PREDICTING AQI WITH MACHINE LEARNING

Presentation By:
Jirapon Kluaymaingarm
Napassorn Sripum
Aye Nyein Thu
Patsakorn Tangkachaiyanunt

ABSTRACT

Objective:
- Predict PM2.5 levels in Bangkok using machine learning.

Data Used:
- Pollutant Levels: PM2.5
- Meteorological Factors: Temperature, Humidity, latitude, longitudes, sea level, number of car, number of tree
- Timestamp Factor : Day of week, month, hour, time step

Models Compared:
- Linear Regression (Baseline)
- Lasso Regression
- Ridge Regression
- Decision Trees
- Random Forest
- XGBoost

Impact:
Machine learning can improve air quality forecasting. A grid-based mapping system could enhance public awareness & policymaking.

INTRODUCTION

- Air pollution is a major public health concern, especially in urban areas with fluctuating PM2.5 levels.
- Most research focuses on deep learning, overlooking a broader range of predictive models.
- Few studies focus on AQI prediction in Bangkok, despite its severe seasonal air pollution. Most prior research targets other Asian cities
- Developed a machine learning-based model for predicting AQI in Bangkok, incorporating environmental & meteorological factors.

CONTRIBUTION

- **Comparative Model Evaluation** – Assess multiple ML models (Linear Regression (Baseline), Lasso Regression, Ridge Regression, Decision Trees, Random Forest, XGBoost) to determine the best PM2.5 predictor.
- **GSI & Environmental Integration** – Incorporate geospatial, environmental, and demographic factors for improved accuracy.
- **Bangkok-Specific Forecasting** – Develop a localized AQI model tailored to Bangkok's unique pollution patterns.
- **Key Predictor Analysis** – Identify and rank critical PM2.5 factors to guide policy and urban planning.

RESEARCH QUESTION

01 First Question

Which model predicts PM 2.5 the most accurately and precisely?

02 Second Question

What are the features that influence the results of PM2.5 prediction the most?

03 Third Question

How accurately does the general public interpret and use data visualizations to improve their awareness of health issues?

LITERARY REVIEWS

Previous research has extensively explored the use of machine learning models for PM2.5 prediction, particularly in South Asia, where air pollution is a significant environmental concern

01 Literary Review

Ganguli et al., 2025 compared statistical models such as ARIMA and SARIMA with deep learning approaches like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), demonstrating that deep learning models consistently outperform traditional methods

02 Literary Review

Mohammadi et al., 2024 highlighting the effectiveness of LSTM in handling complex meteorological factors, including temperature, humidity, and wind speed

LITERARY REVIEWS

Previous research has extensively explored the use of machine learning models for PM2.5 prediction, particularly in South Asia, where air pollution is a significant environmental concern

03 Literary Review

Shuvo et al., 2021 developed spatiotemporal PM2.5 prediction models, addressing the challenges posed by sparse air quality monitoring networks in Nepal and Bangladesh

04 Literary Review

Sookchaiya & Phimoltares (2022) compared various machine learning models for predicting PM2.5 and PM10 levels in Chiang Mai Province but did not integrate their findings with spatial visualization tools for decision-making

GSI REVIEWS

Geographic and Spatial Information (GSI) improves PM2.5 prediction by integrating location-based factors, satellite-derived data, and advanced modeling techniques to enhance accuracy for urban planning and public health management.

05 Literary Review

Alam et al., 2023 research on air quality forecasting in Bishkek, Kyrgyzstan, has explored various machine learning models such as CatBoost, LightGBM, XGBoost, and Random Forest, demonstrating their effectiveness in predicting pollutants like PM2.5, PM10, and CO

06 Literary Review

Lin & Li (2021) developed an LSTM-based air quality prediction model that incorporated land-use patterns, traffic density, and meteorological conditions, leading to improved forecasting accuracy

07 Literary Review

Di et al. (2016) used satellite-derived AOD data alongside meteorological and land use variables to estimate ground-level PM2.5 concentrations in the Southeastern United States

GSI REVIEWS

Geographic and Spatial Information (GSI) improves PM2.5 prediction by integrating location-based factors, satellite-derived data, and advanced modeling techniques to enhance accuracy for urban planning and public health management.

08 Literary Review

Kermani et al. (2024) utilized AOD, meteorological data, and land-use information to refine PM2.5 predictions, showing that geospatial factors remain critical for air quality forecasting.

09 Literary Review

Zhang & Wang (2022) explored air quality prediction using Random Forest models in China, prioritizing model accuracy over practical implementation through geospatial mapping

10 Literary Review

Kumar & Singh (2023) expanded on this by conducting a comparative study of machine learning models for air quality prediction in smart cities, emphasizing the need for interactive, grid-based AQI mapping systems to aid decision-making for urban planners and policymakers.

REVIEW ON JAPAN'S AIR QUALITY RESEARCH

Japanese air quality research provides critical insights into long-term pollutant reductions, emission control effectiveness, and the need for advanced predictive models to address transboundary pollution and secondary pollutant formation.

11 Literary Review

Yamagami et al. (2021) studied long-term PM2.5 trends from 2003 to 2018 in Nagoya City, revealing a 53% reduction in PM2.5 levels due to stringent diesel emission regulations and changes in secondary aerosol formation. However, their findings also emphasized the impact of transboundary pollution from China, particularly in sulfate (SO_4^{2-}) and nitrate (NO_3^-) concentrations.

12 Literary Review

Ito et al. (2021) provided a comprehensive 30-year air quality analysis in Japan, highlighting a 26–30% decline in PM2.5 and 55–65% reduction in suspended particulate matter (SPM) mainly due to emission regulation, yet noting that ozone (O_3) levels continue to rise due to secondary particle formation and transboundary pollution.

