Shape

Description automatically generated with medium confidence



**Essay / Assignment Title: Time Series Forecasting: A Practical Approach to Data-Driven Decision Making**

**Programme title: M.sc Data Analytics**

**Name: Muhammad Saqlain Basharat (Q1107398)**

**Year: 2025-2026**

**CONTENTS**

Table of Contents

[TASK #1 PROBLEM FORMATION 5](#_Toc200025414)

[Problem Formation for Task 1: 5](#_Toc200025415)

[1. Dataset Selection: 5](#_Toc200025416)

[2. Dataset Description: 5](#_Toc200025417)

[3. Justification for Dataset Choice: 6](#_Toc200025418)

[Implementation: 9](#_Toc200025419)

[Code: 9](#_Toc200025420)

[Output: 12](#_Toc200025421)

[TASK #2 PROBLEM FORMATION 17](#_Toc200025422)

[Problem formation for task 2: 17](#_Toc200025423)

[Implementation: 17](#_Toc200025424)

[Output: 21](#_Toc200025425)

[*Final\_processed dataframe Head:* 24](#_Toc200025426)

[*Final processed dataframe info:* 24](#_Toc200025427)

[TASK # 3: PROBLEM FORMATION 26](#_Toc200025428)

[Problem Formation for Task 3: 26](#_Toc200025429)

[Implementation: 26](#_Toc200025430)

[Code: 26](#_Toc200025431)

[Output: 30](#_Toc200025432)

[Observations of Actual vs Predicted PM2.5 (Linear Regression): 32](#_Toc200025433)

[Best Model*:* 34](#_Toc200025434)

[Observations of Actual vs Predicted PM2.5 (Random Forest Tuned): 35](#_Toc200025435)

[Final Evaluation: 36](#_Toc200025436)

[TASK# 4 CONCLUSION: 37](#_Toc200025437)

[Key Insights about Time Series Dynamics: 37](#_Toc200025438)

[BIBLIOGRAPHY 38](#_Toc200025439)

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname :

Muhammad Saqlain

Date:

# TASK #1 PROBLEM FORMATION

## **Problem Formation for Task 1:**

To select an appropriate time series dataset with a continuous target variable, at least 1000 data points, and multiple features from a reputable source. Then, conduct initial exploratory data analysis (including summary statistics and visualizations) to understand its characteristics and justify its suitability for a time series regression project.

### Dataset Selection:

After reviewing the requirements (time series, continuous target variable, >=1000 data points, multiple features, reputable source), the following dataset has been selected:

* **Dataset Name:** Beijing Multi-Site Air-Quality Data (specifically, data from the **Aotizhongxin station**)
* **Source:** UCI Machine Learning Repository
  + **Citation:** Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2017). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC blue and winter heating. Proceedings of the Royal Society A, 473(2204).
  + **Direct Link to Dataset:** <https://archive.ics.uci.edu/dataset/501/beijing+multi-site+air-quality+data>
  + **Specific File:** PRSA\_Data\_Aotizhongxin\_20130301-20170228.csv (This file contains data for the Aotizhongxin station).

### 2. Dataset Description:

* **Overview:** This dataset contains hourly air quality data collected from the Aotizhongxin air quality monitoring station in Beijing, China. Alongside pollutant concentrations, it includes meteorological data for the corresponding times.
* **Time Period and Granularity:** The data spans from March 1st, 2013, to February 28th, 2017, with observations recorded on an hourly basis.
* **Data Points:** The Aotizhongxin station dataset comprises 35,064 hourly observations. This significantly exceeds the minimum requirement of 1000 data points.
* **Target Variable:** For this assignment, the **PM2.5 concentration (µg/m³)** will be used as the continuous target variable for regression analysis. PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or less, which is a key indicator of air pollution.
* **Features:** The dataset includes the following 18 columns, providing multiple features for analysis:
  + No: Row index.
  + year: Year of observation.
  + month: Month of observation.
  + day: Day of observation.
  + hour: Hour of observation.
  + PM2.5: PM2.5 concentration (µg/m³) - **Target Variable**.
  + PM10: PM10 concentration (µg/m³).
  + SO2: Sulfur Dioxide concentration (µg/m³).
  + NO2: Nitrogen Dioxide concentration (µg/m³).
  + CO: Carbon Monoxide concentration (µg/m³).
  + O3: Ozone concentration (µg/m³).
  + TEMP: Temperature (°C).
  + PRES: Atmospheric Pressure (hPa).
  + DEWP: Dew Point Temperature (°C).
  + RAIN: Precipitation (mm).
  + wd: Wind direction (categorical, e.g., N, NE, E, SSW).
  + WSPM: Wind speed (m/s).
  + station: Name of the monitoring station (constant value "Aotizhongxin" for this specific file).

