Sections

CS7641 Endterm Report

Group 22

Air Quality and Airborne Virus Transmission (COVID-19)

Introduction/Background

This project examines the impact of air quality on COVID-19 transmission and severity worldwide. Our aim is to enhance predictive models of airborne virus infection by integrating air quality data.

While studies like Arora et al.[3] used historical COVID data for forecasting, we build on this by integrating air quality data, following approaches by Ku et al.[5] and Qiu et al.[6], who successfully linked weather and air quality to respiratory and cardiovascular patient rates. This motivates our use of air quality data to predict COVID spread and severity.

We use **two** key datasets: **Google's COVID-19 Open Data**[2], which includes **~12.5M** records of cases and deaths from **20,000+** locations *(01/01/2020–09/15/2022)* along with demographic, economic, and vaccination data for those locations; and **OpenAQ's** data[4] on particulate matter and hazardous gases from **10,000** monitors across **93** countries. These datasets will be correlated by time and location.

Problem Definition

COVID-19 became the second leading cause of death globally in 2021[7], with rapid transmission overwhelming healthcare systems[8]. Accurate forecasting models are crucial for efficient resource allocation, yet existing models largely rely on historical data and often overlook other key factors. Given air quality's well-established links to respiratory illnesses[1], we propose a machine learning model that integrates air quality data to predict high-severity COVID-19 cases, aiming to improve model accuracy and resource distribution in high-risk areas.

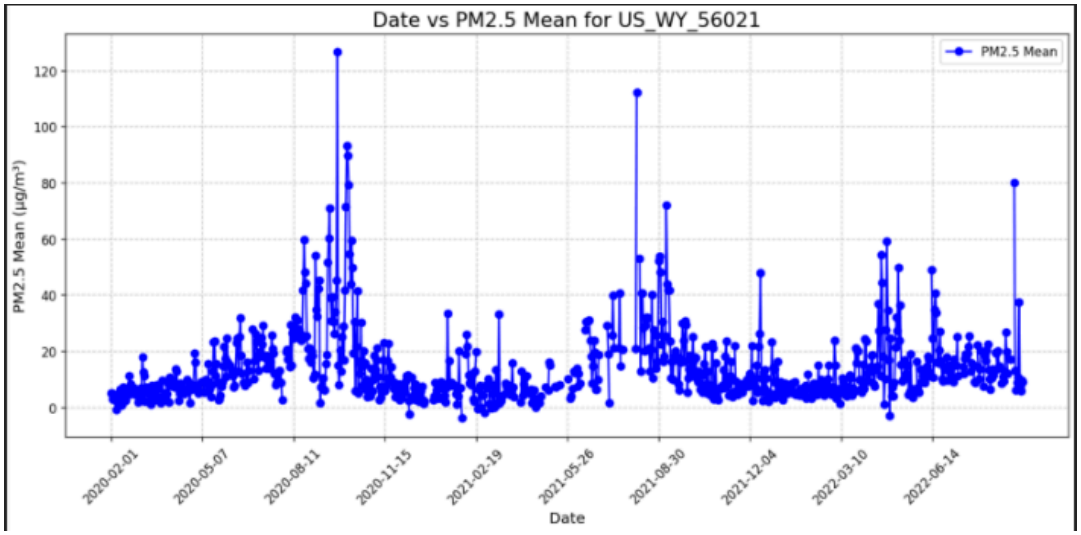
Methods

Data Preprocessing

To create the air quality dataset, we first attempted to query each possible monitor location from the **OpenAQ API** and collect the daily measurements from each of the sensors in these locations for the years from *2020 and after*. However, due to bugs and limitations in the OpenAQ API, this was not possible. We did EDA of the **Google Covid** dataset to analyze the distribution of different available features. To limit the scope of API queries so that they would not crash the API, we gathered air quality data on a county-by-county basis, limiting to the United States. This also allowed us to match the location formatting in the COVID dataset. This data was collected by first iterating through a list of every US county's Federal Information Processing Series (**FIPS**) code[9] and retrieving the coordinates of the county using Python's **GeoPy API**[10]. These coordinates were then used to query for all air monitor sensors within a *25* km radius of the coordinates. For each of the found sensors, queries were run to gather the sensor's daily measurements including minimum, q02, q25, median, q75, q98, maximum, average, and standard deviation from *January 1, 2020, and after*.

