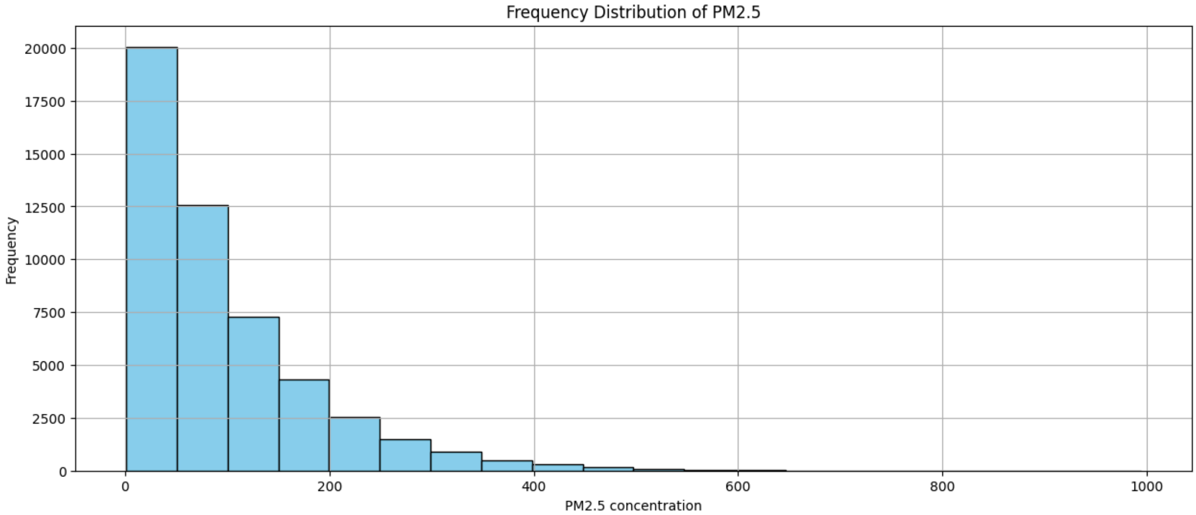
Cind 820: Big Data Analytics Initial Results

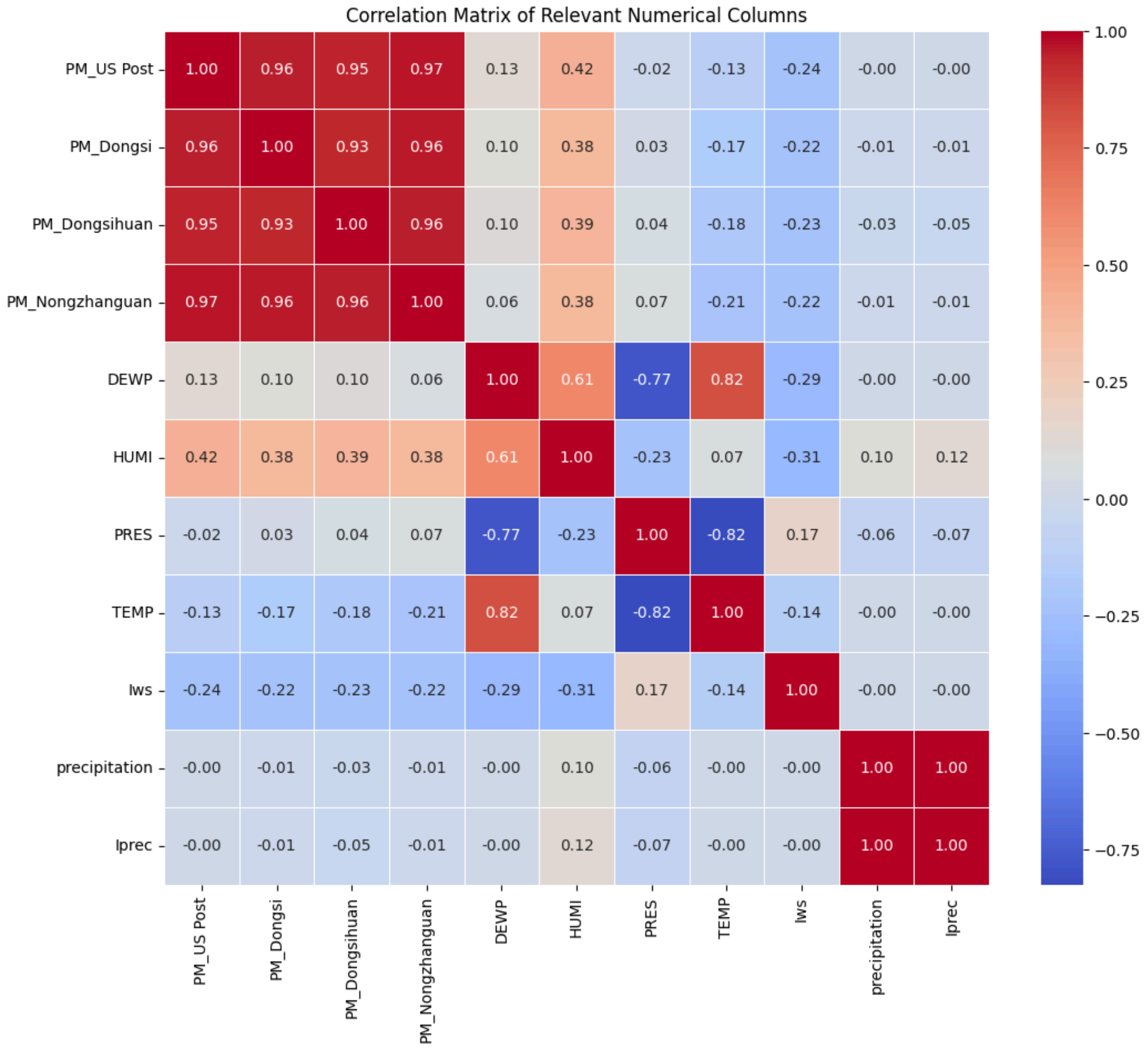
Adam Bruno
Student # 500904101
Dr. Tamer Abdou
Dr.Ceni Babaoglu