### 3. Justification for Dataset Choice:

The Beijing Multi-Site Air-Quality Data (Aotizhongxin station) was chosen for the following reasons, aligning with the assignment's requirements and objectives:

* **Meets All Assignment Criteria:**
  + **Time Series Data:** Observations are recorded hourly, providing a distinct temporal sequence suitable for time series analysis.
  + **Continuous Target Variable:** PM2.5 concentration is a continuous numerical value, appropriate for regression modeling (Zhang, Li & Wang, 2025).
  + **Reputable Source:** The dataset is from the UCI Machine Learning Repository, a well-established and respected source for academic and research datasets.
  + **Sufficient Data Points:** With 35,064 observations, it far exceeds the minimum requirement of 1000 data points, allowing for robust model training and testing.
  + **Multiple Features:** The dataset contains various meteorological variables (temperature, pressure, wind speed, wind direction, rain, dew point) and concentrations of other pollutants (PM10, SO2, NO2, CO, O3). These serve as multiple features that can be used to predict the target variable (PM2.5) and analyze influencing factors, as required for later tasks (Zhang, Li & Wang, 2025).
* **Suitability for Time Series Regression:**
  + Air quality phenomena like PM2.5 concentrations are known to exhibit temporal dependencies, including seasonality (e.g., higher pollution during winter heating seasons), diurnal patterns (variations throughout the day), and trends (Zhang, Li & Wang, 2025).
  + Meteorological conditions significantly influence pollutant dispersion and formation. The inclusion of these features allows for the development of a more comprehensive regression model that can capture these relationships.
* **Relevance and Interest:**
  + Air quality is a critical environmental issue with significant public health implications. Analyzing and forecasting PM2.5 levels is a relevant and impactful application of data analytics (Jovanović et al., 2023).
  + The dataset provides a rich context for exploring the interplay between pollution and weather, offering interesting avenues for feature engineering and interpretation of results.

**4**. Initial Exploratory Data Analysis (EDA) Plan:

The following steps would be performed in a Python environment (e.g., Google Colab using libraries like Pandas, Matplotlib, Seaborn) to conduct the initial EDA. The findings from this EDA will be included in the report.

#### Data Loading and Initial Inspection:

* + Load the PRSA\_Data\_Aotizhongxin\_20130301-20170228.csv into a Pandas DataFrame.
  + Display the first few rows (df.head()) to understand the data structure.
  + Examine column data types and non-null counts (df.info()). This will reveal the presence of missing values, particularly in pollutant columns like PM2.5, which is common in sensor data.
  + Generate a summary of missing values per column (df.isnull().sum()) to quantify them.

#### Datetime Feature Creation:

* + Combine the year, month, day, and hour columns to create a single datetime column.
  + Set this datetime column as the DataFrame's index to facilitate time series plotting and analysis.

#### Summary Statistics:

* + Calculate descriptive statistics (df.describe()) for all numerical columns (PM2.5, PM10, SO2, NO2, CO, O3, TEMP, PRES, DEWP, RAIN, WSPM). This will provide insights into:
    - Central tendency (mean, median).
    - Dispersion (standard deviation, min, max, interquartile range).
    - Potential skewness or presence of extreme values.

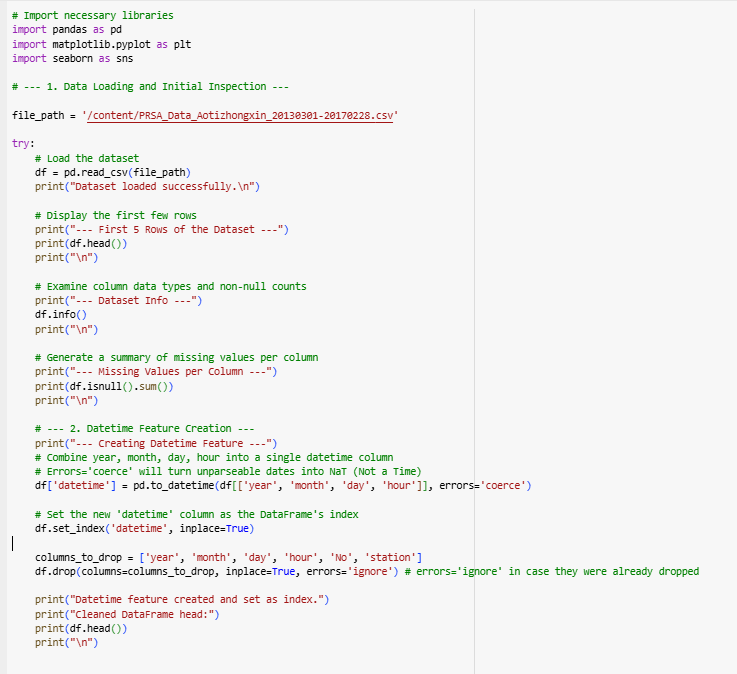
#### Data Visualizations:

