alpacanonymous / **capstone** Public

0 stars 0 forks

Star | Stop ignoring

Code Issues Pull requests Actions Projects Wiki Security

main

alpacanonymous pdfs and notebook ... 28 seconds ago 44

View code

README.md

Forecasting NYC Air Quality Based on Four Major Pollutants

Overview

This project analyzes air pollution data of four major gas pollutants--ground-level ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2)--and creates time series models to forecast future air quality in New York City.

Note: *For ease of reference, I will use O3, CO, NO2, and SO2 when naming the pollutants, knowing full well that technically they are not accurate chemical formulas.*

Business Problem

Air pollution is a huge problem for everyone. According to the Environmental Defense Fund (EDF), air pollution is currently the biggest environmental risk of premature death. It is highly linked to cardiovascular and respiratory disease and worsens symptoms of susceptible populations.

Not only is air pollution bad for public health, it's also bad for the economy. Air pollution costs the US roughly 5% of its annual GDP in damages ($790 billion in 2014). The highest costs come from premature deaths. A study by Anthony Heyes, Matthew Neidell, and Soodeh Saberian even suggests that air pollution affects the stock market.

Air pollution also exacerbates the race-class divide. Racial and ethnic minorities are exposed to higher levels of air pollution, especially in highly segregated neighborhoods. Urban areas are more polluted than rural areas, which is where there are denser populations of minorities.

Decreasing air pollution would benefit public health and the economy and contribute to a more equitable society.

(source)

Data Understanding

The data for this project was collected from the US Environmental Protection Agency (EPA). The EPA provides open-source pre-generated data files on air pollution dating back to 1980. I gathered the daily summary data for the years 2000-2021. Each pollutant had its own dataset of daily records per year, totalling 88 individual datasets for this project. Each dataset had the same 29 features. The target variable is the Air Quality Index (AQI) score. I chose to focus on four major gas pollutants.

Air Quality Index (AQI)

The AQI was developed by the EPA to provide a simple, uniform way to report daily air quality conditions across all recorded pollutants. The national standard is set at 100, meaning that this is the score at which the EPA deems air quality to be safe for most of the population. After that, there is increased risk of illness for sensitive groups up until hazardous conditions above 300. For each day, each pollutant records an AQI value, which may vary, but the final AQI chosen is the one that reports the highest AQI value. For example, if the AQI of O3 is 98, the AQI of CO is 74, and the AQI of NO2 is 103, the AQI of that day will be reported as 103.

Ground-level Ozone (O3)

- concentration measured in ppb
- commonly known as smog
- formed from combustion of fossil fuels
- short-term exposure: chest pain, coughing, throat irritation
- long-term exposure: decreased lung function, COPD

Carbon Monoxide (CO)

- concentration measured in ppm
- formed from burning of fossil fuels, mainly by vehicles
- reduces amount of oxygen that can be transported by bloodstream
- short-term exposure: chest pain
- enclosed environment: dizziness, confusion, unconscious, death

Nitrogen Dioxide (NO2)

- concentration measured in ppb
- produced primarily by transportation sector
- can result in development and exacerbations of asthma and bronchitis
- can lead to higher risk of heart disease

Sulfur Dioxide (SO2)

- concentration measured in ppb
- emitted by burning of sulfur-containing fossil fuels
- causes eye irritation, worsens asthma, increases susceptibility to respiratory infections, impacts cardiovascular system
- combined with water, forms sulfuric acid, the main component of acid rain, which then contributes to deforestation

Data Preparation & Analysis

After downloading all 88 required datasets, I concatenated them into their respective pollutant datasets. I kept data only of the 50 US States and DC, dropped rows where the `Pollutant Standard` did not produce `AQI` values, got rid of columns that were either redundant (ie. location) or unnecessary (ie. units) for the purposes of this project, and renamed a few columns for conciseness. I took those and created time series datasets that only contained `Date`, `State`, `County`, `City`, and `AQI`. Finally, I extracted the data pertaining only to NYC to start modeling on a smaller scale (with hopes I had enough time to expand nationwide) and made four more datasets. A total of 16 datasets were created and exported as `.csv` to be imported into my main notebook later.

Many of the resulting `.csv` files were too large to upload onto github with its limit of 100MB, but you can download all the files I used from the EPA site and run my `create_datasets` notebook to get the compiled datasets.

Click here for more details on my data preparation.

Modeling & Forecasting

I chose RMSE (root mean squared error) as my forecast metric. RMSE is easily interpretable and on the same scale as my target variable, AQI.

