scatter Plots
The scatter plots illustrate the relationship between pneumonia deaths and COVID-19 deaths. In the first plot, there's a positive correlation between pneumonia deaths and the proportion of COVID-19 deaths. The second plot shows a similar trend, with higher COVID-19 deaths among older age groups. Males tend to have higher COVID-19 death counts.

3. Geographic Variation in Covid-19 Mortality

The bar plot reveals that California, Texas, and Florida have the highest number of deaths attributed to COVID-19. Following closely are Ohio, Pennsylvania, New York, Michigan, Illinois, and Georgia. Conversely, Vermont, Alaska, and Hawaii have recorded the fewest COVID-19 deaths.

4. Total Deaths, Influenza, Pneumonia, Covid-19 Relationship

a) Model Performance:
The mean squared error (MSE) for the linear regression model is approximately 168,682. The coefficients of the model are 1.09368808 for pneumonia deaths and -5.23325612 for influenza deaths. This suggests that there is a positive relationship between pneumonia deaths and COVID-19 deaths, while there is a negative relationship between influenza deaths and COVID-19 deaths.

b) Heatmap:
The heatmap displays the relationships between various features including COVID-19 Deaths, Total Deaths, Pneumonia Deaths, Pneumonia and COVID-19 Deaths, Influenza Deaths, and Pneumonia, Influenza, or COVID-19 Deaths. Most cells have a colour intensity ranging from 1.00 to 0.96, indicating strong positive correlations. However, the correlations Pneumonia-influenza and Influenza-Pneumonia, Influenza, or COVID-19 Deaths are slightly lower at 0.90. The correlation involving Influenza-Pneumonia, Influenza, or COVID-19 Deaths-influenza is even lower at 0.88.

5. Yearly & Monthly Trends of Covid-19 Deaths

   The two Time series analysis compare actual COVID-19 death data with forecasted values over time. While the models generally track trends, discrepancies exist, especially during high periods. Even though these forecasting are challenging, it informs vital public health decisions to improve accuracy and effectiveness in pandemic response strategies.

6. Influenza Deaths by Sex

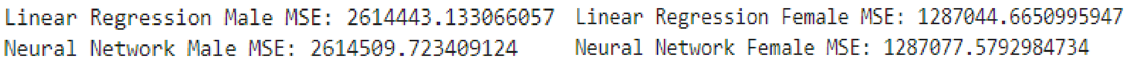
   The Random Forest model accurately predicted influenza deaths for males and females. It has high accuracy, precision, and recall rates, above 95%, indicating its effectiveness in classifying and forecasting influenza deaths based on sex.

References:

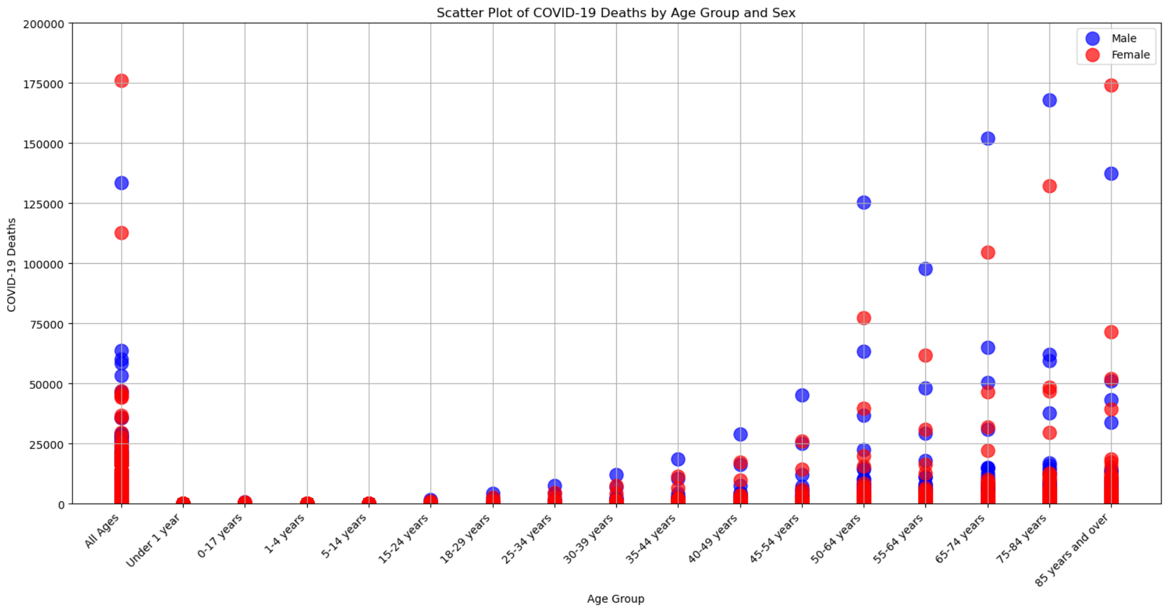
https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state