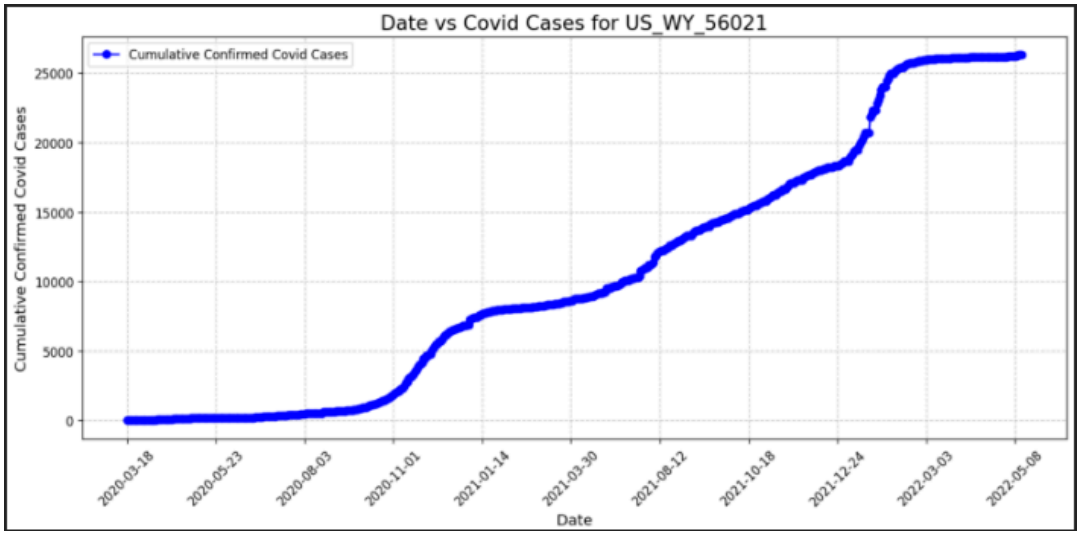
At this point, we picked our target dates to be from *February 1, 2020*, to *September 17, 2022* (the last day of data in the COVID dataset) and began filtering the data. We first removed all sensors that did not have data for at least **80%** of the days between these dates. Next, we removed all counties that did not have at least 1 sensor remaining from the dataset. Then, because sources like the European Environment Agency[11] report that PM2.5 has the greatest effect on human health, we filtered out all counties that did not have at least 1 PM2.5 sensor. Finally, for counties that had multiple sensors for the same air quality metric, we kept only the sensor data from the sensor that had daily values for the highest number of days in our date range.

The final air quality dataset includes **537** US counties from all **50** US states. The metrics it measures include: PM2.5, PM10, SO2, O3, NO2, CO, and BC. PM2.5 and PM10 are measures of the level of particulate matter in the air of size 2.5/10 micrometers or less in diameter. SO2, O3, NO2, and CO are all harmful gases, and BC is black carbon (which is a key component of the PM measurements). This dataset was converted into a csv file with each row representing a single day of measurements for a specific county. This data was merged with the Google COVID dataset by combining rows with the same date and FIPS county id.



