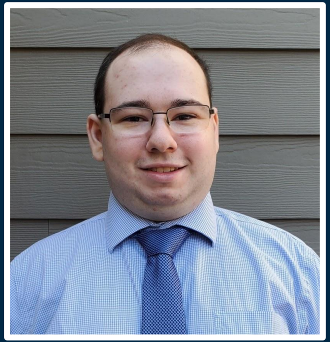
Cohort 40 Capstone

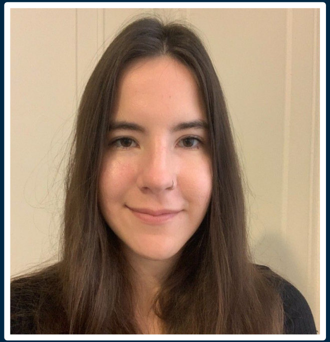
Emissions Experts

Meet Our Team

James (Finn) McSweeney
Indiana University
B.S. in Information Systems



Daniel Rose
Syracuse University
B.S. in Computer Engineering



Rachel Walter
New York University
B.A. in Psychology and English

Isabelle Hyppolite
Hofstra University
B.A. in Speech Pathology

Project Overview

Project Breakdown

- Exploring multiple datasets on emissions, air quality, pollutants and industries as well as using real time data from an air quality API

Process

- Clean and transform raw datasets
- Use Kafka to create a pipeline with a producer and consumer
- Analyze data to create insights
- Create machine learning model

Data Used

- Global Population and GDP Data
- Emissions by fuel types
- OpenWeather API
- Pollutant Levels in the US

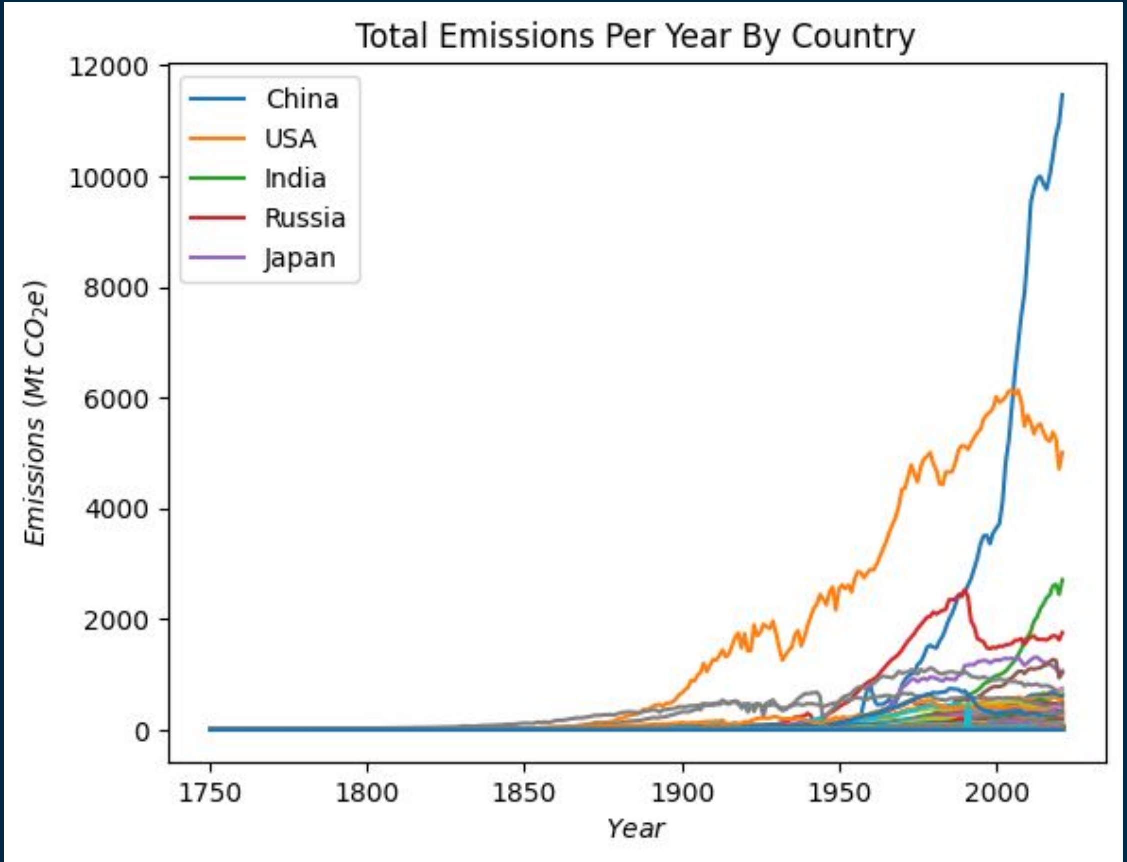
Goal

- Find correlations between our datasets and find out what are the driving factors of poor air quality and increasing emissions