Appendices:

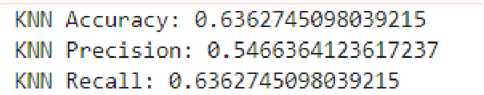
1.



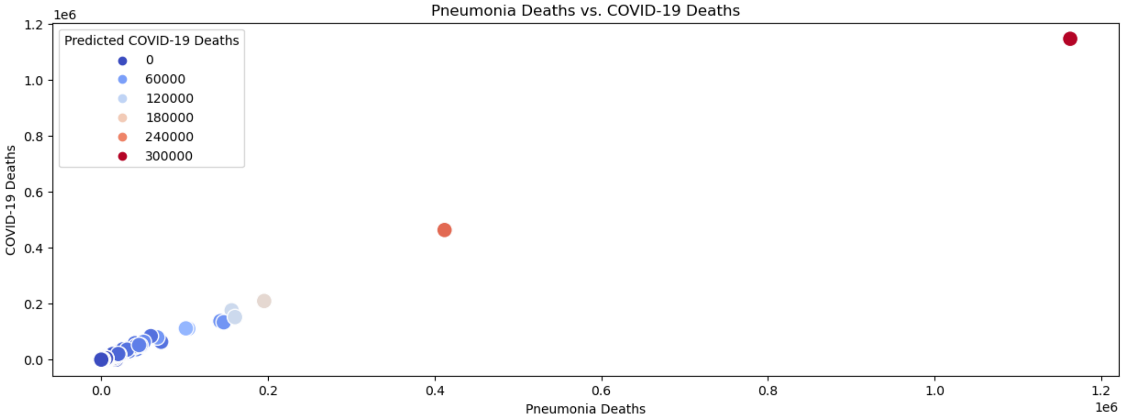
```
Linear Regression Male MSE: 2614443.133066057  Linear Regression Female MSE: 1287044.6650995947
Neural Network Male MSE: 2614509.723409124     Neural Network Female MSE: 1287077.5792984734
```

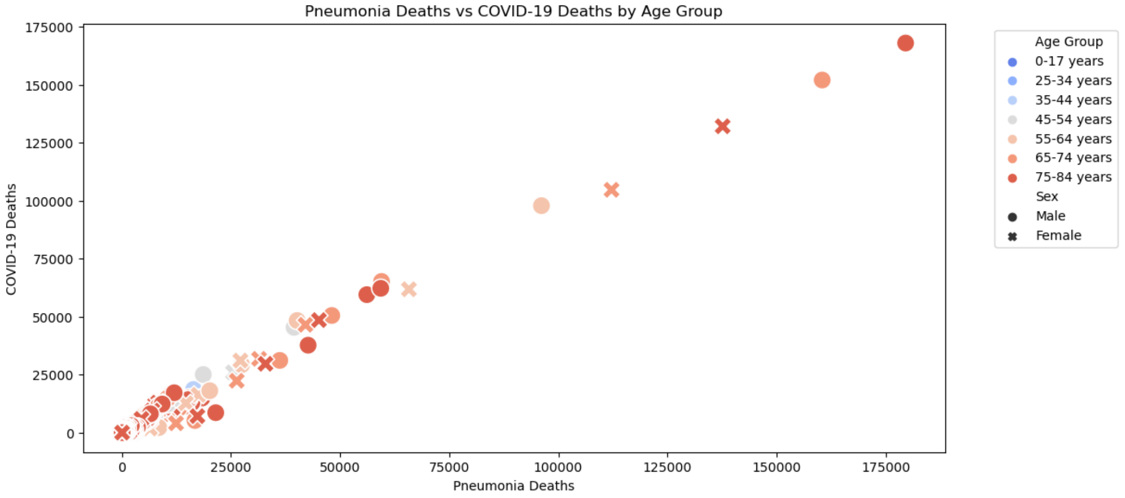


2.