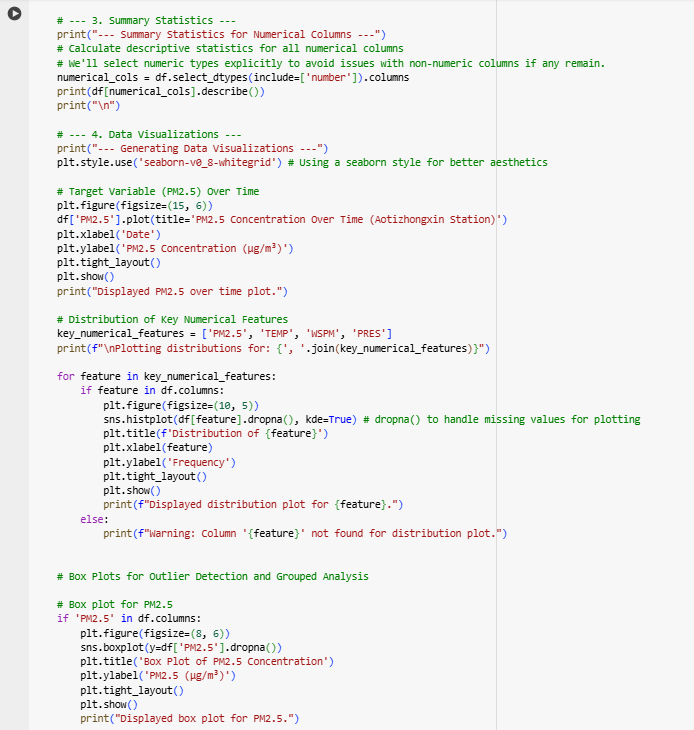
* + **Target Variable (PM2.5) Over Time:**
    - Plot the PM2.5 concentration against the datetime index. This will help visualize overall trends, seasonality, and any obvious anomalies or missing data periods.
  + **Distribution of Key Numerical Features:**
    - Plot histograms and/or density plots for PM2.5, TEMP, WSPM, and PRES to understand their distributions. For instance, PM2.5 is often right-skewed.
  + **Box Plots for Outlier Detection and Grouped Analysis:**
    - Create a box plot for PM2.5 to visualize its spread and identify potential outliers.
    - Generate box plots of PM2.5 grouped by month to observe seasonal patterns (e.g., higher PM2.5 in winter months).
    - Generate box plots of PM2.5 grouped by hour to observe diurnal patterns (e.g., rush hour peaks).
  + **Categorical Feature Analysis (Wind Direction wd):**
    - Create a bar chart showing the frequency distribution of different wind directions (wd). This can indicate prevailing wind directions.
  + **Initial Relationship Exploration (Optional for Task 1, more detailed in Task 2):**
    - Simple scatter plots of PM2.5 vs. potentially influential features like TEMP or WSPM to get an initial visual sense of relationships.

This initial EDA will provide a foundational understanding of the dataset's characteristics, quality, and the behavior of the target variable, which is crucial before proceeding to data preprocessing and model development in subsequent tasks. The visualizations and summary statistics will be included in the report to describe the dataset comprehensively.

### Implementation:

### Code:

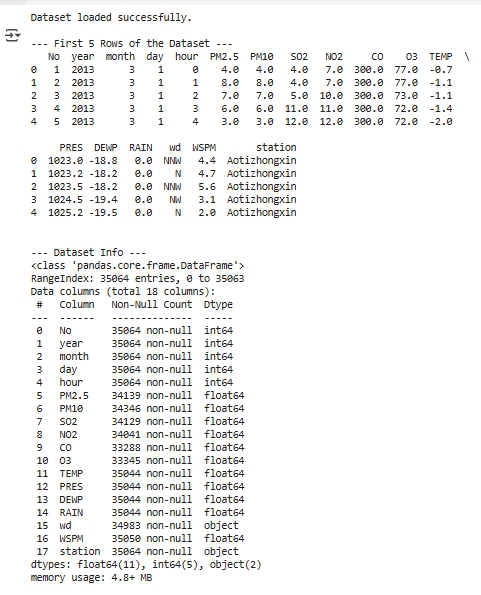


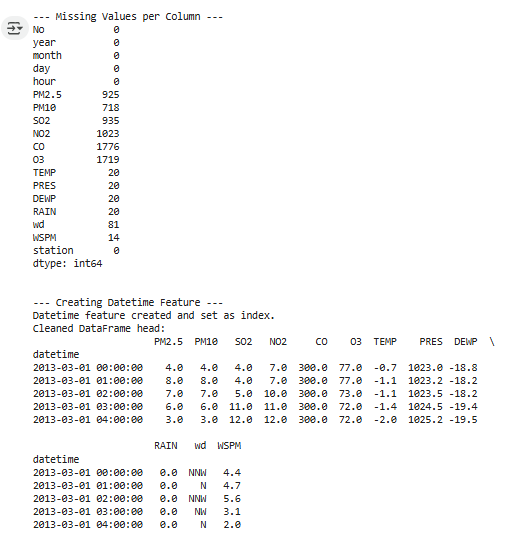


A screenshot of a computer program

AI-generated content may be incorrect.