Data

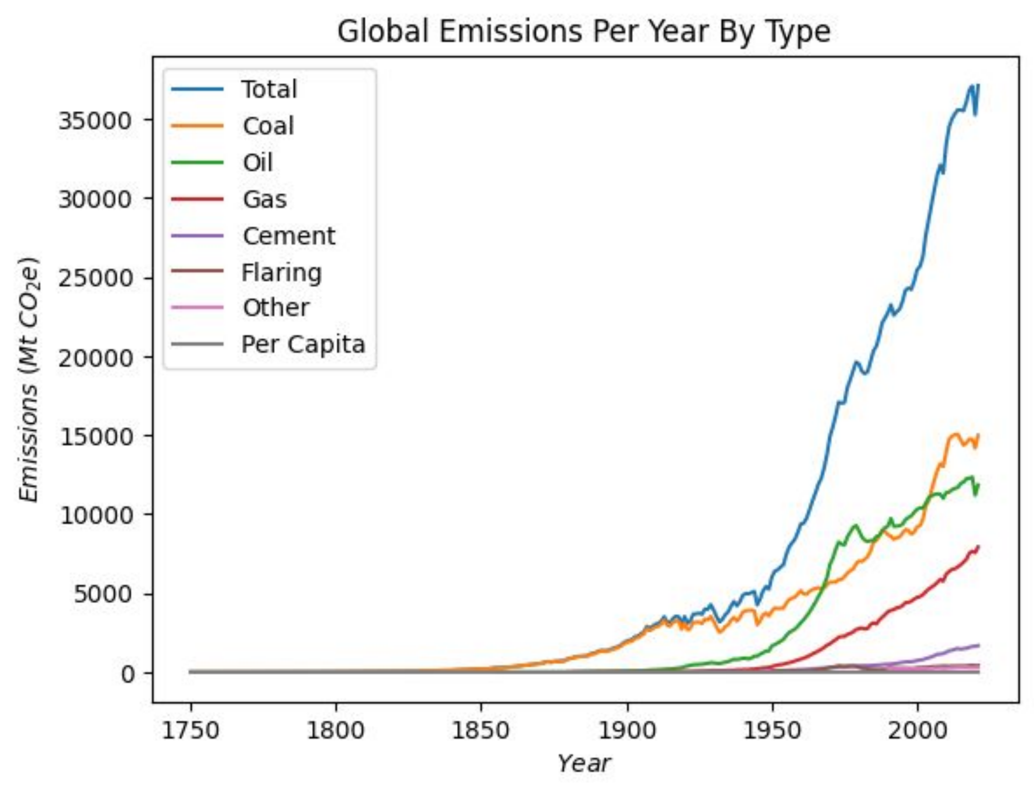
The primary data sets that were analyzed are:

1. Emissions by Country - Provides data for global emissions on a country level. It contains information on total emission and details on specific emissions such as oil, coal, gas, cement, etc.
2. Air Quality Index - Provides data about how clean or polluted the air is in a given country. The AQI focuses on health effects you may experience after breathing in polluted air. Provides historical data for every country by different parameters such as size of country, density, population growth rate, world population percentage, etc.
3. World Population - Provides historical data for every country by different parameters such as size of country, density, population growth rate, world population percentage, etc.

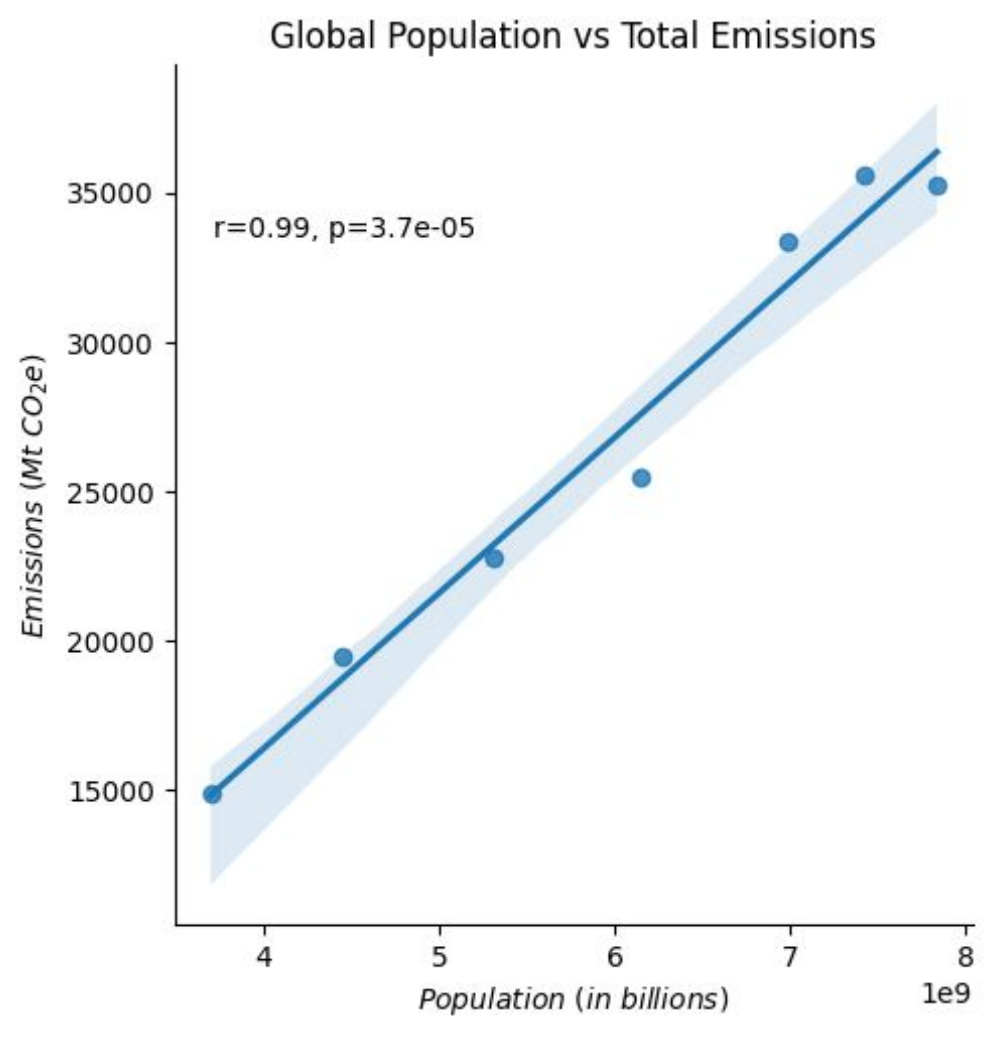


What are the countries with the highest emissions?