Baseline Model

- RMSE: 1.55

Model 1

- RMSE: - -

Model 2

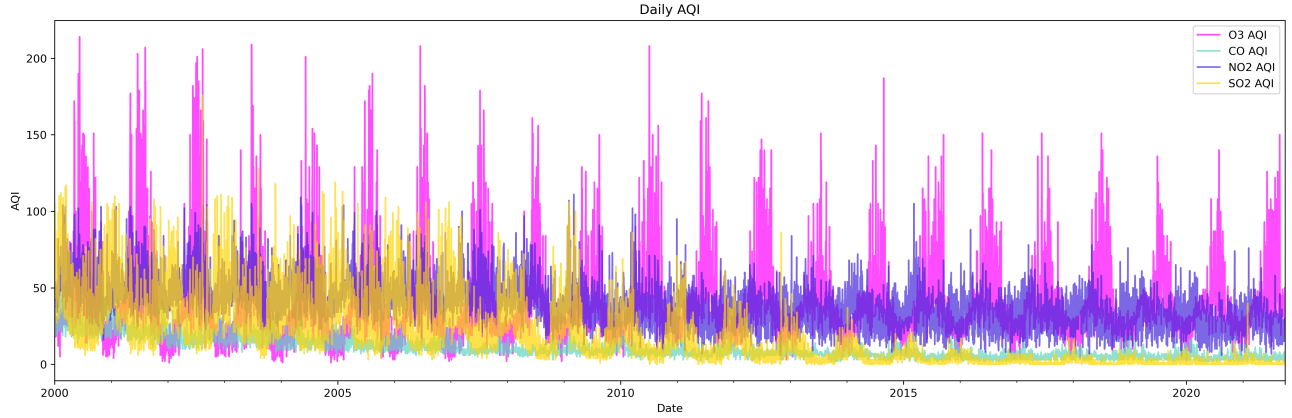
- RMSE: - -

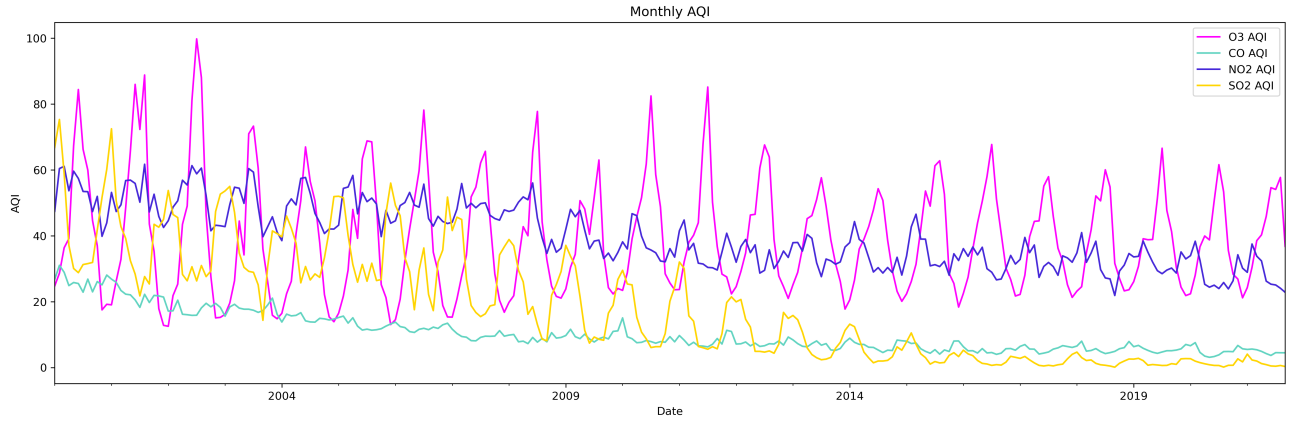
Model 3

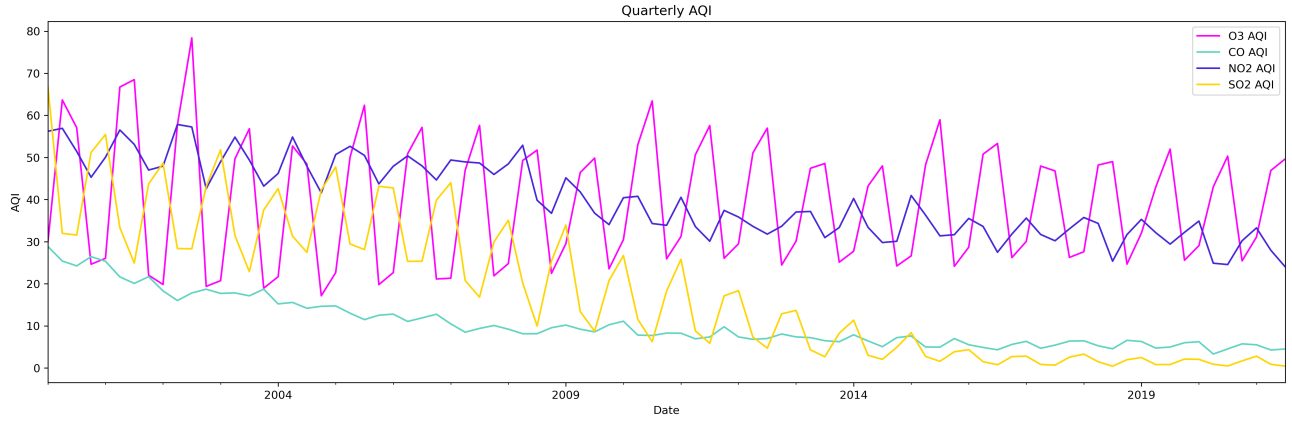
- RMSE: 0.8

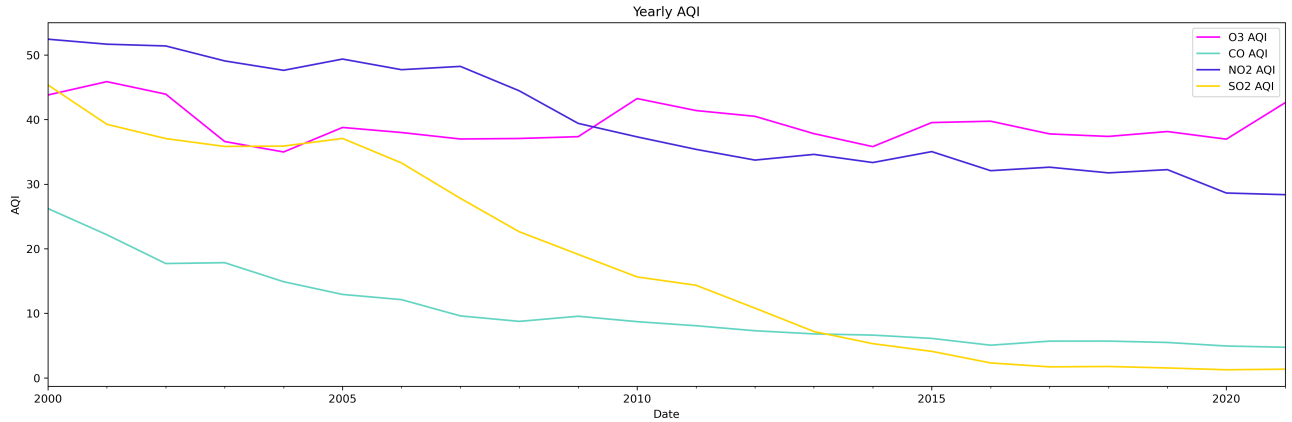
Click here for further details on my iterative modeling approach.

Visualizations









Conclusions

In conclusion, my SARIMA model forecasted air quality in New York City quite well and could even be used in shaping government policy on public health. I would recommend implementing measures to decrease the presence of air pollutants, especially ozone and nitrogen dioxide, as there hasn't been much decrease from 2000. I would also suggest posting air quality forecasts, so that vulnerable populations can plan ahead.

To view my presentation, click here.

Next Steps

Given more time and resources, I would like to explore beyond New York City, modeling for other cities and even seeing how cities compare to suburban or rural areas. Another pollutant I'd like to consider is particulate matter.

In terms of modeling, I would like to see how well a recurrent neural network would perform and venture into vector auto regression for multivariate time series.

Sources

- US EPA
  - EPA AirData Daily Summary Data
  - About AirData Reports
- AirNow
  - Technical Assistance Document for the Reporting of Daily Air Quality (pdf)
- Air Pollution | WHO
- Explore Air Pollution in New York | 2021 Annual Report
- Health Impacts of Air Pollution | Environmental Defense Fund
- The Effect Of Air Pollution On Investor Behavior: Evidence From the S&P 500 (pdf) | A. Heyes, M. Neidell, S. Saberian
- How much does air pollution cost the U.S.? | Stanford University

Repository Structure

```
├── [data]
│    ├── nycCO.csv
│    ├── nycNO2.csv
│    ├── nycO3.csv
│    └── nycSO2.csv
├── [images]
├── [pdfs]
```

```
│   ├── github.pdf
│   ├── notebook.pdf
│   └── presentation.pdf
├── .gitignore
├── README.md
├── create_datasets.ipynb
└── notebook.ipynb
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%