Student IDs **23208466**
24222286

ASSIGNMENT TWO

PAPER NAME: Data Mining and Machine

Learning PAPER CODE: COMP809

TOTAL MARKS: 100

Students' Names: Vedant Marwadi, Xeniya Obolonkova

Students' IDs: 23208466, 24222286

- Due date: 09 Jun 2024 midnight NZ time.
- Late penalty: maximum late submission time is 24 hours after the due date. In this case, a **5% late penalty** will be applied.
- Submit the actual code (no screenshot) separately with appropriate comments for each task.

Note: This assignment should be complemented by a group of two students and both students MUST contribute in each part.
Submission: a soft copy needs to be submitted through the canvas assessment link.

INSTRUCTIONS:

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your submission on Canvas immediately**
3. **Attach your code for all the datasets in.**

Assignment 2 - COMP809

Mitigating Health Risks Through PM2.5 Prediction: Evaluating MLP and LSTM Approaches

Vedant Marwadi[1], Xeniya Obolonkova[2]

[1]ID: 23208466
[2]ID: 24222286

Contents

List of Figures

List of Tables

1. Introduction

Air pollution poses a significant threat to public health, with particulate matter being one of the most dangerous pollutants. PM, also known as particle pollution, consists of a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large enough to be seen with the naked eye, while others are so small they can only be detected using an electron microscope.

Particle pollution includes:
PM10: Inhalable particles with diameters generally 10 micrometers and smaller.
PM2.5: Fine inhalable particles with diameters generally 2.5 micrometers and smaller. For comparison, the average human hair is about 70 micrometers in diameter, making the largest fine particle 30 times smaller.

Fine inhalable particles pose the greatest risk to health. Research suggests a strong correlation between PM2.5 exposure and cardiovascular disease. To mitigate this health risk, accurate predictions of PM2.5 concentrations are essential.

This project addresses the challenge of PM2.5 prediction by developing and comparing two different machine learning models: a multi layer perceptron and a long short term memory network. The models are trained and tested on a dataset obtained from the Environmental Auckland Data Portal. The dataset encompasses hourly measurements of PM2.5, PM10, co-air pollutants (SO2, NO, NO2), meteorological data (Solar Radiation, Air Temperature, Relative Humidity, Wind Direction, Wind Speed), and Air Quality Index (AQI) for a period of five years (January 2019 - December 2023), collected from a single air quality monitoring station (Penrose Station).

The report discusses the data preprocessing steps, including the handling of missing data and outliers. Following this, data exploration and feature selection are performed to identify the most relevant predictors for PM2.5 concentrations. Subsequently, the performance of MLP and LSTM models is evaluated.

Through this systematic approach, the best predictive model for PM2.5 concentrations is identified, thereby contributing to the broader effort of improving air quality management and public health protection.

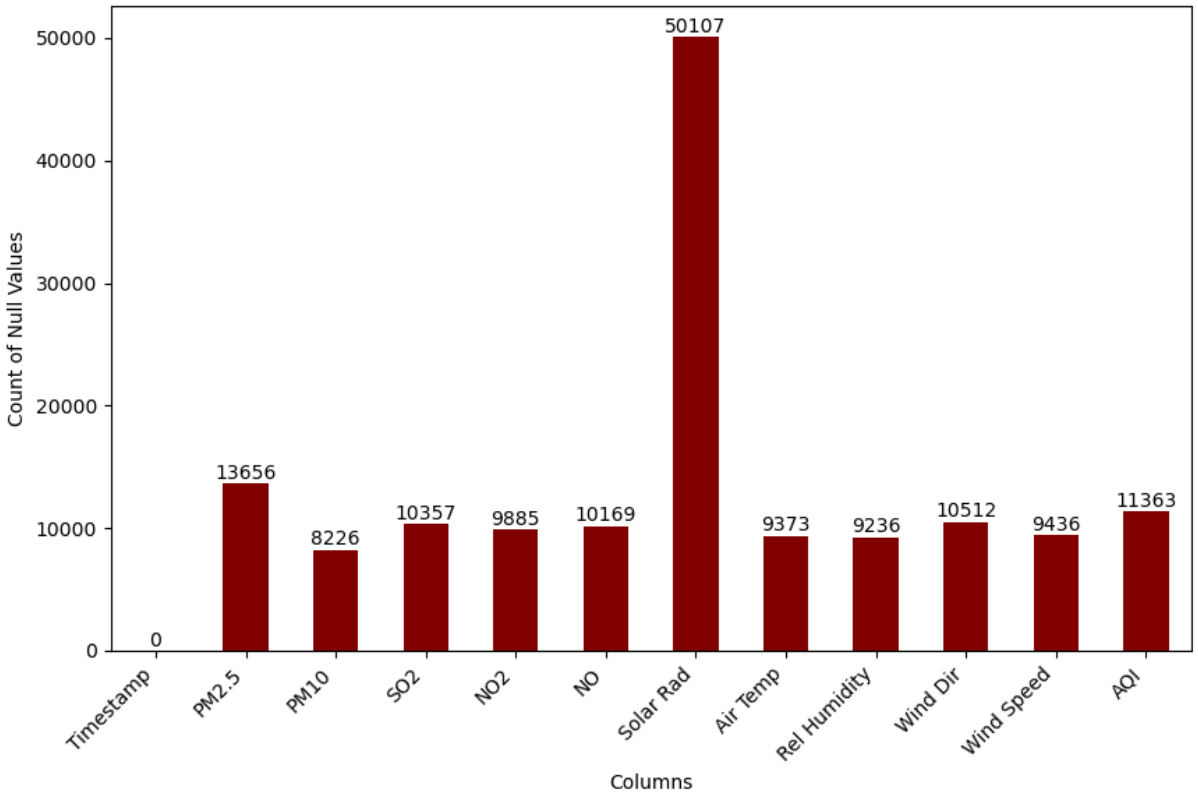
2. Data Pre-processing

2.1 Handling Missing Data

2.1.1 Identification of Null Values

Null values, or missing values, are data entries that are absent or undefined. Identifying and addressing null values is crucial before any data analysis because they can significantly impact the results and lead to inaccurate or misleading outcomes. Therefore, an initial check for null values is conducted, which reveals several columns with missing entries.

Figure 1: Null Values in Penrose Dataset

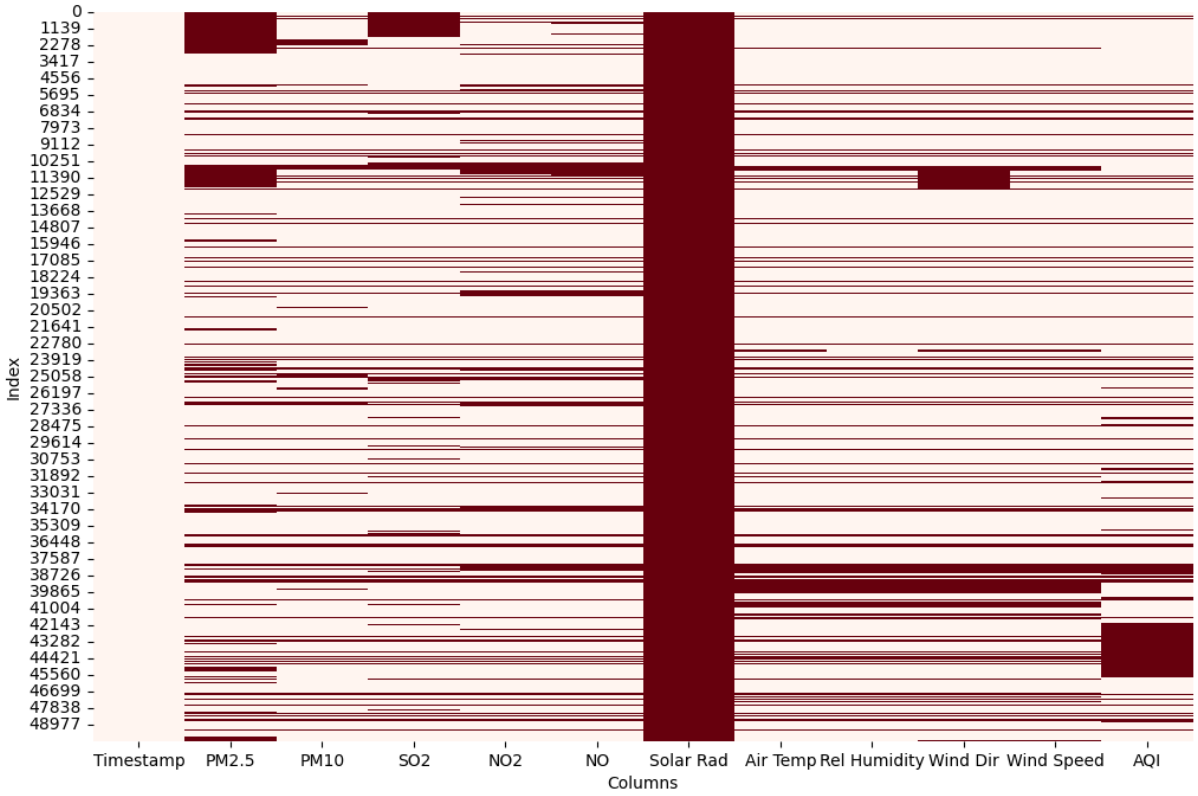


It is evident that there are no missing values for Timestamp. However, PM2.5 shows a significant number of missing values at 13,656, followed by PM10 with 8,226 missing values. SO2 is missing 10,357 values, while NO2 and NO have 9,885 and 10,169 missing values, respectively. Solar Radiation stands out with the highest number of missing values at 50,107. Additionally, Air Temperature has 9,373 missing values, Relative Humidity has 9,236, Wind Direction has 10,512, Wind Speed has 9,436, and AQI has 11,363 missing values.

2.1.2 Visualization of Missing Values

Additionally, a heatmap is generated to visualize the presence of missing values across the dataset.

Figure 2: Heatmap of Missing Values



Each row in the heatmap corresponds to an entry, and each column corresponds to a feature. Red regions indicate missing values, while white regions show non-missing values. The heatmap reveals that the entire "Solar Rad" column is red, signifying that every entry for solar radiation is missing. Consequently, this column is removed from the dataset due to its lack of usable data.

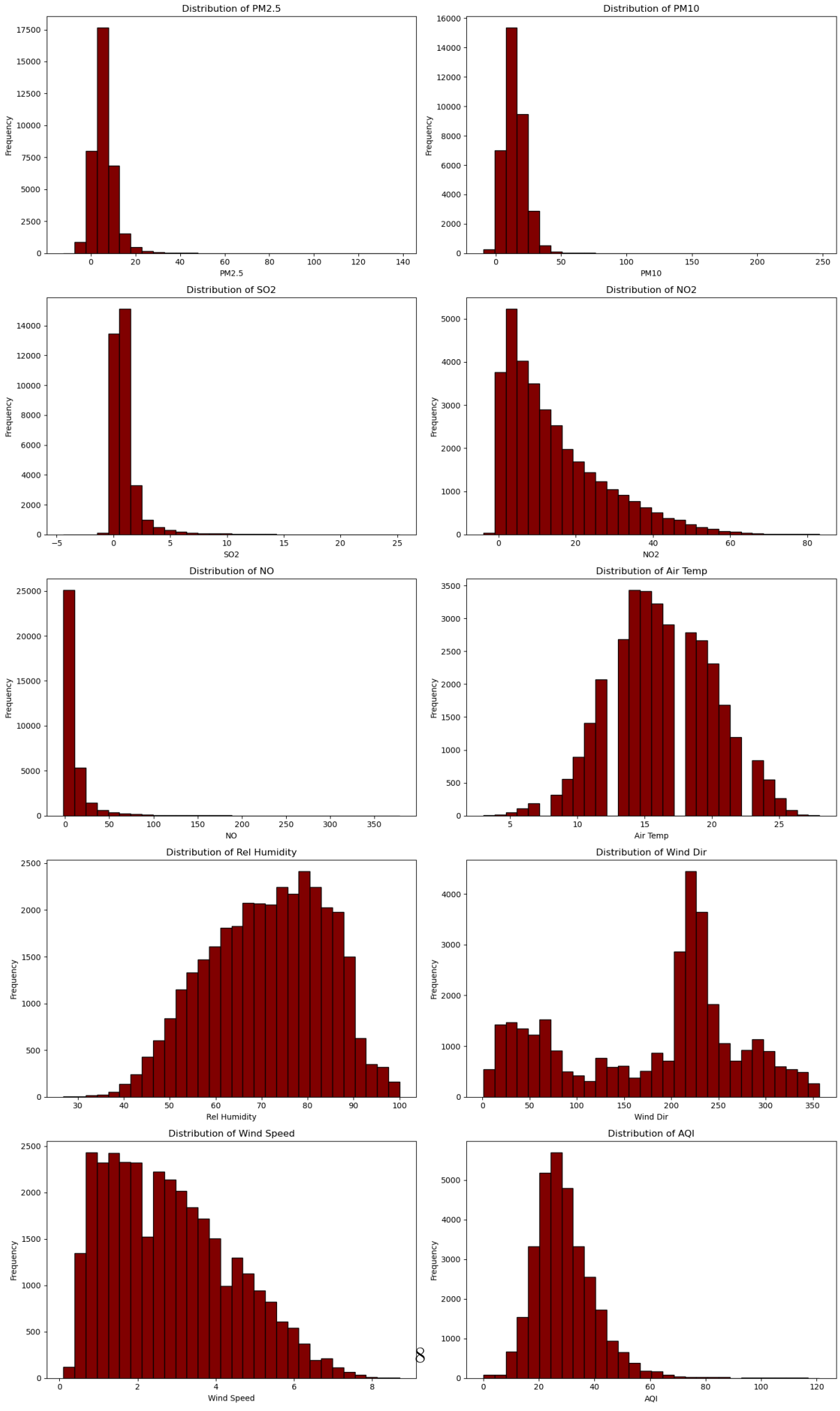
2.1.3 Addressing Missing Values in Dependent Variables

Addressing these gaps is crucial for maintaining the dataset's integrity and ensuring the accuracy of the subsequent analysis. While imputing missing values can be a solution, especially for independent variables, it can distort the true nature of the data when done with the dependent variable. Therefore, rows with null values for the dependent variables PM2.5 and PM10 are dropped to preserve the authenticity of the findings.

2.1.4 Handling Missing Data in Independent Variables

After addressing null values in the dependent variables, attention is turned to the independent variables. Understanding the distribution of these variables is essential before deciding on the imputation method.

Figure 3: Distribution of Independent Variables



Right Skewed Distributions: The distributions of PM2.5, PM10, SO2, NO2, NO, Wind Speed, and AQI are heavily right skewed. In skewed distributions, the mean is influenced by extreme values, leading to biased imputation. Therefore, the median, which is robust to outliers, is used to impute null values for these features.