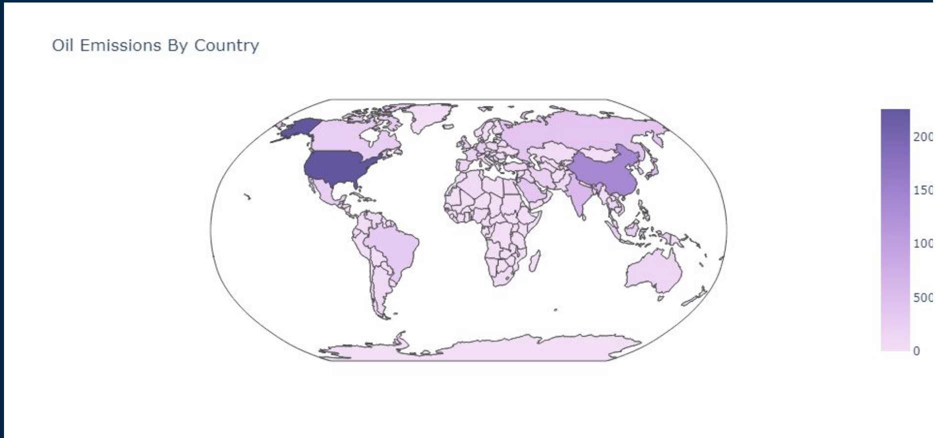
How has the production of emissions changed over time?

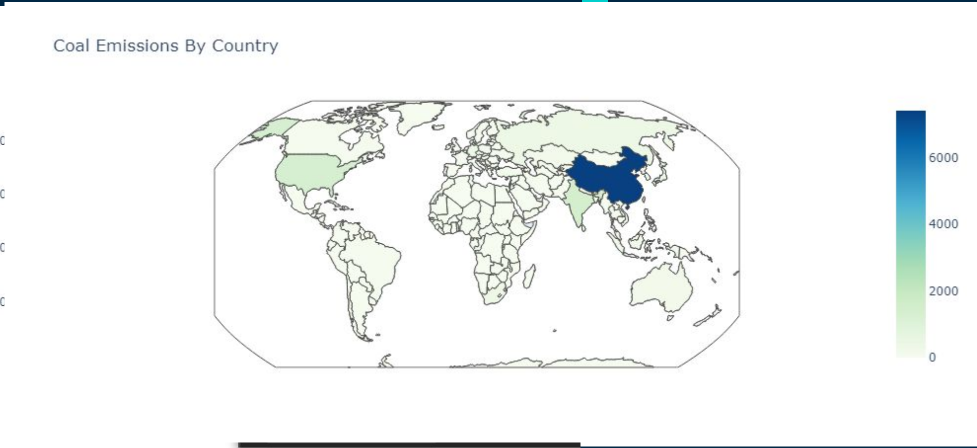


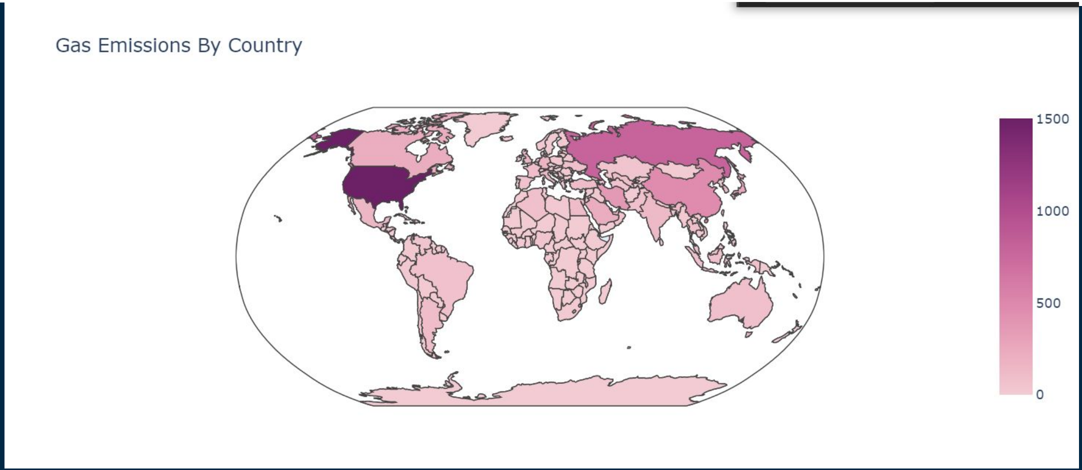
Is an increase in population over time correlated with growing emissions?



Distribution of Emissions







Datasets

1. Emissions by Unit and Fuel Type (Industry) – Information on Carbon Dioxide, Methane and Nitrous Oxide emissions from facilities of different industries.

