Project Proposal

Course: *CMSC 6950 -- Fall 2023*

Name: *Mohammadmahdi Mirmojarabian* (Student #: 202292549)

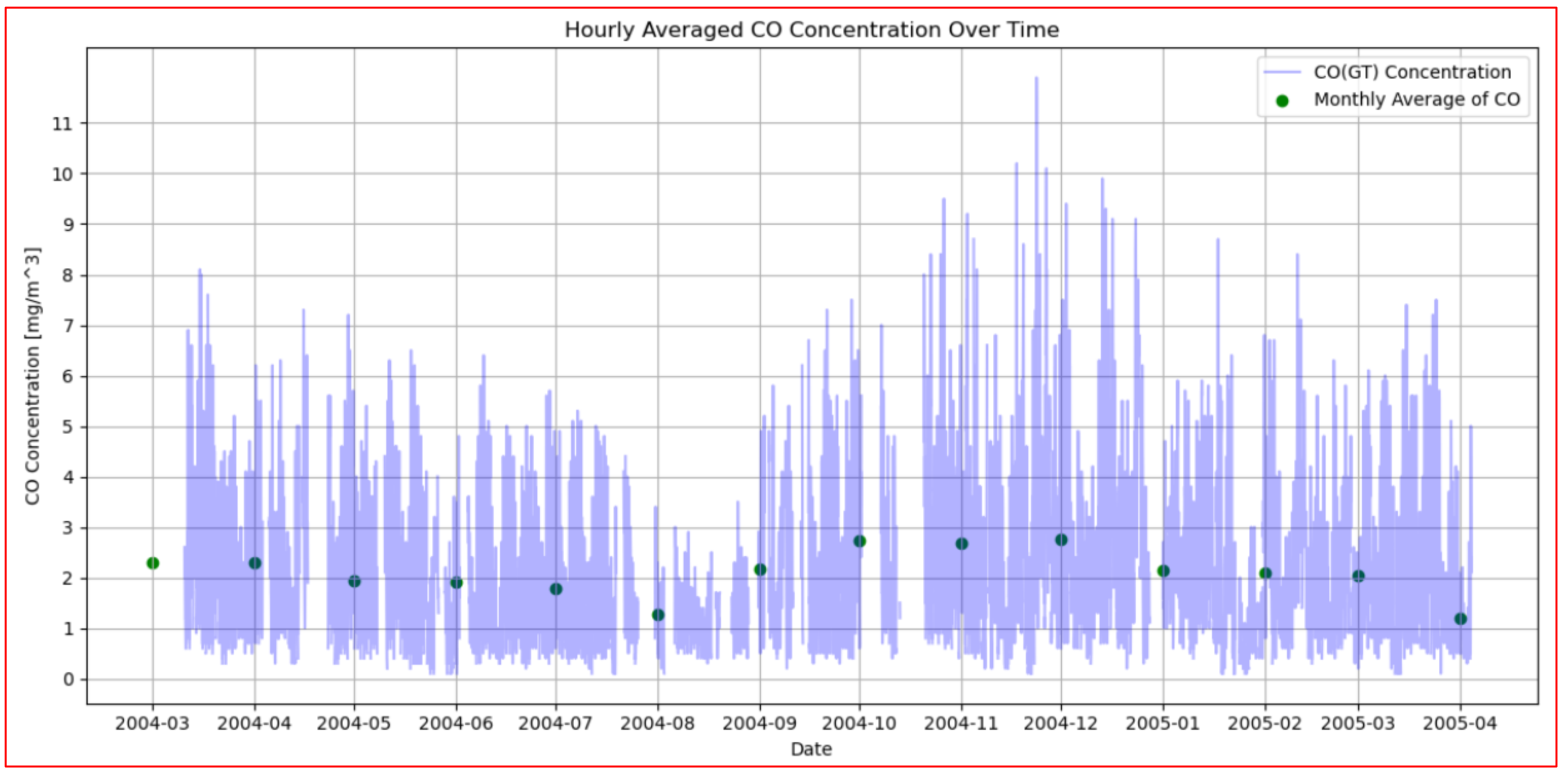
Overview of Dataset

For my project, I intend to use an Air Quality dataset from https://archive.ics.uci.edu/dataset/360/air+quality . These data are the output responses of a gas multisensor device deployed on a field in an Italian city to measure the air quality over time. The dataset contains 9357 rows of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. Data was recorded for a period of one year (Early March 2004 – Early April 2005). Also, a co-located reference certified analyzer provided Ground Truth hourly averaged concentrations for CO, Non-Methanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2). Missing values are labeled with "-200".

