REPORT

“A Deep Learning Analysis Predicting The Future Risk For Lung Cancer In Louisiana Based on Air Pollution.”

April 24, 2024

Shelby Franklin, Tahmina Akter Anondi, Mridula Mavuri

CSC 469/669

Table of Contents

Introduction

Lung cancer is recognized as being one of the leading causes of death worldwide, and air pollution has been identified as a major contributor to its emergence. A study that was conducted in 2023 by the Journal of Thoracic Oncology revealed that air pollution was a pronounced contributor to lung cancer, with an estimated 30% increase in deaths (Berg et al., 2023). Although prior research has been conducted, there are significant gaps within the availability of data, types of data, and spatial vulnerability.

Our study aims to bridge the gap, by examining lung cancer vulnerability parish-wide in Louisiana, we focused on identifying the most vulnerable areas and predicting the number of people that could be affected. To broaden our analysis we included additional risk factors such as, COPD, obesity, smoking, uninsured individuals, and poverty, which have conveyed to aid in an increase of developing lung cancer.

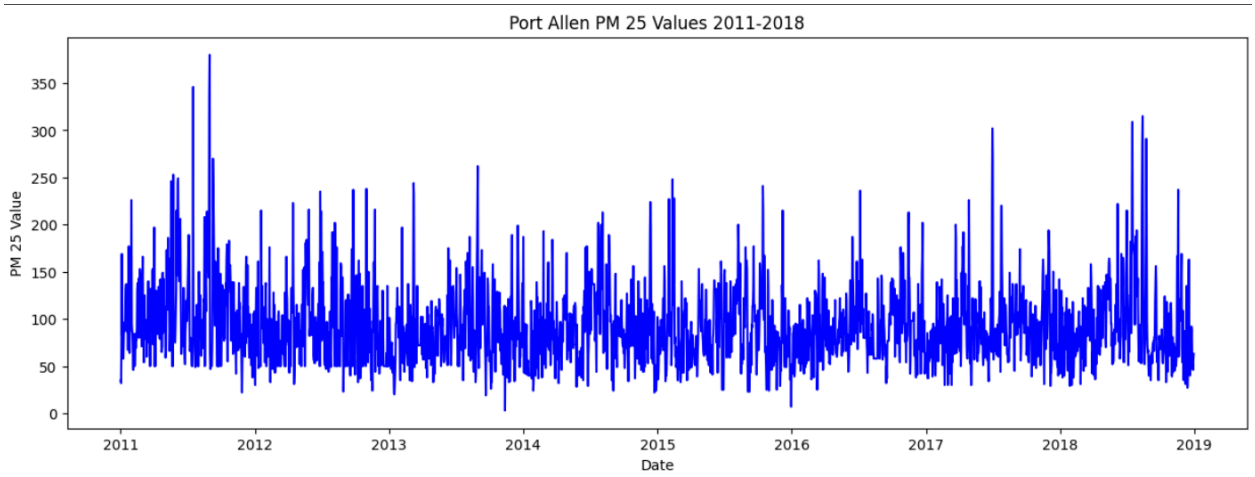
We've implemented several deep learning methods to predict and display the risk of lung cancer using several datasets from 2010-2021 based on Louisiana parish data. Specifically, we predicted PM2.5 values for the next 12 months in 1D utilizing Long Short-Term Memory(LSTM), Convolutional Neural Network(CNN), and mixed models. We also utilized LSTM to predict a 3D model for PM2.5 for the next month. For Spatial Correlation Analysis, we investigated the spatial correlation between lung cancer and PM2.5 values generating a map of the Louisiana parishes highlighting the comparison of the two over the span of 10 years. Moran's I values were also calculated for each year to compare spatial autocorrelation, clustering, and variation. Lastly, a correlation and health factors impact study was conducted highlighting the correlation of COPD, and air pollution using multivariate analysis. The multivariate analysis was also utilized to analyze the impact of various health factors on lung cancer rates, aiming to understand their influence on their lung cancer incidence.

User Requirements

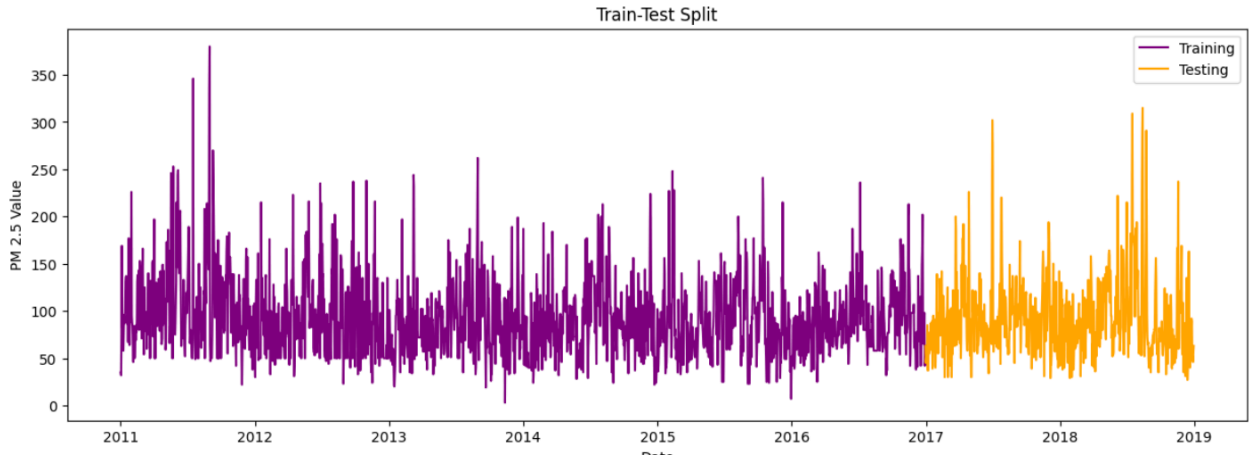
1. User requirements include Google Collab and Jupyter for compiling our code.
2. Microsoft Excel was utilized as well for processing our data.

Screenshots and Outputs

PM2.5 value for PORT Allen(2011-2018)

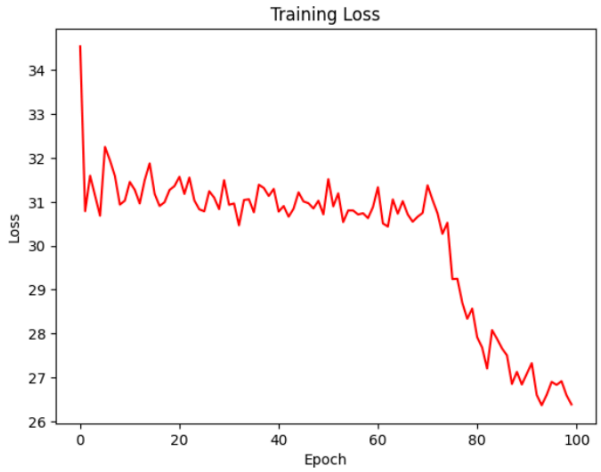


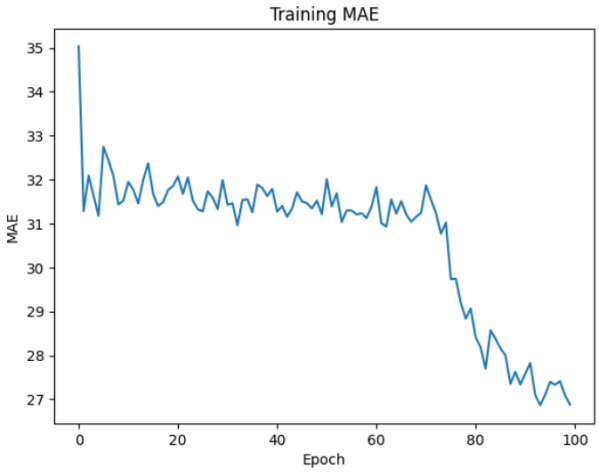
Train Test Split



Training Loss and MAE for LSTM(1D) Model

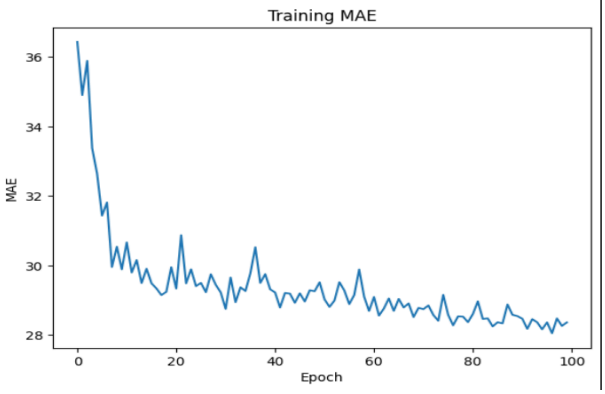
LSTM MODEL





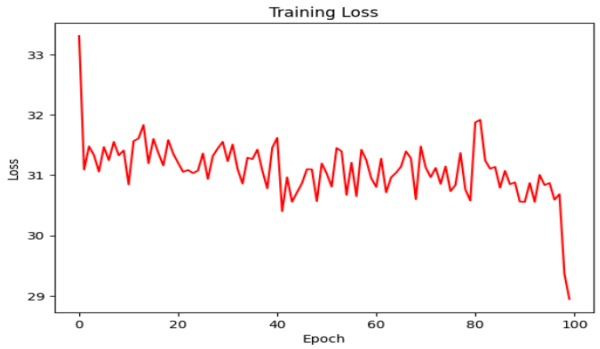
Training Loss and MAE for CNN(1D) Model

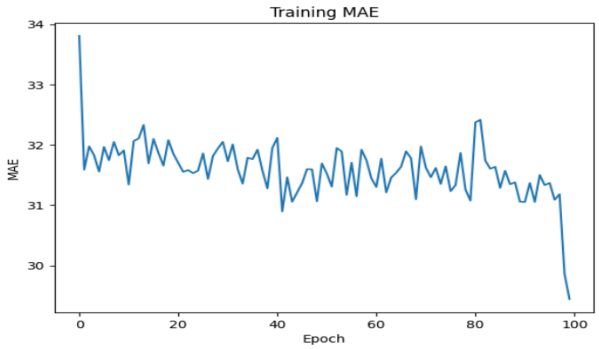
CNN MODEL



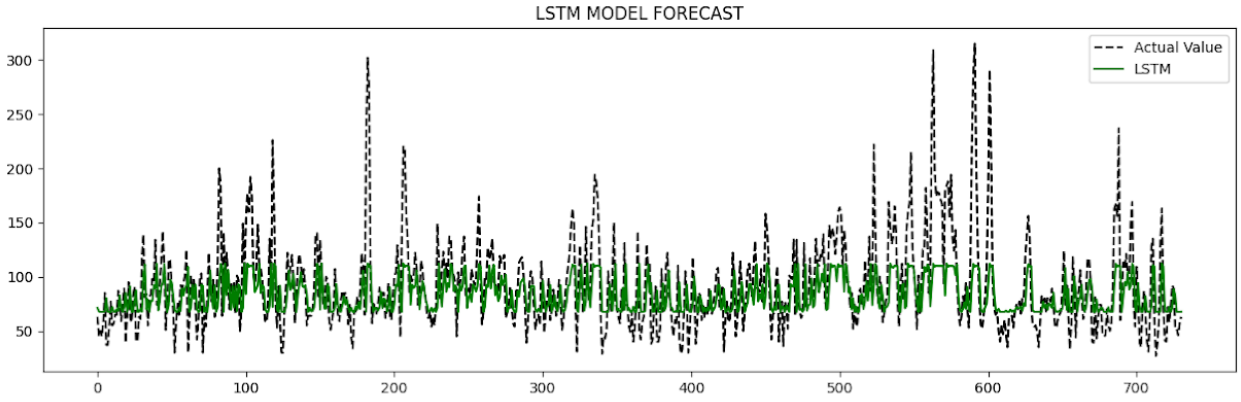
Training Loss and MAE for both CNN and LSTM

