Data Engineering Course

Abstract

The goal of this project was to use methods to gather and store large data in an efficient manner. I was able to gather large data from the Environmental Protection Agency (EPA) weather data API and store it in a SQLite database effectively. Then, using Streamlit (an app for deploying web pages), I was able to create an interactive dashboard for users to be able to view some simple plots regarding my data.

Design

My project design follows a format of data gathering →data processing → data storage → data deployment/presentation.

Data

The dataset I gathered consisted of over 25 million rows with 29 unique columns of which I selected 6 to look at.

Algorithms

There were no major algorithms for this project. I mostly used tools like matplotlib and plotly to visualize the data. They worked well even regarding the large size of the data.

Tools

- Pandas for data manipulation
- Matplotlib and Plotly for plotting visuals
- SQLite for data storage
- Google Sheets/Slides for generating URL's for my CSV files and presenting

- Streamlit for displaying data and presenting

Communication

A Google Slide presentation for a walkthrough of my data acquisition and storage process and Streamlit for displaying some of my data and findings using interactive dashboards.