Attributes of Dataset:

➔ Date: (DD/MM/YYYY) ➔ Time: (HH.MM.SS) ➔ CO(GT): True hourly averaged CO concentration [mg/m^3] (reference analyzer) ➔ PT08.S1(CO): PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted) ➔ NMHC(GT): Non Metanic HydroCarbons concentration [µg/m^3] (reference analyzer) ➔ C6H6(GT): True hourly averaged Benzene concentration [µg/m^3] (reference analyzer) ➔ PT08.S2(NMHC): PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted) ➔ NOx (GT): True hourly averaged NOx concentration [ppb] (reference analyzer) ➔ PT08.S3(NOx): PT08.S3 (tungsten oxide) hourly averaged sensor response ➔ NO2(GT): True hourly averaged NO2 concentration [µg/m^3] (reference analyzer) ➔ PT08.S4(NO2): PT08.S4 (tungsten oxide) hourly averaged sensor response ➔ PT08.S5(O3): PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted) ➔ T: Temperature [°C] ➔ RH: Relative Humidity (%) ➔ AH: Absolute Humidity

First Plot of The Data:



In this initial plot, we observe the year-long distribution of carbon monoxide (CO) concentrations, spanning from March 10, 2004, to April 4, 2005. Additionally, the graph presents calculated monthly averages for CO concentration, represented as green data points on the first day of each month. This enables us to compare the average CO concentration across the entire year swiftly. Upon closer examination, we note that CO levels remain relatively stable at around 2 mg/m³ during the summer months. However, in August, there is a noticeable decline, followed by a gradual increase in September, reaching a peak in October, November, and December. Subsequently, CO concentrations revert to the typical average of 2 mg/m³ in January.

Planned Next Steps:

The statistics and analyses I may want to perform on this dataset: **1**- Calculate basic descriptive statistics for each attribute, such as mean, median, mode, standard deviation, minimum, maximum, and quartiles. This provides a general overview of the dataset. **2**- Explore how air quality parameters (e.g., CO, NO2, Benzene) change over time. I can create time series plots to visualize trends and seasonality. I can analyze the dataset to identify seasonal variations in air quality and determine whether air quality worsens or improves during specific months or seasons. **3**- Do data cleaning. Identify the extent of missing data (indicated by "-200"). Calculate the percentage of missing values for each attribute. Decide on a strategy to handle missing data, whether through imputation or exclusion (filling NaNs using a centrality measurement/Interpolation/the decided Regressor + Data Scalers). **4**- Plot the correlation matrix to examine the relationships between different attributes. For instance, I can check how CO levels correlate with temperature or humidity. High correlations may indicate dependencies between variables. **5**- Identify and analyze extreme values or outliers in the data. I can use box plots, scatter plots, or statistical tests (calculating z-score and IQR for each attribute) to detect and understand outliers. **6**- Examine the distribution of each attribute using histograms, probability density plots, and box plots. This helps us understand the data's underlying distribution, which can be necessary for modeling and hypothesis testing. **7**- Investigate whether air quality parameters exhibit diurnal patterns. For instance, I can compare CO levels during different hours of the day to see if there are consistent patterns. **8**- Compare the measurements from the multisensor device with the reference analyzer data (CO, NOx, NO2, etc.). Calculate the mean absolute error or other error metrics to assess the accuracy of the sensors. **9**- Build regression models to predict one air quality parameter based on others. For example, I can predict CO concentrations based on temperature, humidity, and NO2 levels. **10**- Formulate and test hypotheses about the factors affecting air quality. For instance, I can test if there is a statistically significant difference in CO levels between weekdays and weekends. **11**- Utilize machine learning algorithms for more advanced predictive modeling. This could include random forests, support vector machines, or neural networks for forecasting air quality parameters. **12**- Create various types of visualizations, including scatter plots, line plots, bar plots, and heatmaps, to make the data more understandable and facilitate communication of findings.