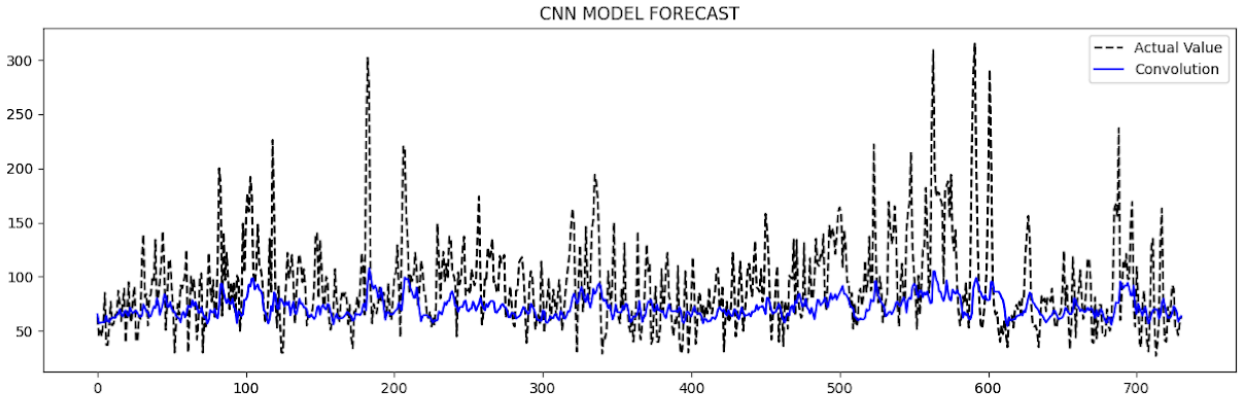
MIXED MODEL

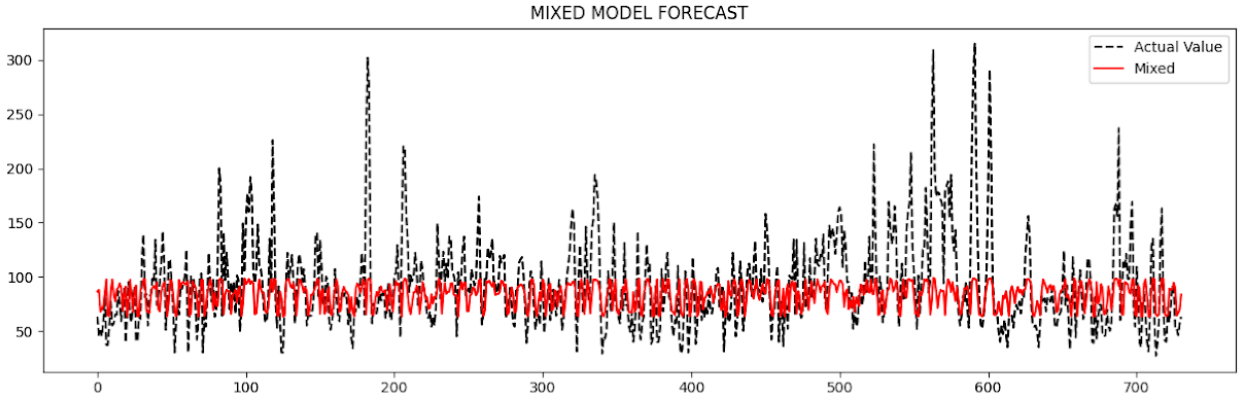


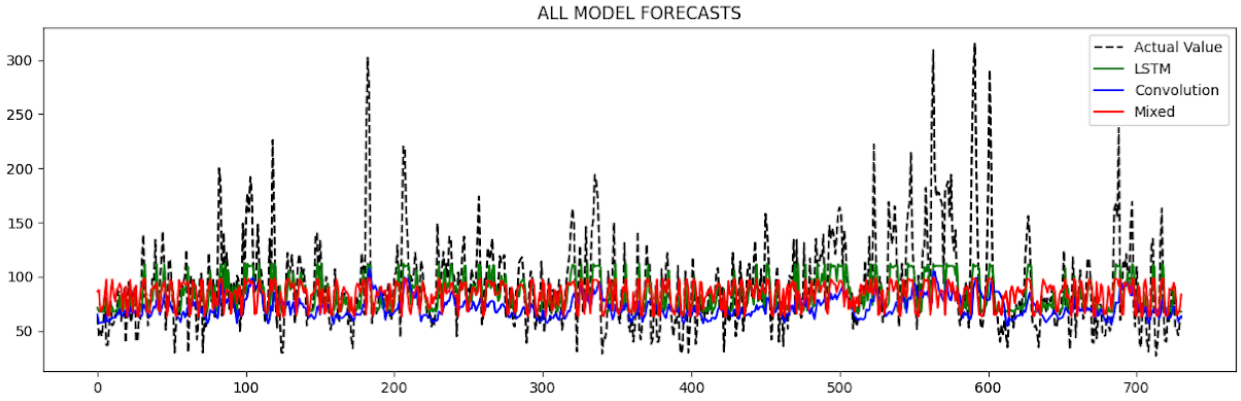


Forecast with LSTM , CNN and Mixed model of them

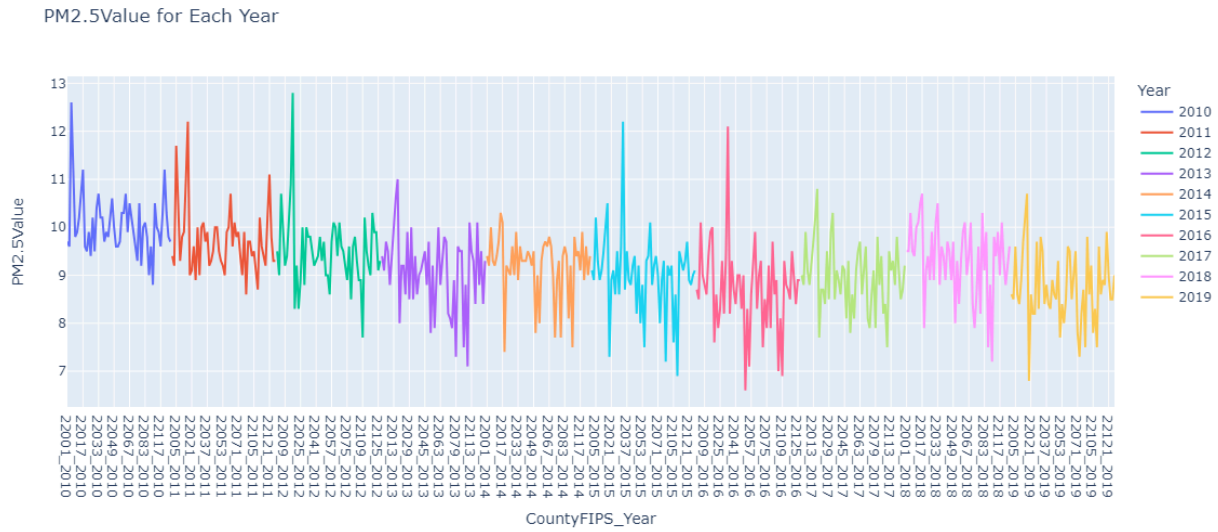


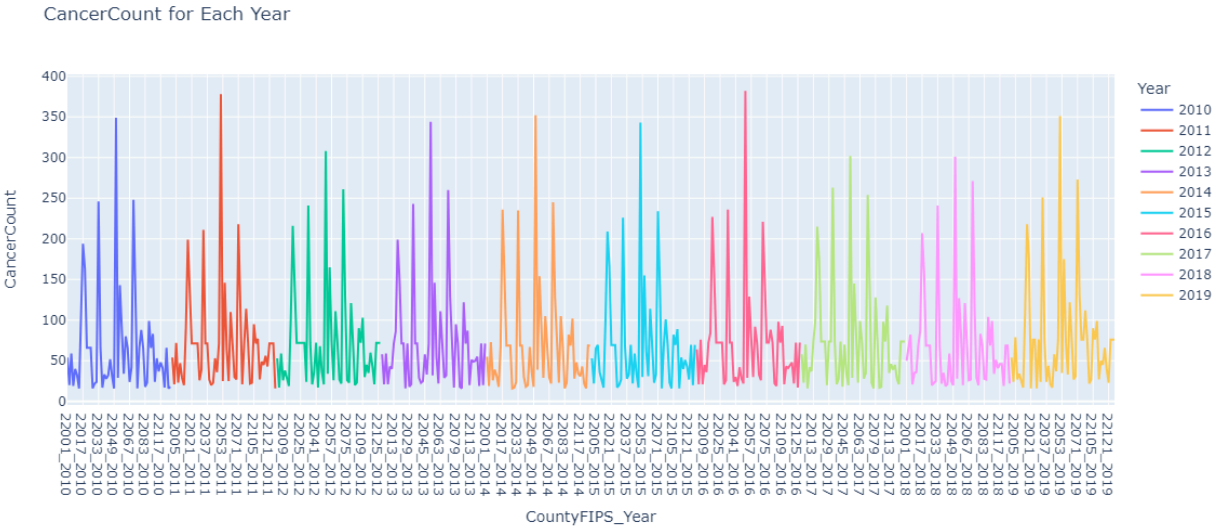


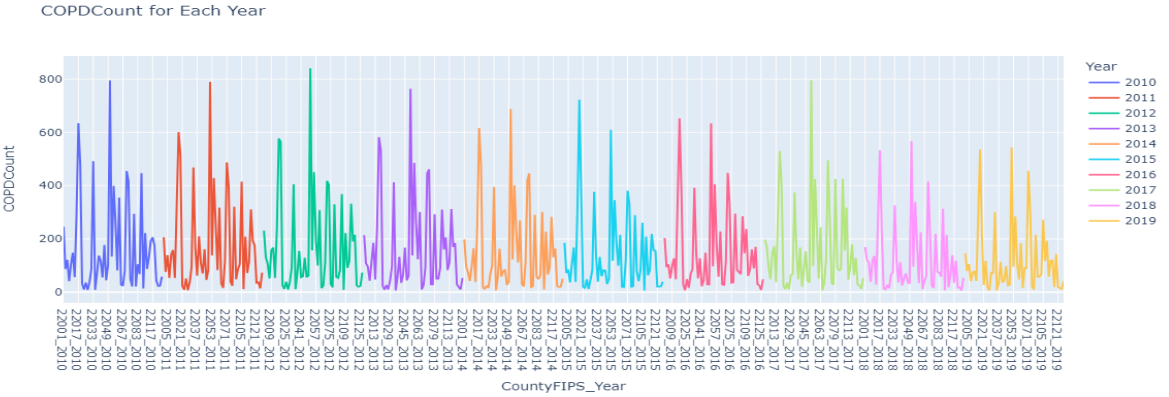




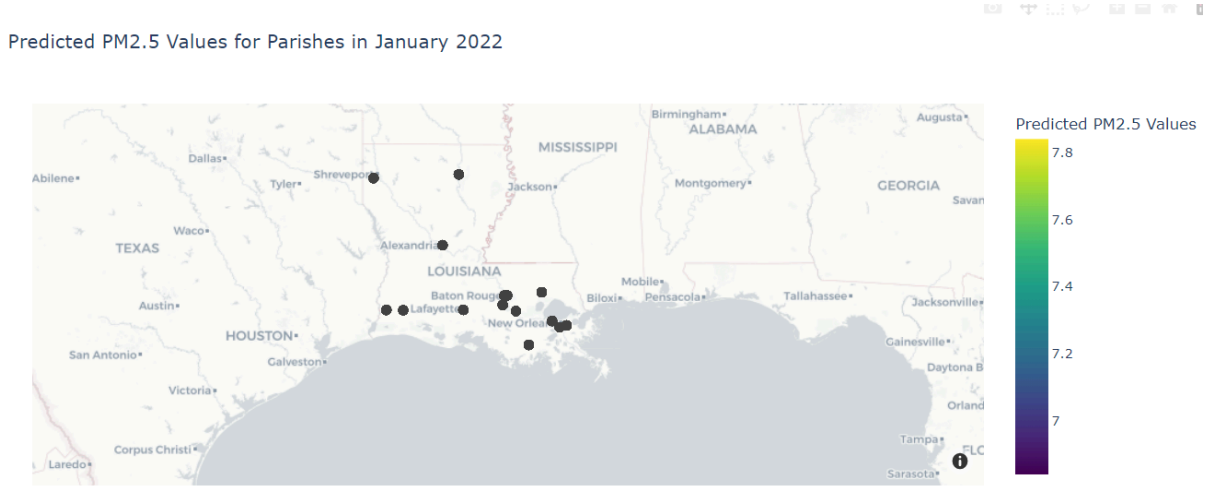