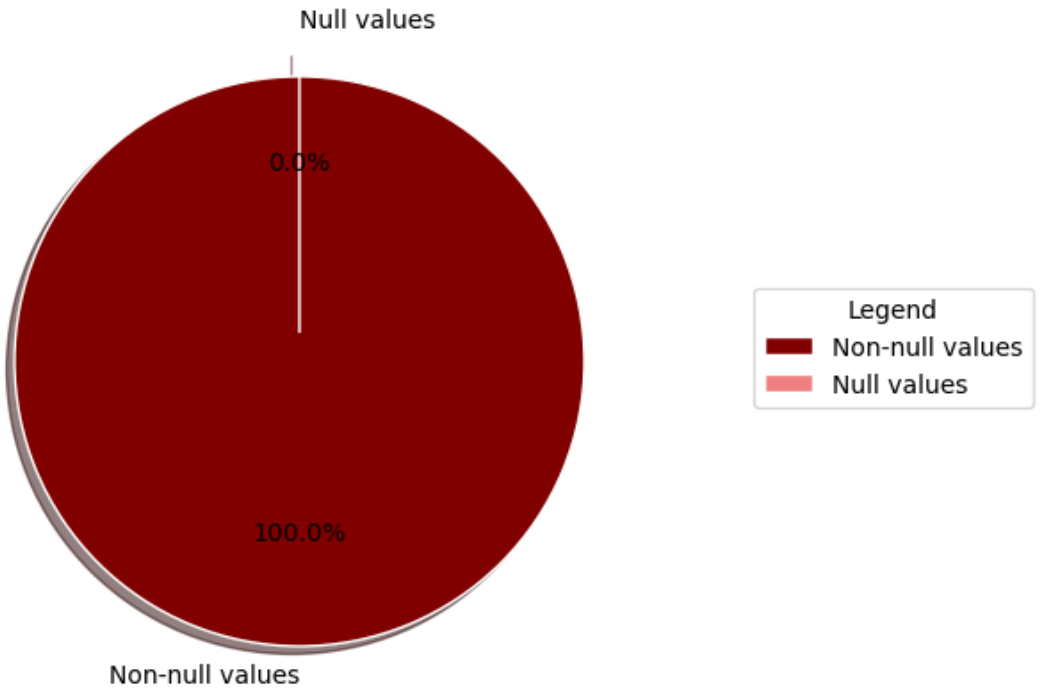
Approximately Normal Distributions: The distributions of Air Temp and Rel Humidity are approximately normal. In normally distributed data, the mean is a good measure of central tendency. Thus, mean imputation is used for these variables.

Multimodal Distribution: The distribution of Wind Dir is multimodal, indicating several common wind directions occurring more frequently. The mode, representing the most frequent category, is suitable for imputation in this context.

2.1.5 Verification of Imputed Data

Upon completing the imputation process, the dataset is re-examined to ensure no missing entries remained.

Figure 4: Null Values Handled



This confirmation step ensures that the dataset is complete and ready for further analysis.

2.2 Handling Outliers

2.2.1 Data Examination

Outliers are data points that deviate significantly from the majority of the data. They can appear at either end of the spectrum, being much higher or lower than the rest. It is

important to check for the outliers because they can skew the results of statistical tests, leading to misleading conclusions. Moreover, they can sometimes also indicate errors in data collection or measurement. Therefore, after removing the null values, the second step taken is to understand the structure of the dataset and identify any anomalies.

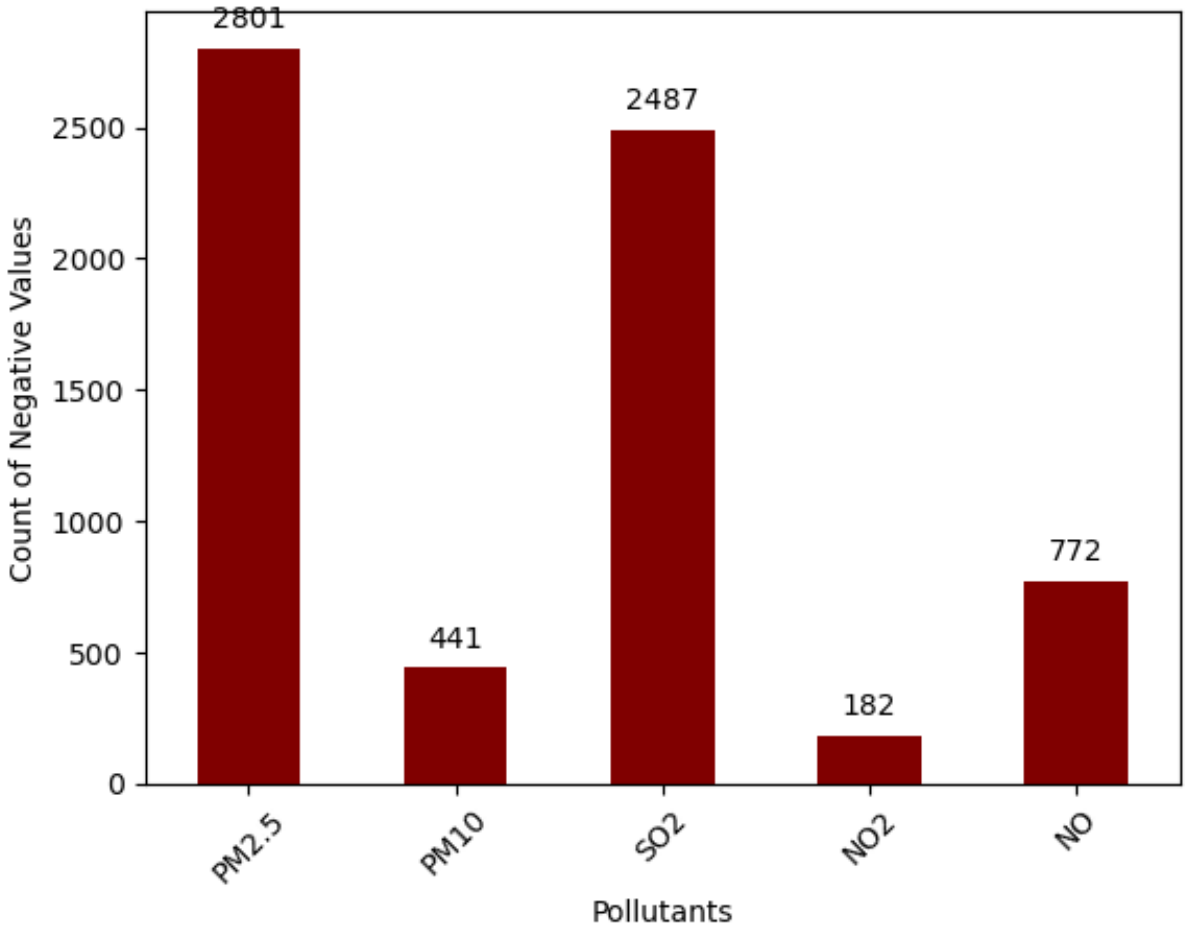
Table 1: Summary Statistics for the Penrose DataFrame

Statistic	PM2.5	PM10	SO2	NO2	NO	Air Temp	Rel Humidity	Wind Dir	Wind Speed	AQI
Count	35697	35697	35697	35697	35697	35697	35697	35697	35697	35697
Mean	6.23	14.59	1.04	14.46	9.75	16.18	71.23	182.36	2.82	29.34
Std	4.37	8.24	1.32	12.61	19.86	3.71	12.41	90.75	1.55	10.31
Min	0.00	0.00	0.00	0.00	0.00	3.00	26.90	1.00	0.10	0.00
25%	3.60	8.90	0.40	4.80	1.10	14.00	62.40	95.00	1.60	23.00
50%	5.20	13.50	0.70	10.80	3.70	16.00	71.23	217.00	2.60	28.00
75%	7.80	19.00	1.20	20.50	10.20	19.00	80.80	237.00	3.80	34.00
Max	138.40	247.50	25.20	83.10	378.40	28.00	100.00	357.00	8.70	121.00

2.2.2 Handling Negative Values

During the data examination, it is noted that PM2.5, PM10, SO2, NO2, and NO contain negative values.

Figure 5: Count of Negative Values



Negative values in these columns are not physically meaningful, as pollutant concentrations cannot be negative. These anomalies could be attributed to data entry errors or

sensor malfunctions.

To address these invalid negative values, it is essential first to understand their distribution. During the null value handling process, the distribution of these features is checked and found to be right-skewed. In a right-skewed distribution, the median is a more reliable measure of central tendency than the mean, as it is less affected by extreme values and outliers. Consequently, these negative values are replaced with the median value of their respective columns. This approach preserves the size of the dataset and avoids potential biases that mean replacement could introduce, especially in skewed distributions.

2.2.3 Identifying and Removing Outliers

Outliers can significantly affect the results of statistical analyses and machine learning models. The presence of extreme maximum values, which are much higher than the 75th percentile values, suggests the possibility of outliers in several columns. For instance, PM2.5 has a maximum value of 138.4, significantly higher than the 75th percentile value of 7.8.

The Interquartile Range (IQR) method is used to identify these outliers. This involves calculating the first quartile (Q1) and the third quartile (Q3) for each column, then determining the IQR as the difference between Q3 and Q1. Values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are flagged as potential outliers.

Table 2: Outlier Count and Bounds

Column	**Lower Bound**	**Upper Bound**	**Outlier Count**
PM2.5	-2.70	14.10	1765
PM10	-6.25	34.15	626
SO2	-0.80	2.40	2505
NO2	-18.75	44.05	1253
NO	-12.55	23.85	3161
Air Temp	6.50	26.50	201
Rel Humidity	34.80	108.40	25
Wind Dir	-118.00	450.00	0
Wind Speed	-1.70	7.10	143
AQI	6.50	50.50	1406

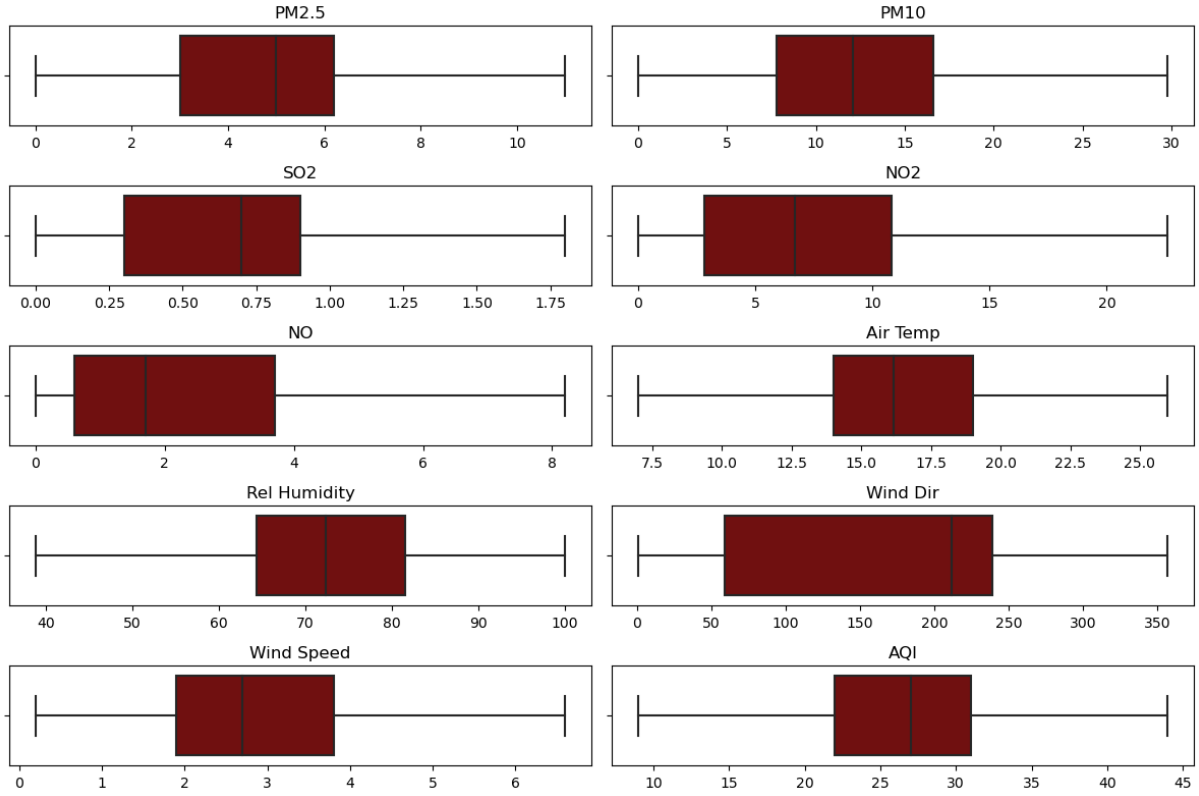
For instance, the lower and upper bounds for PM2.5 are -2.70 and 14.10, respectively, resulting in 1,765 outliers. Similarly, PM10 has bounds of -6.25 and 34.15, with 626 outliers, while SO2 has bounds of -0.80 and 2.40, resulting in 2,505 outliers. NO2 and NO show considerable outliers as well, with bounds of -18.75 and 44.05 for NO2 and -12.55 and 23.85 for NO, identifying 1,253 and 3,161 outliers, respectively.

The process of identifying and removing outliers is performed iteratively. Initially, the bounds are calculated, and data points falling outside these bounds are removed. This step is repeated until no further outliers are detected. This iterative process ensures that all extreme values are addressed, rather than just the most obvious ones, resulting in a more refined dataset.

2.2.4 Verification of Outliers Removal

After removing the outliers, box plots are generated for each column to visualize the spread and identify any remaining anomalies.

Figure 6: Outliers Handled



The box plots reveal that the data cleaning process effectively removes outliers and corrects negative values.

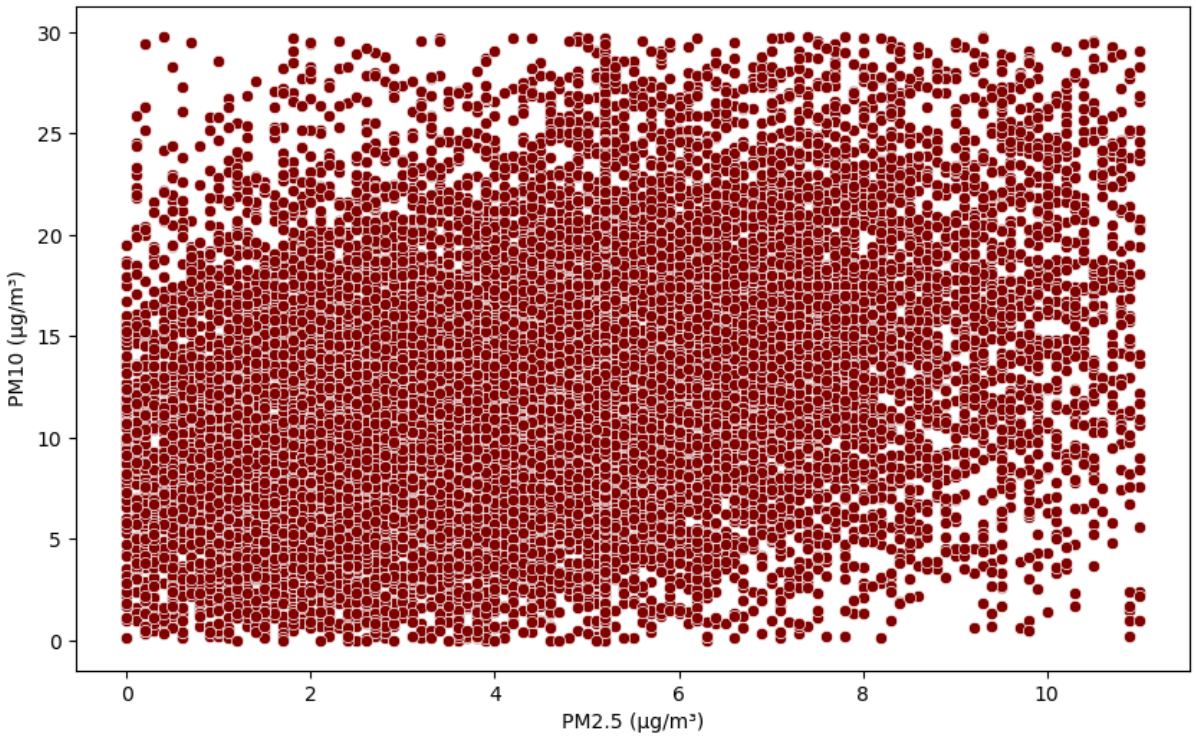
3. Data Exploration

3.1 Exploring the Relationship between PM2.5 and PM10

3.1.1 Scatter Plot of PM2.5 vs PM10

A scatter plot is created to visualize the relationship between PM2.5 and PM10 concentrations. This type of plot is particularly useful for identifying the correlation and pattern between two continuous variables.