REAL API DATA RESEARCH

Previous studies have leveraged real-time data to develop innovative approaches for air quality monitoring, integrating advanced technologies such as machine learning models, satellite retrievals, and low-cost sensor networks to enhance accuracy, coverage, and accessibility of air pollution assessments

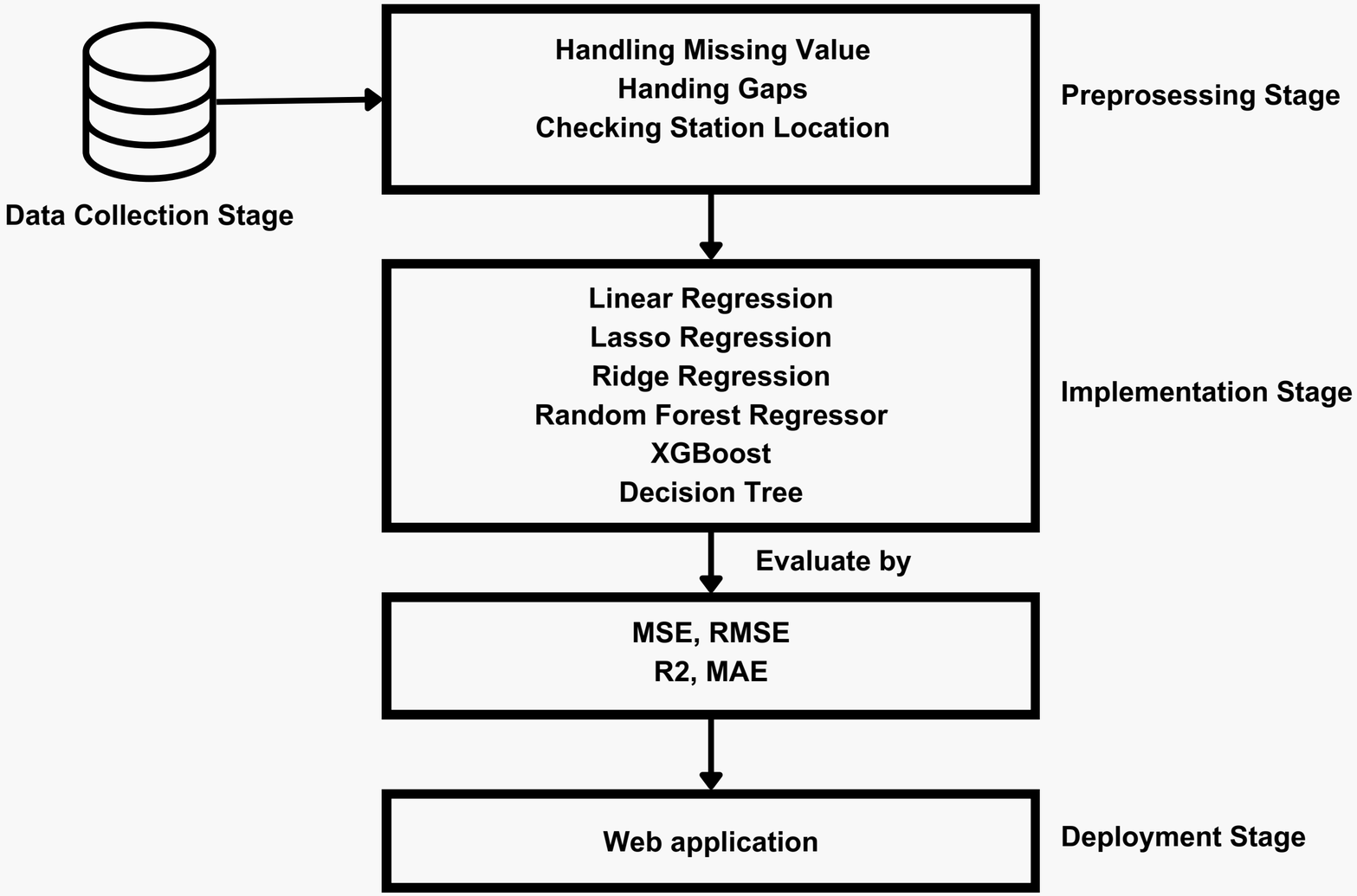
13 Literary Review

Geng et al. (2021) introduced the Tracking Air Pollution in China (TAP) database, which provides near real-time, full-coverage $PM_{2.5}$ data at a 10 km spatial resolution across China since 2000. This database integrates multiple data sources, including ground measurements, satellite retrievals, emission inventories, and chemical transport model simulations, to offer timely updates and historical records of air pollution levels

14 Literary Review

Mendez et al. (2022) utilized low-cost sensors to assess $PM_{2.5}$ concentrations in four South Texan cities along the U.S.–Mexico border. This year-long study demonstrated the feasibility of using affordable sensing technology for localized air quality monitoring, providing valuable data for public health assessments and policy-making.

METHODOLOGY



METHODOLOGY

01 Data collection

Dataset1

- Dataset name : “Time Series Air Quality Data of Bangkok per Hour”
- Period : From January 1, 2024 to Present
- Acquired from : Open Weather API
- Containing : 14841 row and 19 distinct features

Dataset2

- Dataset name : “Factory Data of Bangkok”
- Period : 2020
- Acquired from : The Department of City Planning (BMA CPUD)
- Containing : Factory data of each station in Bangkok.