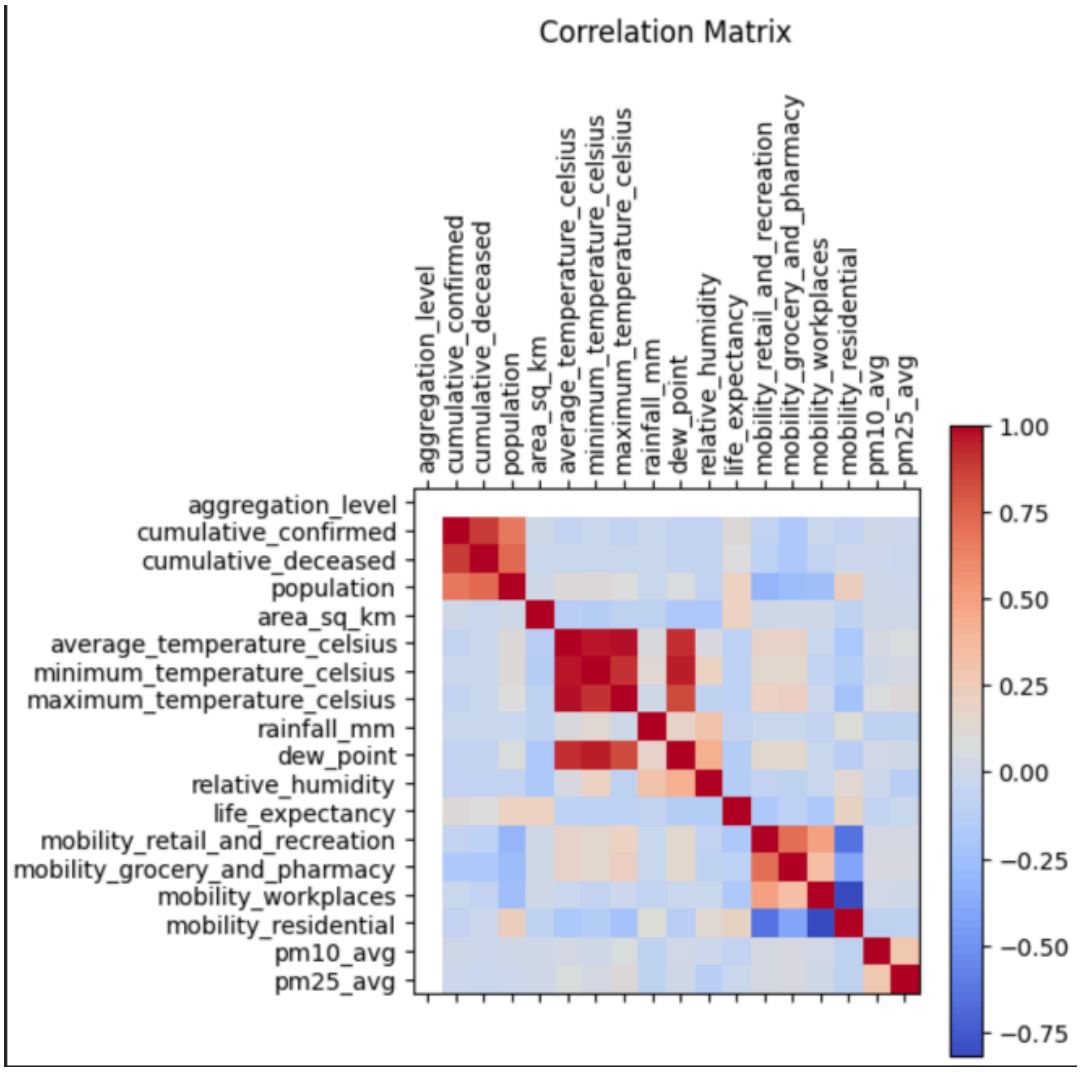
Air Quality Data for 1 county

On combination with the Google COVID dataset, we ended up with **1.1** million data points in total with **762** features which could be broadly classified into demographic features, diseases prevalence features, socioeconomic features, features related to search trends on google, underlying health comorbidities, features on government policies to reduce the spread of COVID, late vaccination statistics, environmental features like temperature, humidity, rainfall, etc., and mobility features.



Cumulative Confirmed Cases Data for 1 county

To reduce the number of features, we performed EDA and tried to analyze important features with high correlations to the Cumulative Confirmed cases. We dropped all rows which did not have the Cumulative Confirmed value. Since we aim to establish the correlation between pollutants such as the PM2.5 metric that have a severe impact, we filtered again to always include the data.
Not every county has measurements for each of the other metrics, but each county is guaranteed to at least have data for the PM2.5 metric. We decided to drop columns that were >20% Nan since the data was too sparse to have meaningful impact for our model. We identified features like mobility rate, life expectancy etc as important and further filtered the dataset to include them. Our final dataset consists of 54744 rows and 116 features.



Correlation Matrix

Dataset Overview

	Dataset Name	Number of Datapoints (Rows)	Number of Features (Columns)	Remarks
0	Air Quality Dataset	516,480	66	Collected using
1	Google COVID-19 Dataset	12,500,000	708	Features like de
2	Initial Merged Dataset	516,480	762	Filtered for US
3	Final Preprocessed Dataset	54,744	116	Final dataset af

Feature Engineering

To capture temporal patterns, we engineered new features by creating lagged values for confirmed COVID-19 cases at intervals of 1, 3, 5, 7, and 14 days, as well as for air quality measures, namely, PM2.5 and PM10, with lags of 1, 3, and 5 days. Additionally, rolling averages for PM2.5 and PM10 are calculated over 3-day and 5-day windows to account for trends in air pollution of longer duration. The numerous demographic features are also coalesced into smaller number of features by redefining the bins in contrast to the ones present in the original dataset, including gender and age-based population ratios, which are derived by normalizing population groups within each location.

ML Algorithms

KNN

The Google COVID-19 Dataset contained varying degrees of NaN values across different features. During the preprocessing phase, we excluded features with more than 80% missing values to ensure data quality. However, several key features with missing values remained essential for our modeling efforts. To address this, we employed a K-Nearest Neighbors (KNN) imputation strategy, utilizing five nearest neighbors to identify the most similar data points for each sample. This approach was used to impute missing values and prepare the dataset for subsequent modeling.

PCA

In addition to imputation, we employed Principal Component Analysis (PCA) as a dimensionality reduction and exploratory data analysis technique to gain insights into the relationships between the dataset's features and cumulative confirmed COVID-19 cases. PCA allowed us to identify the most significant components that capture the variance within the data, providing a clearer understanding of the underlying patterns and interdependencies among features.

Regression Supervised Models

To build our final regression model on the preprocessed data, we tried different regression models such as XGBoost regressors, LightGBM regressors, RandomForestRegressor, ARIMA models to do a rudimentary analysis. We picked XGBoost Regressor and Random Forest Regressor for further analysis because of their strength in handling complex datasets with non-linear relations and multicollinearity among features.

XGBoost Regressor is an advanced gradient boosting algorithm which builds an ensemble of decision trees (weak learners) sequentially and minimizes the squared loss. Random Forest employs a similar mechanism by constructing a set of multiple decision trees where each decision tree is trained on different subsets and different features of the data. This lets the model generalize better and prevents overfitting for training dataset.

SARIMAX Model

The SARIMAX or Seasonal Autoregressive Integrated Moving Average Exogenous model is a derivative of the ARIMA model that incorporates both seasonality and exogenous features. This model was considered due to the commonality and previous success of its use in the forecasting problems that we observed during literature review and was ultimately chosen among many tested models due to its superior performance on our dataset. The SARIMAX model was specifically chosen over the base ARIMA model for two reasons. The first is that strong seasonal trends were observed in our COVID dataset where winter months often saw a spike in cases and deaths and summer months saw transmission rates stalling. The SARIMAX model can capture this yearly trend and incorporate it into its forecasting decisions. The second reason for the SARIMAX model is that for our experiment of testing whether air quality data affects the accuracy of COVID forecasting models, we need to be able to include features outside of the feature we are predicting. These outside features are the X or exogenous part of the SARIMAX model that influence its forecasting of the main feature.

