Project and Data Management Plan

[Github](https://github.com/RahulS77/7PAM2002-0509-2024-Final-Project-AQI)

[Colab](https://colab.research.google.com/drive/1bawbYw6zvh4edpAffMt_mFcxD0R3MbOw?usp=sharing)

[Dataset from USEPA](https://aqs.epa.gov/aqsweb/airdata/download_files.html) and [API docs](https://aqs.epa.gov/aqsweb/documents/data_api.html#param) which were used to obtain data

Project Overview

* Title: Short-Term Air Quality Index (AQI) Forecasting for Los Angeles Using Hybrid Machine Learning Models
* Summary: This project predicts the 7-day AQI for Los Angeles using EPA AQS data (2014–2024), focusing on pollutants (PM2.5, PM10, Ozone) and meteorological factors (temperature, wind, humidity). I’ll use a hybrid ARIMA-LSTM model, to forecast AQI and interpret key drivers using SHAP(SHapley Additive exPlanations).
* Research Question: Can a hybrid ARIMA-LSTM model accurately predict the 7-day AQI for Los Angeles, and what are the key drivers of AQI in this region?
* Objectives:
  + Collect and preprocess EPA AQS data for Los Angeles (2014–2024).
  + Perform EDA to identify trends and correlations in AQI and predictors.
  + Develop and evaluate a hybrid ARIMA-LSTM model for 7-day AQI forecasting.
  + Use SHAP to interpret the model and identify key AQI drivers.
  + Compare performance with a CatBoost model.
* Literature References:
  + 2024 Seoul Study: Hybrid ARIMA-LSTM for AQI forecasting (similar model).
  + 2021 Beijing Study: CNN-LSTM with spatial features (potential extension).
  + 2023 Indian Study: CatBoost with SHAP for AQI prediction (backup model, interpretability).

Project Plan (Timeline)

* Week 1 (May 26 – June 1): Finalized data collection script, fetched US EPA data (Completed)
* Week 2 (June 2 – June 8): Literature search, EDA, PDM plan.
* Week 3 (June 9 – June 15): Preprocess data (merge CSVs, compute composite AQI, handle missing data), commit to GitHub.
* Week 4 (June 16 – June 22): Implement ARIMA-LSTM model, train on 2014–2023, test on 2024.
* Week 5 (June 23 – June 29): Evaluate ARIMA-LSTM (RMSE, MAE), implement CatBoost as a backup model.
* Week 6 (June 30 – July 6): Perform SHAP analysis to interpret key AQI drivers.
* Week 7 (July 7 – July 13): Draft final report, create visualizations (e.g., forecast plots, SHAP plots).
* Week 8 (July 14 – July 20): Prepare presentation, finalize report, submit project.