Dataset3

- Dataset name : “Statistic data of Bangkok”
- Period : 2024
- Acquired from : Strategy and evaluation department
- Containing : Population statistics, including total population, population density, and green space ratio

OpenWeather

สำนักการวางผังและพัฒนาเมือง
CITY PLANNING AND URBAN DEVELOPMENT DEPARTMENT

กทม สำนักยุทธศาสตร์และประเมินผล
STRATEGY AND EVALUATION DEPARTMENT

METHODOLOGY

01 Data collection

20 Feature

Variable	Explanation	Unit of measure
PM2.5	Particulate matter with an aerodynamic diameter less than or equal to a nominal 2.5 micrometers	µg/m3
Temperature	The measure of air temperature in the atmosphere	Celcius (°C)
RH	Relative Humidity – the percentage of moisture in the air	Percent (%)
NO2	Nitrogen Dioxide – a major air pollutant primarily from vehicle emissions	Parts per billion (ppb)
CO	Carbon Monoxide – a toxic gas primarily from incomplete combustion	Parts per million (ppm)
Latitude	Angle between the straight line in the certain point and the equatorial plane	Degrees
Longitude	Angle pointing west or east from the Greenwich Meridian	Degrees
Sea level	An average surface level of one or more among Earth's coastal bodies of water	meter (m)
Traffic Level	Traffic scale level with in that area	1-10
Number of Factory	Number of factory with in that area	Factory (units)
Factory Area	Area of the given factory space	Square metre (m^2)
Green Space	Density of greenspace per population	Square metre per person
Month	12 Different months in a year	Months (mo.)
Day	Seven Different day in a week	Days (d)
Hour	A unit of time in a day	Hour (h)
Population	Total number of people living in a given area.	People (person)
population density	Number of people per unit area, typically measured to understand urbanization levels.	square meter per people (m^2/people)
household	Total number of households in the given area, where a household refers to individuals living together in one dwelling unit.	Households (units)
greenspace	The total area covered by parks, forests, or other vegetated land within a given area.	square meter per people (m^2/population)
season	The weather patterns and temperature variations throughout the year.	Hot/ Rainy/ Cool season.

Table 1 Notation with the explantation and relative unit of measures

METHODOLOGY

02 Data Preprocessing

Performed these steps to prepare the data for predictive modeling :

- **Acquire Data:** Data were acquired via an API and also through a CSV file. Both sources were converted into a dataframe format to facilitate further processing.

- **Merging Data:** We integrated datasets originating from disparate sources to assemble a comprehensive and cohesive dataset, ensuring uniformity in data handling and analysis.

- **Data Mapping:** We enhanced the dataset's utility by mapping each season to the appropriate months.

- **Data Encoding:** To facilitate statistical analysis, we converted all categorical string data related to seasons, which were defined in the mapping section, into numerical formats.

- **Data Extraction:** Using the dt accessor, we extracted critical time-based elements such as the day of the week, month, and hour from the 'datetime' column.

METHODOLOGY

02 Data Preprocessing

Performed these steps to prepare the data for predictive modeling :

- **Handling Missing Values:** To ensure data completeness, missing PM2.5 values were first interpolated using a linear interpolation method within each monitoring station, preserving local trends. Any remaining missing values were then filled using a backward fill (bfill) approach, ensuring that no gaps remained in the dataset. Additionally, the data were sorted by latitude, longitude, and timestamp to maintain temporal consistency before interpolation.

- **Handling Gaps:** To ensure a complete and continuous dataset, a full range of timestamps was generated using hourly intervals from the minimum to the maximum recorded datetime in the dataset. Next, of all possible station locations (latitude, longitude) and timestamps was created to form a comprehensive reference grid. To align the existing dataset with this grid, each station's latitude and longitude were combined into tuples for efficient merging. The dataset was then merged with the full datetime-station grid, filling in missing timestamps for each station. Finally, the latitude and longitude values were extracted from the merged tuples, and redundant columns were dropped to maintain dataset integrity.

- **Checking Station Location:** Since we acquired data from API sources, we need to figure out which station the data were pulled from to match with the exact GSI (longitude and latitude).

METHODOLOGY

03 Implementation

Utilized several specialized Python libraries, notably Statsmodels and Scikit-learn

Workspace :

- CPU: Corei7 13700HX
- GPU: RTX 4060 laptop 8GB VRAM
- RAM: 24GB

Dataset :

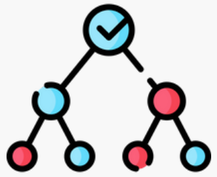
	%	Sample
Training Set	80	X
Testing Set	20	X

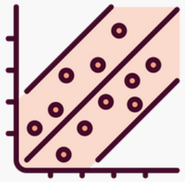
METHODOLOGY

03 Implementation

Model :

	Model name	Evaluation Matrix
Regession	Linear Regression Lasso Regression Ridge Regression Decision Trees Random Forest XGBoost	R^2 score Mean Absolute Error (MAE) Mean Squared Error (MSE) Root Mean Squared Error (RMSE)



METHODOLOGY

04 Deployment

Using : Flask-based web application

Providing :