This model was trained and tested on the COVID data of individual counties from the dataset. The task of this model was to forecast either the number of cumulative confirmed cases or the number of cumulative deceased cases into the future. The exogenous features incorporated into the model without air quality

data included the date, new confirmed cases, cumulative confirmed cases, new deceased cases, and cumulative deceased cases. The exogenous features incorporated into the model with air quality data included all of these as well as PM10 max and average daily levels and PM25 max and average daily levels. The date was converted into an integer representing the number of days since 02-01-2020 allowing for the model to account for gaps in the dataset. The PM10 and PM25 data was converted into weekly rolling averages of the max and average daily values with a one-day lag. The new confirmed cases and new deceased cases were converted into weekly rolling averages with a seven-day lag and cumulative confirmed cases and cumulative deceased cases variables were both given with a seven-day lag. All rows with missing data either from the original dataset or due to the preprocessing of the data were simply dropped.

Four models were trained and extensively tuned for each county individually. SARIMAX models have seven tunable parameters: p, d, q, P, D, Q, and s. p represents the number of lagged series we use for the autoregressive aspect of the model. d represents how many times the data must be differenced in order to make it stationary. q represents the number of lagged forecast errors to be included in the moving average aspect of the model. P, D, and Q represent the same thing but for the seasonal component of the model. s represents the seasonality period (we chose 12 for monthly periods). Each model was tuned based on which combination of these hyperparameters produced the least forecasting error calculated using root mean squared error.

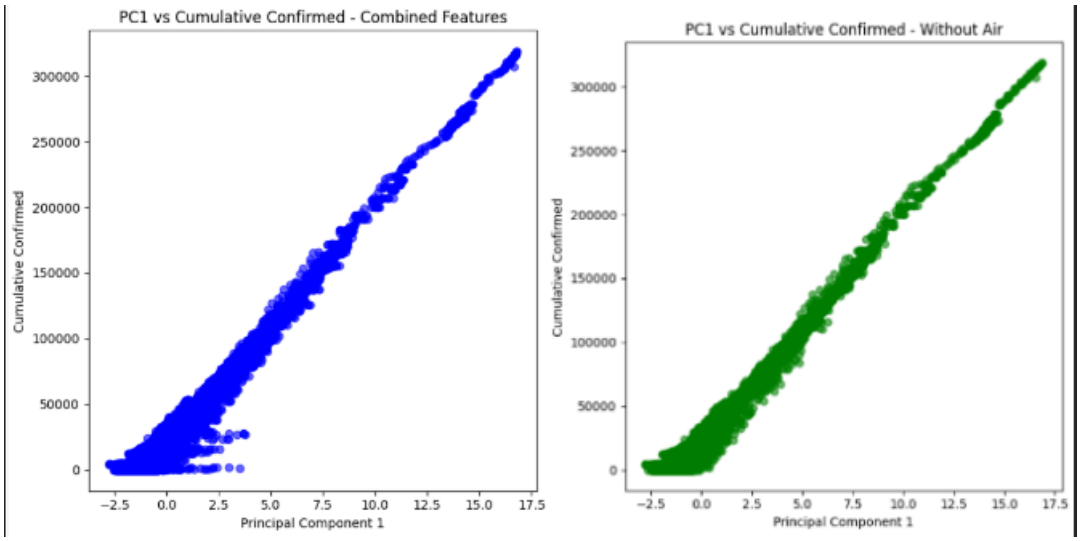
Results and Discussion

We used KNN to fill in the missing values in the dataset. The statistics for the same are as follows:

Feature Missing Values Overview

	Feature Name	Missing Values before KNN	Missing Values after KNN
0	new_deceased	319	0
1	cumulative_deceased	318	0
2	average_temperature_celsius	757	0
3	maximum_temperature_celsius	757	0
4	minimum_temperature_celsius	758	0
5	rainfall_mm	757	0
6	dew_point	757	0
7	relative_humidity	757	0
8	mobility_retail_and_recreation	1,701	0
9	mobility_residential	3,301	0

We observed a linear trend of the most significant PCA feature with the number of confirmed cases, especially at larger values as can be seen in the graphs below that contrast between PCA done on all features (with air_quality data) and without air_quality data.



