Progress Report: Air Quality Dataset Analysis

**1. Dataset Description**

The UCI Air Quality dataset captures hourly air pollution and weather data from a monitoring station in an Italian city. It includes measurements of pollutants like NO2, CO, and C6H6, as well as environmental factors such as temperature and humidity. Spanning a year (March 2004 - April 2005), the dataset features 9,358 records across 15 variables with some missing data (-200) and sensor-specific challenges like drift and cross-sensitivity.

**2. Exploratory Data Analysis (EDA)**

**Achieved EDA:**

1. Data Cleaning - Follow standard procedures for handling each variable. This includes printing the total missing values for each column and generating a frequency table to show the distribution of missing data across rows.
2. Matrix Pairwise Correlation between variables and Heatmap visualization
3. Provide an overview of the statistical summary of the dataset, including measures like mean, median, standard deviation, and percentiles.

**Planned EDA:**

Felix: Hybrid Modeling: Combined SARIMAX model for time series forecasting by exogenous covariances (CO,...) with outputs from SAPRC BioChemistry simulation model to predict NO2.

Duc: Neural nets (Multi layer perceptrons) for pollutant concentration prediction based on sensor data and environmental conditions

Khang: Pairwise Correlation(Between sensor responses(PT), true concentrations(GT) and environmental variables), Cross-Sensitivity(how sensors respond to non-target gasses), Hypothesis Testing, AQI Calculation(from CO, NOX and SAPRC output- PM10)

**3. Tentative Analysis Questions**

1. Time Series Model + Knowledge Domain Model

Is the NO₂ series stationary? Which predictors (temperature, humidity, NOx, CO, C6H6) and SAPRC-derived chemical reaction outputs correlate strongly with NO₂? How do NO₂ levels change across seasons? Is there a seasonal autocorrelation pattern (e.g., weekly peaks)? How well does a hybrid model (SARIMAX + SAPRC simulation outputs) predicting NO₂ levels compare to a SARIMAX only approach?

Duc: does a deep learning architecture perform better than traditional machine learning techniques for this dataset?

Khang: Sensor Calibration

How do sensor responses correlate with true pollutant concentrations, and how strong are the associations between environmental variables and these concentrations?

Is there a bias between sensor(PT) and true measurements(GT)?

Do sensor residuals depend on environmental factors?

How does the AQI vary over time, what are the predominant pollutants contributing to air quality?

**4. Planned Methods**

Felix:

1. Data Preprocessing & Stationarity Transformation for ARIMA model.

2. Basic SARIMAX Model Development: SARIMA(p, d, q)(P, D, Q, s) with Exogenous Predictors

4. Integration of SAPRC BioChem Simulation follows Y\_t ​= SARIMA(p,d,q) + β\*X\_t ​+ ε\_t​

5. Use Hybrid SARIMAX model fitting for trend analysis and forecasting future pollution levels.

Duc: Predicting target variables: CO(GT), C6H6(GT), NOx(GT), NO2(GT)

- Involve normalizing data and PCA to reduce dimension before training

- Design several simple neural nets with 4 output nodes, choose the best one out of them with MSE as loss function.

Khang:

- Pairwise Correlation: Pearson/Spearman correlations and visualize using a heat map.

- Cross-Sensitivity: Calculate correlations between sensor responses (PT) and non-target true concentrations(GT)

- Perform paired t-tests to compare sensor readings with reference measurements.

- Use EPA breakpoints to calculate AQI for CO, NO2, and O3, visualize trends and categorize health risk levels.