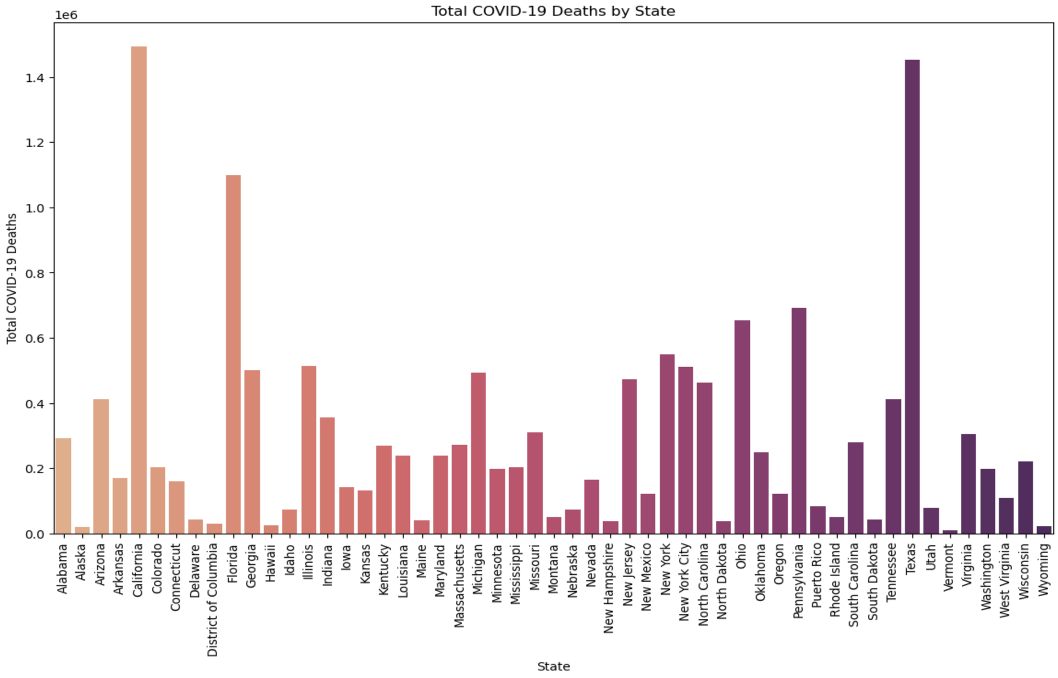


```
KNN Accuracy: 0.6362745098039215
KNN Precision: 0.5466364123617237
KNN Recall: 0.6362745098039215
```

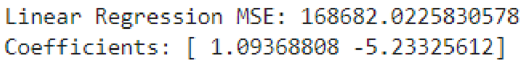

Pneumonia Deaths vs. COVID-19 Deaths
Predicted COVID-19 Deaths
● 0
● 60000
● 120000
● 180000
● 240000
● 300000
COVID-19 Deaths
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Pneumonia Deaths
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6


Pneumonia Deaths vs COVID-19 Deaths by Age Group
Age Group
● 0-17 years
● 25-34 years
● 35-44 years
● 45-54 years
● 55-64 years
● 65-74 years
● 75-84 years
Sex
● Male
✖ Female
COVID-19 Deaths
175000
150000
125000
100000
75000
50000
25000
0
Pneumonia Deaths
0
25000
50000
75000
100000
125000
150000
175000