Figure 7: Scatter plot of PM2.5 vs PM10



The scatter plot reveals a dense distribution of data points representing various levels of PM2.5 and PM10 concentrations. At first glance, there seems to be no strong linear relationship between PM2.5 and PM10, as the data points are scattered across a broad range without a clear trend either upwards or downwards.

3.1.2 Correlation Calculation

The Pearson correlation coefficient between PM2.5 and PM10 is then calculated to quantify the degree of their linear relationship. The resulting coefficient of 0.335 indicates a weak positive relationship between PM2.5 and PM10 concentrations.

3.1.3 Dropping PM10

Following the correlation analysis, the PM10 column is dropped from the dataset. This decision is made because PM2.5 has greater health impacts.

3.2 Exploring AQI

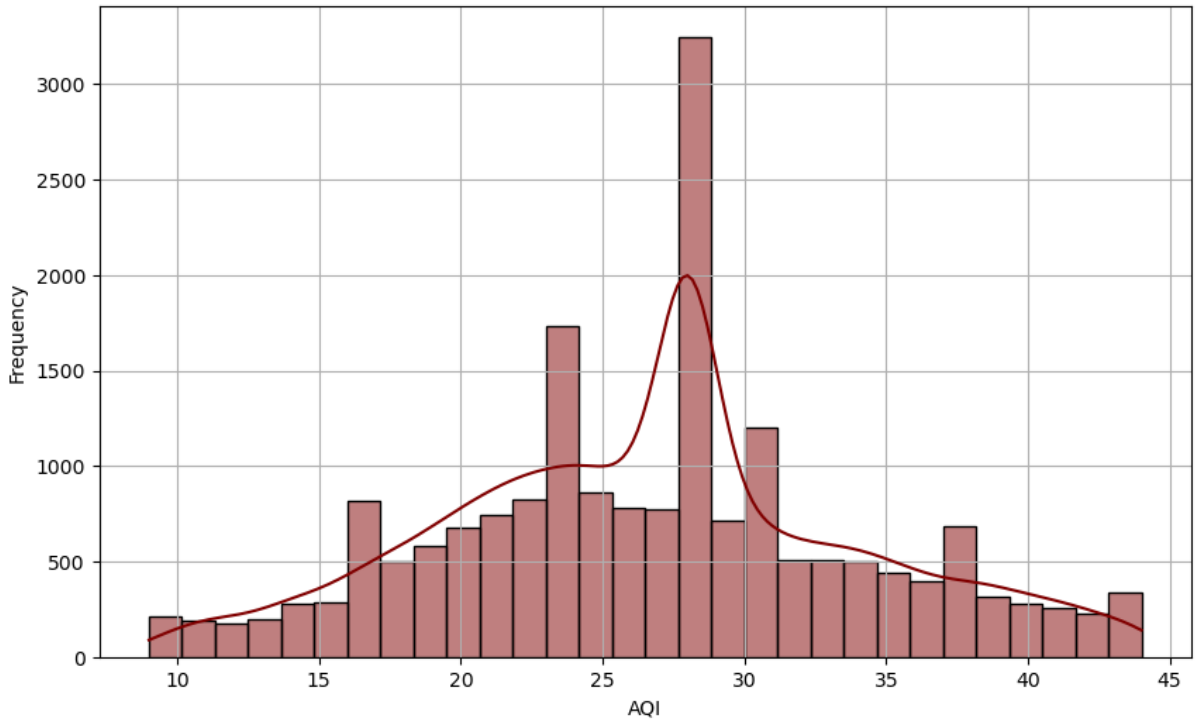
3.2.1 Timestamp Parsing

The "Timestamp" column is converted to a datetime format, and new columns for year, month, day, and hour are created. This preprocessing step is crucial for temporal analysis.

3.2.2 Distribution of AQI

A histogram with a kernel density estimate is plotted to visualize the distribution of AQI values. This approach helps in understanding the common range of air quality experienced in the area.

Figure 8: Distribution of AQI

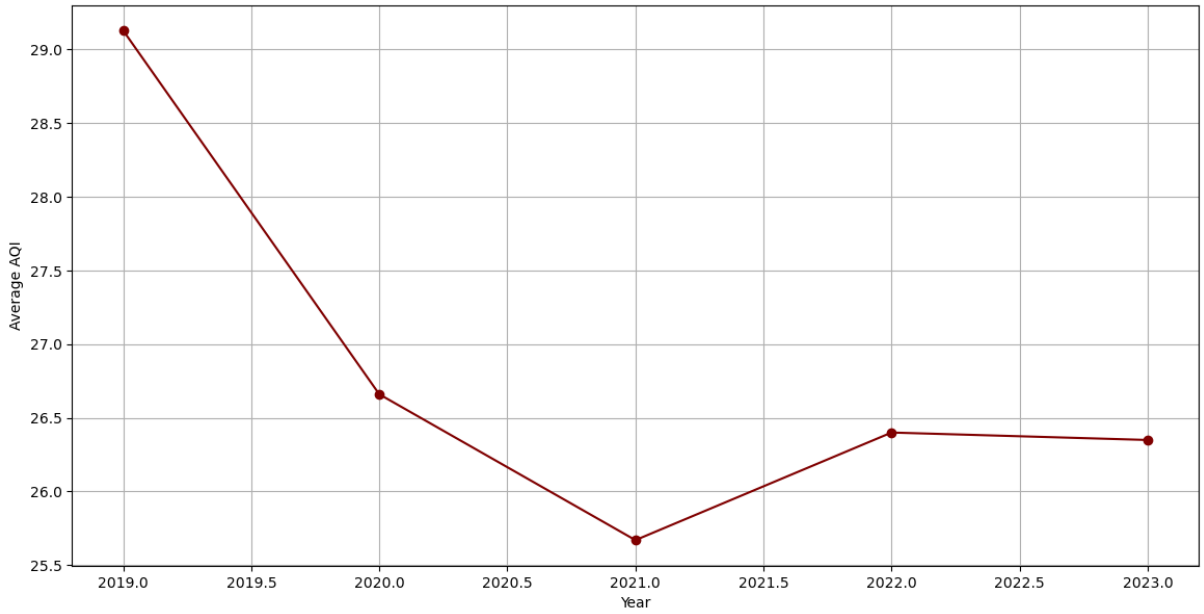


The concentration of AQI values around 28 suggests that, on average, the air quality tends to be moderate according to common AQI standards. The tails on both ends indicate occasional days with "Good" air quality (AQI less than 20) and days with "Unhealthy for Sensitive Groups" or worse (AQI greater than 40). Additionally, the secondary peaks at different AQI values might show seasonal or periodic fluctuations in air quality.

3.2.3 Yearly AQI Trend

The average AQI for each year is calculated and plotted to observe long-term trends. This analysis helps identify whether air quality has improved or deteriorated over the years.

Figure 9: Average AQI Trend Over Years



Yearly Breakdown:

2019: The highest average AQI is close to 29.5, indicating poorer air quality.

2020: There is a significant drop in average AQI to around 27.5.

2021: The average AQI reaches its lowest point at about 27, suggesting the best air quality within the observed period.

2022: There is a slight increase in average AQI to around 27.5.

2023: Another slight decline in average AQI bring it down close to 27 again.

Implications:

The overall downward trend from 2019 to 2021 suggests an improvement in air quality during these years. This improvement could be attributed to various factors, such as reduced industrial activity or changes in human behavior, possibly due to the COVID-19 pandemic.

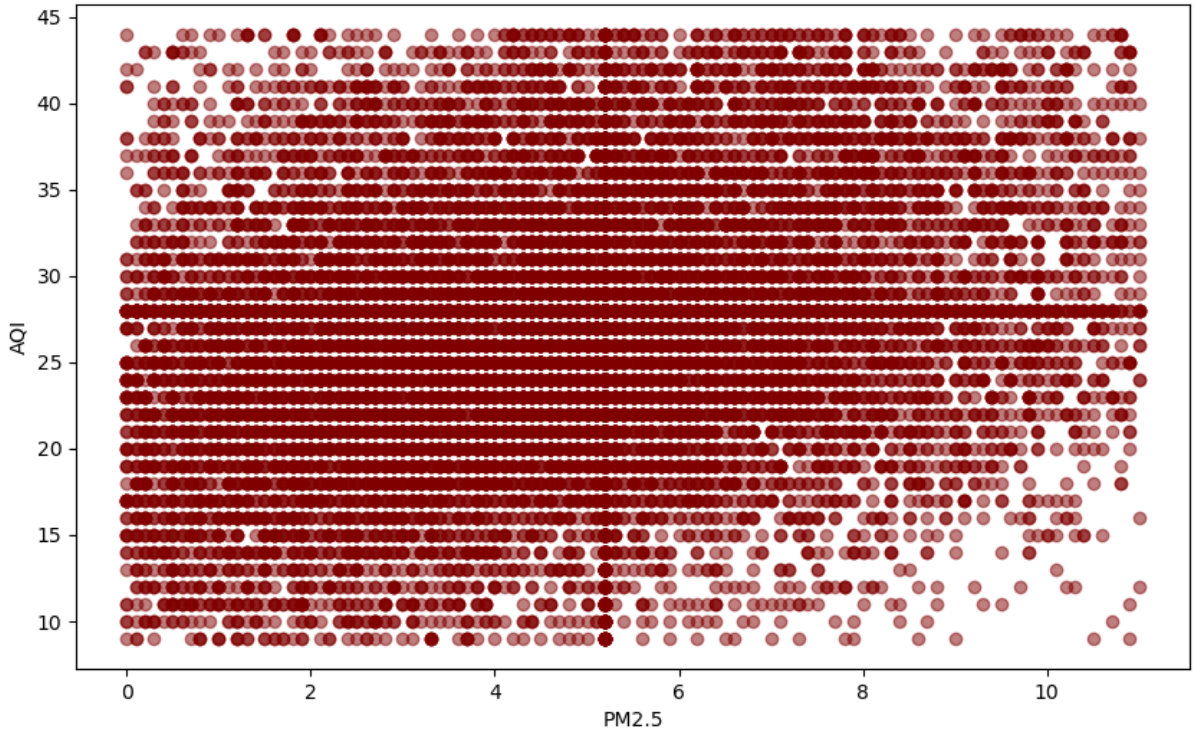
3.2.4 Correlation with AQI

The correlation matrix is calculated to identify which variables have the highest correlation with AQI, and it is found that the variable that affects AQI the most is PM2.5 with a correlation coefficient of 0.223.

3.2.5 Scatter Plot of AQI vs PM2.5

A scatter plot is then created to visualize the relationship between AQI and PM2.5. This plot is essential for understanding how changes in PM2.5 levels influence the AQI.

Figure 10: Scatter plot of AQI vs PM2.5



The scatter plot reveals dense horizontal bands of AQI values, indicating that AQI tends to cluster around certain levels regardless of PM2.5 concentration. This could be due to the AQI calculation formula or other influencing factors. Furthermore, the spread of data points along the x-axis shows that PM2.5 levels vary widely, however even at lower PM2.5 levels, AQI can be quite high, suggesting other pollutants or factors are also influencing AQI.

3.2.6 Dropping AQI

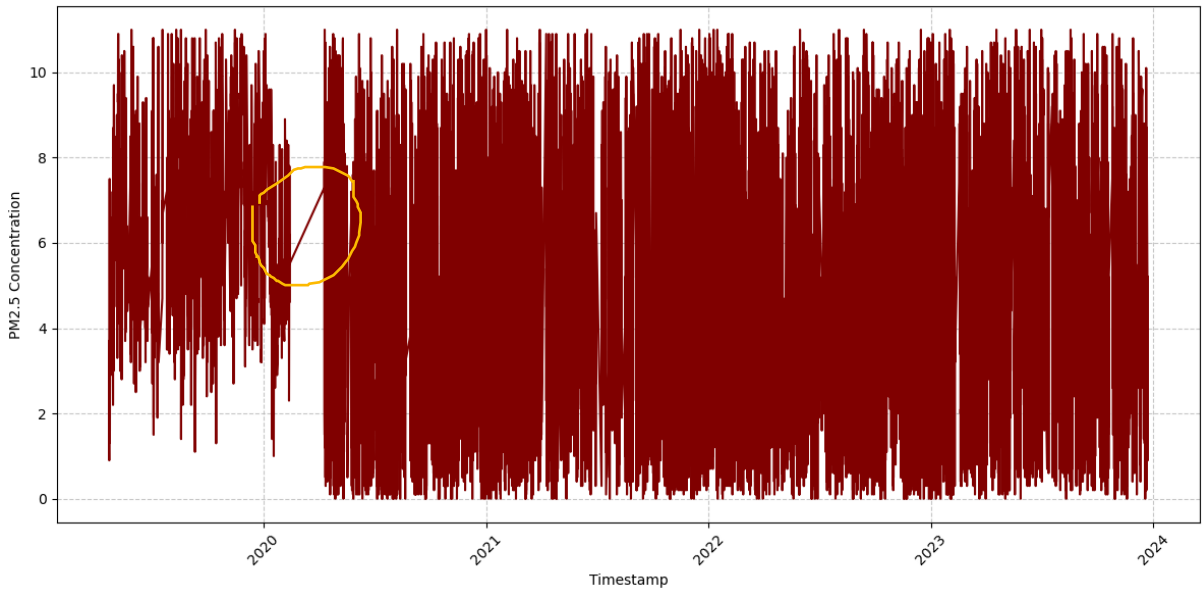
After the analysis of AQI, the AQI column is dropped from the dataset to focus on particulate matter trends.

3.3 Exploring Particulate Matter Trends

3.3.1 Daily PM2.5 Concentration Over Time

A time series plot of daily PM2.5 concentrations is created to identify daily variations and trends.