Plotting pm2.5, cancer and COPD to find out the relationship



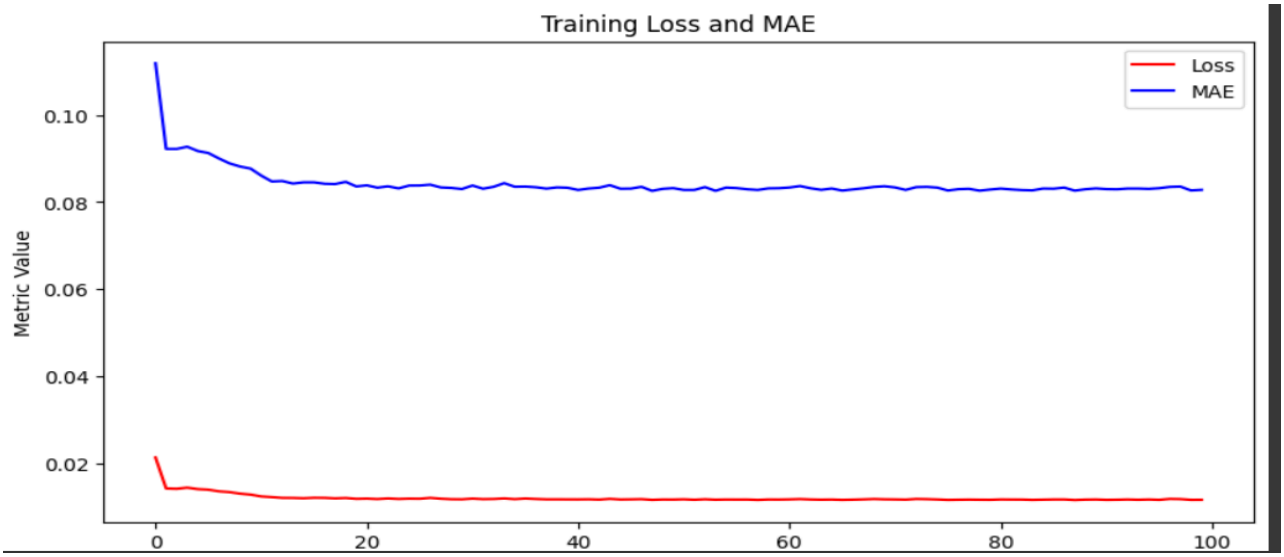




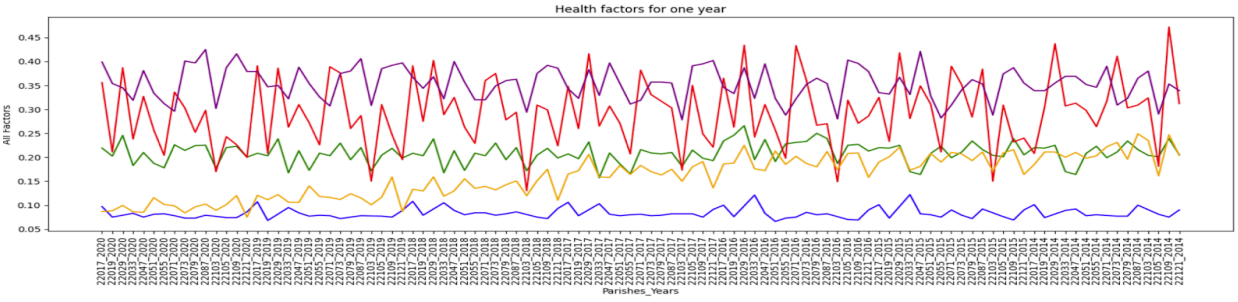
Prediction for next 1 month(January) with LSTM(3D input and Output)

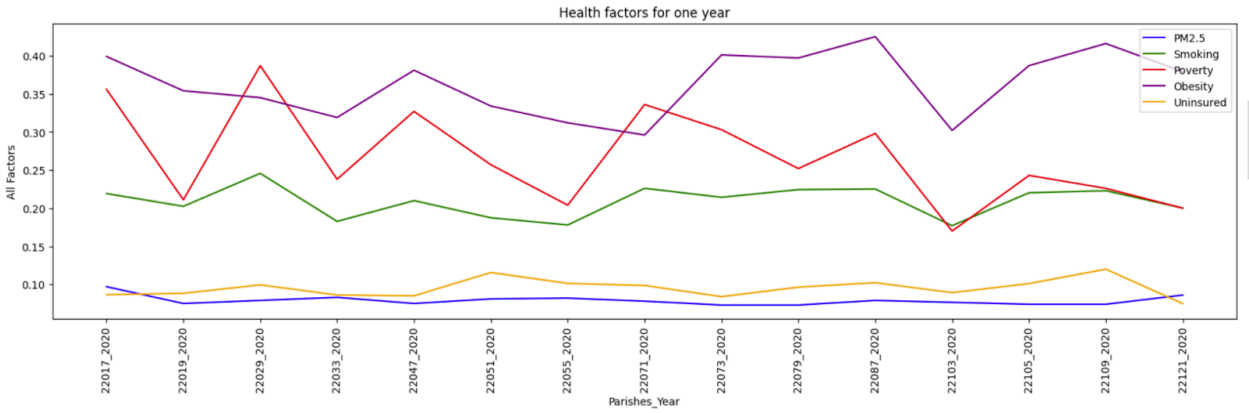


Training loss and MAE for LSTM (3D)



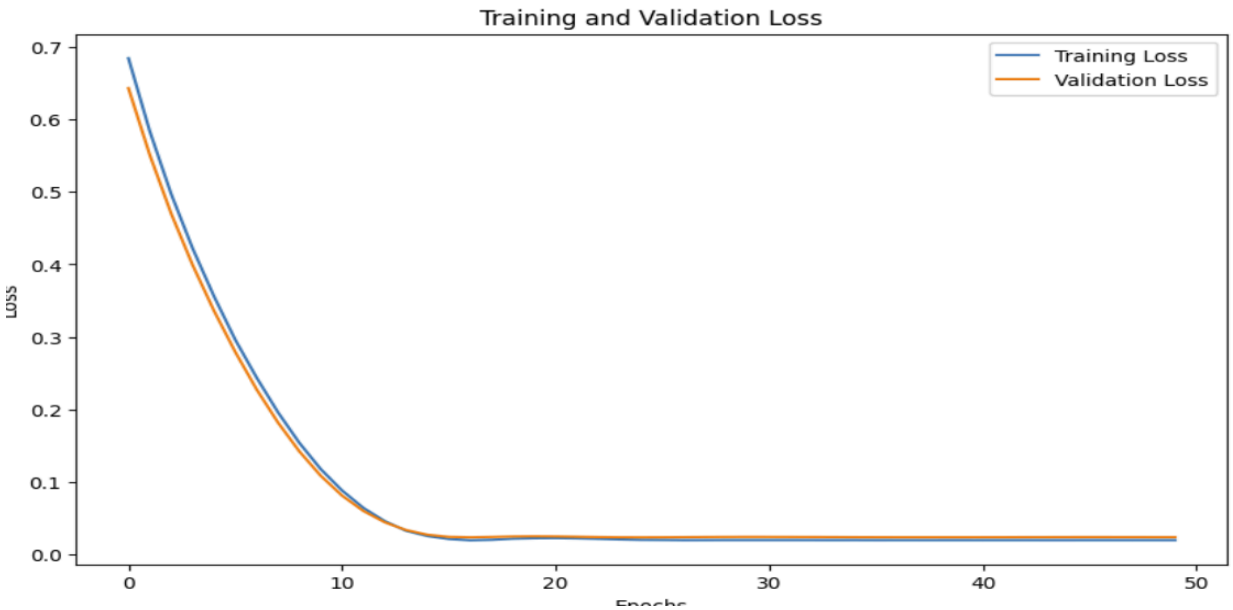
Multivariate Regression of Other Factors of Lung Cancer