Interface

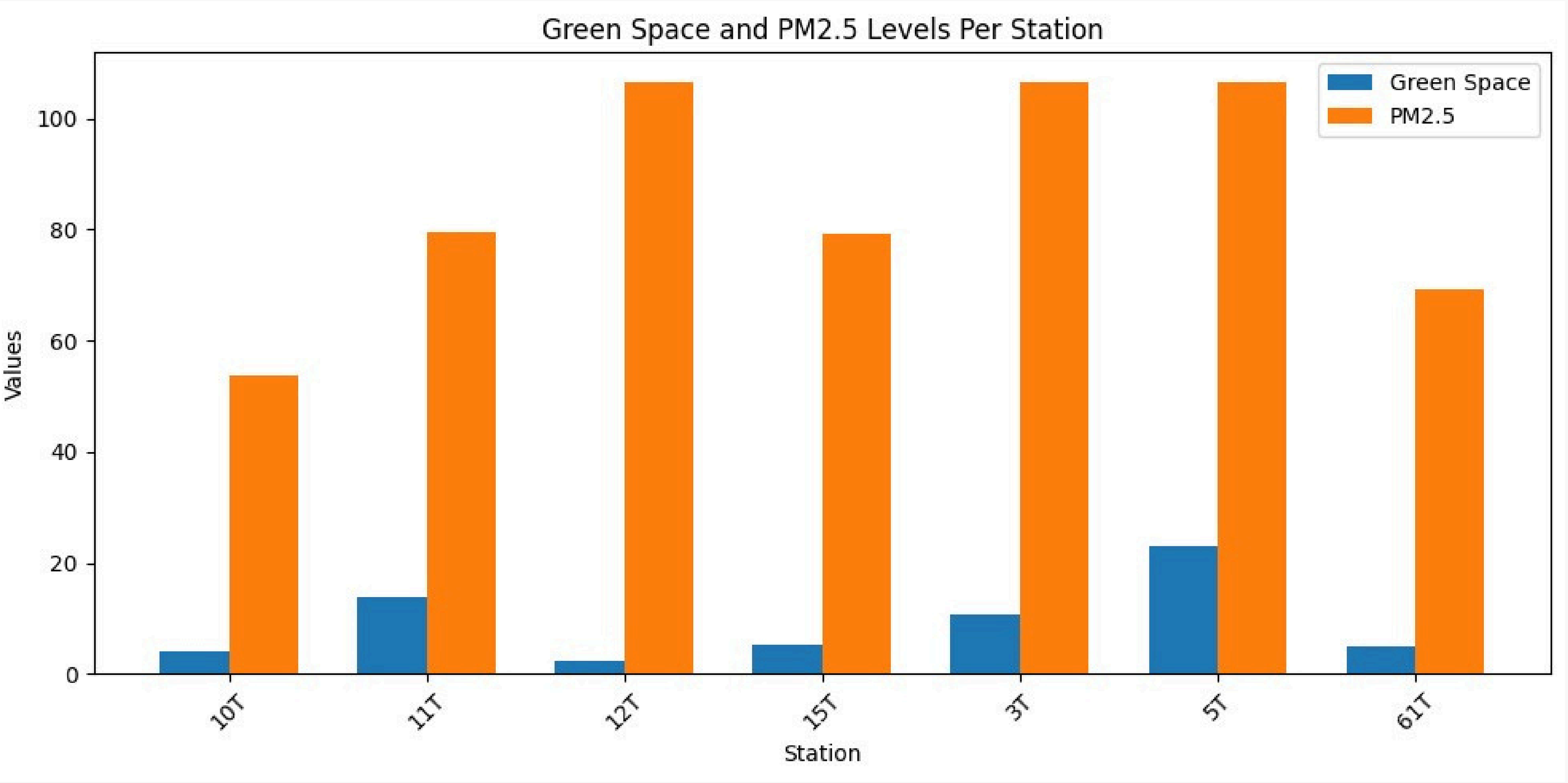
RESULT AND DISCUSSION

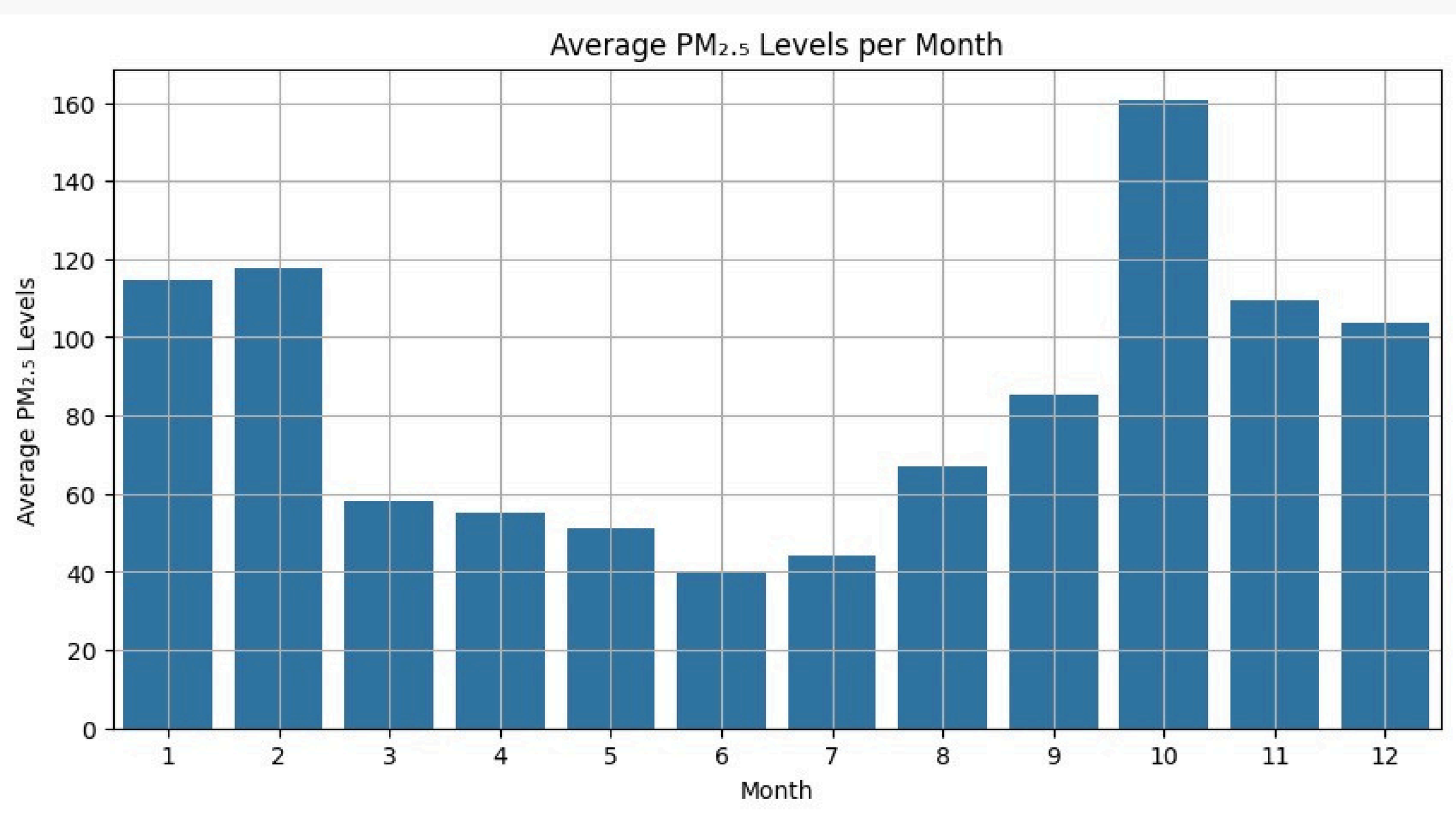
01 Variable importance

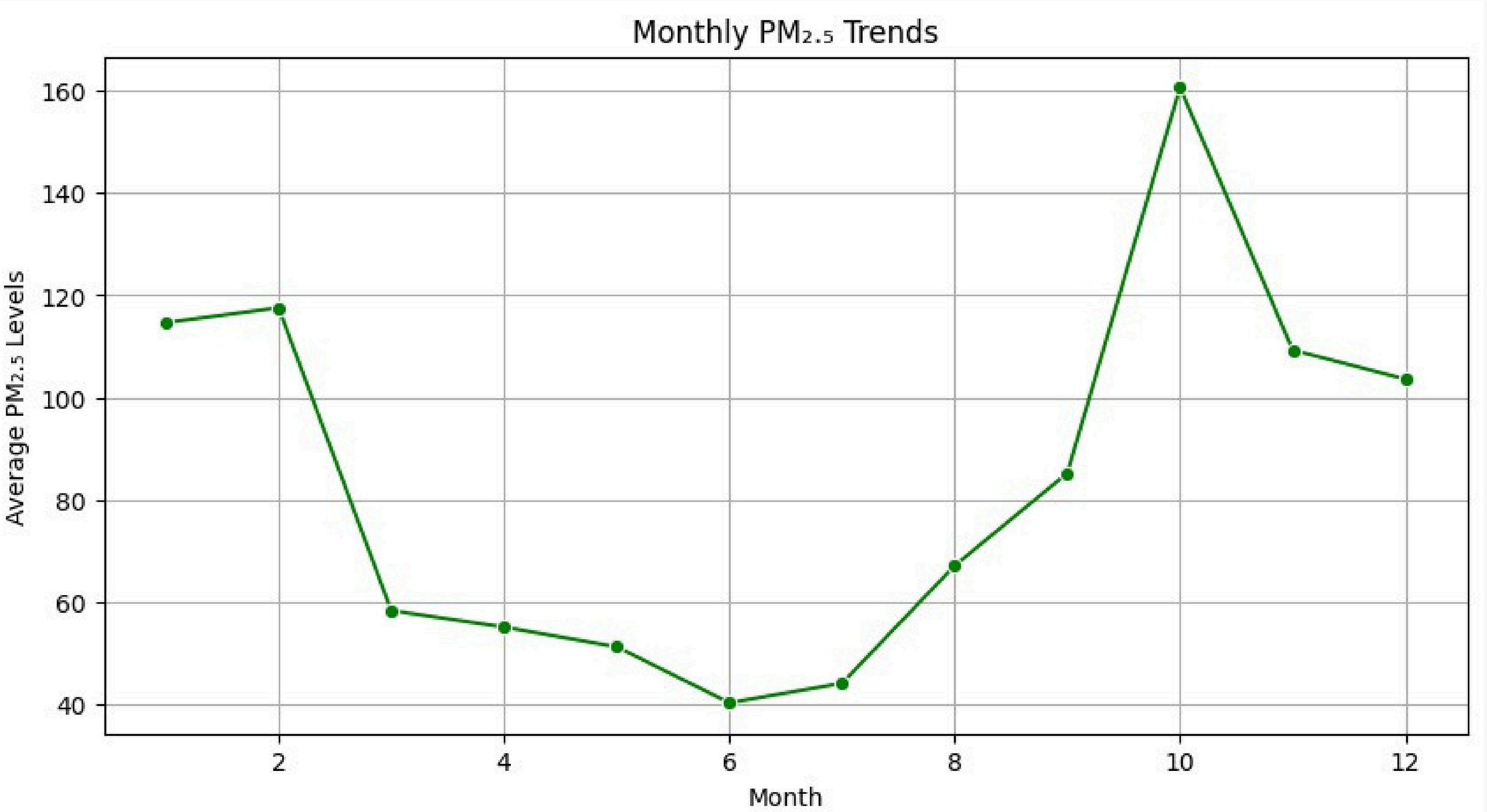