Variation of Predicted Value with 1st Principle Component (with/without Air Quality)

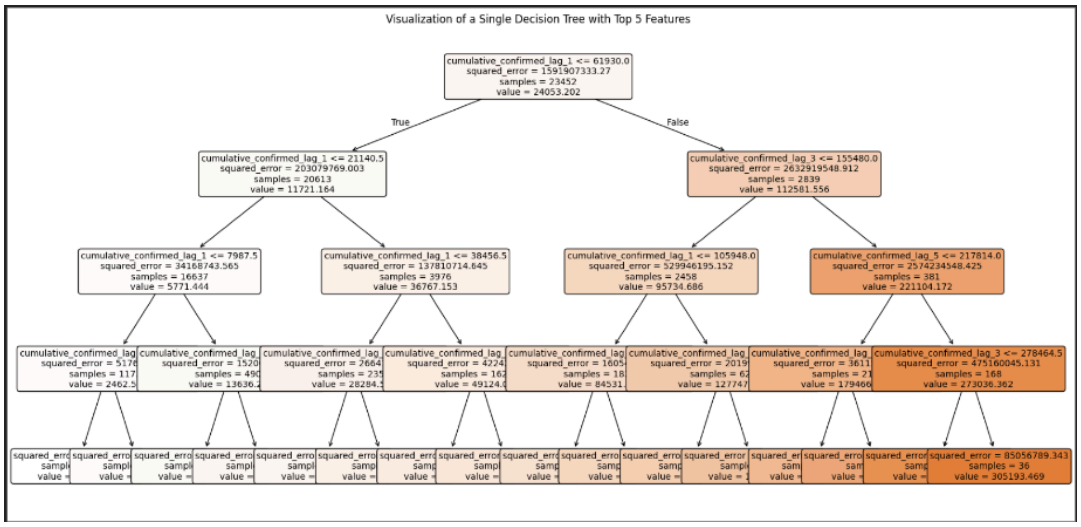
Our analysis explored both regression for USA as a whole and county specific models to predict COVID-19 levels of case counts based on air quality and demographic data. Below, we detail the outcomes of each approach, illustrated with relevant visualizations to enhance our findings.

Regression Model Results

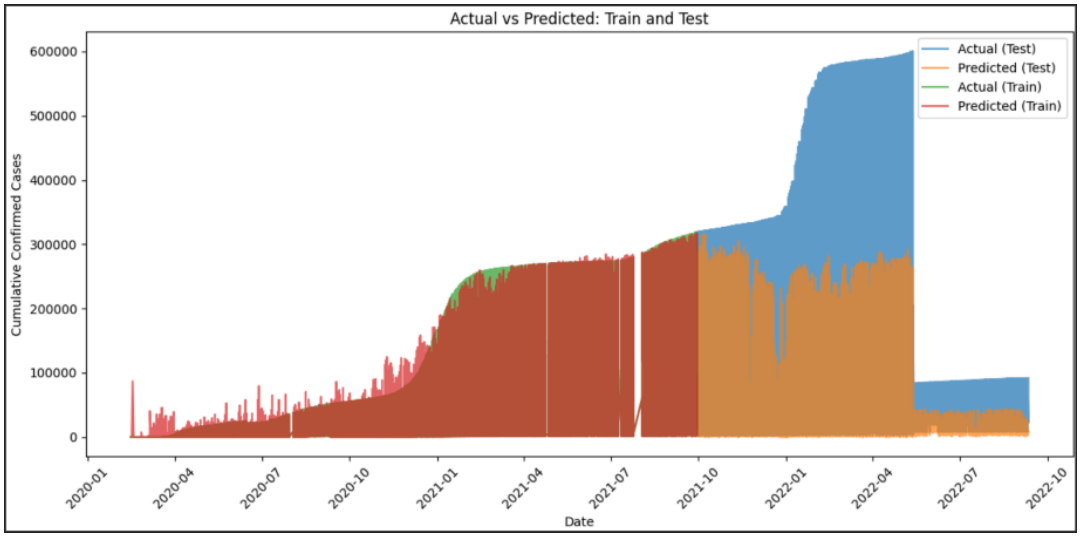
Since we were dealing with time-series data, to ensure that future values do not impact the prediction values, we split the train and test data based on a date chosen as cutoff. To find the cutoff date, we sorted the data on the 'date' column and chose the data that gave a roughly 70:30 split for training vs test data. The initial regression model aimed to predict cumulative COVID-19 case counts using temporal, demographic, and air quality features. However, as illustrated in the figure below, the model struggled to capture the sudden peaks in COVID-19 cases accurately, particularly during periods of rapid increase.

	Model	Root Mean Square Error (RMSE)	Mean Absolute Percentage Error (MAPE)
0	XGBoost Regressor (Without AQ)	34,393.2766	0.0321
1	XGBoost Regressor (With AQ)	34,491.6366	0.0323
2	RandomForestRegressor (With AQ)	23,744.9918	0.0130
3	RandomForestRegressor (Without AQ)	23,608.5980	0.0130