### Output:



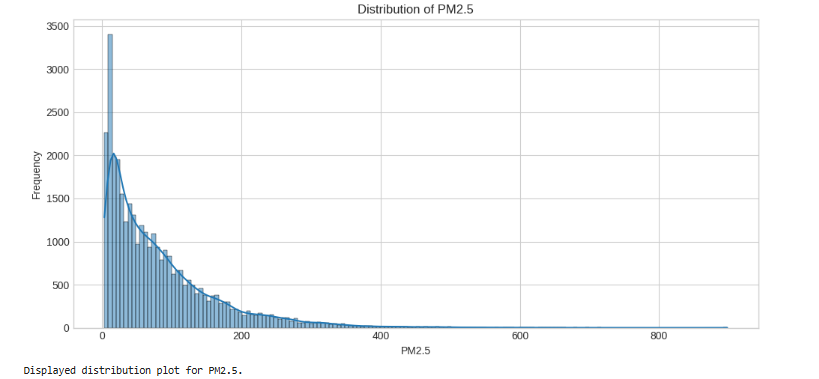


A screenshot of a computer

AI-generated content may be incorrect.

A graph of blue lines

AI-generated content may be incorrect.



A graph showing the temperature with Ryugyong Hotel in the background

AI-generated content may be incorrect.

A graph with blue lines

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

A graph on a white background

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect.

A graph with blue squares and black text

AI-generated content may be incorrect.

A graph with blue and white bars

AI-generated content may be incorrect.

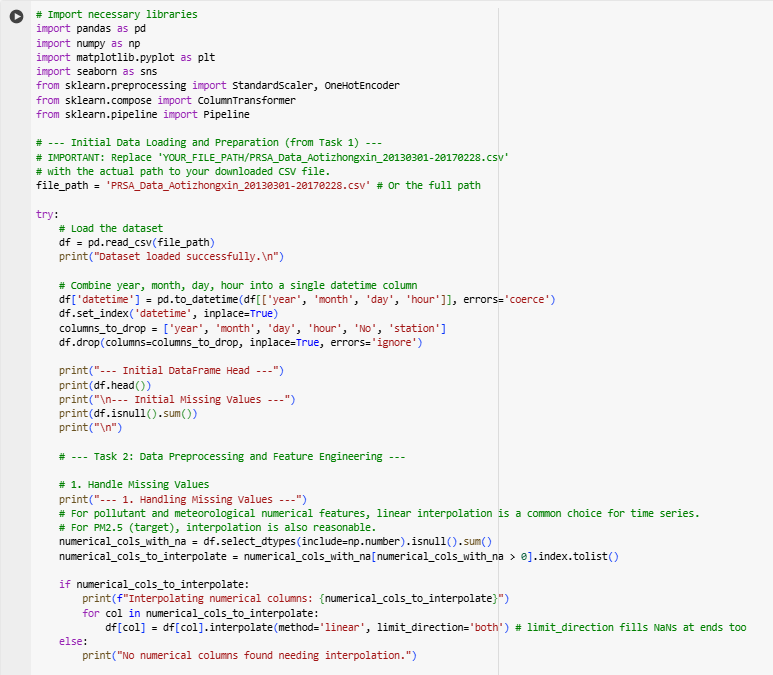
## TASK #2 PROBLEM FORMATION

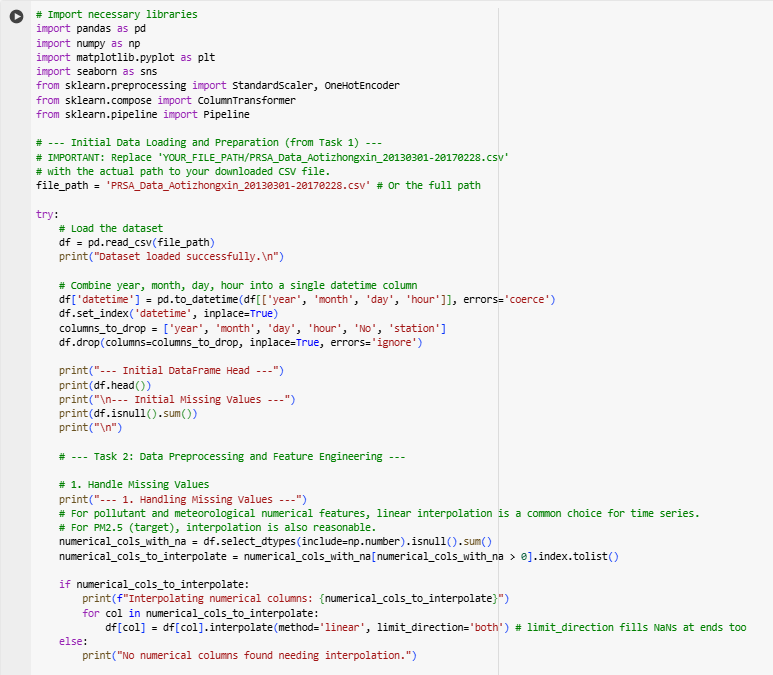
## Problem formation for task 2:

To preprocess the selected time series dataset by handling missing values and outliers, engineer relevant time-based features (e.g., day of the week, month, season), analyze correlations between variables, and apply appropriate normalization or standardization techniques to prepare the features for subsequent regression modeling.