RESULT AND DISCUSSION

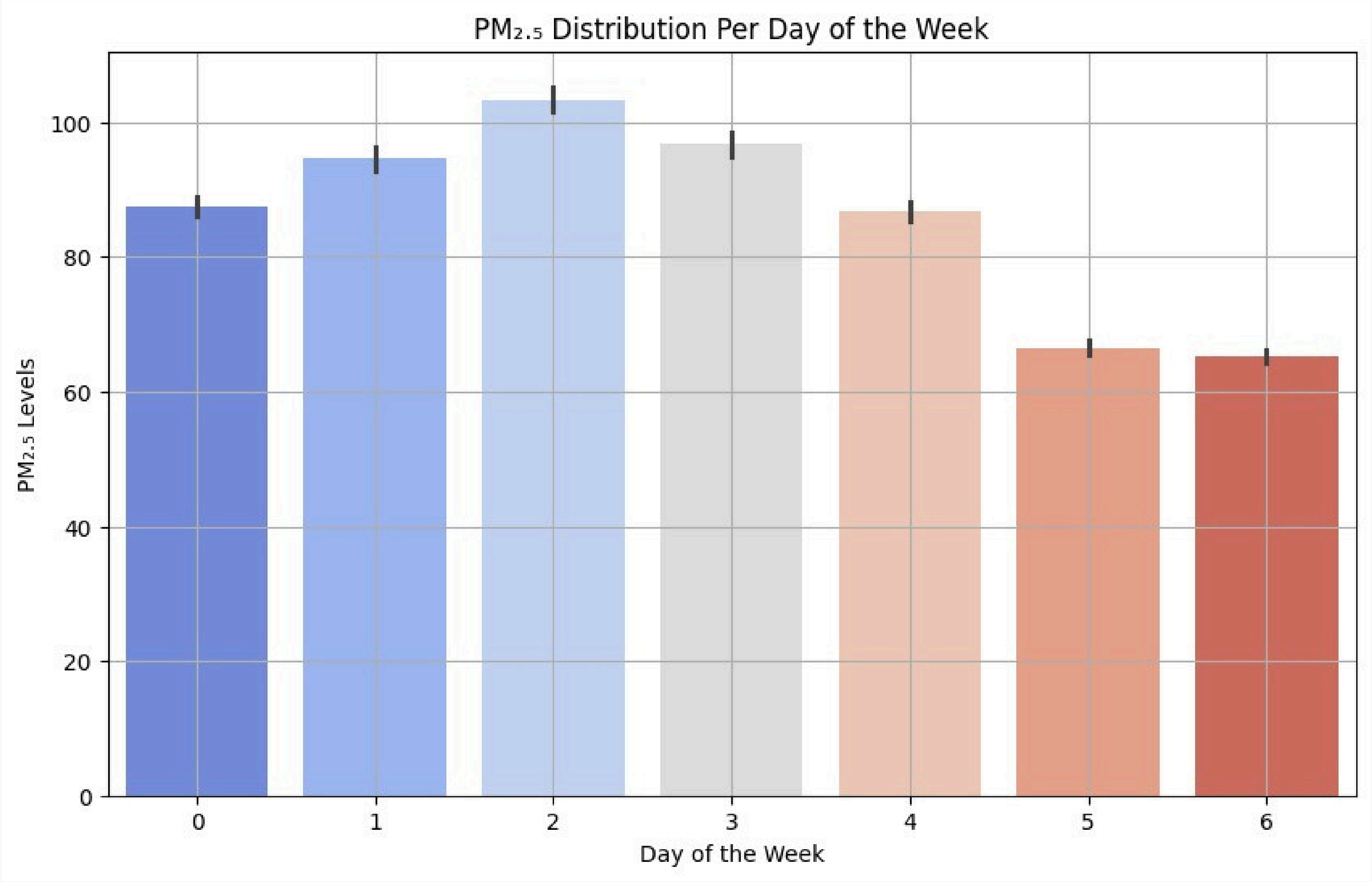
02 Finding and Analysis

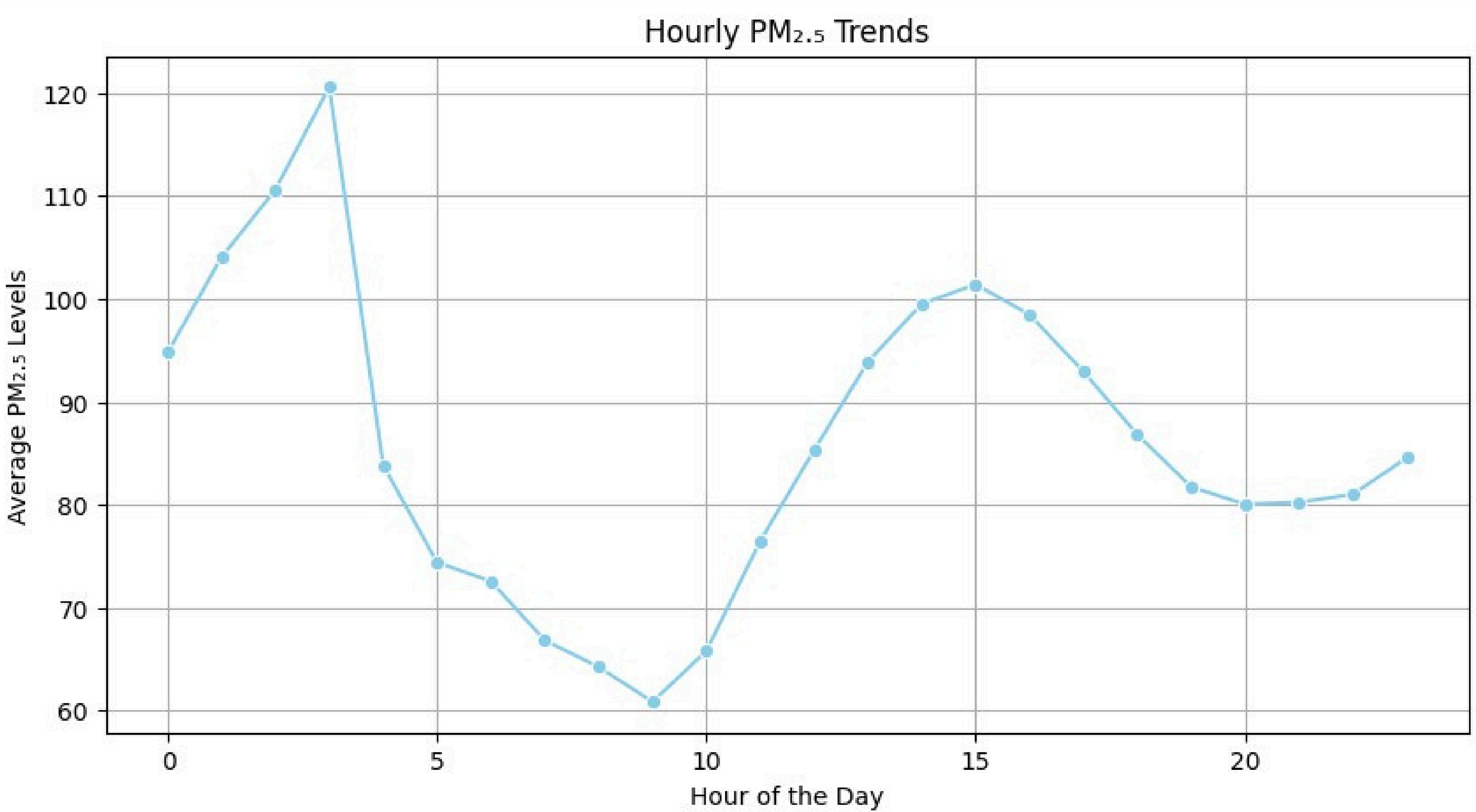
- Weekend has lower PM2.5 level compared to weekdays
  - Wednesday has the highest PM2.5 among weekdays
- October Highest Level of PM2.5
- Cool season has the highest PM2.5 levels