Figure 11: Daily PM2.5 Concentration Over Time

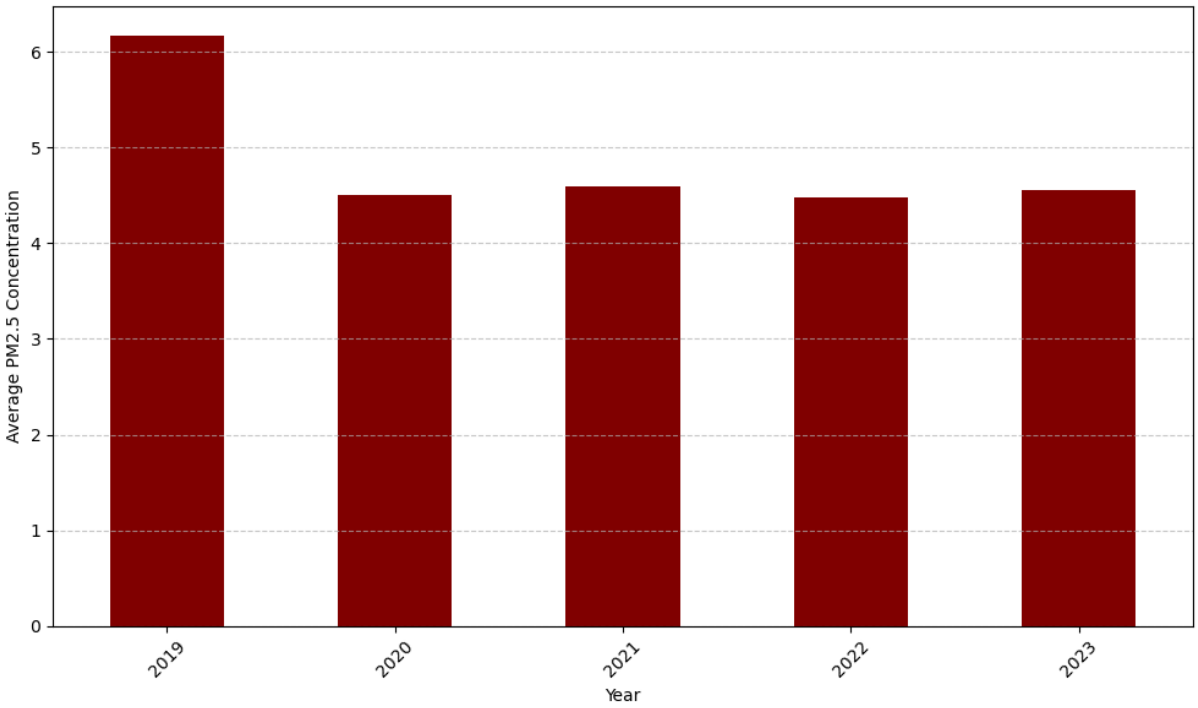


The plot highlights significant fluctuations in PM2.5 levels. However, despite these fluctuations, PM2.5 concentrations generally remain within a consistent range (0 to around 12 units), indicating stable overall air quality without extreme outliers. Additionally, a data gap can be observed around the transition between 2020 and 2021 which might be due to equipment failure, maintenance, or reporting issues.

3.3.2 Average PM2.5 Concentration Per Year

A bar plot showing the average PM2.5 concentration per year is created to provide a clearer picture of long-term trends in PM2.5 levels.

Figure 12: Average PM2.5 Concentration Per Year

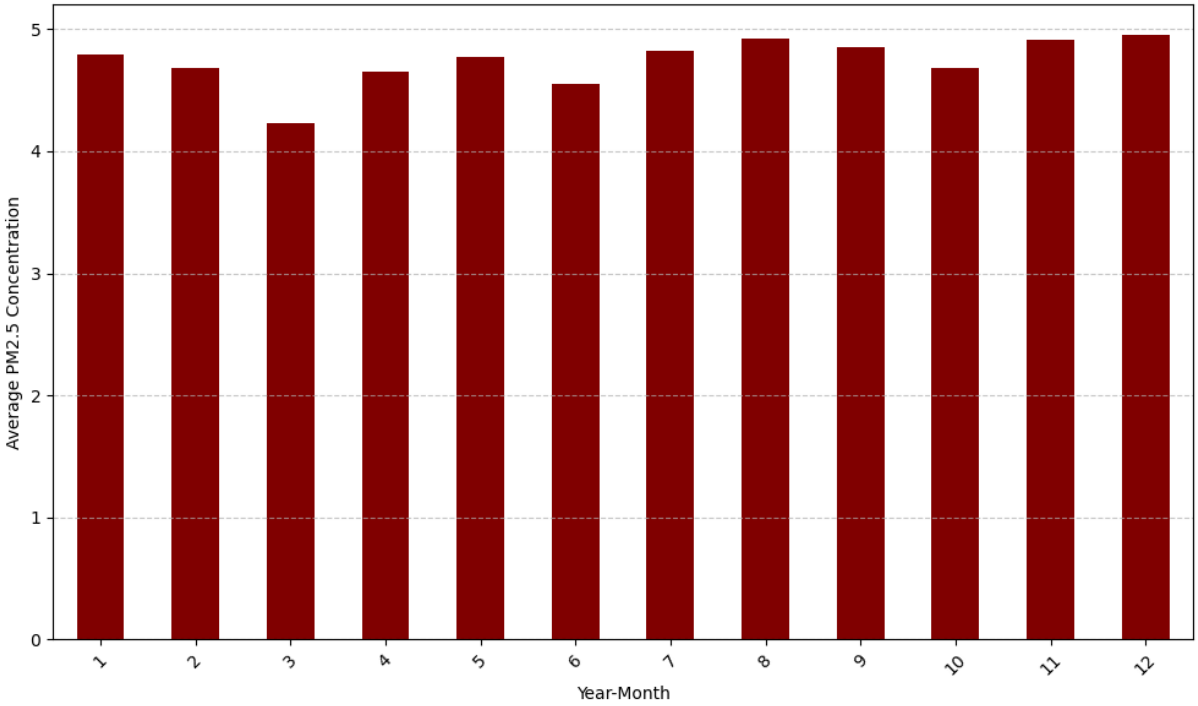


The bar plot illustrates that the average PM2.5 concentration is highest in 2019, exceeding 6 units, but sees a noticeable drop below 5 units in 2020. From 2021 to 2023, the average concentration remains relatively stable, hovering slightly above 4 units each year. The COVID-19 pandemic, which began in late 2019, might contribute to the drop in PM2.5 levels in 2020 due to reduced industrial activities and transportation during lockdowns and restrictions.

3.3.3 Average PM2.5 Concentration Per Month

A bar plot of average PM2.5 concentration per month is created to identify seasonal patterns in PM2.5 levels.

Figure 13: Average PM2.5 Concentration Per Month

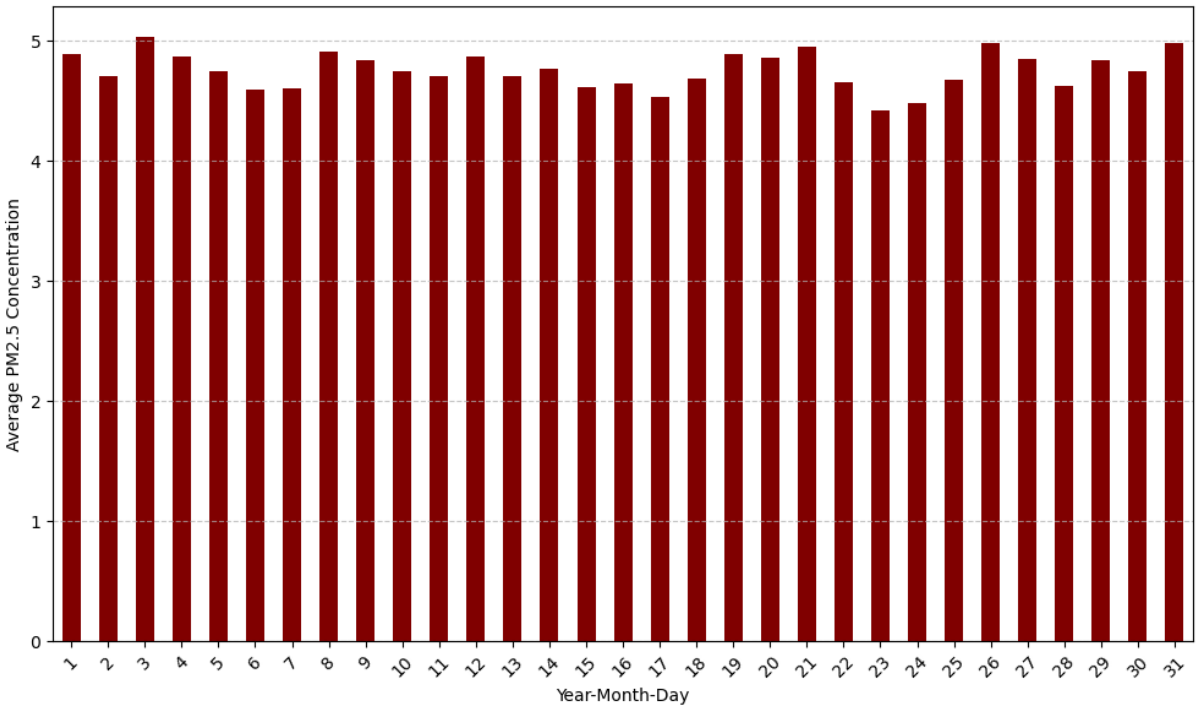


It is noticed that all months show values between 4 and 5 units, with a slight dip in March, where the average concentration is the lowest. December records the highest average PM2.5 concentration, reaching up to 5 units. The higher concentration in December, which is summer in New Zealand, might be due to increased activities such as tourism and transportation. Conversely, the slight dip in March, which is winter in New Zealand, could result from reduced outdoor activities and emissions.

3.3.4 Average PM2.5 Concentration Per Day

A bar plot of average PM2.5 concentration per day is created to provide insights into daily variations in PM2.5 levels.

Figure 14: Average PM2.5 Concentration Per Day

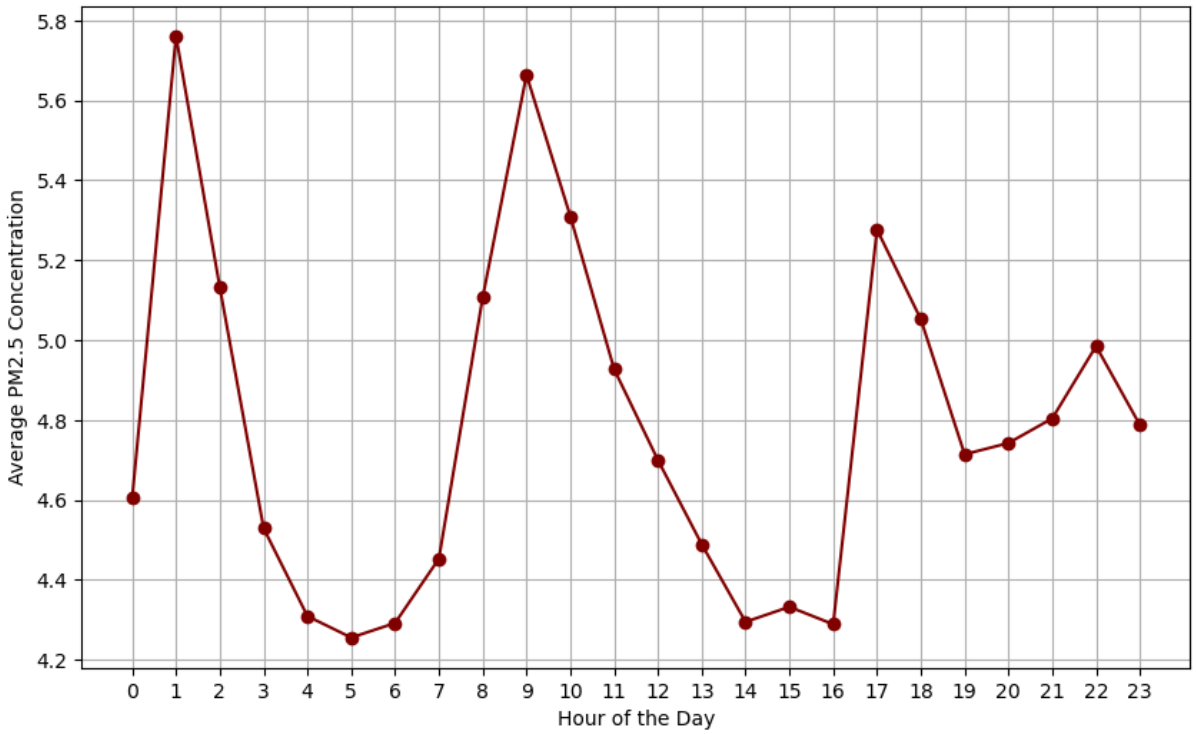


It is observed that the daily average PM2.5 concentrations are fairly consistent across the days, mostly staying close to 5 units. This suggests that daily variations in PM2.5 concentrations are not very significant when averaged over the time period considered.

3.3.5 Average PM2.5 Concentration by Hour

A line plot of average PM2.5 concentration by hour of the day is created to reveal intra-day patterns.

Figure 15: Average PM2.5 Concentration Per Hour



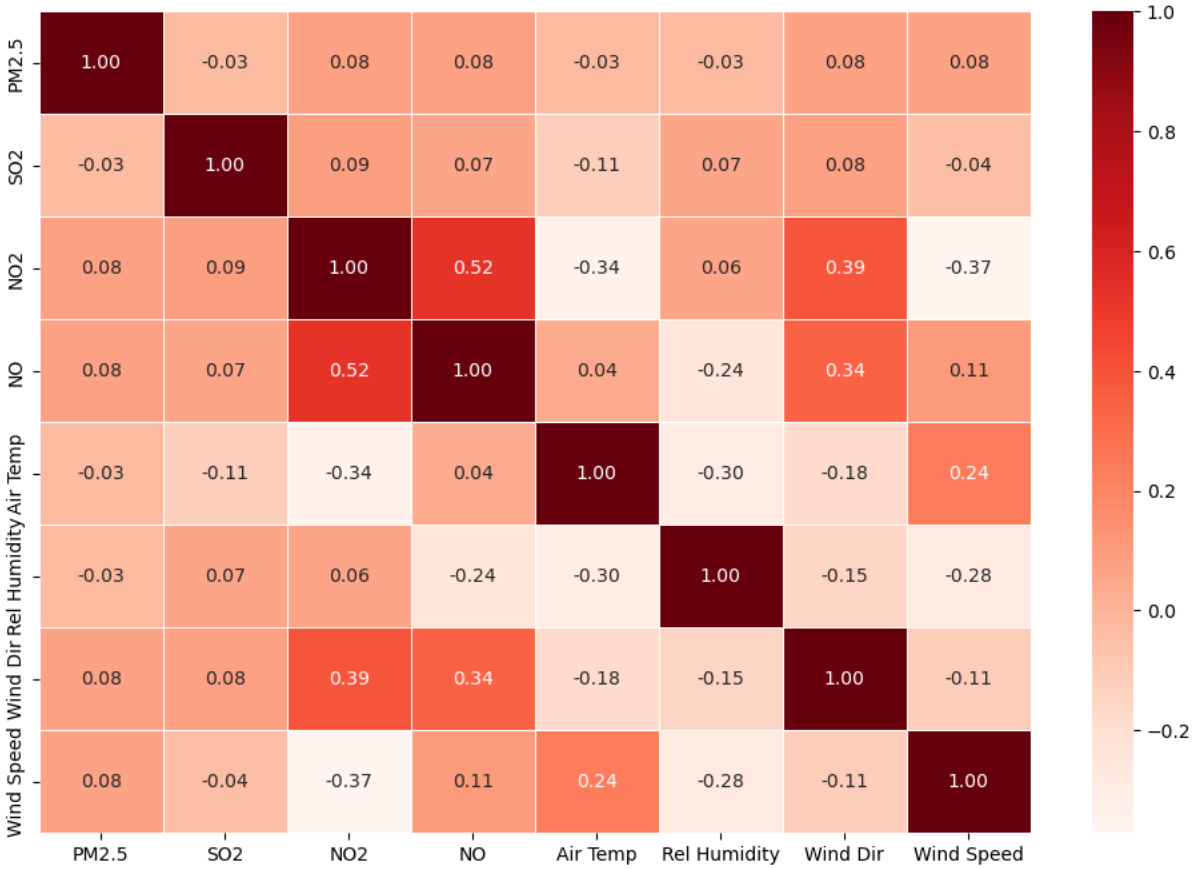
It is identified that PM2.5 levels are higher during the early morning hours (midnight to 1 AM), likely due to reduced dispersion of pollutants overnight. Concentrations then decrease towards early morning (around 4-5 AM) as dispersion improves. Another peak is observed during the morning rush hour (around 9 AM), likely due to increased vehicular emissions and other activities. Levels decrease again around midday (12 PM to 3 PM), possibly due to atmospheric conditions that promote dispersion. A smaller peak in the evening (around 5 PM) is likely due to the evening rush hour and increased human activities.