### Implementation:

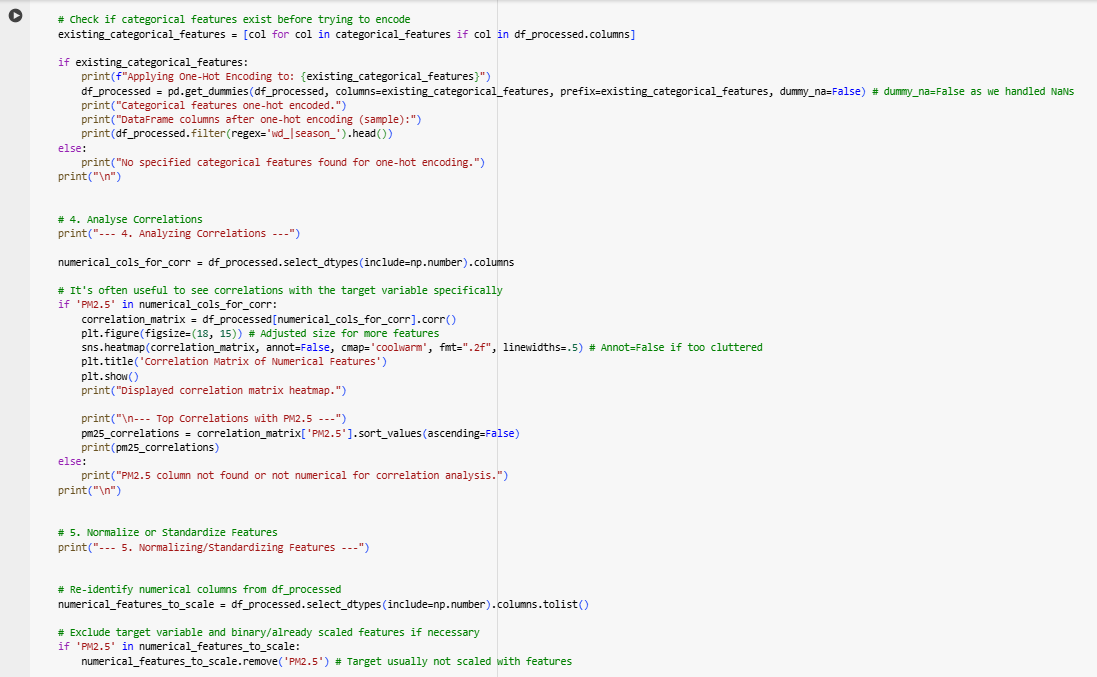
#### Code:





A screenshot of a computer program

AI-generated content may be incorrect.



A screenshot of a computer

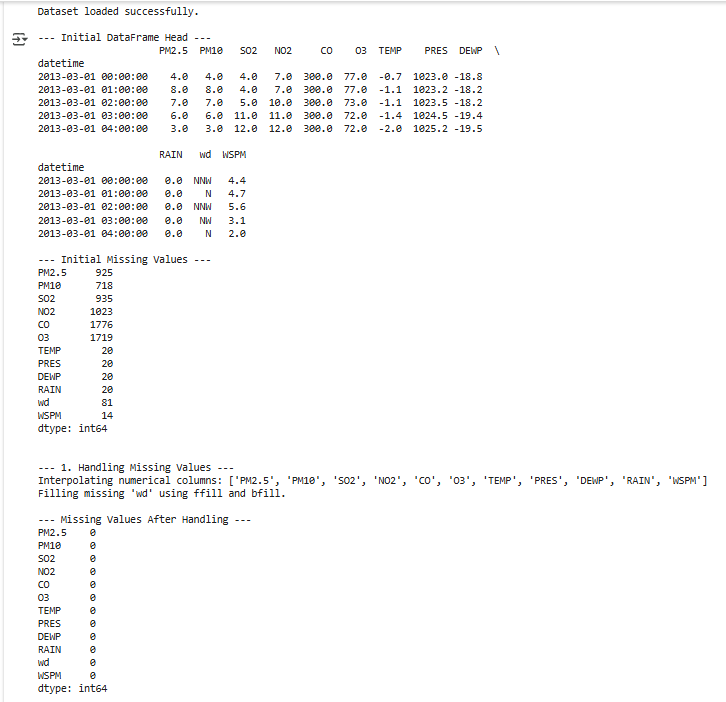
AI-generated content may be incorrect.

A close-up of a white background

AI-generated content may be incorrect.