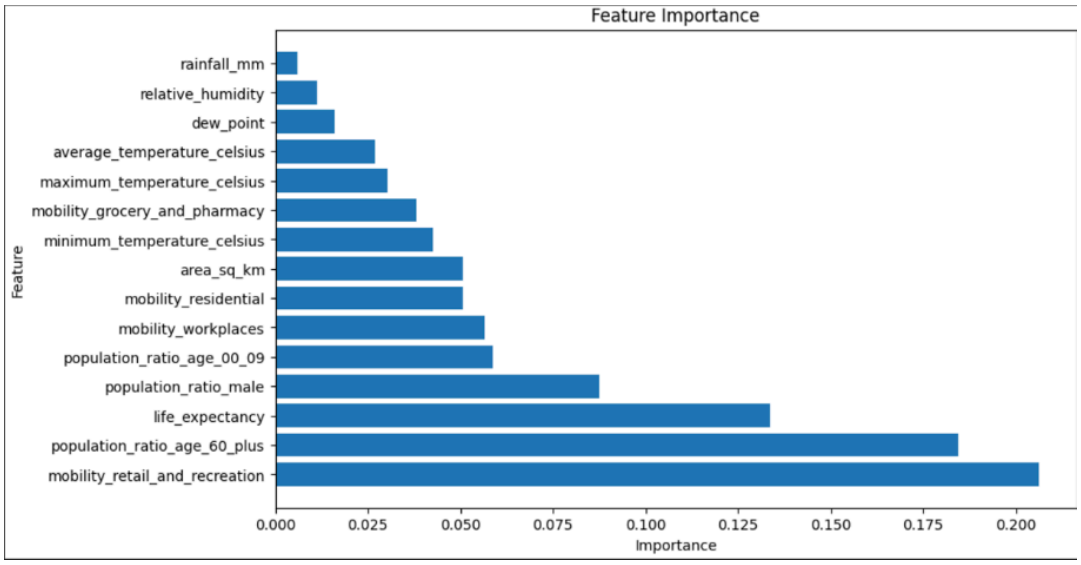
These metrics underscore the model's limited ability to handle rapid fluctuations. While it could track general trends, the error values indicate significant errors, especially around peak case periods caused by new COVID variants (which are difficult to predict with given feature set). The inclusion of AQ data introduced noise into the model, as its relationship with COVID-19 transmission is complex and confounded by many other socio-environmental factors. For instance, AQ impacts may vary significantly between urban and rural areas, leading to less generalizable patterns. These errors were likely driven by the model's inability to account for factors like county-specific policies, testing rates, or reporting inconsistencies that were not adequately captured in the available feature set. While this approach is more generalizable and suitable for broad cases, this did not provide us with the expected results. This prompted us to explore training models for specific counties rather than a generalizable model.



Random Tree Splitting on the 5 most important features



Trend Captured by the regressor



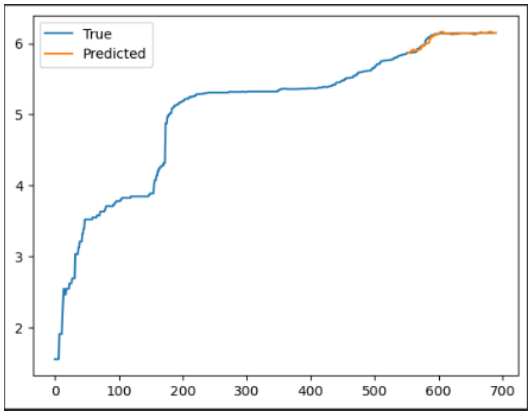
Feature Importance for RandomForestRegressor

SARIMAX Model Results

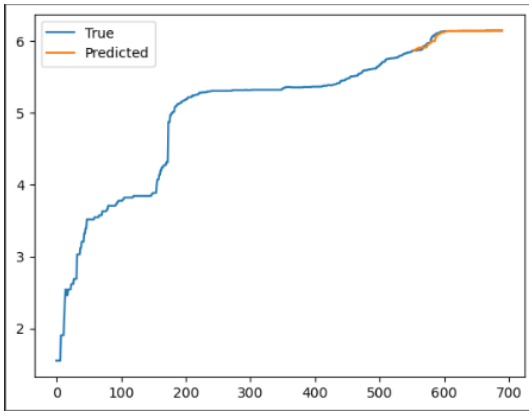
Four models were trained for each of the 50 counties: one to forecast the number of cumulative cases with air quality data included, one to forecast the number of cumulative cases without air quality data included, one to forecast the cumulative number of deceased with air quality data included, and one to forecast the cumulative number of deceased without air quality data included.

For the models forecasting the number of cumulative cases, the models trained with air quality data outperformed the models trained without air quality data in only 12 out of the 50 counties. For the models forecasting the number of deceased cases, the models trained with air quality data outperformed the models trained without air quality data in only 14 out of the 50 counties. Therefore, for most counties, including the air quality data was actually a dampening factor on model accuracy. An interesting observation to make is that the counties that performed better with the air quality data than without had, on average, higher variance in air quality data. For the models forecasting the number of cumulative cases, the counties that performed better with the air quality data had an average variance of 1.2903 in air quality whereas the models that performed worse with the air quality data had an average variance of 1.0542. For the models forecasting the number of cumulative deceased, the counties that performed better with the air quality data had an average variance of 1.1892 in air quality whereas the models that performed worse with the air quality data had an average variance of 1.0804.