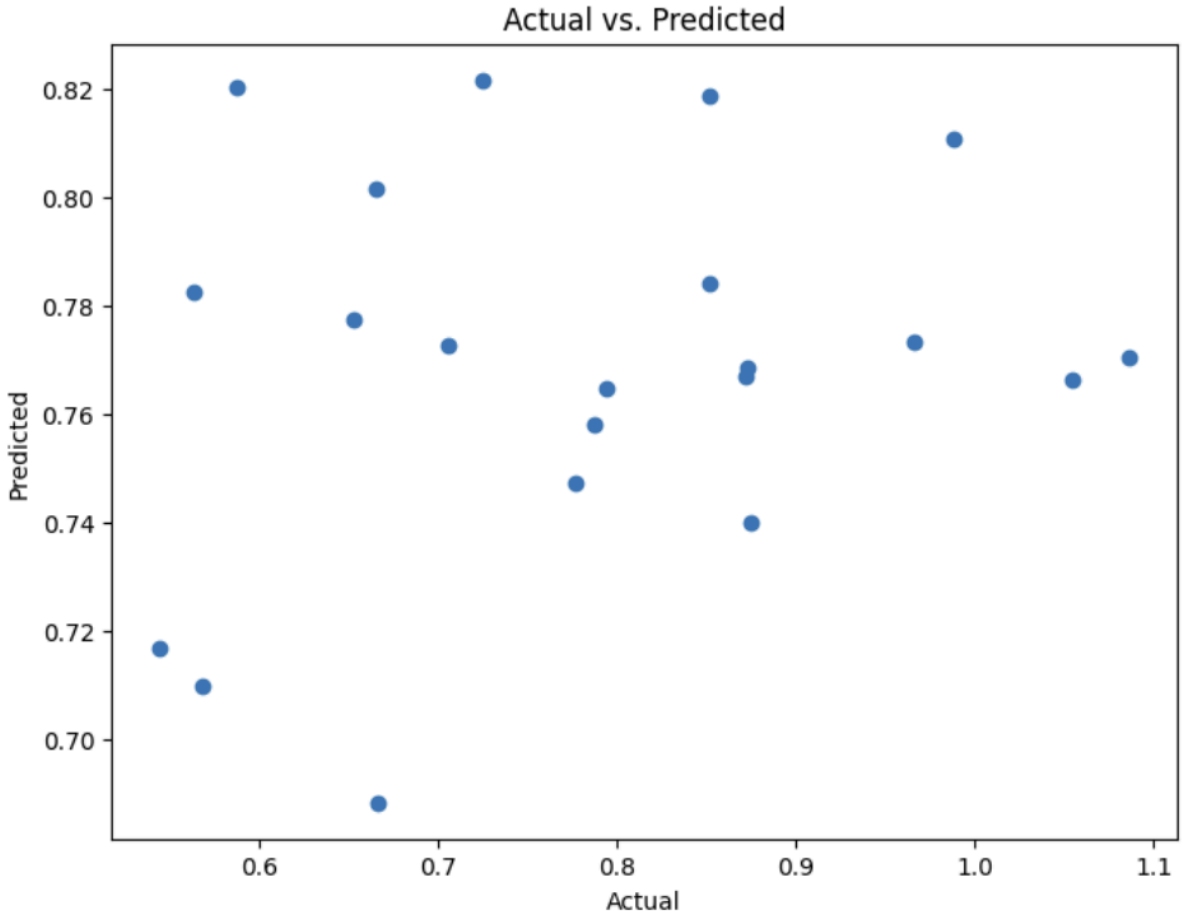





```
# Printing sample predicted and actual values for comparison
for i in range(5):
    print(f"Predicted: {predictions[i][0]}, Actual: {y_test.iloc[i]}")
```

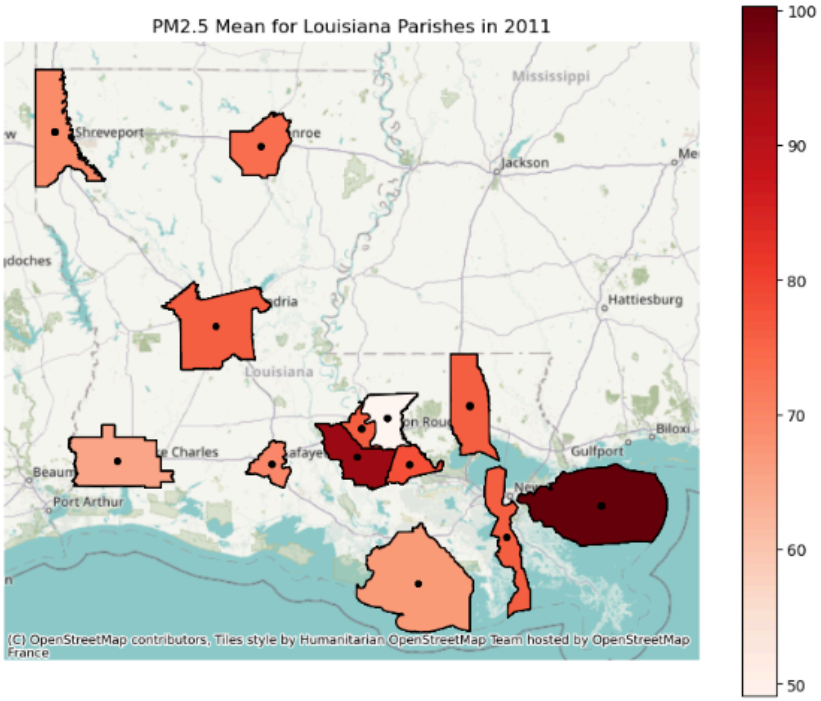
```
1/1 [==============================] - 0s 27ms/step - loss: 0.0239 - mae: 0.1296
1/1 [==============================] - 0s 101ms/step
Predicted: 0.7842105627059937, Actual: 0.851483927
Predicted: 0.7398312091827393, Actual: 0.874412004
Predicted: 0.8187766075134277, Actual: 0.851970181
Predicted: 0.7725936770439148, Actual: 0.705521669
Predicted: 0.76854008436203, Actual: 0.87335535
```

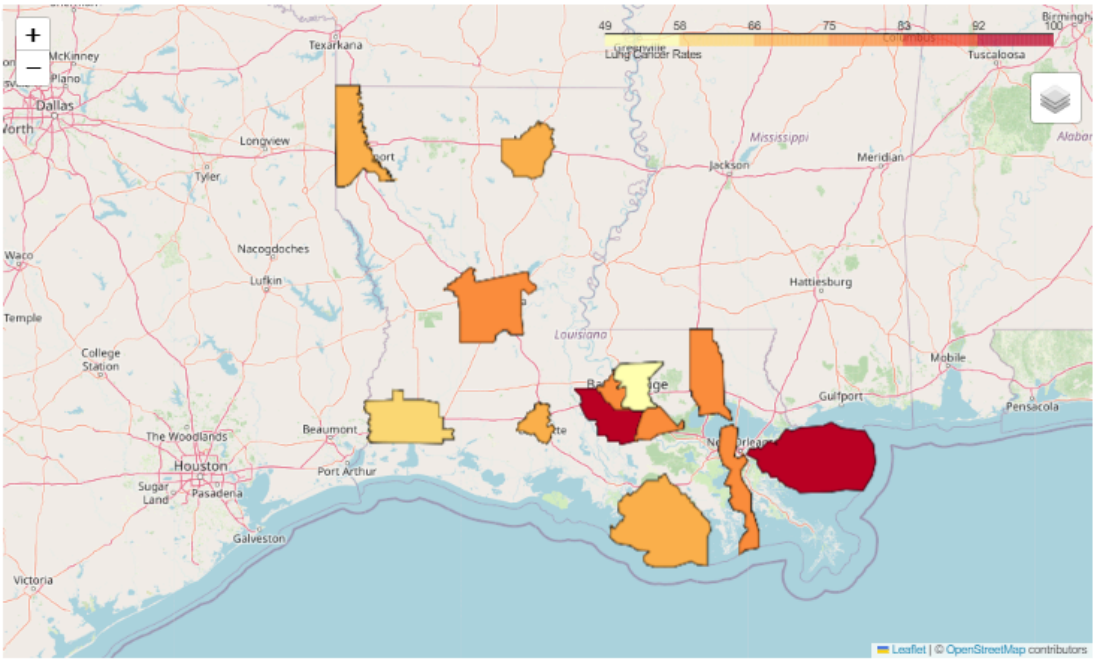

Training and Validation Loss
Training Loss
Validation Loss
Loss
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50


Actual vs. Predicted
Predicted
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.6
0.7
0.8
0.9
1.0
1.1
Actual

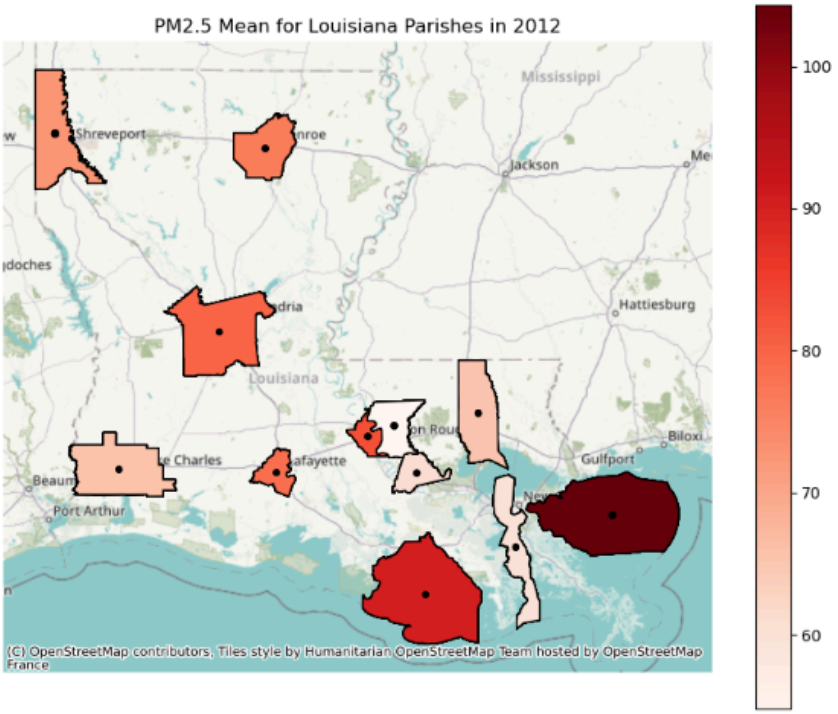
Spatial Correlation and Moran's I Calculation for 2011-2020