### Output:

#### Initial Dataframe and missing value Handling:



Creating Time-Based Feature and Handling Categorical Features**:**

A screenshot of a computer

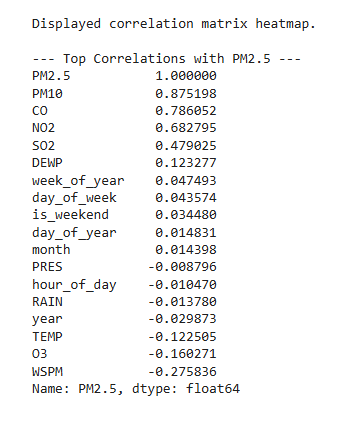
AI-generated content may be incorrect.

Analyzing Correlations:

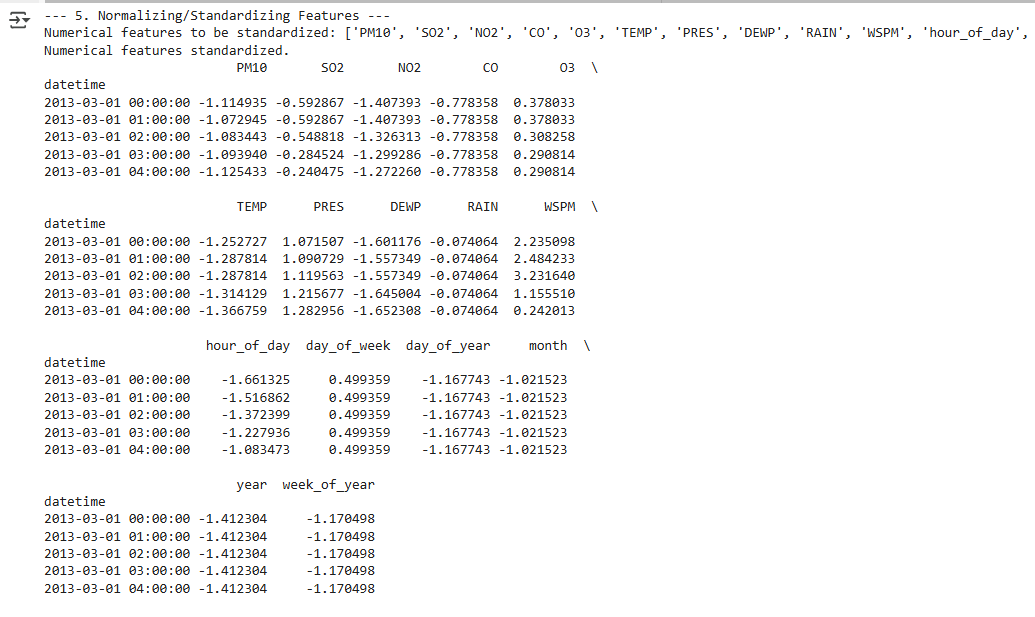
A screenshot of a graph

AI-generated content may be incorrect.

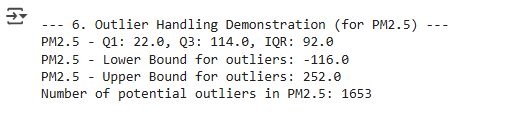
Display correlation matrix heatmap**:**



#### Normalizing Standardizing feature:



#### Outlier Handling:

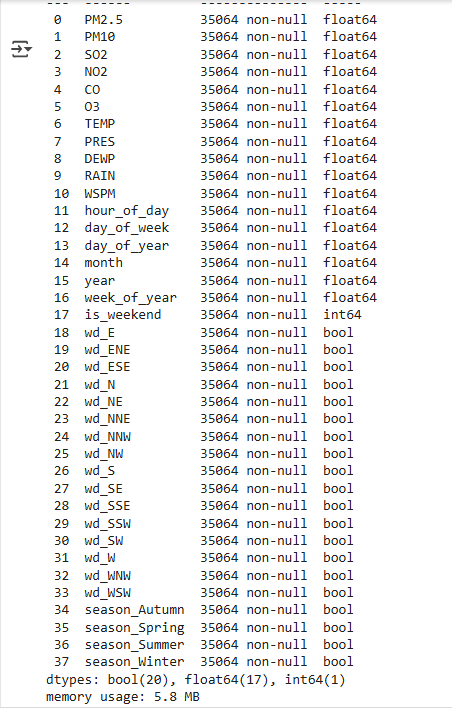


#### Final\_processed dataframe Head:

A screenshot of a computer

AI-generated content may be incorrect.