	Facility Id	FRS Id	Facility Name	City	State	Primary NAICS Code	Reporting Year	Industry Type (subparts)	Industry Type (sectors)	Unit Name	Unit Type	Unit Reporting Method	Unit Maximum Rated Heat Input (mmBTU/hr)	Unit CO2 emissions (non-biogenic)	Unit Methane (CH4) emissions
0	1012147	NaN	17Z Gas Plant - Chevron	McKittrick	CA	211130	2018	C,NN,W	Natural Gas and Natural	CP-03.00	OCS (Other combustion	Tier1/2/3	30.0	3304.7	1.50
			Chevron USA Inc.						Gas Liquids Suppliers,...	03.00	combustion source)				
4	1012147	NaN	17Z Gas Plant - Chevron USA Inc.	McKittrick	CA	211112	2016	C,NN,W	Natural Gas and Natural Gas Liquids Suppliers,...	CP-03.00	OCS (Other combustion source)	Tier1/2/3	30.0	9922.2	4.75

	City	Reporting Year	Industry Type (sectors)	Unit CO2 emissions (non-biogenic)	Unit Methane (CH4) emissions	Unit Nitrous Oxide (N2O) emissions
0	McKittrick	2018	Petroleum	3304.7	1.50	1.788
1	McKittrick	2017	Petroleum	9106.1	4.25	5.066
2	Brooklyn	2021	Power Plants	23434.5	11.00	11.920
3	Brooklyn	2020	Power Plants	25233.9	13.50	14.900
4	Brooklyn	2019	Power Plants	19780.8	9.25	11.920

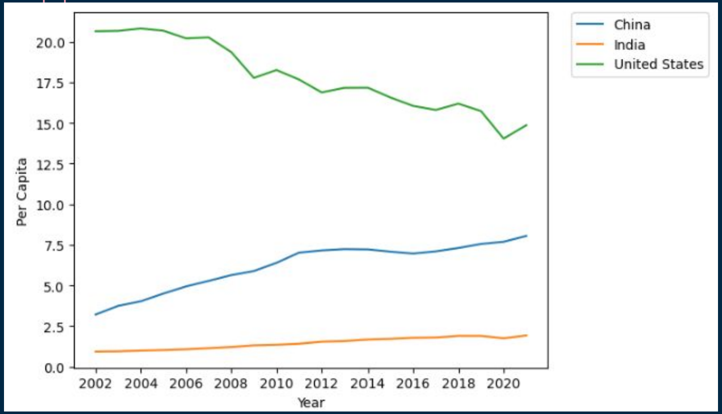
Datasets

1. Daily PM10/PM2.5 Speciation – Provides daily information on the level of PM10/PM2.5 particles in the air using an arithmetic mean on a particular day.
2. Daily NO2 Criteria Gas Summary Data – Provides daily information on the mean level of Nitrogen Dioxide in a given city on a particular day.

	Latitude	Longitude	Parameter Name	Sample Duration	Date Local	Units of Measure	Arithmetic Mean	1st Max Value	1st Max Hour	AQI	Local Site Name	State Name	City Name
0	33.553056	-86.81										Alabama	Birmingham
1	33.553056	-86.81										Alabama	Birmingham
2	33.553056	-86.81										Alabama	Birmingham
3	33.553056	-86.81										Alabama	Birmingham
4	33.553056	-86.815	dioxide (NO2)	1 HOUR	0:00	billion	16.493750	29.4	18	27	Birmingham	Alabama	Birmingham

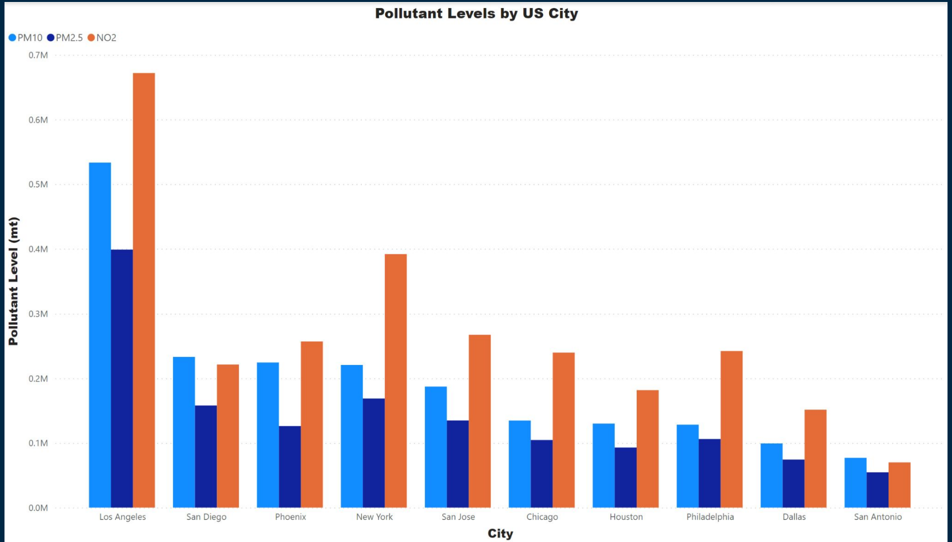
	City Name	State Name	Parameter Name	Nitrogen dioxide Levels	AQI	Year
0	Birmingham	Alabama	Nitrogen dioxide (NO2)	8.785318	18.601671	2022
1	Phoenix	Arizona	Nitrogen dioxide (NO2)	15.770163	30.327928	2022
2	Buckeye	Arizona	Nitrogen dioxide (NO2)	7.716651	17.268868	2022
3	Tucson	Arizona	Nitrogen dioxide (NO2)	7.772277	17.983003	2022
4	Marion	Arkansas	Nitrogen dioxide (NO2)	6.014533	13.961749	2022

Question #1: Given the United States high emissions per capita, what regions have the highest levels of pollutants?