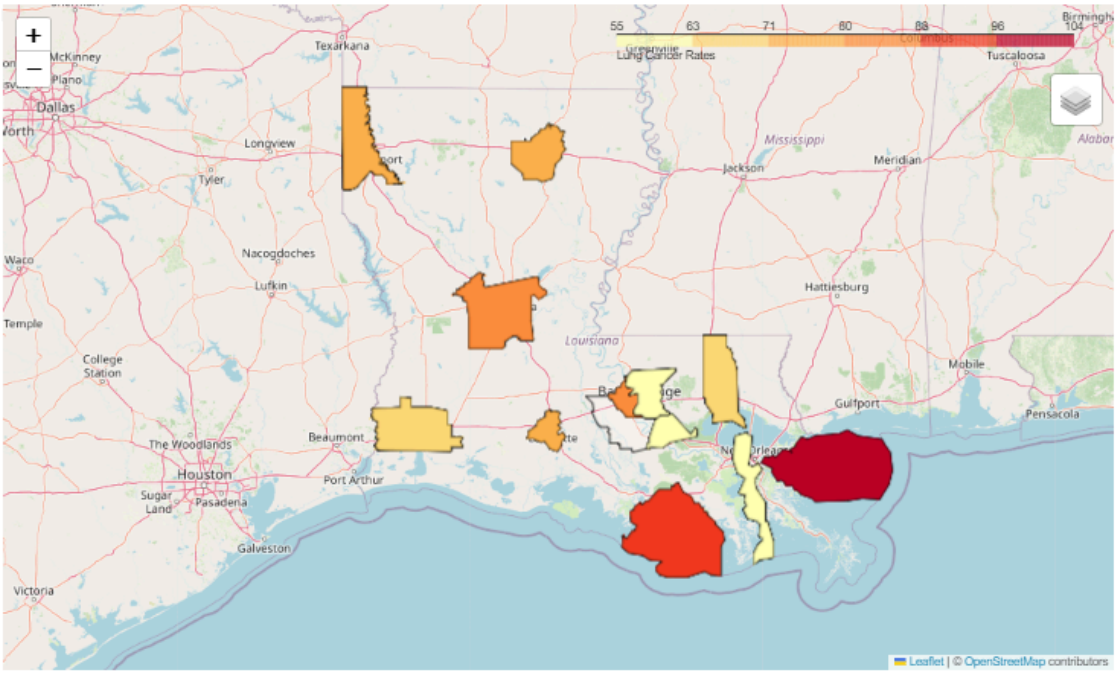
```
Moran's I for PM2.5 in 2011: 0.0892
Moran's I p-value for PM2.5 in 2011: 0.0010
Moran's I for Lung Cancer Rates in 2011: 0.3878
Moran's I p-value for Lung Cancer Rates in 2011: 0.0010
```



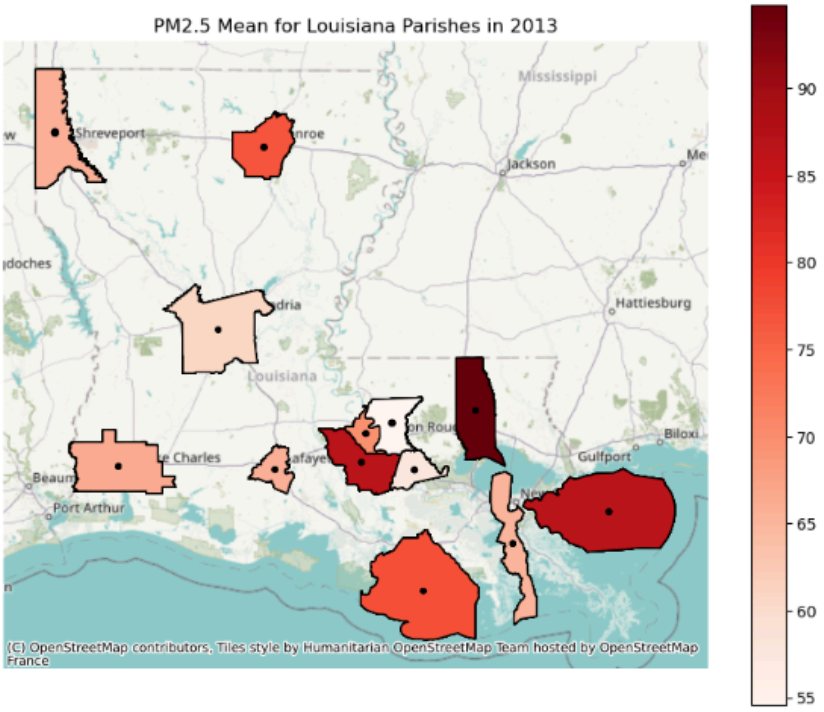


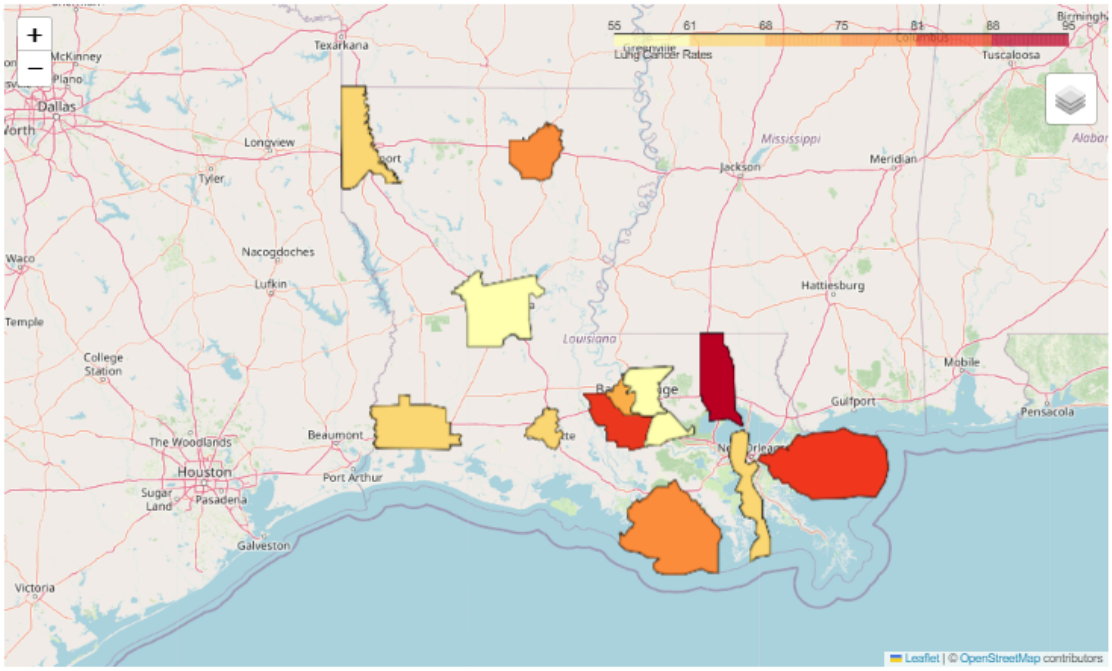
```
Moran's I for PM2.5 in 2012: 0.2339
Moran's I p-value for PM2.5 in 2012: 0.0010
Moran's I for Lung Cancer Rates in 2012: 0.8358
Moran's I p-value for Lung Cancer Rates in 2012: 0.0010
```