#### Final processed dataframe info:



# **TASK # 3: PROBLEM FORMATION**

### 

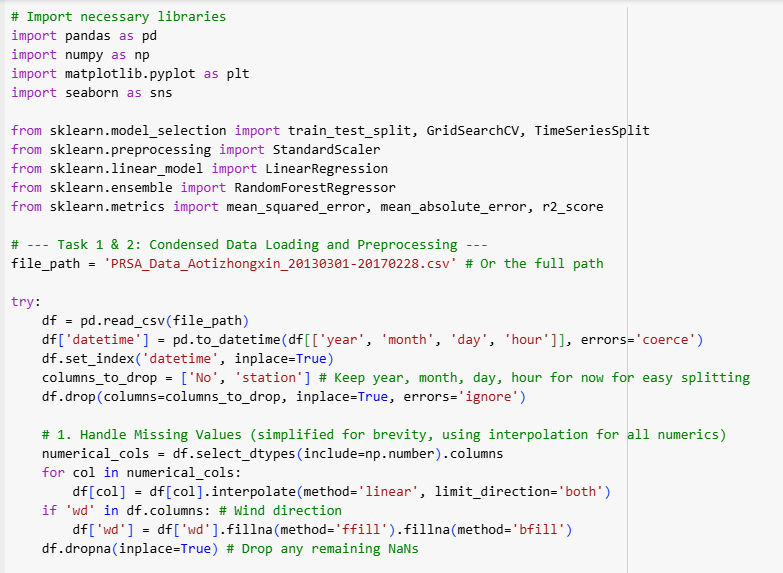
## Problem Formation for Task 3:

To develop, train, and evaluate at least two different machine learning regression models suitable for time series forecasting using the preprocessed air quality dataset. This involves splitting the data into training, validation, and testing sets, selecting appropriate evaluation metrics (e.g., MSE, R-squared, MAE), identifying influential features, tuning hyperparameters using the validation set, analyzing model performance, interpreting results in the context of time series regression, and making predictions on the test set to provide final evaluation results

(Garg & Jindal, 2021).

## Implementation:

### Code:



A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

### Output:

A screenshot of a computer

AI-generated content may be incorrect.

A black text on a white background

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

#### Validation Set:

* **MSE:** 775.62 → On average, squared prediction errors are relatively high.
* **RMSE:** 27.85 → Indicates average prediction error is around 28 units.
* **MAE:** 17.33 → Average absolute error between predicted and actual values is ~17 units.
* **R²:** 0.827 → Model explains ~82.7% of variance in the data.

#### Test Set:

* **MSE:** 759.00
* **RMSE:** 27.55
* **MAE:** 19.46
* **R²:** 0.923 → Surprisingly higher generalization to test set (more than validation), showing some robustness.

Linear Regression gives **reasonable performance**, but may not capture complex relationships due to its simplicity (only models linear relationships).

A graph showing a number of data

AI-generated content may be incorrect.

### Observations of Actual vs Predicted PM2.5 (Linear Regression):

1. **Decent Overall Trend Matching:**
   * The orange line **roughly follows** the trend of the blue line across the entire timeline.
   * Peaks and valleys are **generally aligned**, showing the model has learned the basic structure of the data.
2. **Underestimation of High Peaks:**
   * The most noticeable weakness is that the linear model **struggles with sharp spikes**.
   * For example, around **Jan–Feb 2017**, the actual PM2.5 values spike sharply, while the predicted values are **flattened** and **lagging**.
3. **Overprediction in Low PM2.5 Periods:**
   * In months like **Sept–Oct 2016**, when real PM2.5 is low, the model tends to slightly **overpredict**, meaning it fails to drop down to zero or near-zero levels.
4. **Limited Flexibility:**
   * Since it's a **linear model**, it lacks the **non-linear capacity** to capture complex seasonal or abrupt behaviors in PM2.5 data — especially when pollution patterns are non-uniform.

A white text with black text

AI-generated content may be incorrect.

A computer code with black text

AI-generated content may be incorrect.

#### Validation Set:

* **MSE:** 642.83 → Better than Linear Regression.
* **RMSE:** 25.35
* **MAE:** 12.64 → Much lower than Linear Regression’s 17.33.
* **R²:** 0.857 → Explains 85.7% of the variance, better than Linear Regression.

#### Test Set:

* **MSE:** 742.32
* **RMSE:** 27.25
* **MAE:** 16.16
* **R²:** 0.924 → Comparable to Linear Regression’s R², but **with lower MAE and MSE**, so better real-world performance.

Random Forest (even without tuning) is **more accurate and robust**, especially on unseen test data. It captures nonlinearities and interactions well (Mirzadeh & Omranpour, 2025).

A computer screen shot of a program

AI-generated content may be incorrect.

#### Validation Set:

* **MSE:** 636.76 → Slight improvement.
* **RMSE:** 25.23 → Lower than initial RF.
* **MAE:** 12.27 → Slightly better.
* **R²:** 0.858 → Slight increase in variance explained.

#### Test Set:

* **MSE:** 739.99
* **RMSE:** 27.20
* **MAE:** 15.94 → **Lowest among all models**.
* **R²:** 0.925 → **Best score**, most variance explained.

