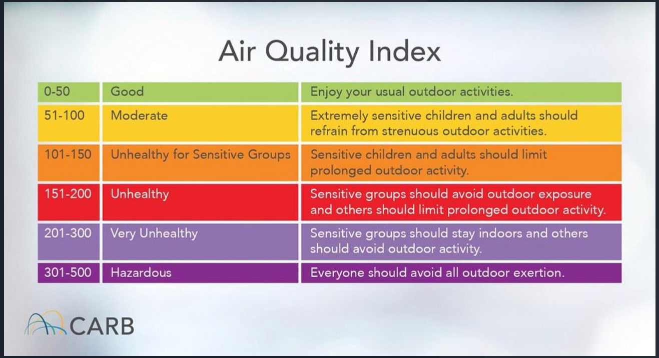
Air Quality Data Pipeline



Air Quality Index

0-50	Good	Enjoy your usual outdoor activities.
51-100	Moderate	Extremely sensitive children and adults should refrain from strenuous outdoor activities.
101-150	Unhealthy for Sensitive Groups	Sensitive children and adults should limit prolonged outdoor activity.
151-200	Unhealthy	Sensitive groups should avoid outdoor exposure and others should limit prolonged outdoor activity.
201-300	Very Unhealthy	Sensitive groups should stay indoors and others should avoid outdoor activity.
301-500	Hazardous	Everyone should avoid all outdoor exertion.

Erik Paulson

The purpose

The purpose of my project was mainly to try and gather a large amount of data and store it in an effective way that doesn’t require a huge upfront resource (tons of ram or high end GPU/CPU for example)

The secondary objective of this project was to try and see which states and which counties had the best air quality in the past five years by checking a few major pollutants.

Data Gathering

- I used the EPA Air Quality API to gather data from outdoor air quality testing sites in the United States from 2015 to 2020.

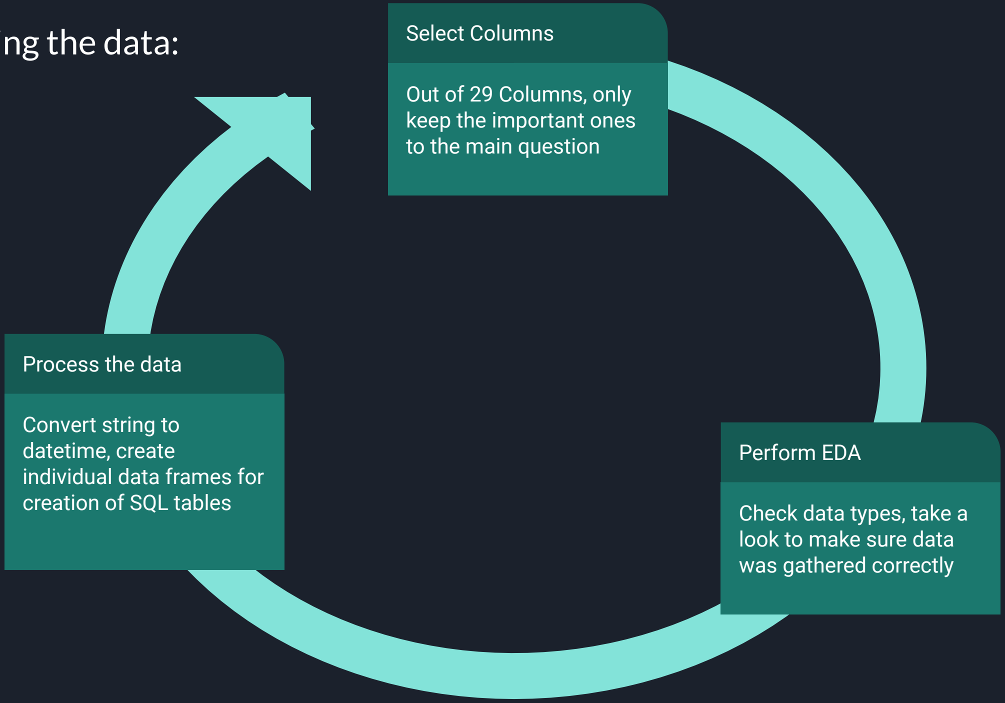
- The data consists of approximately 25M rows and 29 columns, of which I focus on 6. The 6 I kept have to do with quantity measurements, states, counties, date, and pollutant.

- There may not be data from each state for each year.

UNITED STATES • ENVIRONMENTAL PROTECTION AGENCY •

Data Processing

After gathering the data:



Select Columns

Out of 29 Columns, only keep the important ones to the main question

Perform EDA

Check data types, take a look to make sure data was gathered correctly

Process the data

Convert string to datetime, create individual data frames for creation of SQL tables

Data Storage

To effectively store the data, I used a SQLite database.

After gathering a specific pollutant's data, I used python to create a CSV to update/place the data in my database.

Each pollutant was given a table in my SQLite database, with data about the pollutant from all states that gave data.

Data Deployment

The main way I deployed my data is through an app called Streamlit. Using this app allowed me to display preliminary findings about the data such as:

- States with the highest concentration of pollutants

Using cool graphs from plotly so each plot is interactive.

Link to Streamlit page

- https://swedes5-data-engineering-project-app-nu3qn8.streamlitapp.com/

Streamlit

Conclusion and Future Work

In conclusion, it appeared that one specific state does not have vastly superior air quality compared to any others. (At least at the air quality testing sites)

Some future work could be continued analysis of state air quality and instead of looking at average concentrations per state/county look at trends from year to year or month to month to see if air quality is improving or getting worse over time.