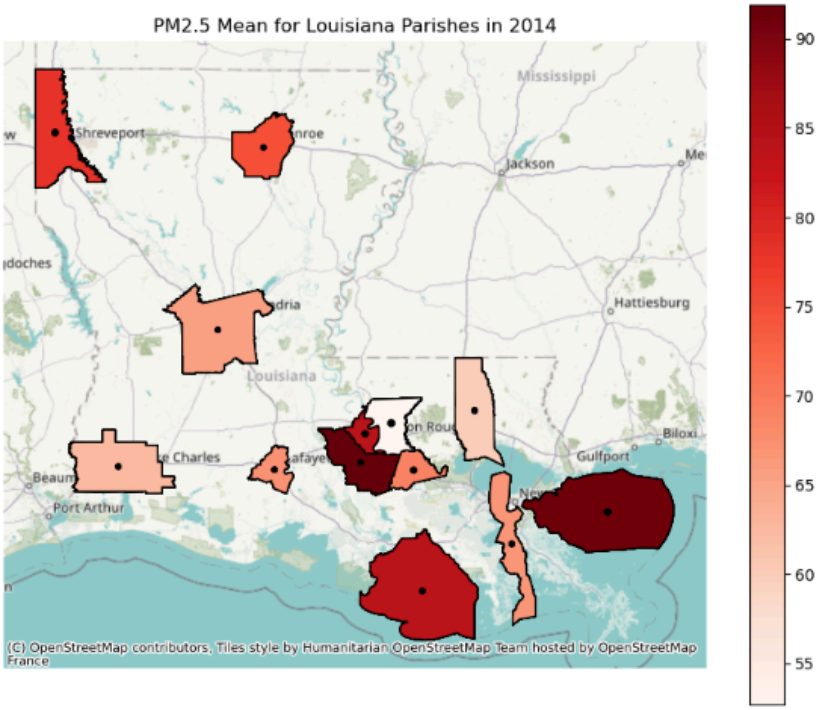


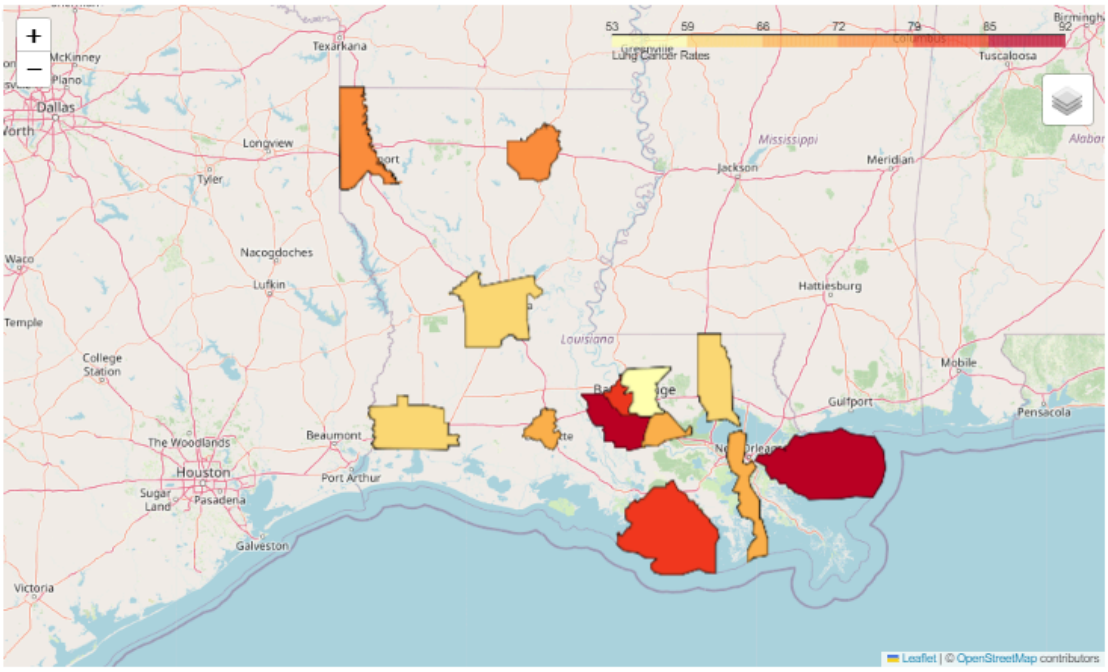
```
Moran's I for PM2.5 in 2013: 0.2303
Moran's I p-value for PM2.5 in 2013: 0.0010
Moran's I for Lung Cancer Rates in 2013: 0.7177
Moran's I p-value for Lung Cancer Rates in 2013: 0.0010
```



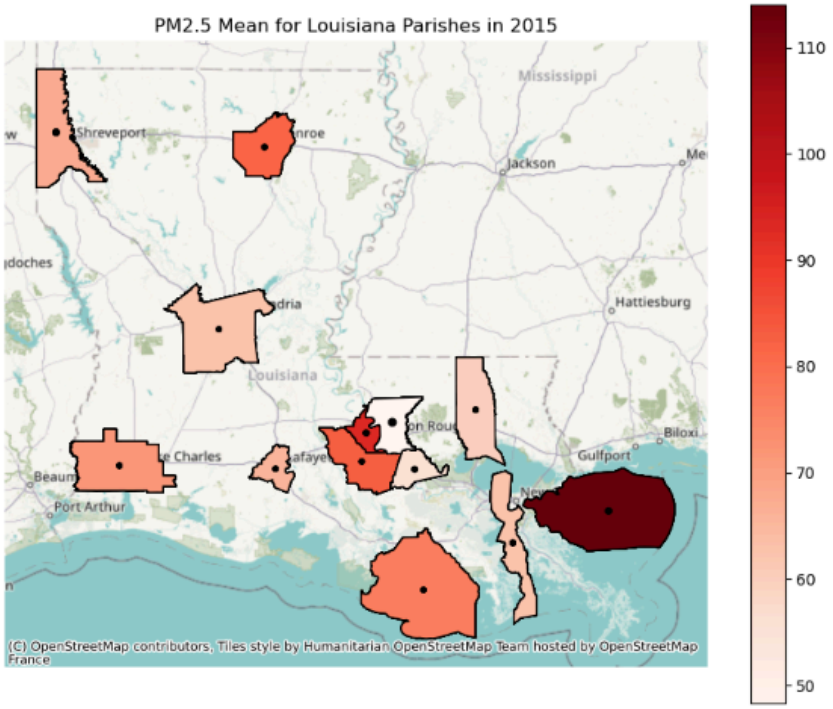


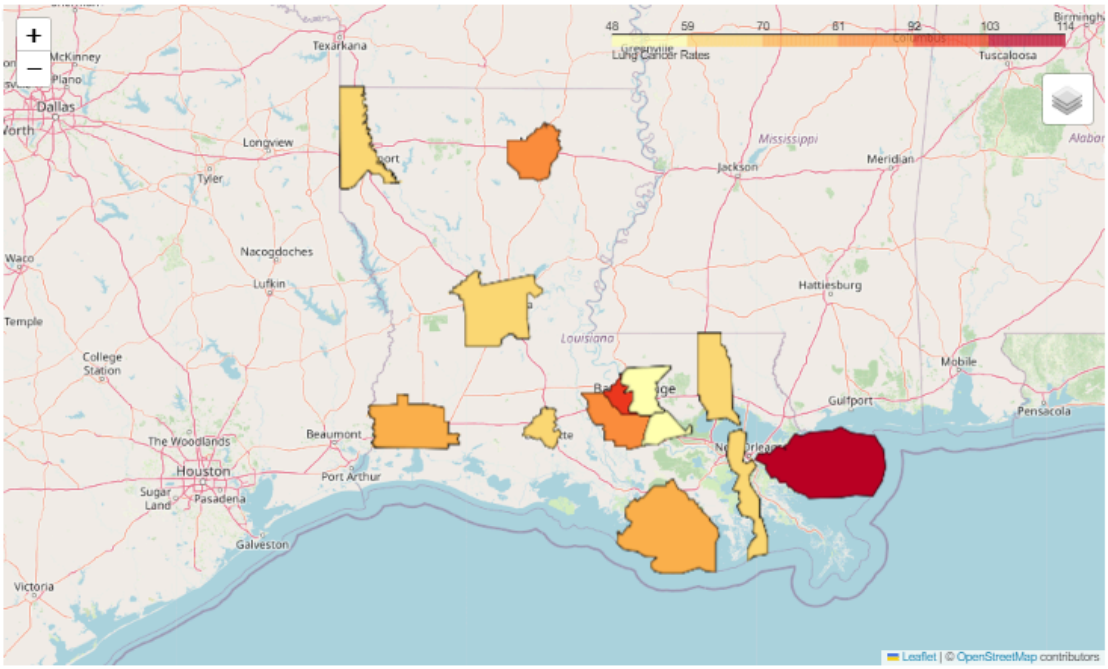
```
Moran's I for PM2.5 in 2014: 0.1568
Moran's I p-value for PM2.5 in 2014: 0.0010
Moran's I for Lung Cancer Rates in 2014: 0.4695
Moran's I p-value for Lung Cancer Rates in 2014: 0.0010
```



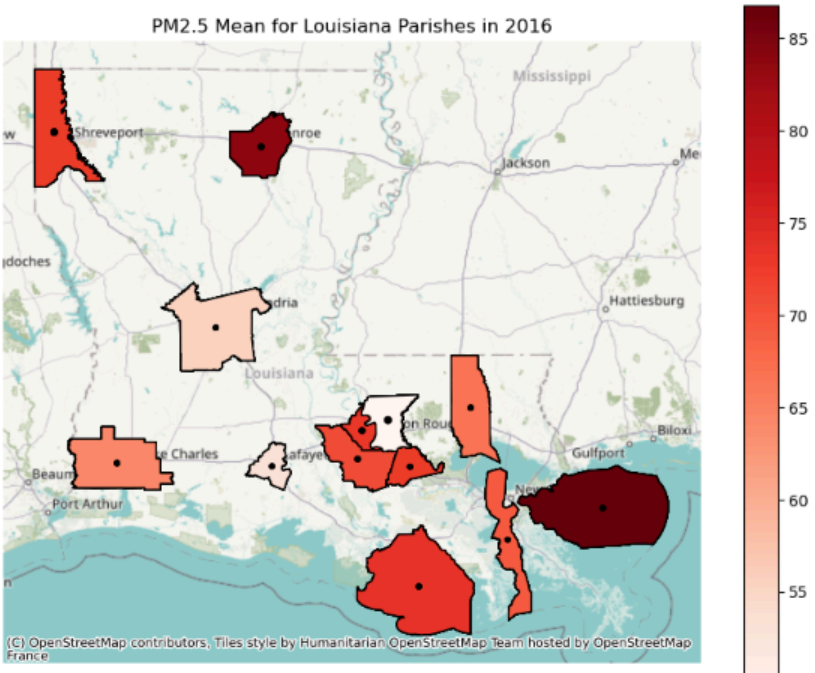


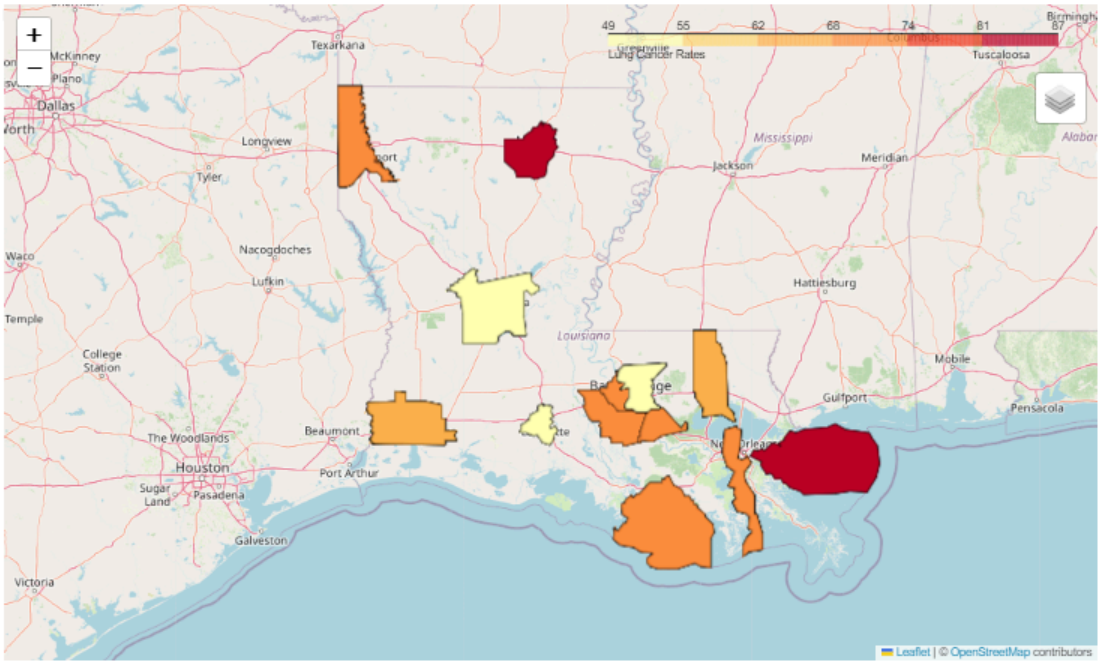
```
Moran's I for PM2.5 in 2015: 0.1456
Moran's I p-value for PM2.5 in 2015: 0.0010
Moran's I for Lung Cancer Rates in 2015: 0.6232
Moran's I p-value for Lung Cancer Rates in 2015: 0.0010
```