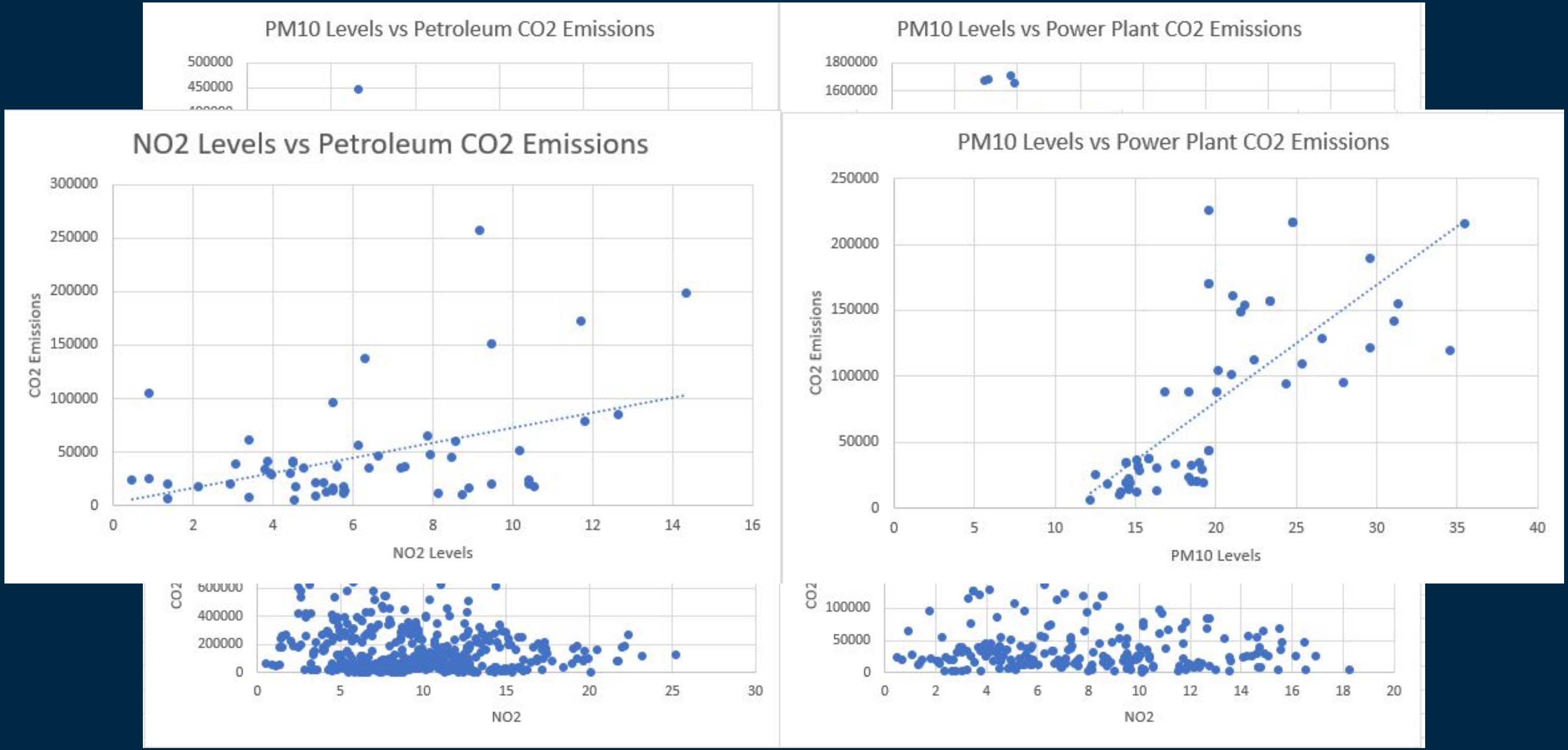


Findings

- United States, while less populated than India and China, has high emissions per capita
- Los Angeles is highest in all categories
- New York and Philadelphia have higher than average Nitrogen Dioxide emissions



Question #2: What industries have the biggest impact on pollutants

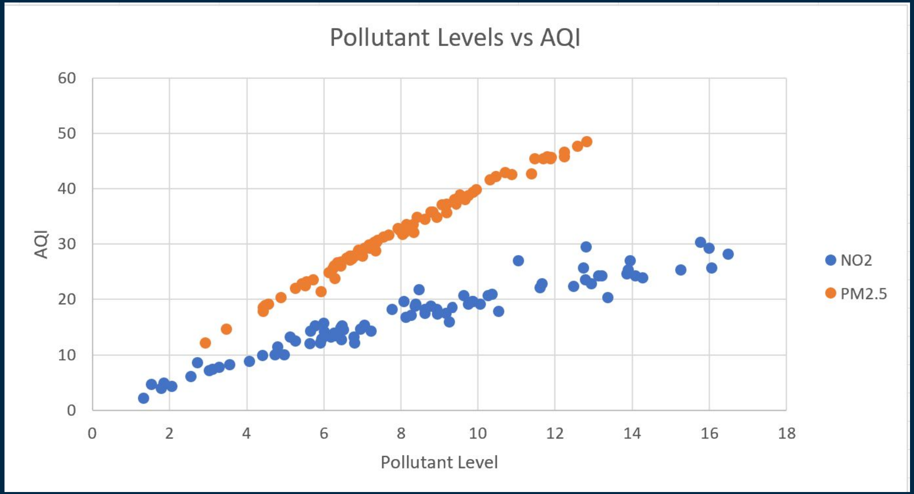


Why this matters

- These pollutants have a direct impact on the air quality index in these cities.