Example Where Air Quality Data Improved Cases Forecast Model

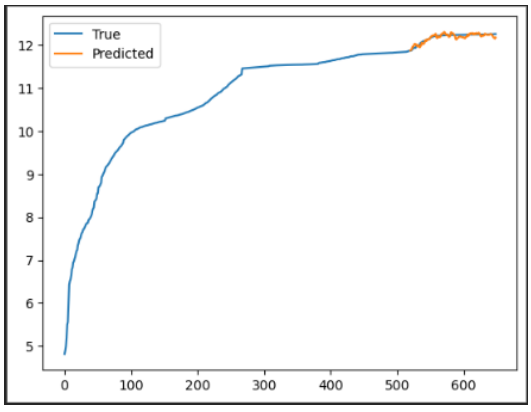


US_KS_20075 County with AQ Data

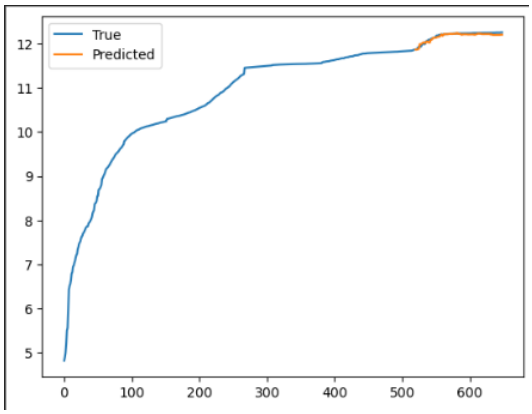


US_KS_20075 County without AQ Data

Example Where Air Quality Data Did Not Improve Cases Forecast Model

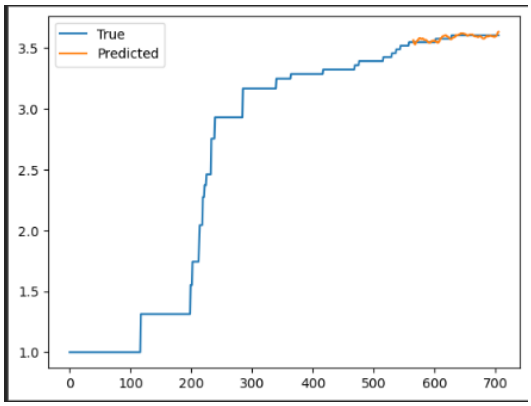


US_GA_13135 County with AQ Data

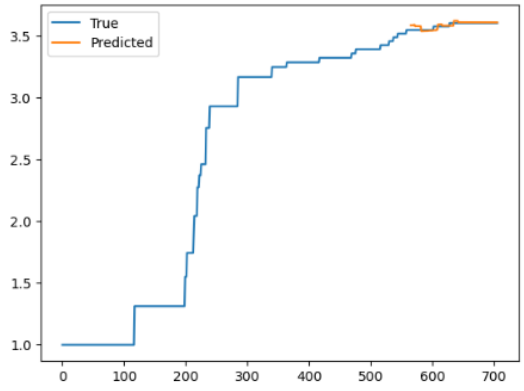


US_GA_13135 County without AQ Data

Example Where Air Quality Data Improved Deaths Forecast Model

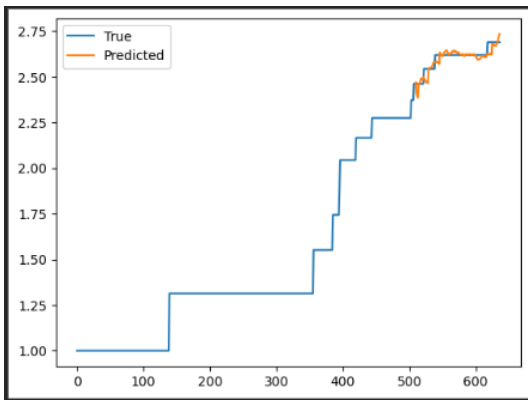


US_KS_20171 County with AQ Data

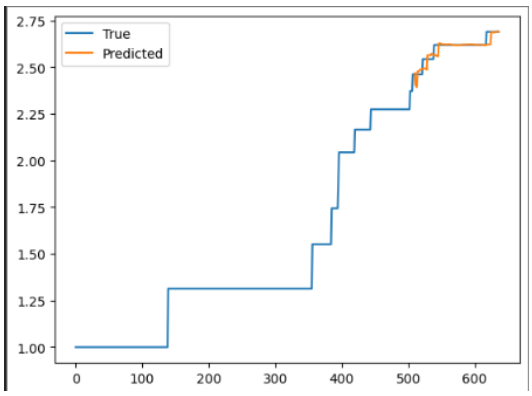


US_KS_20171 County without AQ Data

Example Where Air Quality Data Did Not Improve Deaths Forecast Model



US_KS_20171 County with AQ Data



US_KS_20171 County without AQ Data

Overall, each of the SARIMAX models performed quite well in every county. The average mean absolute percentage error (MAPE: |true – predicted| / true) for forecasting cumulative cases with air quality data over all counties was ~0.23% error and the average root mean squared error (RMSE) for these models was ~0.0286. The MAPE for forecasting cumulative cases without air quality data over all counties was ~0.197% error and the average RMSE for these models was ~0.0241. The MAPE for forecasting cumulative deceased with air quality data over all counties was ~0.462% error and the RMSE for these models was ~0.0272. The MAPE for forecasting cumulative deceased without air quality data over all counties was ~0.405% error and the RMSE for these models was ~0.0239.