November 10, 2025

Data set Analysis:

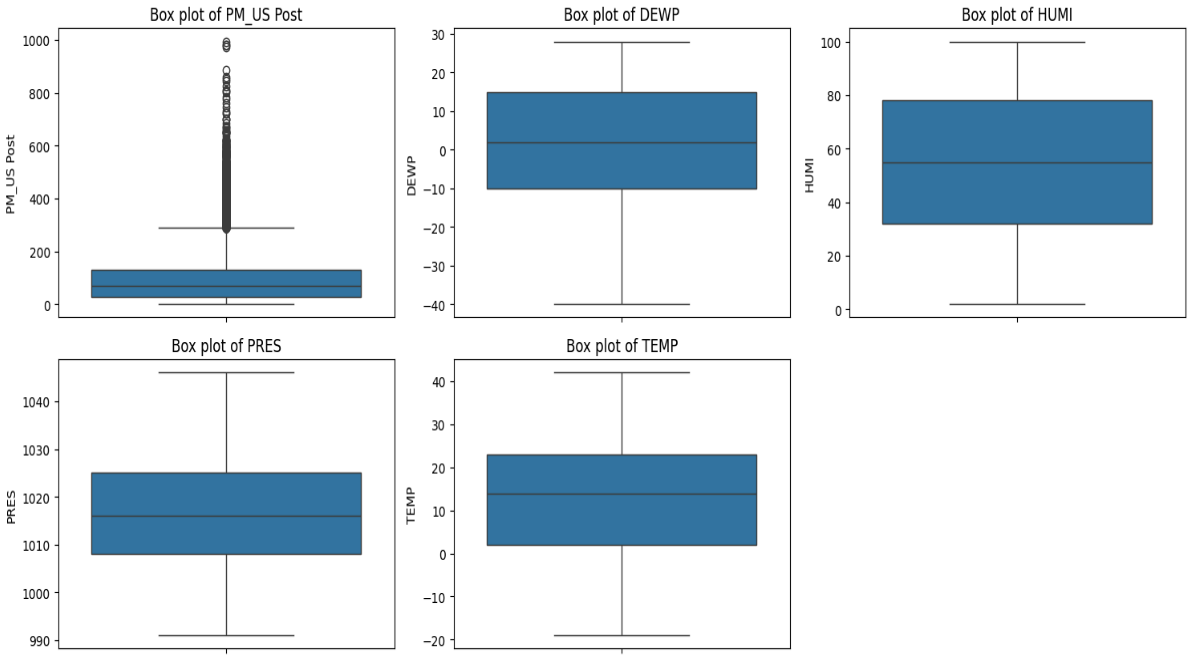
This project phase aims to breakdown the dataset, highlighting distinct features, patterns and trends. A summary of key dataset attributes will be examined to establish a comprehensive outline of the datasets scope. Deeper understanding of datasets' characteristics prove vital in performing accurate preprocessing techniques as well as building reliable models. Visualizations will be employed to determine possible patterns, trends and outliers. Ultimately, the initial insight into the dataset will highlight potential inferences. The key dataset feature that will be primarily focused on within this project is PM2.5 concentration. This denotes the fine particle matter reading from neighbouring station sensors. Additionally, the dataset features various meteorological features as well as time attributes. This benefits the project's ability to infer potential air quality degradation connection with changes of season or time. Visualizations provided within the literature review indicate the frequency of PM2.5 within the dataset. Graphing the frequency distribution ultimately illustrates that the dataset is heavily skewed to the right. This indicates that a class imbalance exists within the dataset and as result may impact performance. Ultimately, the graph's findings establish potential implications for models favouring lover PM classes. Alternatively, this indicates misleading accuracy metrics as well as poor majority class performance. Additionally, the correlation matrix showed notable relations between PM measurements and meteorological factors. Positive correlation favoured particle matter with dew point and humidity readings representing approximately 0.68 and 0.19 respectively. Further modeling will prove to reaffirm the dataset analysis and the initial results found.