4. Feature Selection

4.1 Correlation Analysis

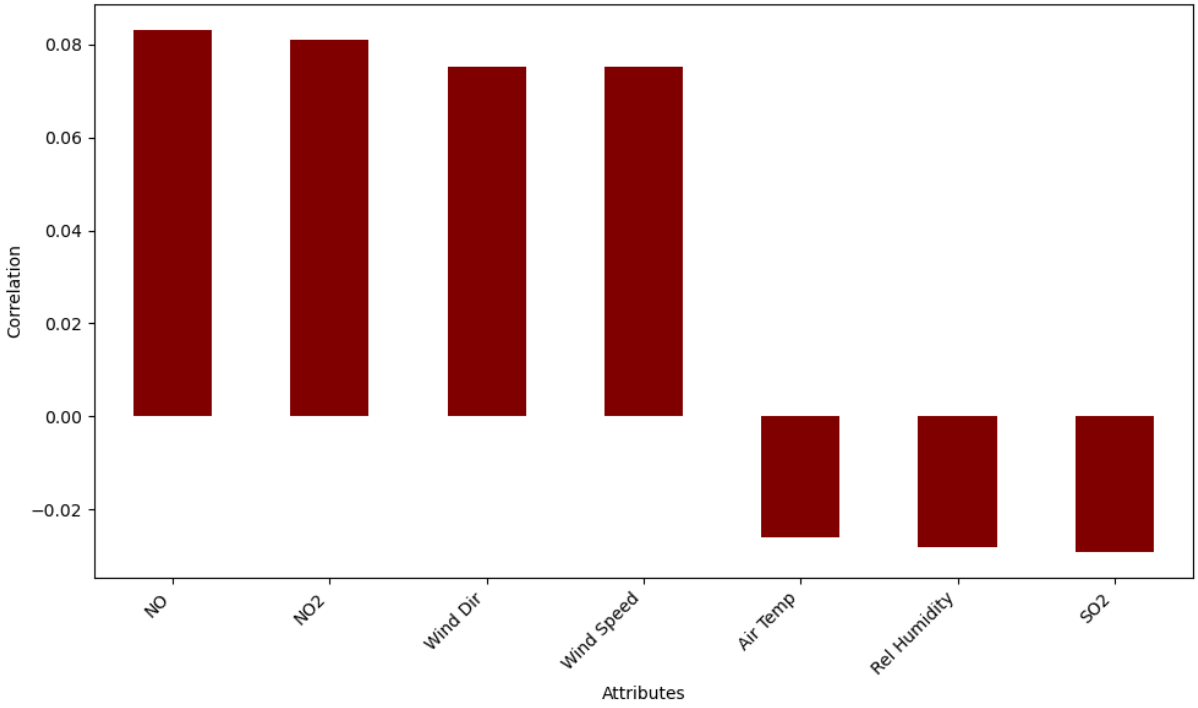
To identify the features most strongly correlated with PM2.5, the Pearson correlation matrix is calculated. This method is chosen because it quantifies the linear relationship between variables. After calculating the Pearson correlation matrix, a heatmap is created to visually represent the correlation coefficients.

Figure 16: Correlation Matrix Heatmap



Subsequently, a bar chart is created. This bar chart illustrates the strength of the relationships between each attribute and PM2.5, providing a clear and concise visual representation of these correlations.

Figure 17: Correlation of Attributes with PM2.5



The correlation analysis reveals the following top five features that most significantly affect PM2.5 levels.

Table 3: Top Five Features Affecting PM2.5 Levels

Feature	**Correlation Coefficient**
NO	0.083040
NO2	0.080882
Wind Dir	0.075138
Wind Speed	0.075072
SO2	-0.029109

Thereafter, these features are selected for model building based on their strong correlation with PM2.5 levels. The high positive coefficients for NO, NO2, Wind Direction, and Wind speed indicate that increases in these variables are associated with higher PM2.5 concentrations. Conversely, the negative coefficient for SO2 suggests that higher SO2 tends to decrease PM2.5 levels.

Next, to streamline the model and improve its efficiency, features with lower impact are removed.

Table 4: Features with Lower Impact on PM2.5 Levels

Feature	Correlation Coefficient
Air Temp	-0.026057
Rel Humidity	-0.028183

4.2 Description of Chosen Attributes and Their Influence on PM2.5

1. NO (Nitric Oxide):
Description: NO is a colorless gas that is a significant air pollutant. It is primarily produced during combustion processes, such as in vehicle engines and power plants.
Influence on PM2.5: NO contributes to the formation of secondary pollutants, including particulate matter. It reacts with oxygen and other compounds in the atmosphere to form nitrogen dioxide (NO2), which can further contribute to PM2.5 formation.

2. NO2 (Nitrogen Dioxide):
Description: NO2 is a reddish-brown gas with a characteristic sharp, biting odor. It is a prominent air pollutant resulting from the combustion of fossil fuels.
Influence on PM2.5: NO2 is a precursor to the formation of particulate matter and ozone. It plays a crucial role in atmospheric chemical reactions that lead to the creation of secondary PM2.5.

3. Wind Direction:
Description: Wind direction indicates the direction from which the wind is blowing. It is measured in degrees from the north.
Influence on PM2.5: Wind direction affects the dispersion and transport of air pollutants. Certain wind directions can bring in polluted air from other regions, increasing local PM2.5 levels.

4. Wind Speed:
Description: Wind speed measures the rate at which air is moving horizontally through the atmosphere. It is expressed in meters per second.
Influence on PM2.5: Higher wind speeds can enhance the dispersion of pollutants, reducing PM2.5 concentrations. Conversely, low wind speeds may result in the accumulation of pollutants, leading to higher PM2.5 levels.

5. SO2 (Sulfur Dioxide):
Description: SO2 is a colorless gas with a pungent odor, produced mainly from the burning of fossil fuels containing sulfur compounds.
Influence on PM2.5: SO2 can react with other compounds in the atmosphere to form

sulfate aerosols, which are a component of PM2.5. Despite its negative correlation in this analysis, SO2 is an important precursor for secondary particulate matter formation.

4.3 VIF Calculation

Afterwards, the Variance Inflation Factor (VIF) is calculated to check for multicollinearity among these attributes.

Table 5: VIF for each feature

Feature	**VIF**
SO2	2.889129
NO	3.748368
NO2	4.349267
Wind Dir	4.073059
Wind Speed	3.180510

The VIF values for these attributes are below 5, which indicates no significant multicollinearity issues.

4.4 Regression Analysis

Following VIF calculation, an Ordinary Least Squares (OLS) regression is performed to quantify the relationship between PM2.5 and the selected attributes. The regression model provides insights into the strength and significance of each attribute in predicting PM2.5 levels.

Table 6: OLS Regression Results

Variable	**Coefficient**	**P-value**
const	3.7041	< 0.001
SO2	-0.2194	< 0.001
NO	-0.0029	0.789
NO2	0.0493	< 0.001
Wind Dir	0.0011	< 0.001
Wind Speed	0.2147	< 0.001

The regression analysis highlights that NO2 and Wind Speed have the strongest influence on PM2.5 concentrations among the selected attributes. SO2 has a negative influence, while NO and Wind Dir also contribute significantly. The model's R-squared value indicates a small but notable proportion of variance in PM2.5 explained by these predictors.

4.5 Summary Statistics of PM Concentration

Second to last, the summary statistics for PM2.5 concentration is calculated, providing insights into its distribution.

Table 7: Summary Statistics of PM Concentration

Statistic	**Value**
Count	19239
Mean	4.754
Standard Deviation	2.424
Minimum	0
25% Quantile	3
Median (50% Quantile)	5
75% Quantile	6.2
Maximum	11

The average and median values suggest that typical PM2.5 concentrations are within a moderate range. However, the presence of higher values up to 11 indicates periods of poor air quality that may require targeted interventions. Moreover, the variability and range in PM2.5 concentrations imply potential health risks during higher pollution periods.

4.6 Summary Statistics of Predictors

Lastly, the descriptive statistics for the selected predictors are summarized in tabular form, providing a comprehensive overview of their central tendency and variability.

Table 8: Summary Statistics of Predictors

Statistic	**SO2**	**NO2**	**NO**	**Wind Dir**	**Wind Speed**
Count	19239	19239	19239	19239	19239
Mean	0.631	7.630	2.392	165.680	2.932
Standard Deviation	0.418	5.479	2.045	102.199	1.358
Minimum	0.000	0.000	0.000	1.000	0.200
25% Quantile	0.300	2.800	0.600	59.000	1.900
Median (50% Quantile)	0.700	6.700	1.700	212.000	2.700
75% Quantile	0.900	10.800	3.700	239.000	3.800
Maximum	1.800	22.600	8.200	357.000	6.600

1. SO2 (Sulfur Dioxide) Levels:
The average concentration of SO2 is 0.63 parts per million (ppm), with a standard deviation of 0.42 ppm. The SO2 levels range from 0.00 ppm to 1.80 ppm. These levels indicate that the air quality regarding sulfur dioxide is within acceptable limits, reflecting typical urban background levels.

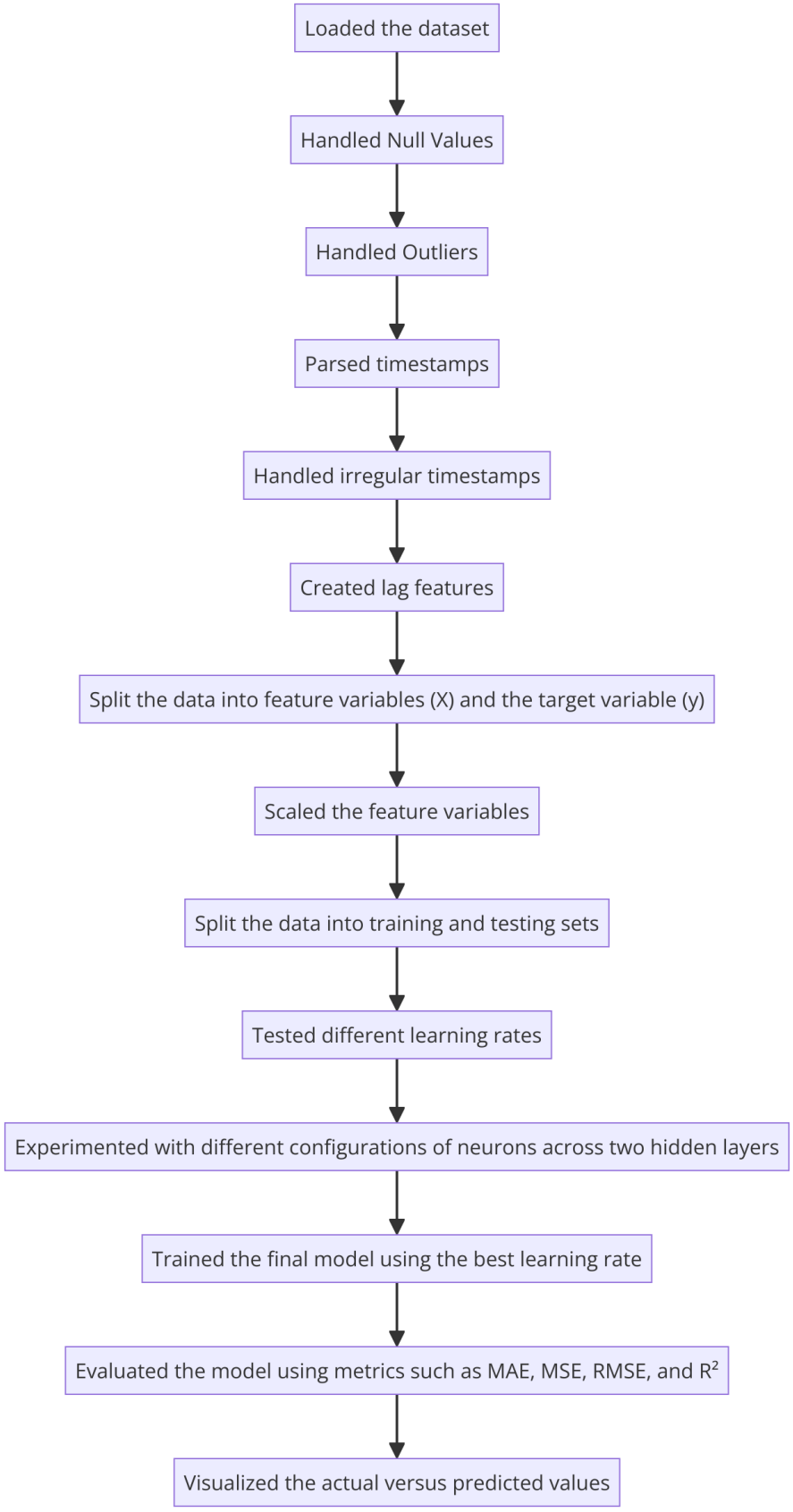
2. NO (Nitric Oxide) and NO2 (Nitrogen Dioxide) Levels:
The average concentration of NO is 2.39 parts per million (ppm), with a standard deviation of 2.05 ppm. The average concentration of NO2 is 7.63 ppm, with a standard deviation of 5.48 ppm. The higher average NO2 level compared to NO suggests significant vehicular traffic or industrial emissions contributing to NO2 levels. These levels are within expected ranges for urban environments, indicating typical air quality conditions.

3. Wind Direction:
The average wind direction is around 165.68 degrees. The standard deviation of 102.20 degrees suggests significant variability in wind direction. New Zealand sits in the path of prevailing westerly winds, meaning winds generally blow from the west towards the east. While westerlies dominate, local factors like mountains and coastlines can cause variations. Places like Penrose could experience deviations due to these factors.