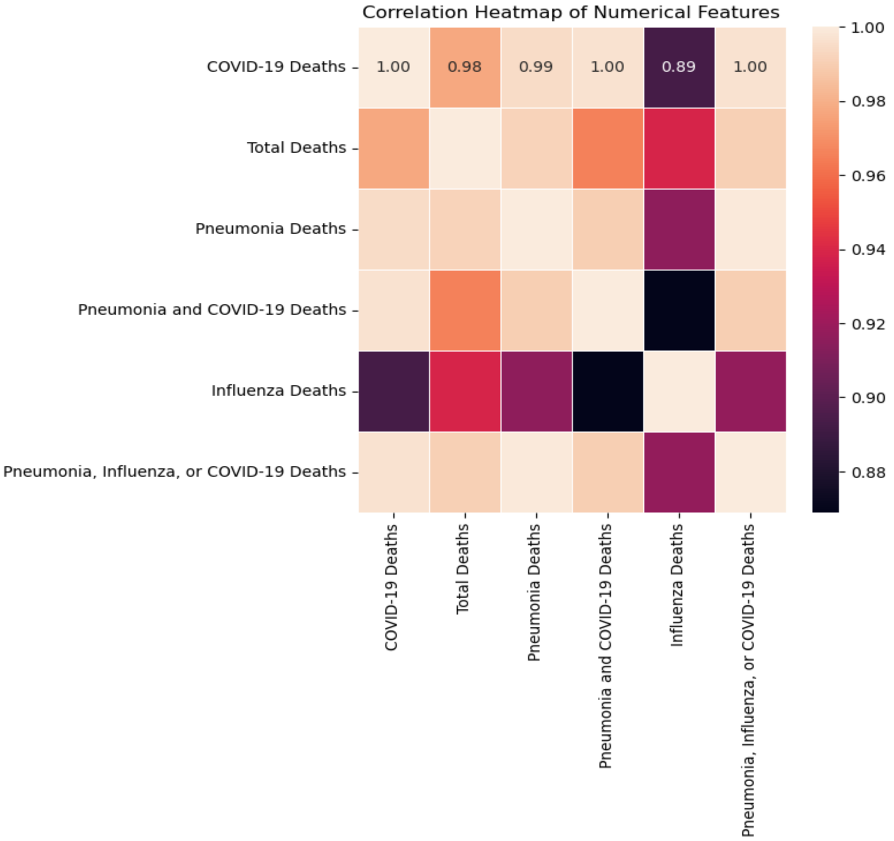
3.


Total COVID-19 Deaths by State
1e6
Total COVID-19 Deaths
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
New York City
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

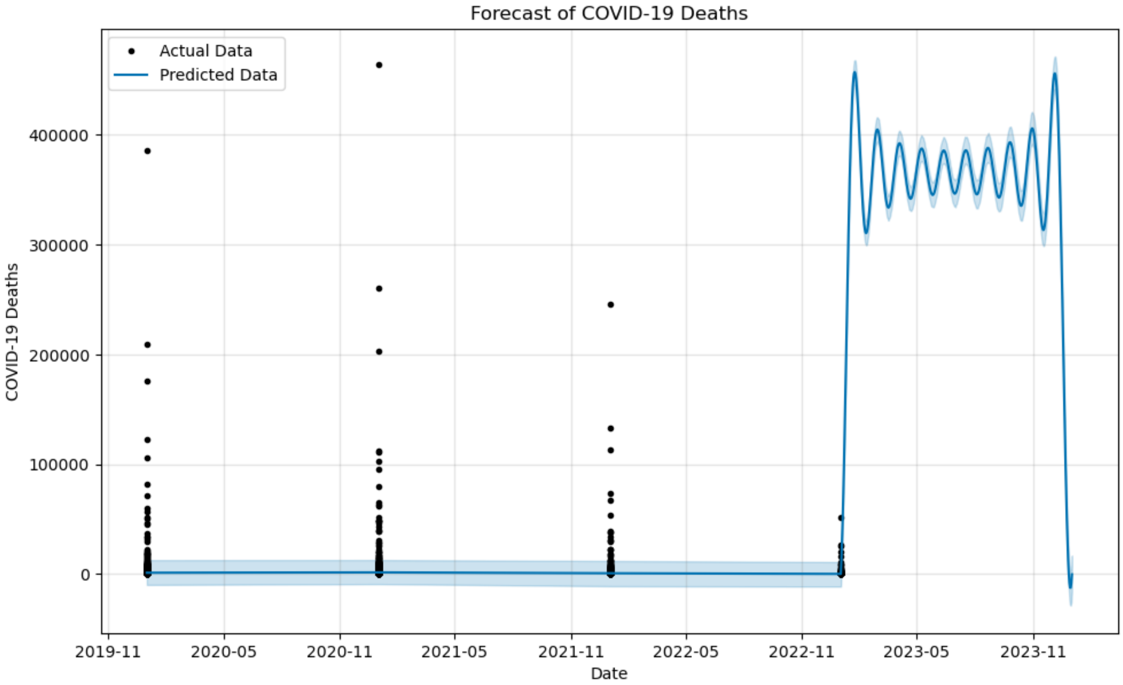
4.