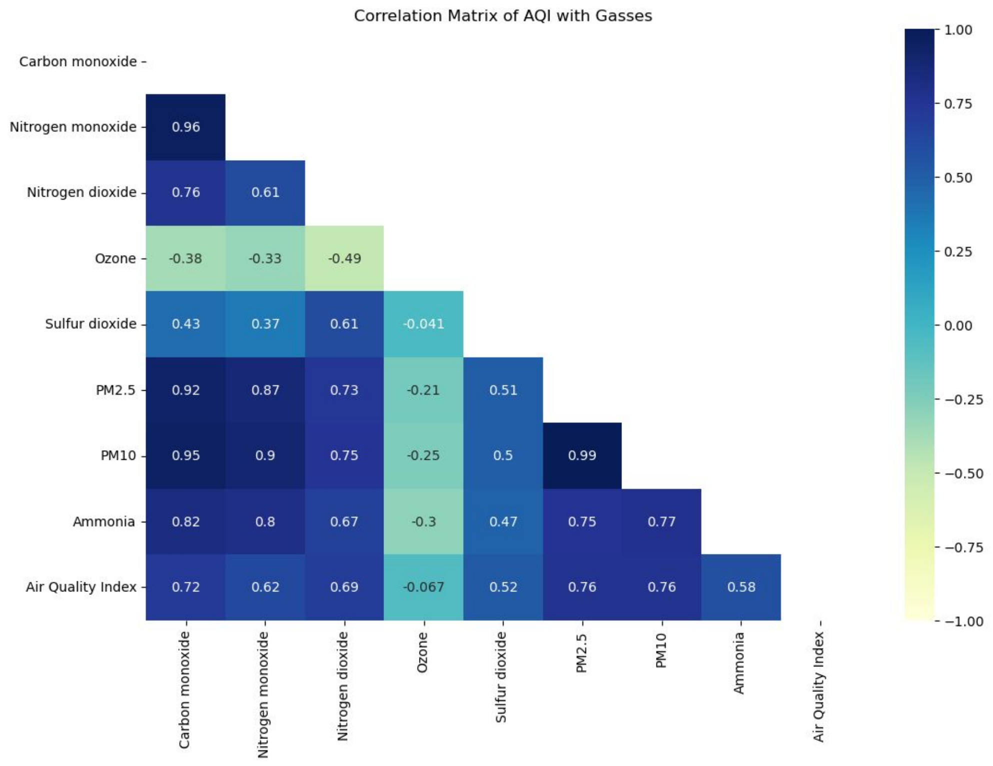
- Although particulate matter affects the AQI more than nitrogen dioxide, both of these pollutants are harmful and contribute to increasing air quality indexes across the county

- If emission levels continue to rise, especially in populated cities, the air quality will only get worse causing unhealthy living conditions, especially those classified in sensitive groups



Machine Learning

- Our aim was to predict a value relating to air quality, so we decided upon PM2.5
- PM2.5 is particulate matter smaller than 2.5 microns
- PM2.5 has a .76 correlation with AQI, the highest besides PM10
- It is considered more dangerous than PM10



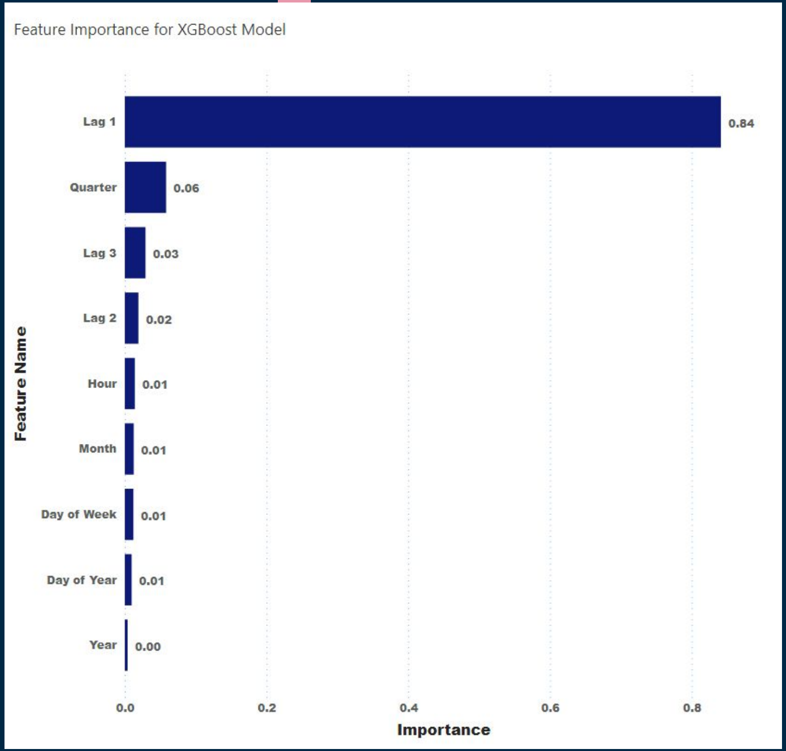
Time Series Data

- The challenge with our data was that it was all time series data
- The traditional models we learned about in class need to be modified to predict time series data values
- So, we decided to create two models and compare their performance
- Our traditional model:
  - XGBoost
- Our model that is meant to analyze time series data:
  - LSTM