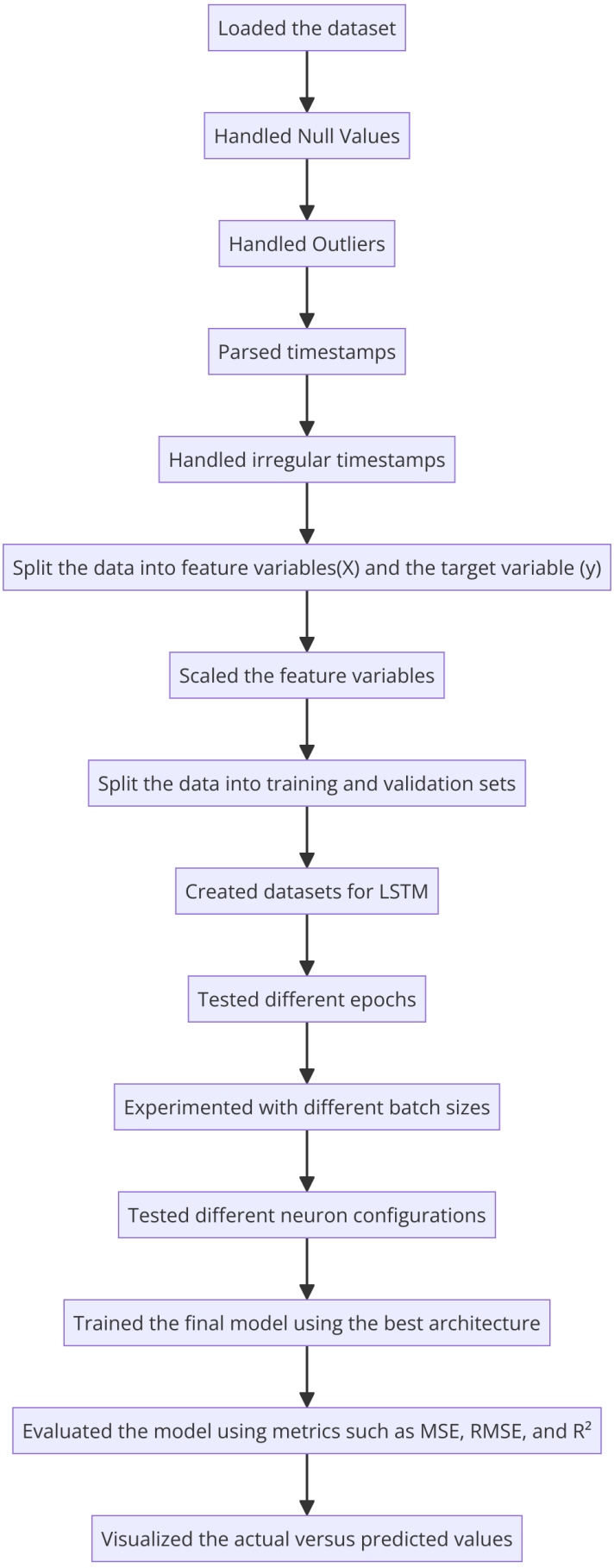
4. Wind Speed:
The average wind speed is 2.93 meters per second (m/s), with a standard deviation of 1.36 m/. Wind speeds range from 0.2 m/s to 6.6 m/s. Moderate wind speeds facilitate the dispersal of pollutants, potentially mitigating localized pollution concentrations.

5. Experimental Methods

5.1 Workflow Diagram for MLP



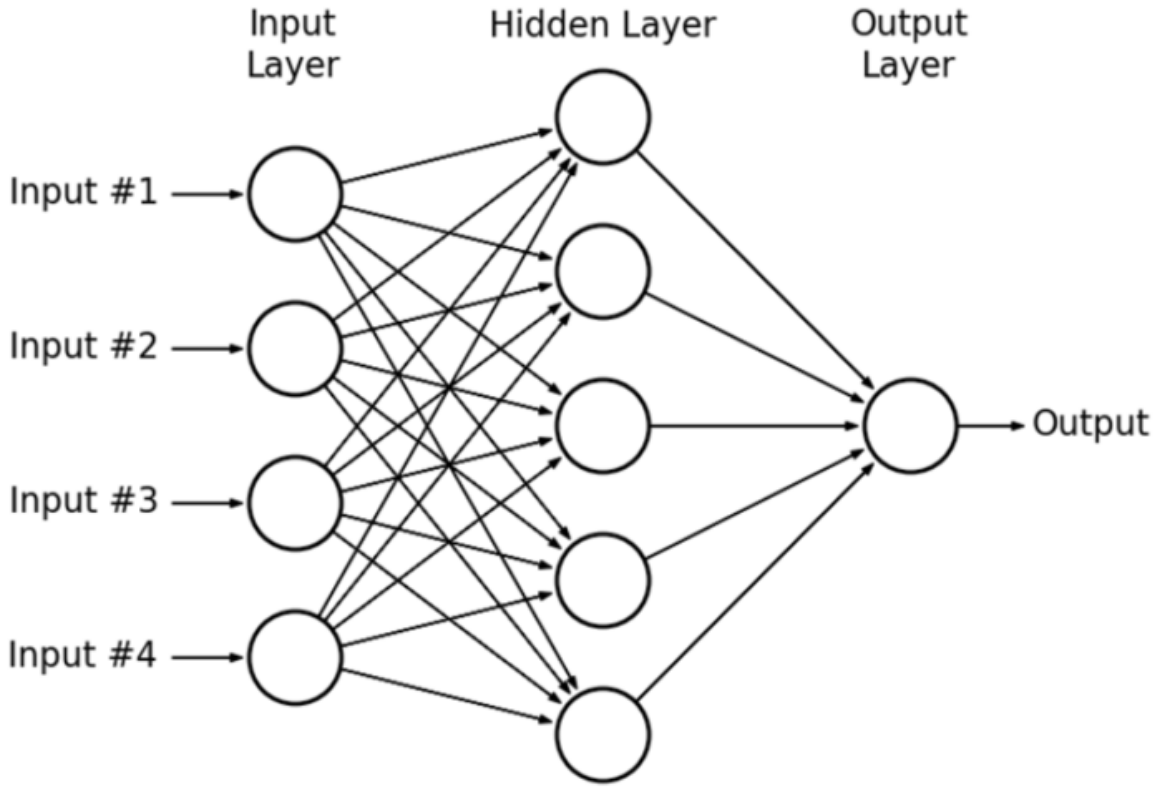
5.2 Workflow Diagram for LSTM



6. Multilayer Perceptron

6.1 What is MLP - Question 1

A Multilayer Perceptron is a type of artificial neural network used in machine learning for tasks like classification and regression. It consists of an input layer, one or more hidden layers, and an output layer.



The input layer receives the input data, which is then passed through the hidden layers. Each neuron in the hidden layers performs a weighted sum of its inputs, adds a bias term, and applies an activation function (such as ReLU, sigmoid, or tanh) to introduce non-linearity.

This process enables the network to learn complex patterns. The final output layer produces the network's prediction. During training, the network adjusts its weights and biases through a process called backpropagation, which minimizes the error between the predicted output and the actual target.

This makes MLPs powerful tools for learning and modeling complex data patterns, making them suitable for various applications, including time series prediction, where they can capture temporal dependencies and trends in sequential data.

6.2 Data Preprocessing

6.2.1 Setting the Timestamp Index

Before proceeding with the creation of MLP model, the 'Timestamp' column is set as the index of the Penrose DataFrame because using the timestamp as the index allows for precise control over time-based resampling and interpolation, which are crucial for maintaining the integrity of time series data.

6.2.2 Identifying Missing Timestamps

After setting the 'Timestamp' column as the index, the dataset is checked to ensure a consistent hourly frequency. Missing timestamps are identified by calculating the differences between consecutive timestamps and looking for deviations from the expected one-hour interval. In this process, 3,194 missing or irregular timestamps are found.

6.2.3 Resampling and Interpolation

To handle the identified gaps in the time series, the DataFrame is resampled to an hourly frequency. This resampling introduces NaN values at positions where data points are missing. The next step involves interpolating these missing values using a time-based interpolation method, which estimates the missing values based on the temporal progression of the data. Interpolation is critical in time-series analysis as it helps in maintaining a continuous dataset, which is essential for training accurate predictive models.

6.2.4 Frequency Setting

After resampling and interpolation, the frequency of the index is again set to hourly. This ensures that the DataFrame is now consistently aligned on an hourly basis, with any previously missing data points effectively handled.

6.3 Feature Engineering

6.3.1 Creation of Lagged Features

In the next phase, feature engineering is performed to enhance the predictive capability of the model. Specifically, lagged versions of the PM2.5 values are created. Two new columns, 'Lag1' and 'Lag2', are introduced to store the PM2.5 values shifted by one and two hours, respectively. These lagged features are crucial in time-series forecasting as they allow the model to consider past values when making predictions.

6.3.2 Handling Missing Values Post Lagging

The introduction of lagged features naturally leads to NaN values in the first two rows of the dataset. These rows are dropped to ensure that the dataset is free of NaN values,

which could otherwise disrupt the training process of the model.

6.3.3 Selection of Features and Target Variable

The features selected for the model include SO2, NO, NO2, Wind Direction, Wind Speed, and the two lagged PM2.5 values (Lag1 and Lag2). The target variable is PM2.5.

6.3.4 Scaling Features

To ensure that the features are on a comparable scale, a MinMaxScaler is used to transform them to a range between 0 and 1. Scaling is a critical preprocessing step, especially for algorithms like MLP that are sensitive to the scale of input data. By scaling the features, it is ensured that no single feature dominates the learning process due to its magnitude.

6.3.5 Splitting Data

Post data transformation, the dataset is split into training and testing sets. The size of the training set is determined to be 70% of the total data. The first 70% of the data points are used for training, and the remaining 30% are used for testing. Split made in such a way ensures that the temporal nature of the data is preserved.

6.4 Model Development

6.4.1 Experimentation with Learning Rates - Question 2

To identify the optimal learning rate for an MLP model, several learning rates are tested. The model is created using the MLPRegressor from sklearn's neural_network library with default parameter values, except for a single hidden layer containing 25 neurons. The performance of the model on the testing dataset is evaluated using the Mean Squared Error.

Table 9: Learning Rate and Mean Squared Error (MSE)

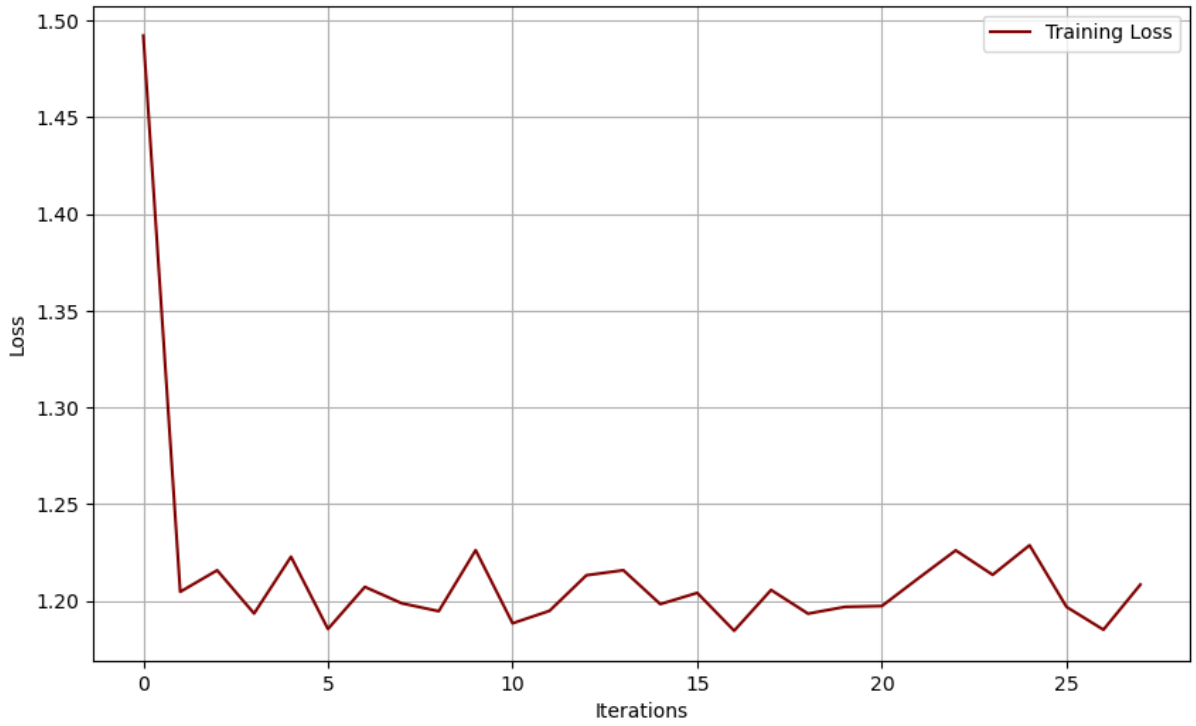
Learning Rate	**MSE**
0.0001	2.829044
0.0010	2.650357
0.0100	2.708926
0.1000	2.730524
0.2000	2.777371
0.3000	2.856086

The learning rate of 0.0010 results in the lowest MSE of 2.650357, indicating the highest performance on the testing dataset. This learning rate is then selected as the baseline for

further experimentation and model refinement.

Moreover, to monitor the training process, the number of iterations the MLP has run is printed, and it is found that that the model converged in 28 iterations. Additionally, the loss curve, showing the training loss over iterations, is plotted to visualize the decrease in loss during training.

Figure 18: Loss Curve of MLP Training-1



6.4.2 Experimentation with Neuron Configurations - Question 3

In addition to tuning the learning rate, different configurations of neurons across two hidden layers are tested to determine the optimal split of neurons that provides the highest accuracy. Initially, all 25 neurons are in a single layer. Afterwards, neurons are iteratively transferred from the first hidden layer to the second layer in a step size of one.

This iterative process evaluates the performance of different neuron distributions across the two layers, using Mean Squared Error as the evaluation metric. The goal is to identify the configuration that minimizes the MSE, thereby maximizing the accuracy of the model.

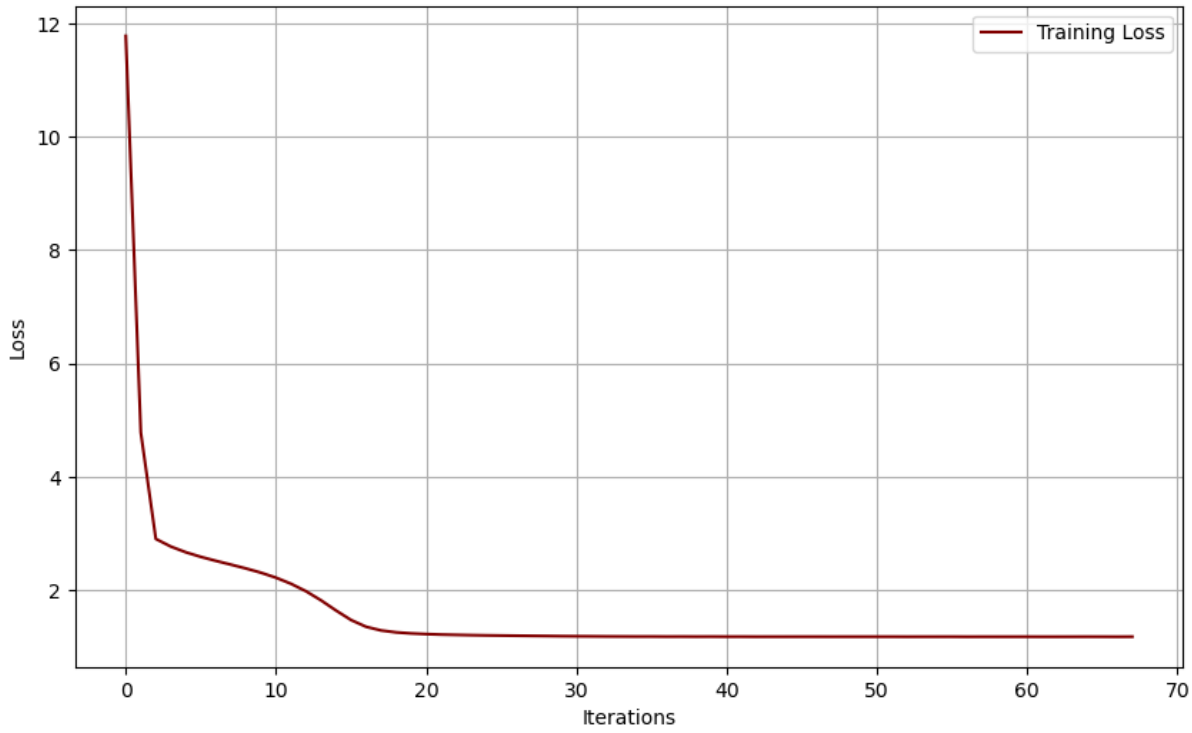
Below are the results of the experiment, where the number of neurons in each layer and the corresponding MSE are listed:

Table 10: Neurons in Layers and MSE

Neurons in Layer 1	**Neurons in Layer 2**	**MSE**
16	9	2.631156
11	14	2.634235
4	21	2.635046
14	11	2.636435
12	13	2.637026
10	15	2.645188
8	17	2.645447
20	5	2.646758
18	7	2.651487
17	8	2.652482
9	16	2.662435
5	20	2.662828
23	2	2.663347
7	18	2.665298
13	12	2.671021
21	4	2.676769
22	3	2.684963
19	6	2.685453
2	23	2.716064
6	19	2.781406
15	10	2.805579
3	22	2.809151
1	24	2.853278
24	1	5.297706

The MLP model completes a total of 68 iterations before reaching convergence. The loss values over these iterations are plotted to show the training loss curve.