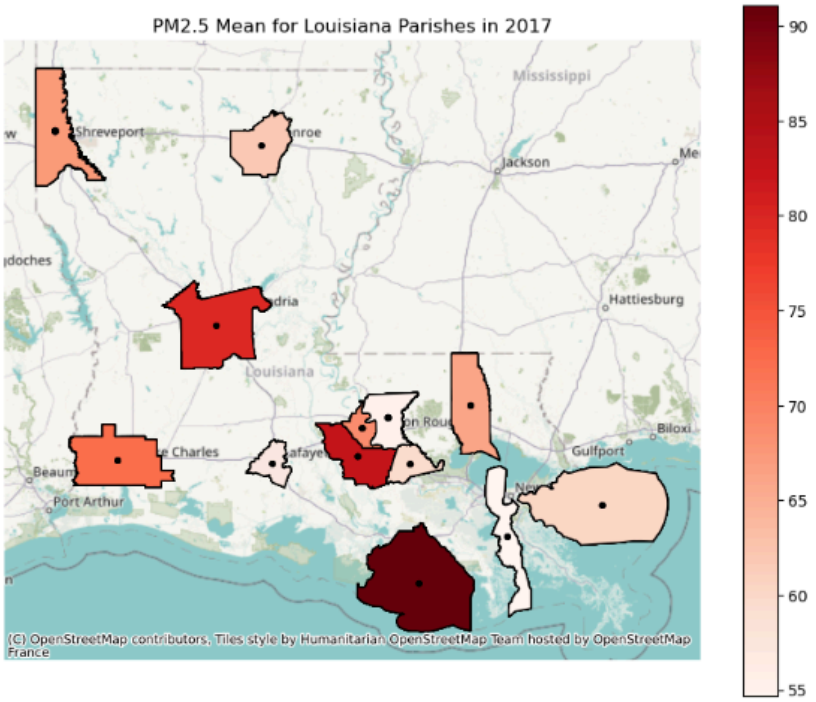


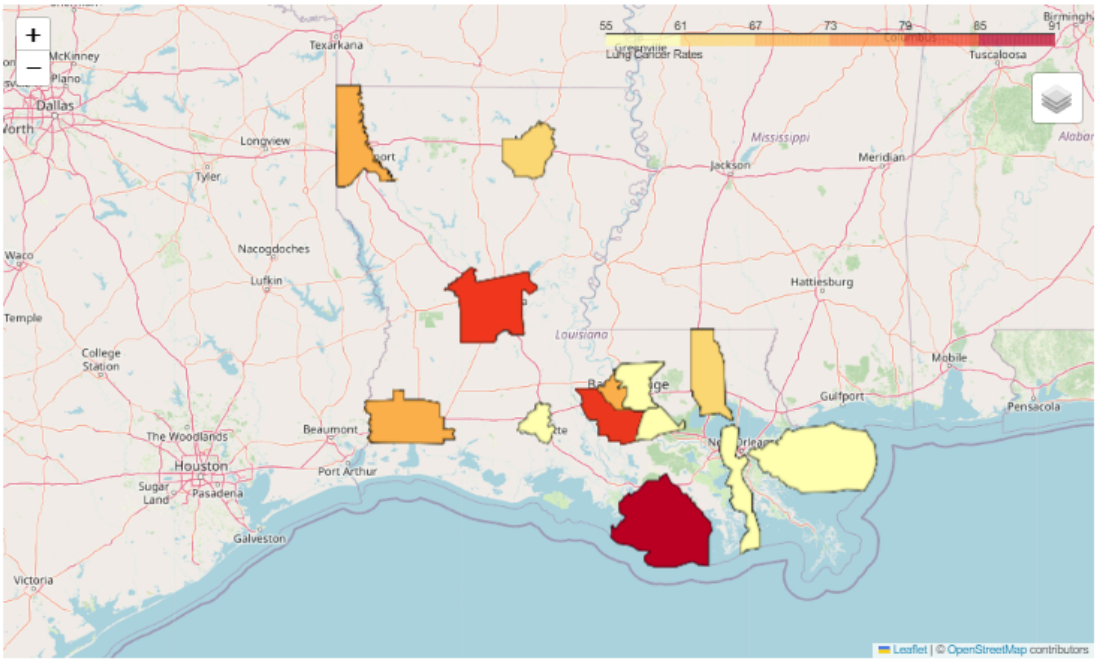
```
Moran's I for PM2.5 in 2016: 0.3067
Moran's I p-value for PM2.5 in 2016: 0.0010
Moran's I for Lung Cancer Rates in 2016: 0.6349
Moran's I p-value for Lung Cancer Rates in 2016: 0.0010
```



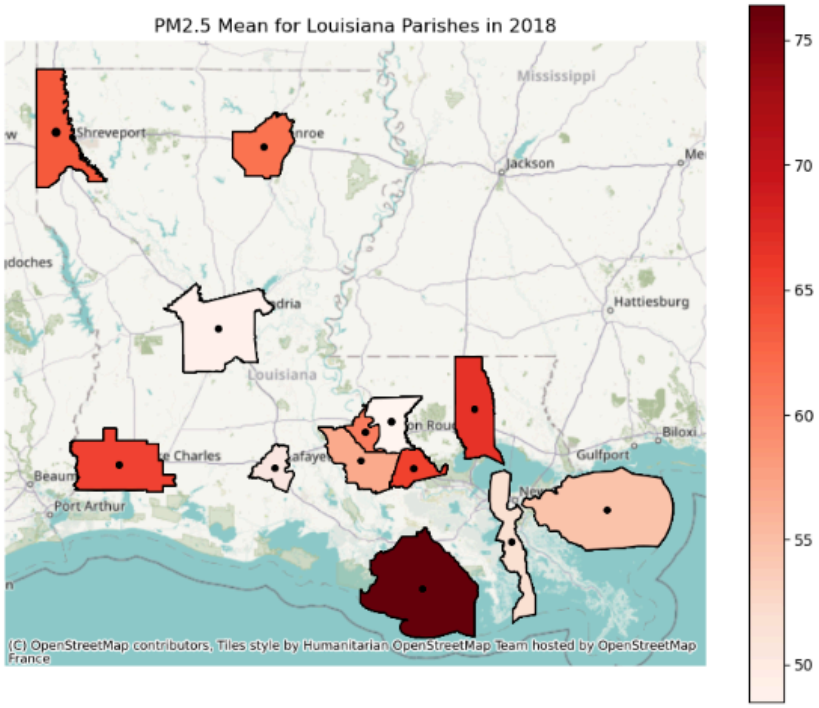


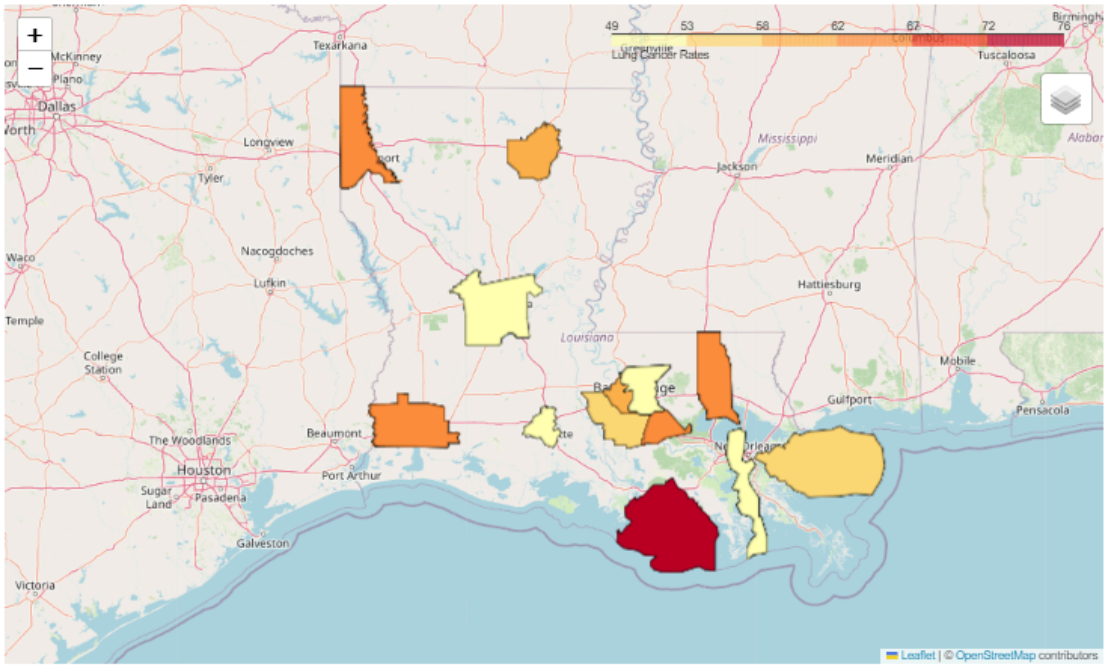
```
Moran's I for PM2.5 in 2017: 0.1772
Moran's I p-value for PM2.5 in 2017: 0.0010
Moran's I for Lung Cancer Rates in 2017: 0.6984
Moran's I p-value for Lung Cancer Rates in 2017: 0.0010
```