	Cumulative Cases Forecast	Avg MAPE (Cumulative Cases)	Avg RMSE (Cumulative Cases)	Cumulative Deaths Forecast	Avg MAPE (Cumulative Deaths)	Avg RMSE (Cumulative Deaths)
0	AQ Data	0.23%	0.0286	AQ Data	0.462%	0.0272
1	No AQ Data	0.197%	0.0241	No AQ Data	0.405%	0.0239

In comparison of the SARIMAX models, we can see that including the air quality data causes a slight decrease in performance on average. The models forecasting the number of cumulative cases experienced an average of about 16% less error without the air quality data and models forecasting the number of cumulative deaths experienced an average of about 12.24% less error without the air quality data. One thing to notice is that in forecasting cumulative deaths, including the air quality data both causes less average increase in error and performs better in more individual counties compared to adding air quality data to models forecasting cumulative cases. This may indicate that air quality data has a more significant impact on the severity of COVID cases than it does on the transmission rates of COVID.

Next Steps

While the SARIMAX model is giving the best performance, it has multiple limitations. It takes a long time to train due to its complexity. It is also trained specific to a county rather than a generalized model having limited scope. In future work, we aim to explore alternative spatio-temporal models that can effectively balance generalizability and computational efficiency. Models such as Dynamic Spatio-Temporal Graph Neural Networks (DST-GNNs) or Bayesian Hierarchical Models could be promising candidates, as they inherently capture spatial and temporal dependencies across regions. By incorporating region-level interactions, these models could provide generalized predictions without compromising accuracy in specific counties.

References

[1] "Ambient Air Pollution," World Health Organization, 2024. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/ambient-air-pollution#:~:text=Worldwide%2C%20ambient%20air%20pollution%20is,26%25%20of%20respiratory%20infection%20deaths. *[Accessed 2 October 2024]*

[2] "Covid-19 Open Data," Google, 2022. [Online]. Available: https://health.google.com/covid-19/open-data/raw-data. *[Accessed 2 October 2024]*

[3] P. Arora, H. Kumar and D. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," Chaos, Solitons & Fractals, vol. 139, October 2020.

[4] "OpenAQ Docs," OpenAQ, 2024. [Online]. Available: https://docs.openaq.org/using-the-api/quick-start. *[Accessed 2 October 2024]*

[5] Y. Ku, S. B. Kwon, J.-H. Yoon, S.-K. Mun and M. Chang, "Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors," Clinical and Experimental Otorhinolaryngology, vol. 15, pp. 168-176, 2022.

[6] H. Qiu, L. Luo, Z. Su, L. Zhou, L. Wang and Y. Chen, "Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure," BMC Medical Informatics and Decision Making, vol. 20, no. 83, 2020.

[7] "The top 10 causes of death," World Health Organization, 7 August 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. *[Accessed 2 October 2024]*

[8] P. Sandu, A. B. Shah and e. al., "Emergency Department and Intensive Care Unit Overcrowding and Ventilator Shortages in US Hospitals During the COVID-19 Pandemic, 2020-2021," Public Health Reports, vol. 137, no. 4, pp. 796-802, 2022.

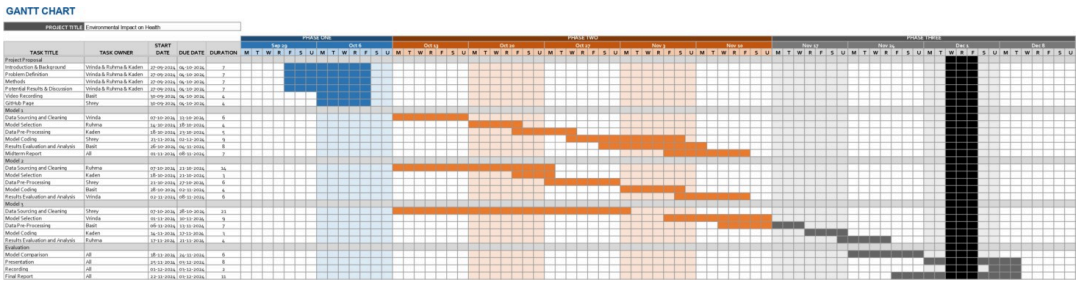
[9] Kjhealy, “GitHub – kjhealy/fips-cods,” GitHub, 2016. https://github.com/kjhealy/fips-codes [Accessed October 28, 2024].

[10]“Welcome to GeoPy’s documentation! — GeoPy 1.21.0 documentation,” geopy.readthedocs.io. https://geopy.readthedocs.io/en/stable/ [Accessed October 28, 2024].

[11] European Environment Agency, “How air pollution affects our health,” European Environment Agency, May 25, 2023. https://www.eea.europa.eu/en/topics/in-depth/air-pollution/eow-it-affects-our-health [Accessed October 27, 2024].

Gantt Chart

Viewable Link: https://1drv.ms/x/s!AjIpKh4JXZx7guJLep8WmZ-M57zytA?e=JdmBRG



Gantt Chart

Contributions

Name	Proposal Contributions
Vrinda Narayan	Data Collection, Visualizations, Cleaning, Report
Ruhma Mehek Khan	Model training, Visualization, Report
Kaden Stillwagon	SARIMAX Model, Report
Basit Khan	Data Preprocessing, Report, Video
Shrey Wadhawan	GitHub/Streamlit, Data Preprocessing, Model Training, Report