date	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi
12/1/2020 5:00	373.84	1.5	43.87	8.49	6.86	9.31	11.75	1.3	2
12/1/2020 6:00	343.8	1.16	37.7	9.39	7.09	8.43	10.47	1.09	1
12/1/2020 7:00	337.12	1.79	35.99	6.35	7.21	8.55	10.8	1.08	1
12/1/2020 8:00	337.12	3.38	34.96	3.09	7.63	8.92	11.57	1.08	1
12/1/2020 9:00	340.46	5.87	33.59	1.16	8.23	9.62	12.64	1.09	1
12/1/2020 10:00	347.14	8.83	33.59	0.37	9.18	10.6	13.9	1.08	2
12/1/2020 11:00	370.5	13.08	33.93	0.07	10.01	11.92	15.37	1.14	2
12/1/2020 12:00	447.27	23.02	35.64	0	10.49	15.04	19.41	1.69	2
12/1/2020 13:00	507.36	30.85	37.01	0.16	10.97	16.85	21.72	2.06	2
12/1/2020 14:00	500.68	30.4	34.96	1.65	11.09	15.64	20.16	1.98	2
12/1/2020 15:00	467.3	24.36	34.27	3.71	11.09	13.76	17.67	1.82	2
12/1/2020 16:00	407.22	13.86	32.22	11.18	9.89	9.9	12.62	1.38	1
12/1/2020 17:00	393.87	12.29	29.13	19.31	10.01	6.93	8.92	1.33	1
12/1/2020 18:00	377.18	12.52	25.02	22.35	10.01	4.57	6.03	1.24	1

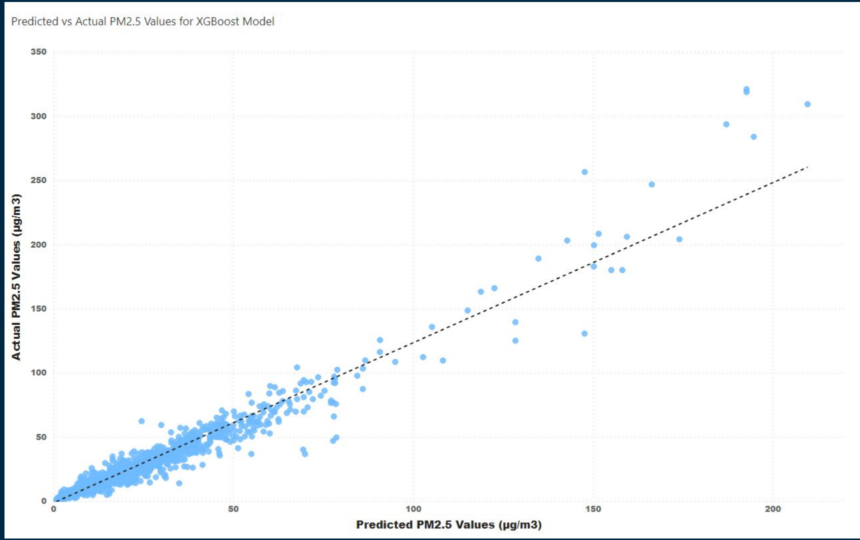
XGBoost

- To make our XGBoost model more suited to predicting time series data, we added time features
- As you can see, “Lag 1” was the most important feature to our model
- We also modified hyperparameters like “n_estimators,” “learning_rate,” and “early_stopping_rounds”



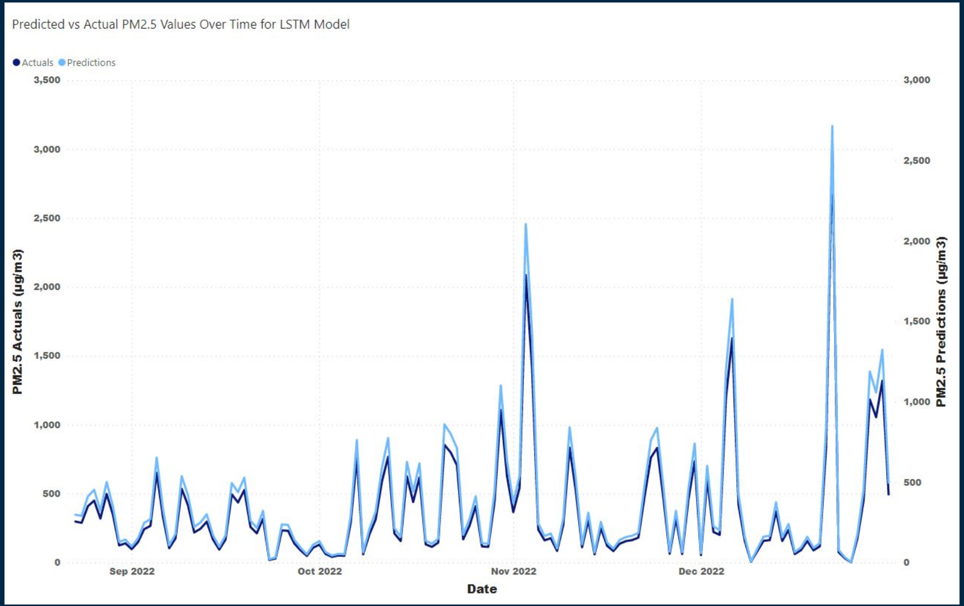
The XGBoost Model's Performance

- Before adding features, the r^2 value for the test set was .12
- After, the model's r^2 values were .94 on the training set and .90 on the test set
- The Mean Squared Error (MSE) for the test set was 40.26



LSTM

- For our LSTM model, we did not have to add any features to improve performance
- The model already takes into account past values when it makes its predictions
- Instead, we modified hyperparameters like learning rate, early stopping, and the number of epochs