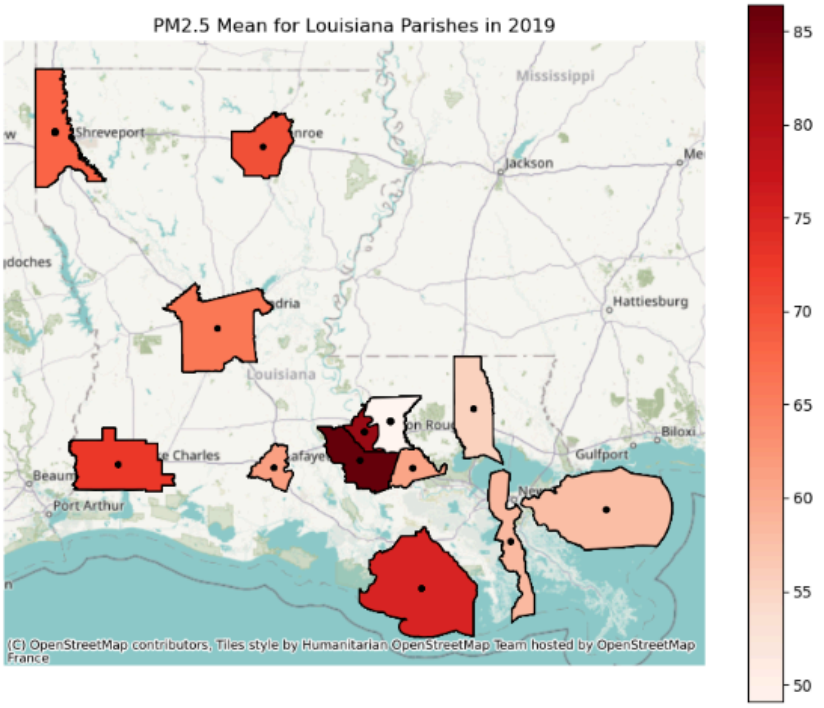


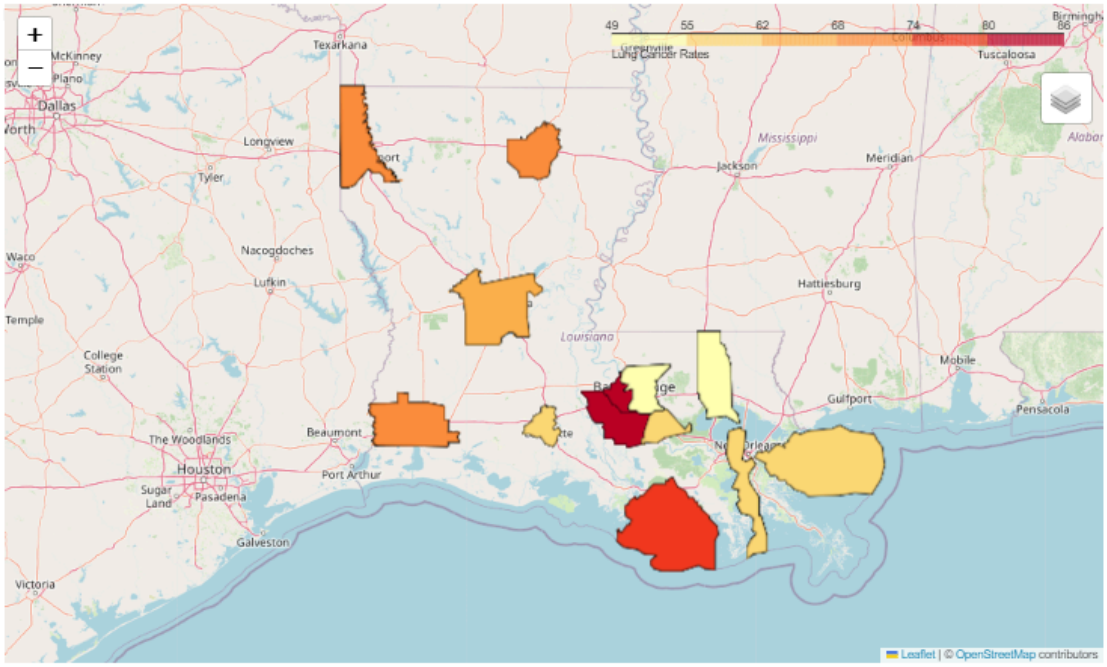
```
Moran's I for PM2.5 in 2018: 0.0321
Moran's I p-value for PM2.5 in 2018: 0.0560
Moran's I for Lung Cancer Rates in 2018: 0.7915
Moran's I p-value for Lung Cancer Rates in 2018: 0.0010
```



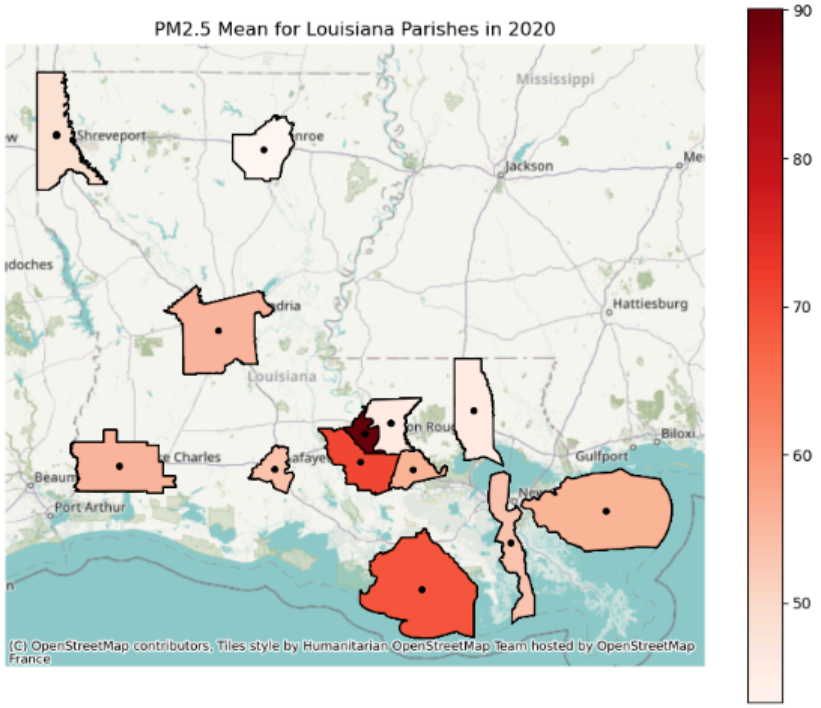


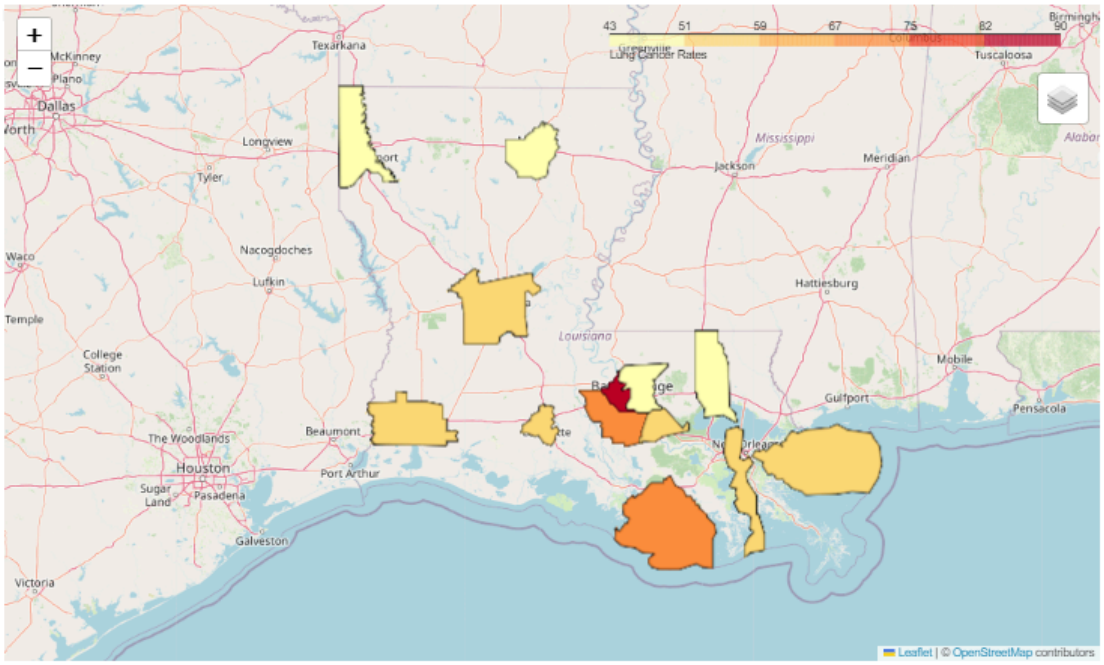
```
Moran's I for PM2.5 in 2019: 0.2154
Moran's I p-value for PM2.5 in 2019: 0.0010
Moran's I for Lung Cancer Rates in 2019: 0.3316
Moran's I p-value for Lung Cancer Rates in 2019: 0.0010
```





```
Moran's I for PM2.5 in 2020: 0.0474
Moran's I p-value for PM2.5 in 2020: 0.0210
Moran's I for Lung Cancer Rates in 2020: 0.3794
Moran's I p-value for Lung Cancer Rates in 2020: 0.0010
```





Functionality

The methods that we were able to implement ran without error. A user must download our necessary .IPNYB and Excel files to view our models.

Algorithms and Methods

To carry out our project we utilized the following steps:

1. Data Preparation
2. Model Definition
3. Model Compilation
4. Model Training
5. Model Evaluation
6. Making Predictions
7. Result Visualization

We also utilized the following methods to analyze our data:
Long Short-Term Memory(LSTM) - 1D and 3D
Convolutional Neural Network(CNN)
Mixed LSTM and CNN
Multivariate Regression for analysis of Cancer , COPD, and PM 2.5
Multivariate Regression for other factors
Moran's I Calculation:
Spatial Correlation:

Limitations and Future Work

We faced limitations in predicting lung cancer incidence rates using PM2.5 values due to discrepancies in the data sources. The lung cancer data was available yearly, whereas the PM2.5 data was not available daily. There was a struggle to convert the PM2.5 data by averaging, which did not meet our expected results. This also led to difficulty in obtaining daily lung cancer data for an accurate correlation. We also encountered issues

when conducting a multivariate analysis to explore the correlation between lung cancer, COPD, and PM2.5 values. The results were deemed to be not as satisfactory as we thought they would be.
Our future work would include finding more data in order to predict lung cancer for future years to identify the risk for Louisiana. We would also like to incorporate predicting lung cancer for the entire United States, with the necessary resources and data availability. We also did not get a chance to work with the GRU and RNN models, which can be possible methods to work on in the future.

User Instructions

Users will open up the .IPNYB file inside of Google Collab or Jupyter. Users should also download the necessary datasets in order to run the code. After opening the notebook and uploading the necessary files the user should import the necessary libraries needed to run the code. After the user has imported the libraries they can run the code snippets and view the output for the code, including multivariate, LSTM, CNN. The Spatial Correlation and Moran's I Calculation section includes an interactive dashboard where users can interact with the slider and change the years from 2011 through 2020 and view the 2 maps one for PM2.5 and one for Lung Cancer which is an interactive map.

Source Code(GitHub Links)

Shelby- https://github.com/scf9199/Spatial-Correlation-and-Moran-s-I-Calculation-for-LA-Parishes-2011-202
Tahmina- https://github.com/TahminaAnondi/DeepLearningProject
Mridula- https://github.com/Mridula96/DeepLearning