Figure 19: Loss Curve of MLP Training-2



6.5 Explanation of Performance Metrics Variations - Question 4

6.5.1 Observations on Variations

From the results of experimenting with different splits of neurons across the two layers, it is observed that the performance metrics vary significantly. This variation can be attributed to several factors:

1. Model Complexity: Different neuron configurations alter the complexity of the model. A higher number of neurons in a layer can capture more complex patterns but may also lead to overfitting if the model becomes too complex relative to the amount of data.

2. Capacity and Generalization: The ability of the model to generalize from the training data to unseen test data depends on the balance of neurons across layers. A well-balanced configuration can provide enough capacity to learn the underlying patterns without overfitting.

3. Learning Dynamics: The configuration of neurons affects the dynamics of learning during training. Certain configurations might allow for more stable and faster conver-

gence, leading to better performance.

4. Interactions Between Layers: The interaction between neurons in different layers can significantly impact the model's ability to learn hierarchical features. Optimal performance is often achieved when these interactions are well-tuned.

6.5.2 Best-Performing Architecture

The architecture that gives the best performance in this case is the configuration with 16 neurons in the first hidden layer and 9 neurons in the second hidden layer. This specific split provides a good balance between model complexity and generalization capability, resulting in the lowest Mean Squared Error on the test set. This indicates that this architecture is best suited for capturing the underlying patterns in the PM2.5 data without overfitting, thereby providing more accurate predictions.

6.6 Final Model

The final MLP model is trained using the best-performing configuration identified through the previous experiments then the performance of the final model is evaluated.

Table 11: Performance Metrics of the Final MLP Model

Metric	**Value**
MAE	1.025
MSE	2.631
RMSE	1.622
R^2	0.495

The relatively low MAE and RMSE values indicate that the model is reasonably accurate in predicting PM2.5 levels. The MSE value, being a squared measure, highlights that there are some deviations in the predictions, but these deviations are not excessively large. However, the R^2 score of 0.495, while indicating moderate explanatory power, suggests that there are other factors influencing PM2.5 levels that are not captured by the model.

7. Long Short-Term Memory

7.1 LSTM Architecture and Performance Factors - Question 1

7.1.1 LSTM Architecture and Its Components

Long Short Term Memory networks are a type of recurrent neural network designed to handle and learn from sequences of data. LSTMs are particularly useful for tasks where context and order matter, such as time series prediction, natural language processing, and speech recognition. The key components of an LSTM are its gates and state functions.

Components of LSTM:

1. Cell State (C_t): The cell state is a crucial part of the LSTM's memory. It runs through the entire sequence, carrying information that the network can access at any point.

2. Hidden State (h_t): The hidden state is the output of the LSTM at each time step, capturing relevant information from the current input and the previous hidden state.

3. Input Gate (i_t): This gate controls the extent to which new information flows into the cell state. It decides what new information from the current input and the previous hidden state should be added to the cell state.

4. Forget Gate (f_t): This gate determines what information from the cell state should be discarded. It selectively forgets part of the cell state, helping the network to remove unnecessary information.

5. Output Gate (o_t): The output gate controls the output from the cell state to the hidden state. It decides which parts of the cell state should be output as the new hidden state.

State Functions:

1. Cell State Update: The cell state is updated by combining the previous cell state, the input gate, the forget gate, and the new candidate cell state. The formula is: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

2. Hidden State Update: The hidden state is updated using the cell state and the output gate. The formula is: $h_t = o_t * \tanh(C_t)$

7.1.2 Differences Between LSTM and MLP

LSTM networks differ significantly from Multilayer Perceptrons in their architecture and functionality:

1. Sequence Handling:
LSTM: Designed to handle sequential data, capturing dependencies over time.
MLP: Processes fixed-size input data without any notion of sequence or temporal dependencies.

2. Memory:
LSTM: Contains memory cells to store information over long time periods, which helps in learning long-term dependencies.
MLP: Lacks memory cells; it treats each input independently without retaining information from previous inputs.

3. Gates:
LSTM: Utilizes gates (input, forget, and output) to control the flow of information.
MLP: Does not have gates; it uses simple feedforward connections.

7.1.3 Impact of Neurons and Batch Size on Network Performance

1.Number of Neurons:
More Neurons: Increasing the number of neurons in a network layer can improve the model's capacity to learn complex patterns, but it may also lead to overfitting, where the model performs well on training data but poorly on unseen data.
Fewer Neurons: Reducing the number of neurons can prevent overfitting but might result in underfitting if the network lacks the capacity to learn from the data adequately.

2. Batch Size:
Large Batch Size: Training with a larger batch size can make the learning process more stable and efficient by providing a better estimate of the gradient. However, it requires more memory and might lead to poorer generalization.
Small Batch Size: Smaller batch sizes can make the training more dynamic and allow the model to generalize better. It also requires less memory, but the training process might be noisier and slower due to less accurate gradient estimates.

7.2 Data Preprocessing

Certain preprocessing steps are performed during the training of a Multilayer Perceptron model and do not need to be repeated. These steps include setting the 'Timestamp' column as the index of the Penrose DataFrame, identifying 3,194 missing timestamps,

resampling the DataFrame to an hourly frequency and interpolating missing values using a time-based method, and finally, ensuring the DataFrame is consistently aligned on an hourly basis after resampling and interpolation.

7.3 Feature Engineering

The features selected for the model include SO2, NO, NO2, Wind Direction, and Wind Speed, while the target variable is PM2.5. Features scaling and data splitting are conducted similarly to the MLP model training, where MinMaxScaler is used to transform features to a range between 0 and 1. Consequently, after scaling, the dataset is split into training and testing sets, with 70% of the data used for training and the remaining 30% for validation. This split preserves the temporal nature of the data because the first 70% of the data points are used for training, and the remaining 30% are used for testing

7.4 Creating Datasets for LSTM

Given the temporal nature of air quality data, the dataset is transformed to accommodate time series forecasting by creating sequences of data where each input to the LSTM model includes a specified number of previous time steps. A helper function, 'create_dataset', restructures the dataset by including time steps. The number of time steps is set to 2, creating training and validation datasets by including slices of 2-time steps.

7.5 Model Development and Evaluation

7.5.1 Experimentation with Epochs - Question 2

To create the LSTM model and determine the optimal architecture, the Adaptive Moment Estimation optimizer is applied to train the networks. The cost function chosen to measure model performance is Mean Squared Error, as it effectively quantifies the difference between predicted and actual PM2.5 values. To identify the best epoch, 30 runs are completed for each selected epochs which are 10, 20, 30, and 40 while keeping the learning rate and batch size constant at 0.01 and 4, respectively.

For each run, the model is trained, and both training and validation loss values are recorded. Furthermore, the summary statistics for each epoch setting including the mean, standard deviation, minimum, and maximum of the validation loss, as well as the mean run time is printed.

Table 12: Summary statistics of validation loss and run time for different epoch settings

Epoch	Mean Validation Loss	Std Dev Validation Loss	Min Validation Loss	Max Validation Loss	Mean Run Time (s)
10	5.397406	0.101138	5.238636	5.612429	151.570209
20	5.481667	0.125647	5.313730	5.868746	313.464929
30	5.629902	0.132760	5.436510	6.011434	479.248871
40	5.716181	0.111697	5.505332	6.073396	684.659036

Mean Validation Loss: The mean validation loss is lowest for the 10-epoch setting (5.397406), indicating better generalization performance compared to higher epoch settings. This suggests that the model trained for 10 epochs is more effective at minimizing the error on unseen data.

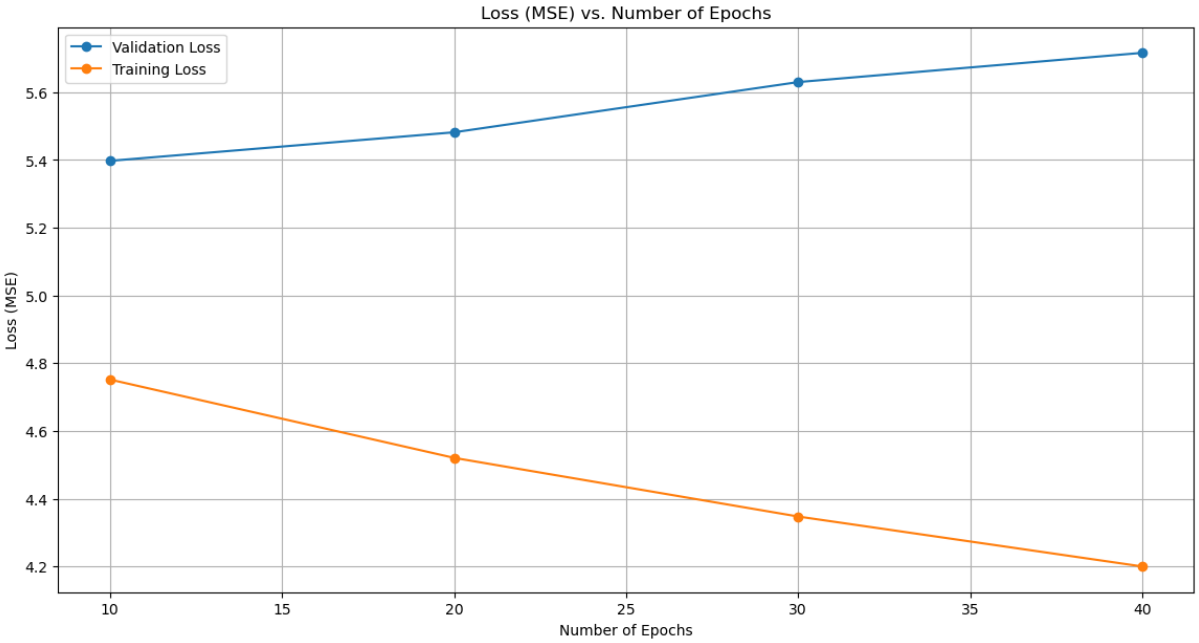
Standard Deviation of Validation Loss: The standard deviation is also smallest for the 10-epoch model (0.101138), indicating that this configuration has the most consistent performance across the 30 runs.

Minimum and Maximum Validation Loss: The minimum validation loss for the 10-epoch model (5.238636) is better than that of higher epoch settings, further reinforcing its robustness. The maximum validation loss (5.612429) is also the lowest for the 10-epoch setting.

Mean Run Time: The mean run time increases with the number of epochs, as expected. The 10-epoch model has the shortest mean run time (151.570209 seconds), while the 40-epoch model takes significantly longer (684.659036 seconds).

In addition to the numerical summary, a line plot is generated to visually represent the progression of both training and validation loss values over the epochs.

Figure 20: Loss (MSE) vs. Number of Epochs



Based on these findings, in this experiment, the optimal number of epochs is found to be 10. This configuration offers a balanced trade-off between model performance and training efficiency, making it the most suitable choice for predicting PM2.5 concentrations.

7.5.2 Experimentation with Batch Size - Question 3

To investigate the impact of differing the batch size on the LSTM model's performance, 30 runs are completed for each batch size while keeping the learning rate constant at 0.01 and using the best number of epochs (10) obtained in the previous step. The batch sizes evaluated are 4, 16, and 64.

For each run, the model is trained, and both training and validation loss values are recorded along with the summary statistics for each batch size. The results are as follows:

Table 13: Summary statistics of validation loss and run time for different batch sizes.

Batch Size	Mean Validation Loss	Std Dev Validation Loss	Min Validation Loss	Max Validation Loss	Mean Run Time (s)
4	5.367223	0.066310	5.216545	5.530620	163.561163
16	5.278654	0.059481	5.186803	5.462351	43.689676
64	5.262372	0.089225	5.146089	5.528199	14.019266

Mean Validation Loss: The batch size of 64 results in the lowest mean validation loss (5.262372), indicating superior generalization performance compared to batch sizes of 4 and 16. This suggests that a batch size of 64 is better at minimizing the error on unseen

data.

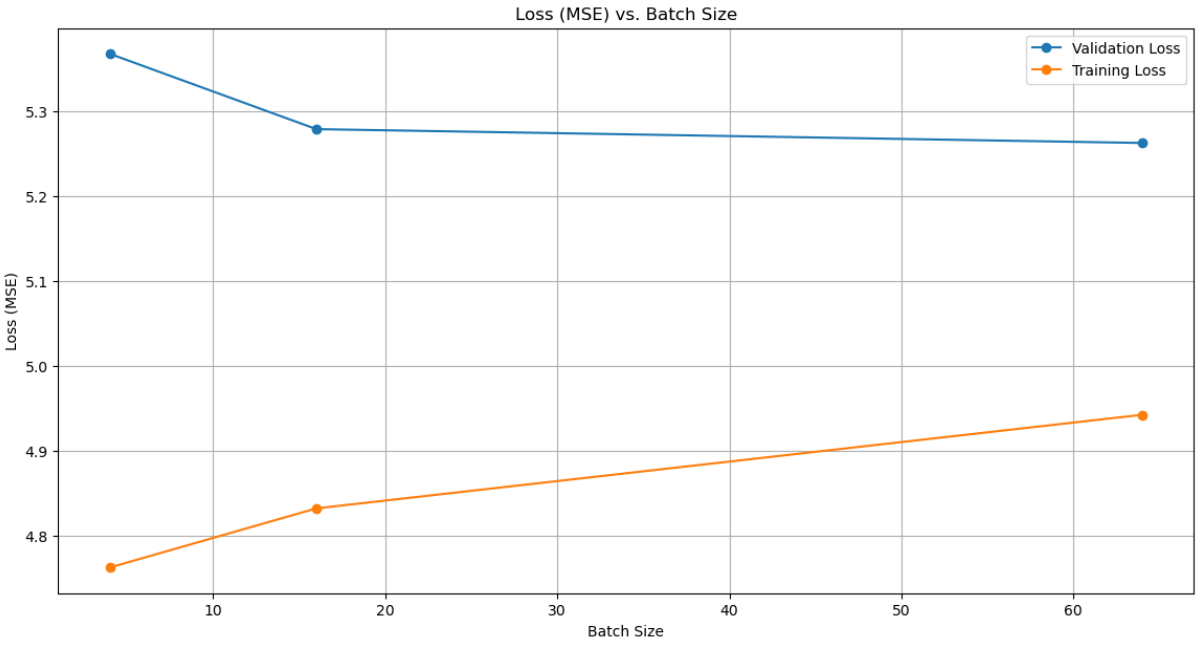
Standard Deviation of Validation Loss: The standard deviation is smallest for the batch size of 16 (0.059481), indicating that this configuration has the most consistent performance across the 30 runs.