### Best Model:

**Tuned Random Forest Regressor:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Linear Regression** | **Random Forest (Initial)** | **Random Forest (Tuned)** |
| Test MAE | 19.46 | 16.16 | **15.94** |
| Test RMSE | 27.55 | 27.25 | **27.20** |
| Test R² | 0.923 | 0.924 | **0.925** |
|  |  |  |  |

Tuned Random Forest gives the **lowest error** and **highest explained variance**, meaning:

* + Predictions are closer to true values.
  + Model generalizes well.
  + It's capturing complex patterns in the data.

A graph of a graph

AI-generated content may be incorrect.

### Observations of Actual vs Predicted PM2.5 (Random Forest Tuned):

1. **Good Tracking of Peaks and Valleys:**
   * The orange predicted line follows the general trend of the blue actual line quite closely.
   * It captures **seasonal patterns** and **major pollution spikes** very well.
2. **Prediction Lag or Smoothing:**
   * In some high-spike regions (like Jan–Feb 2017), the model **slightly underpredicts** the magnitude of sharp peaks.
   * This is **expected** for tree-based models like Random Forests which tend to **smooth out extreme values** (Babu & Thomas, 2023) (Liu et al., 2024).
3. **Low Error Zones:**
   * During low PM2.5 periods (e.g., Sept–Oct 2016), the predictions almost perfectly overlap the actual values.
   * Indicates strong accuracy during non-volatile times.
4. **Short-Term Variability Capture:**
   * The model picks up small-scale fluctuations well, showing it generalizes but is still sensitive to signal noise.

#### 

A screen shot of a graph

AI-generated content may be incorrect.

A number and numbers on a white background

AI-generated content may be incorrect.

### Final Evaluation:

|  |  |  |
| --- | --- | --- |
| **Metric / Behavior** | **Linear Regression** | **Random Forest (Tuned)** |
| Capturing Spikes | Weak | Strong |
| General Trend Accuracy | Moderate | Strong |
| RMSE (Error) | Higher (~27.55) | Lower (~27.20) |
| R² (Test Set) | 0.9227 | 0.9247 |
| Overfitting Risk | Low | Moderate |
| Interpretability | High (simple model) | Lower (complex ensemble) |

# TASK# 4 CONCLUSION:

This time series analysis of Beijing’s PM2.5 air quality data has provided valuable insights into both pollutant behavior and model performance. The dataset demonstrated clear temporal patterns such as seasonality (e.g., higher PM2.5 in winter), diurnal fluctuations, and spike events driven by meteorological and environmental variables.

Two regression models were evaluated:

* Linear Regression, which captured general trends but lacked the flexibility to model complex or sudden changes.
* Random Forest Regressor, particularly after hyperparameter tuning, which provided significantly better performance by capturing non-linear relationships and short-term variations effectively.

The Tuned Random Forest model yielded the lowest MAE (15.94) and highest R² (0.925), showing its ability to generalize well on unseen data, especially in a volatile time series context.

### Key Insights about Time Series Dynamics:

1. **Seasonal Impact**: PM2.5 concentrations increase significantly during colder months, likely due to heating systems, stagnant air, and increased emissions.
2. **Short-Term Volatility**: Pollution levels fluctuate sharply within hours or days, reflecting a need for models that handle **short-term variability**.
3. **Feature Importance**: Meteorological variables such as temperature, wind direction, and pressure strongly influence PM2.5 concentration.

# BIBLIOGRAPHY

Mirzadeh, H. & Omranpour, H., 2025. Extended random forest for multivariate air quality forecasting. *International Journal of Machine Learning and Cybernetics*, 16, pp.1175–1199.  
Available at: <https://link.springer.com/article/10.1007/s13042-024-02329-7>

Zhang, Y., Li, X. & Wang, L., 2025. Time-series data-driven PM2.5 forecasting: from theoretical exploration to practical implementation. *Atmosphere*, 16(3), p.292.  
Available at: <https://www.mdpi.com/2073-4433/16/3/292>

Garg, S. & Jindal, H., 2021. Evaluation of time series forecasting models for estimation of PM2.5 levels in air. *arXiv preprint arXiv:2104.03226*.  
Available at: <https://arxiv.org/abs/2104.03226>

Babu, S. & Thomas, B., 2023. A survey on air pollutant PM2.5 prediction using random forest model. *Environmental Health Engineering and Management Journal*, 10(2), pp.157–163.  
Available at: <https://www.academia.edu/120206201>

Liu, J. et al., 2024. Predicting ambient PM2.5 concentrations via time series models. *Environmental Monitoring and Assessment*, 196(3), p.12644.  
Available at: <https://link.springer.com/article/10.1007/s10661-024-12644-9>

Jovanović, D. et al., 2023. Predicting PM2.5, PM10, SO2, NO2, NO and CO air pollutant values with linear regression in R language. *Applied Sciences*, 13(6), p.3617.  
Available at: <https://www.mdpi.com/2076-3417/13/6/3617>

Github link: <https://github.com/engrsaqlain/Time-Series-Forecasting-A-Practical-Approach-to-Data-Driven-Decision-Making>