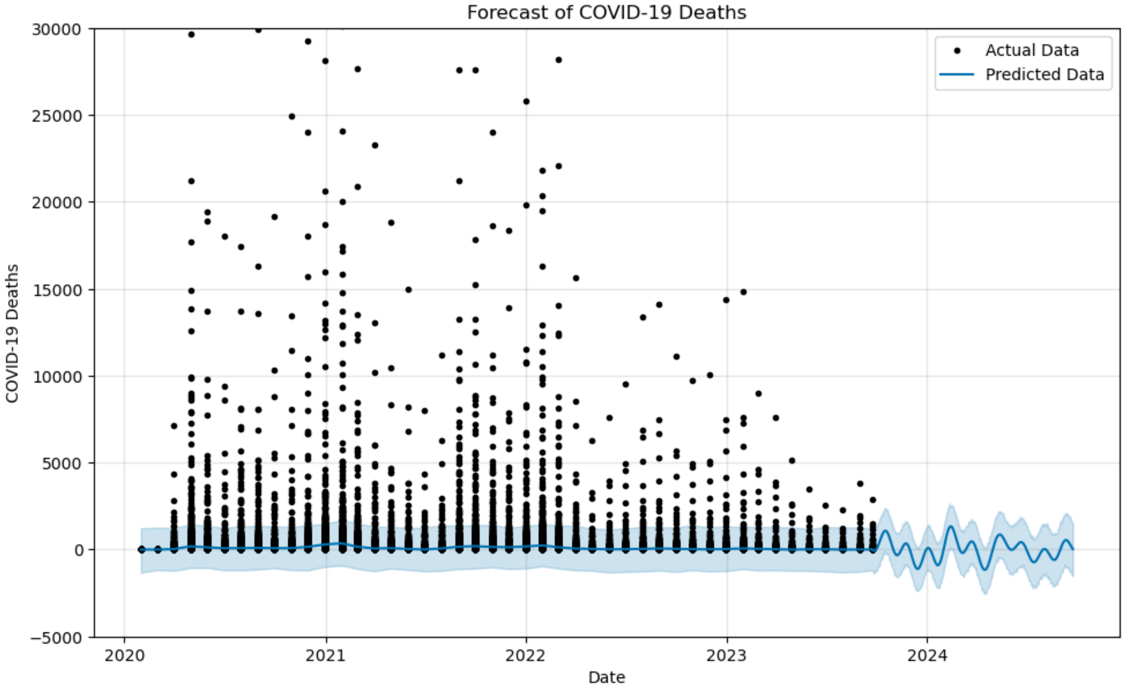


```
Linear Regression MSE: 168682.0225830578
Coefficients: [ 1.09368808 -5.23325612]
```

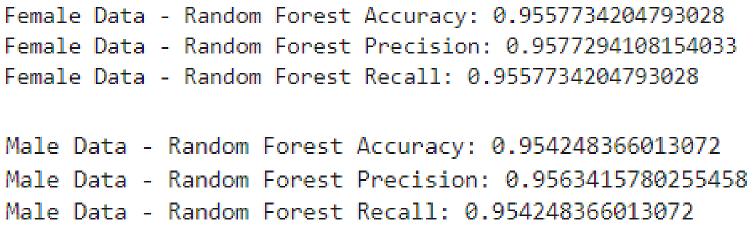


5.





6.


```
Female Data - Random Forest Accuracy: 0.9557734204793028
Female Data - Random Forest Precision: 0.9577294108154033
Female Data - Random Forest Recall: 0.9557734204793028

Male Data - Random Forest Accuracy: 0.954248366013072
Male Data - Random Forest Precision: 0.9563415780255458
Male Data - Random Forest Recall: 0.954248366013072
```