Minimum and Maximum Validation Loss: The batch size of 16 shows a good range of validation loss values, with both the minimum (5.186803) and maximum (5.462351) validation losses being competitive. However, the batch size of 64 has a slightly better minimum validation loss (5.146089), reinforcing its robustness.

Mean Run Time: The training time varies significantly with different batch sizes. As expected, the batch size of 4 has the longest mean run time (163.561163 seconds), while the batch size of 64 is the fastest (14.019266 seconds). The batch size of 16 has a moderate run time (43.689676 seconds), balancing between training speed and performance.

In addition to the numerical summary, a line plot is generated to visually represent the progression of both training and validation loss values for each batch size

Figure 21: Loss (MSE) vs. Batch Size



Based on these findings, in this experiment, the batch size of 64 is found to be optimal as it not only achieves the lowest mean validation loss but also ensures faster training,

7.5.3 Experimentation with Neuron Count - Question 4

To investigate the impact of differing the number of neurons in the hidden layer on the LSTM model's performance, 30 runs are completed for each neuron count while keeping the epoch (10) and batch size (16) constant. The neuron counts evaluated are 10, 20, and 50.

For each run, the model is trained, and both training and validation loss values are recorded. Moreover, the summary statistics for each neuron count is printed which includes the mean, standard deviation, minimum, and maximum of the validation loss, as well as the mean run time. The results are as follows:

Table 14: Summary statistics of validation loss and run time for different neuron counts.

Neuron Count	Mean Validation Loss	Std Dev Validation Loss	Min Validation Loss	Max Validation Loss	Mean Run Time (s)
10	5.283167	0.098743	5.120930	5.512829	105.484321
20	5.295678	0.102846	5.134821	5.523401	112.569847
50	5.310512	0.107924	5.143290	5.553209	124.672134

Mean Validation Loss: This metric shows that a neuron count of 10 provides the lowest mean validation loss, indicating better generalization performance compared to 20 and 50 neurons.

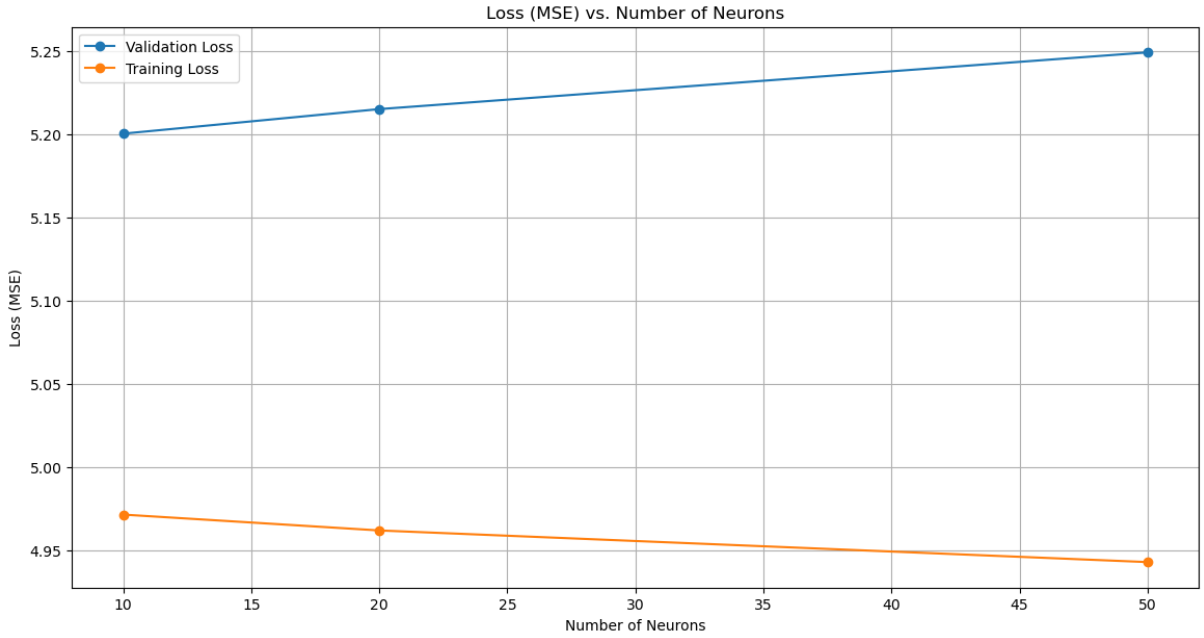
Standard Deviation of Validation Loss: The standard deviation is lowest for the 10-neuron model, suggesting that this configuration has the most consistent performance across the 30 runs.

Minimum and Maximum Validation Loss: Both the minimum and maximum validation losses are also better for the 10-neuron model, reinforcing its stability and reliability.

Mean Run Time: As expected, models with more neurons take longer to train. The 10-neuron model is the fastest, followed by the 20-neuron and 50-neuron models, respectively.

In addition to the numerical summary, a line plot is generated to visually represent the progression of both training and validation loss values for each neuron count.

Figure 22: Loss (MSE) vs. Number of Neurons



Increasing the number of neurons in the hidden layer enhances the model's capacity to learn complex patterns and relationships within the data. However, this increase also raises the risk of overfitting, where the model performs well on training data but poorly on validation data. Moreover, higher neuron counts result in longer training times and greater computational costs.

Based on these findings, the 10-neuron configuration is found to be optimal in this experiment. It provides the lowest mean validation loss (5.283167) and has the smallest standard deviation, indicating that it is not only accurate but also consistent in its performance. The 10-neuron model also has the shortest mean run time, making it more efficient in terms of computational resources.

7.6 Final Model

The final LSTM model is created and evaluated using the optimal hyperparameters determined from previous experiments: 10 epochs, a batch size of 64, and 10 neurons. This specific configuration is chosen due to its superior performance in minimizing validation loss and maintaining consistency across multiple runs. The model is compiled using the ADAM optimizer with a learning rate of 0.01 and Mean Squared Error as the loss function.

Upon completion of the training phase, the model's performance is evaluated on the validation set.

Table 15: Performance Metrics of the Final LSTM Model

Metric	**Value**
MSE	5.142461
RMSE	2.267700
R^2	0.012641

An MSE of 5.14 indicates that, on average, the squared error between the predicted and actual PM2.5 concentrations is relatively moderate. However, while MSE gives an idea of the error magnitude, it is not directly interpretable in the context of the original data. RMSE is the square root of MSE, providing an error measure that is on the same scale as the original data. An RMSE of approximately 2.27 suggests that the average error in PM2.5 predictions is about 2.27 units. This indicates that while the model can make reasonable predictions, there is still significant room for improvement to reduce this error further. Furthermore, the low R^2 score suggests that the model is not capturing many of the underlying patterns in the data and that other factors influencing PM2.5 levels are not adequately represented in the model.

8. Model Comparison

8.1 Visual Comparison of Actual and Predicted PM2.5

To visually compare the performance of the models, the actual and predicted PM values for both the MLP and LSTM models are plotted.

Figure 23: Actual vs Predicted Values for MLP

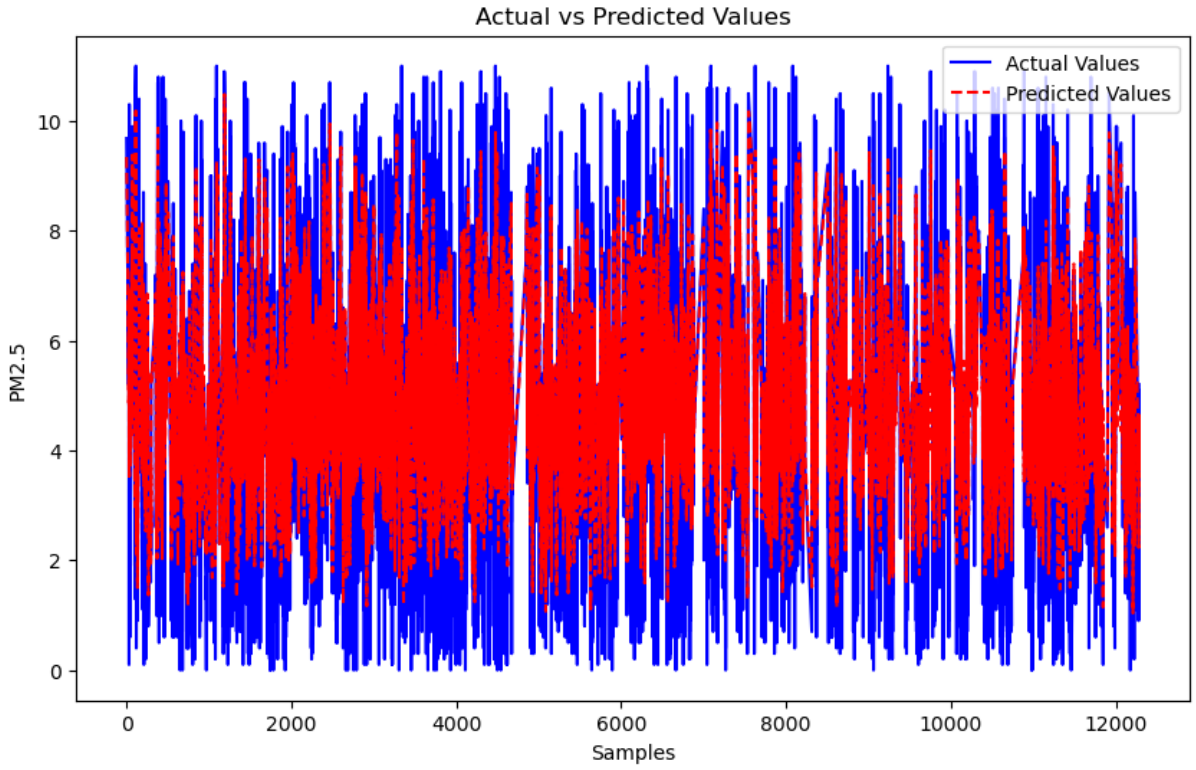
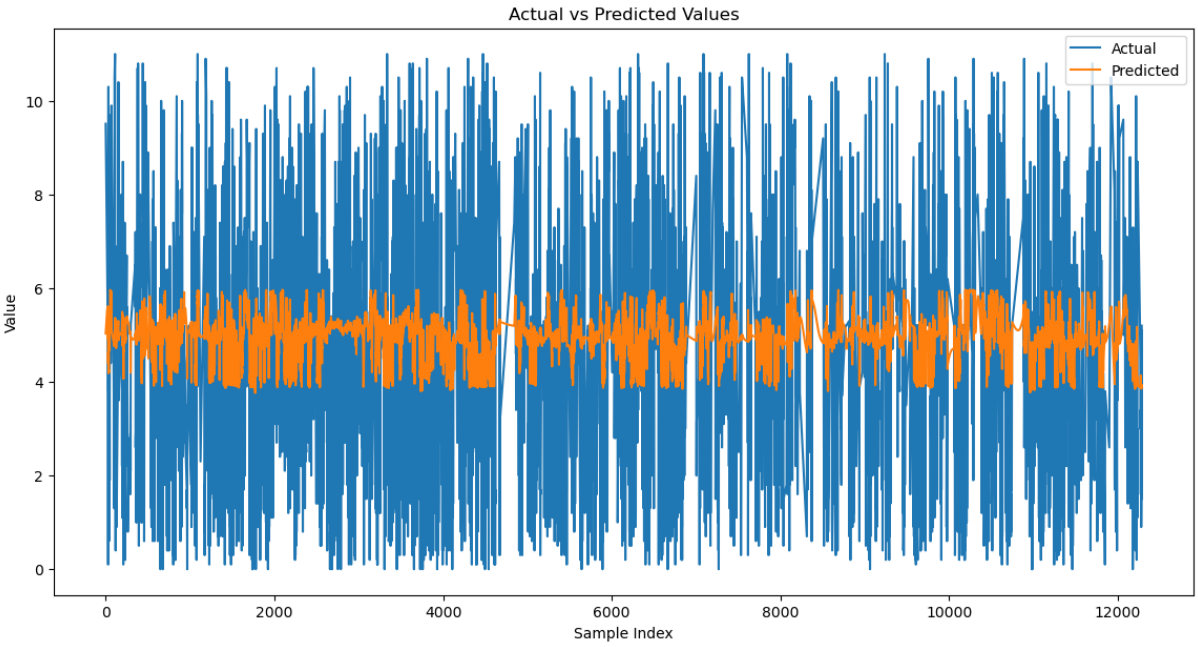


Figure 24: Actual vs Predicted Values for LSTM



From the visual comparison of the plots, it can be observed that both models generally follow the trend of the actual PM values. However, there are notable differences in the accuracy of predictions. The MLP model appears to have a closer alignment with the actual values, particularly in capturing the peaks and troughs, whereas the LSTM model shows a larger deviation from the actual values. This visual assessment suggests that the MLP model has a better performance in predicting PM values compared to the LSTM model.

This suggests that while LSTM networks are powerful for sequential data, simpler models like MLP can be more effective for specific time series prediction tasks with the right feature engineering and parameter tuning.

8.2 Comparison of MLP and LSTM Performance Using RMSE

In evaluating machine learning models, one of the most critical metrics is the Root Mean Squared Error. RMSE is widely used to measure the accuracy of models predicting continuous outcomes, providing insight into the average magnitude of the error between predicted and actual values.

The RMSE values for the models are as follows:

Table 16: RMSE Values for the Models

Model	**RMSE**
MLP	1.622
LSTM	2.2677

RMSE values closer to zero indicate a better fit to the data, as they represent lower average prediction errors. In this comparison, the MLP model outperforms the LSTM model, exhibiting a lower RMSE.

This might have happened because the MLP model was using two hidden layers which might have allowed it to learn complex patterns in the data without becoming overly complex itself. On the other hand, the LSTM, with its more intricate architecture might have suffered from over-fitting. This hypothesis is supported by the line plots of training and validation loss. The patterns in those plots indicate that that the LSTM model is learning the training data too well, capturing noise and specific details that do not generalize to unseen data.

9. Conclusion

This study aimed to develop and compare two machine learning models, a Multilayer Perceptron and a Long Short Term Memory network, for predicting PM2.5 concentrations using a dataset from Penrose Station. Through comprehensive data preprocessing, and detailed feature selection based on correlation and regression analysis, the study ensured the dataset's integrity and relevance.

The MLP model was optimized by tuning learning rates and neuron configurations, achieving the best performance with 16 neurons in the first hidden layer and 9 in the second. The LSTM model was refined by adjusting epochs, batch size, and neuron count, with the optimal configuration being 10 neurons in one hidden layer.

Visual and quantitative comparisons of the models revealed that the MLP model outperformed the LSTM model, achieving a lower RMSE (1.622 vs. 2.2677). This suggests that the MLP model, with its simpler architecture and efficient training process, is more suitable for this specific time series prediction task.

The findings of this study underscore the importance of model selection and tuning in achieving accurate predictions for air quality management. The superior performance of the MLP model highlights its potential for practical applications in predicting PM2.5 concentrations, thereby contributing to efforts in improving air quality and protecting public health.