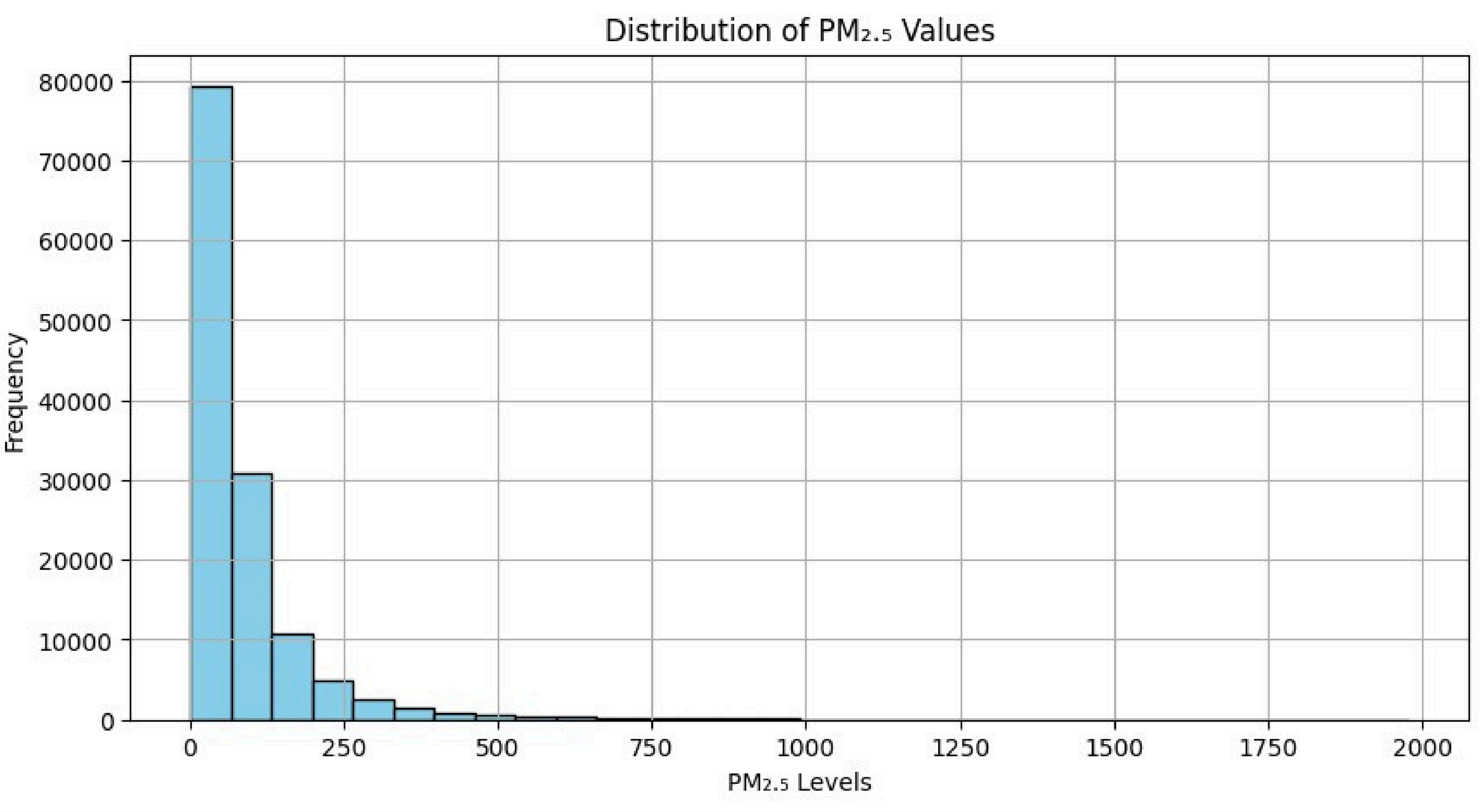

Green Space and PM2.5 Levels Per Station
Green Space
PM2.5
Values
0
20
40
60
80
100
10T
11T
12T
15T
3T
5T
61T
Station