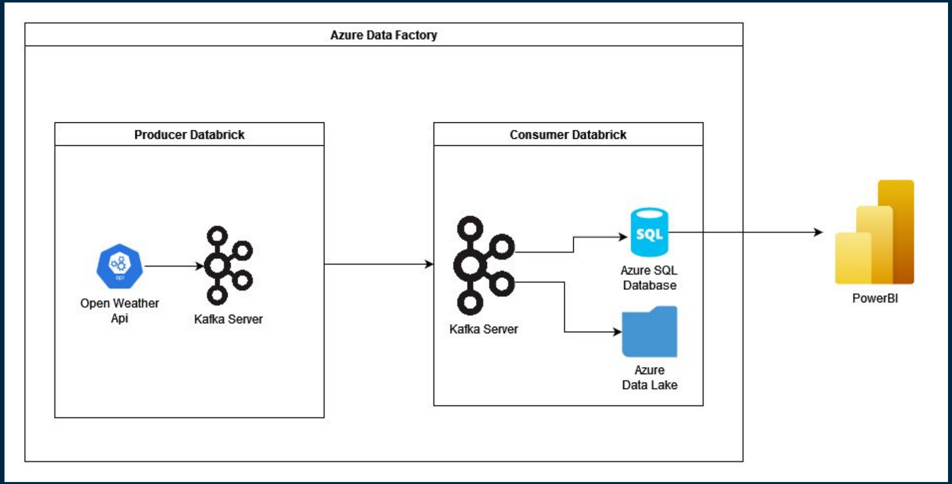
Kafka Pipeline

<u>Producer</u>

- Uses an admin client to create a new kafka topic
- Uses Open Weather API to get the current air quality data from the 10 biggest US cities
- Converts the data into PySpark
- Sends the data over kafka using our new topic

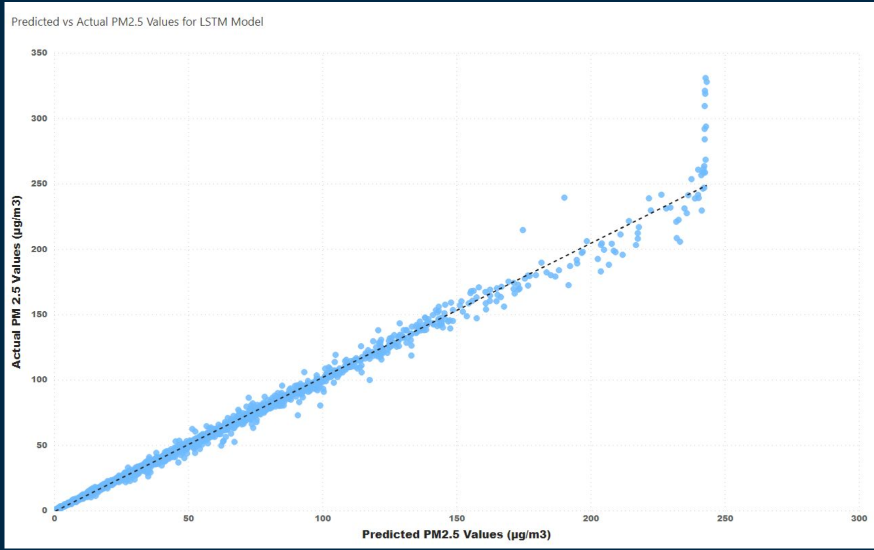
<u>Consumer</u>

- Reads in the data from our kafka topic.
- Decodes the data and turns it into a PySpark dataframe
- Uploads the data to an Azure Data Lake Storage and to an Azure SQL server

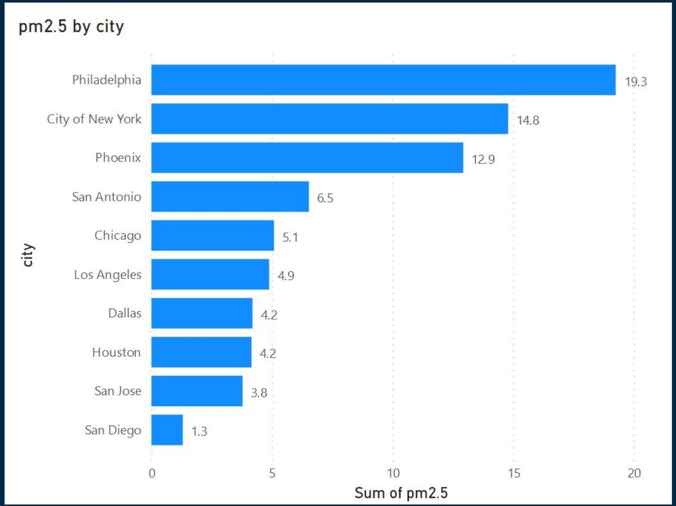


The LSTM Model's Performance

- The model's r^2 values were .99 for the training set, .99 for the validation set, and .98 for the testing set
- The MSE for the test set was 7.37, which is much smaller than the previous model's MSE of 40.26



What are the current air quality levels in US cities?



Dashboard

Data Sources

Banerjee, S. (2022, October 20). *World Population Dataset*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset

Devastator, T. (2023, January 24). *Emissions by country*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022

OpenWeather. (2023). *Air Pollution*. OpenWeather. Retrieved February 1, 2023, from https://openweathermap.org/api/air-pollution

Tas, O. C. (2022, March 19). *World GDP(GDP, GDP per capita, and annual growths)*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths

Wasi, A. T. (2023, January 12). *AQI - air quality index*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/datasets/azminetoushikwasi/aqi-air-quality-index-scheduled-daily-update