Data set Preparation

Data preparation techniques are vital to creating a clean dataset for machine learning models to interpret. Proper data cleaning and scaling practices can have a significant impact on supervised learning model results. The project examines a dataset with readings dating from January 1st 2010 to December 31st, 2015. Within this time frame missing values or faulty reading are properly identified. Absent sensor readings are present and denoted as NaN signifying data cleaning requirements. The dataset contains 52,584 entries as well as 18 columns. From these columns significant missing values are present within various stations potentially demonstrating poor sensor maintenance. This is evident within the project's exploratory data analysis which found that 60% of readings with Dongsihuan station contained missing PM values within the dataset. Initial data cleaning results may deem that these stations be cut from the analysis. As such to remove, the missing values filling techniques were implemented. Specifically, utilising forward fill removed most missing values from the dataset. However, 23 missing values still remained from the PM_US Post station readings indicating the necessity for additional backfilling technique utilization. Additionally, outliers were identified to be present within numerous columns. Although the values were kept as to not bias the data. The initial results will mainly focus on utilization of random forest models to train the test set which offer less sensitivity to outliers. Additionally air quality outliers may be potential readings of severe weather conditions which could heavily alter finding. Use of outlier techniques will have to be observed if impacted further deliverables.



Model Evaluation section

The project aims to utilize classification models to further understand pollutant levels and their adverse effects. As such the initial findings report focuses on using machine learning techniques to determine a model that can be utilized to determine a suitable approach. This report demonstrates the appropriate technique of utilising random forest classifiers on the target feature to train and test machine learning to evaluate pollutant detention across distinct seasons. The features for the distinct model include various meteorological characteristics such as humidity, temperature, etc. Combined wind speed was converted into a numerical format from a categorical format for model training purposes. This was in direct response to receiving error exceptions from the model and after further research it was discovered that the change was necessary. The training and testing split was split into a ratio of 80/20 to allow the model to be trained on a significant portion of the data.

This provides improved evaluation on unseen data, ultimately portraying a more accurate performance estimate. Specifically, the trained model achieved an overall accuracy score of 86% on the testing set. Similarly, the model performed generally well with precision and f1 metrics throughout the four seasons. However, the model struggled to determine high recall scores.This is shown in their 0.79 and 0.78 score in seasons spring and fall comparatively boasting high recall scores in summer and winter at 0.93. Overall the model performed well and was able to indicate varying performance across different seasons. Ultimately this demonstrates that machine learning can be utilised to determine potential growing trends of air pollution and meteorological features across seasons. Further analysis and study is required to develop other machine learning models as well as incorporate potential cross validation. The initial result report has concluded that machine learning principles can be employed to classify categorical seasons within a dataset with metrological and pollutant features.

Github Repository Link:
albruno-droid/CIND820: Data Analytics Air Quality Project

▶ **Initial Result and Code**