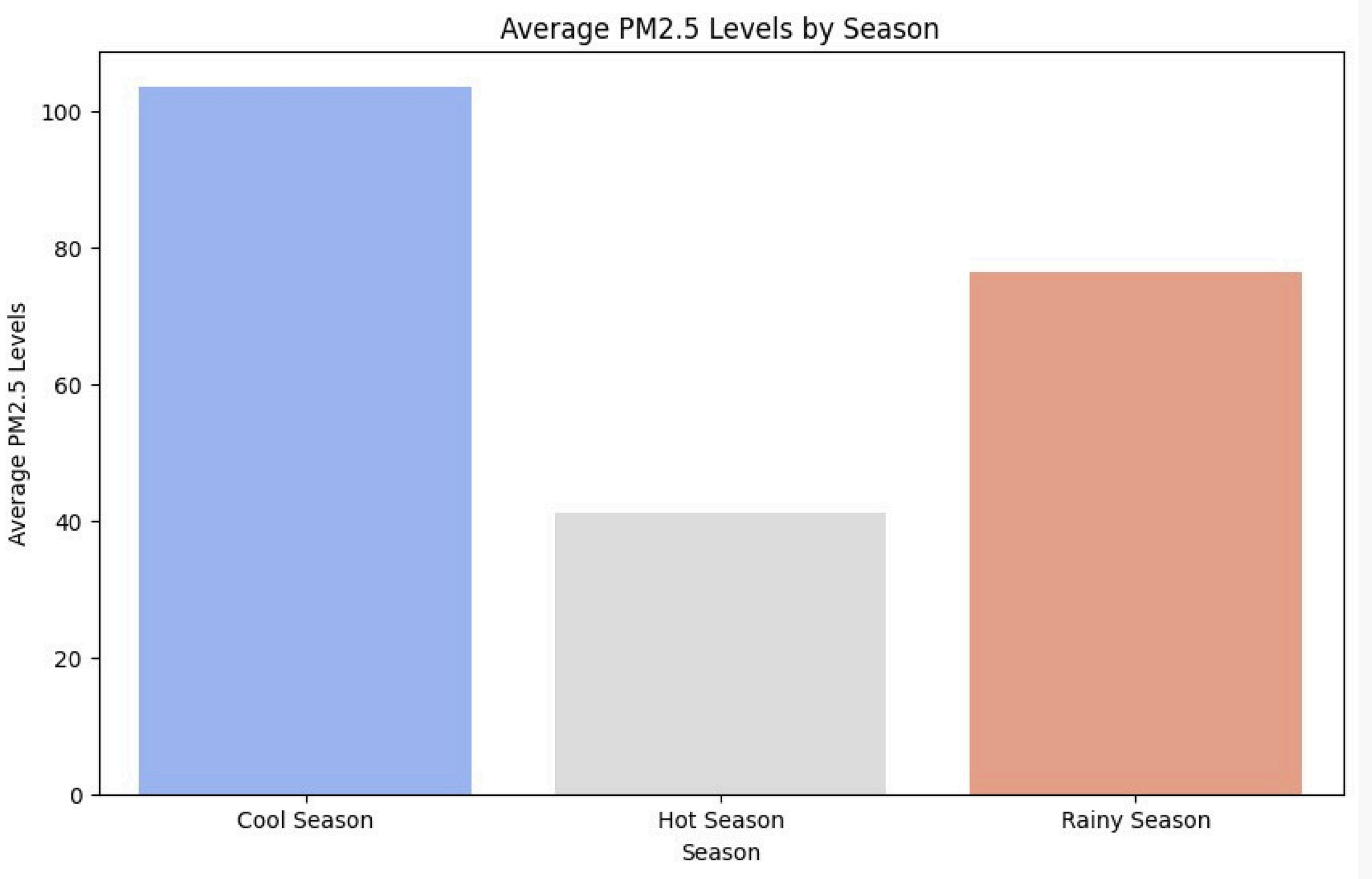

Average PM2.5 Levels per Month
Average PM2.5 Levels
Month
0
20
40
60
80
100
120
140
160
1
2
3
4
5
6
7
8
9
10
11
12


Monthly PM2.5 Trends
Average PM2.5 Levels
160
140
120
100
80
60
40
2
4
6
8
10
12
Month


PM2.5 Distribution Per Day of the Week
PM2.5 Levels
0
20
40
60
80
100
0
1
2
3
4
5
6
Day of the Week


Hourly PM2.5 Trends
Average PM2.5 Levels
120
110
100
90
80
70
60
0
5
10
15
20
Hour of the Day


Distribution of PM2.5 Values
Frequency
80000
70000
60000
50000
40000
30000
20000
10000
0
0
250
500
750
1000
1250
1500
1750
2000
PM2.5 Levels


Average PM2.5 Levels by Season
Average PM2.5 Levels
0
20
40
60
80
100
Cool Season
Hot Season
Rainy Season
Season

RESULT AND DISCUSSION

03 Evaluation of the prediction methods

Interface

SUMMARY PROGRESS

01 Progress So Far

Defined project scope, identified IV & DV, conducted literature review, acquired data (including backup from 2019), and selected Models for AQI prediction. Methodology outlined, covering feature selection and model evaluation.

02 Next Step

Refine the dataset, train and optimize models, and assess performance using R2, MAE, MSE, RMSE

03 Challenges & Consideration

Data availability and quality concerns, computational limitations for complex models, critical feature selection for reliability, and ensuring generalizability for future predictions.

THANK YOU