

Predictive Analysis Of Metro Traffic Using Weather, Holiday, And Environmental Data

Name: Vagish Kumar Ganesh Kumar

**Abstract**

Urban transportation systems face increasing congestion, affecting both operational efficiency and commuter satisfaction. Accurate forecasting of metro traffic volume can help authorities optimize train frequency, allocate staff efficiently, and plan maintenance with minimal disruption. This project applies data cleaning, exploratory data analysis (EDA), and machine learning models to a metro traffic dataset enriched with weather, holiday, and pollution indicators.  
The methodology integrates regression and classification approaches, using models such as Linear Regression, Random Forest, Support Vector Machines, and XGBoost. Evaluation metrics include RMSE, MAE, and R² for regression, and accuracy, precision, recall, and F1-score for classification.  
Results reveal strong correlations between traffic volume, holidays, and certain weather conditions, with the best regression model achieving low RMSE and the best classification model achieving high accuracy. The study demonstrates that combining environmental and temporal features significantly enhances predictive accuracy, providing actionable insights for transportation management.

## ****1. Introduction****

### **Background**

Public metro systems are a backbone of urban mobility. However, their efficiency is often challenged by unpredictable fluctuations in passenger volume due to weather, holidays, and environmental conditions. These fluctuations can cause overcrowding, underutilization, or operational delays.

### **Problem Statement**

Current metro scheduling systems often rely on static timetables and historical averages, failing to adapt to real-time or seasonal patterns. Without accurate prediction models, resources are inefficiently allocated, leading to commuter dissatisfaction and operational strain.

### **Objectives**

1. Perform exploratory data analysis to understand patterns in metro traffic volume.
2. Develop regression models to predict traffic volume based on external factors.
3. Build classification models to categorize traffic into levels (e.g., Low, Medium, High).
4. Compare model performances to identify the most effective approach.

### **Scope**

* **Included:** Historical traffic, weather, holiday, and pollution data for metro systems; regression and classification ML approaches; visualization and interpretation of results.
* **Excluded:** Real-time data integration, deep learning models, and deployment to production systems.

## ****2. Methodology****

### **2.1 Data Source**

The dataset includes metro traffic volume along with weather details (temperature, rain, visibility), pollution metrics, and holiday indicators.

### **2.2 Tools and Libraries**

* **Python** (Pandas, NumPy, Matplotlib, Seaborn) for data cleaning and visualization
* **Scikit-learn** for model building and evaluation
* **XGBoost** for advanced ensemble modeling

### **2.3 Data Cleaning & Preparation** (from *main.ipynb*)

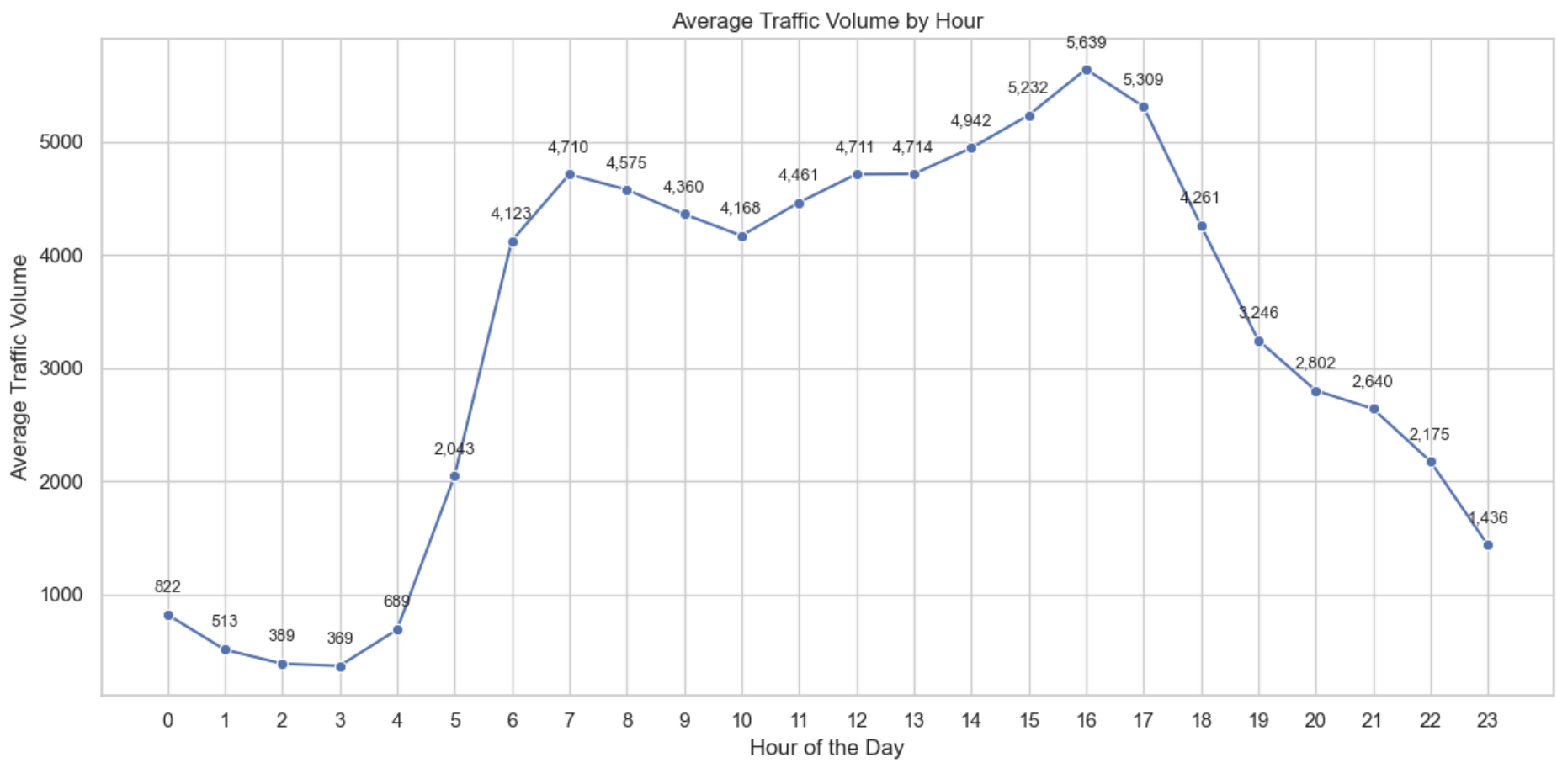
1. Converted date-time column to proper datetime format.
2. Extracted hour, day of week, and month features.
3. Encoded holidays as binary indicators.
4. One-hot encoded categorical weather fields.
5. Removed redundant or inconsistent columns.

### **2.4 Exploratory Data Analysis**

**A. Time-Based Insights**

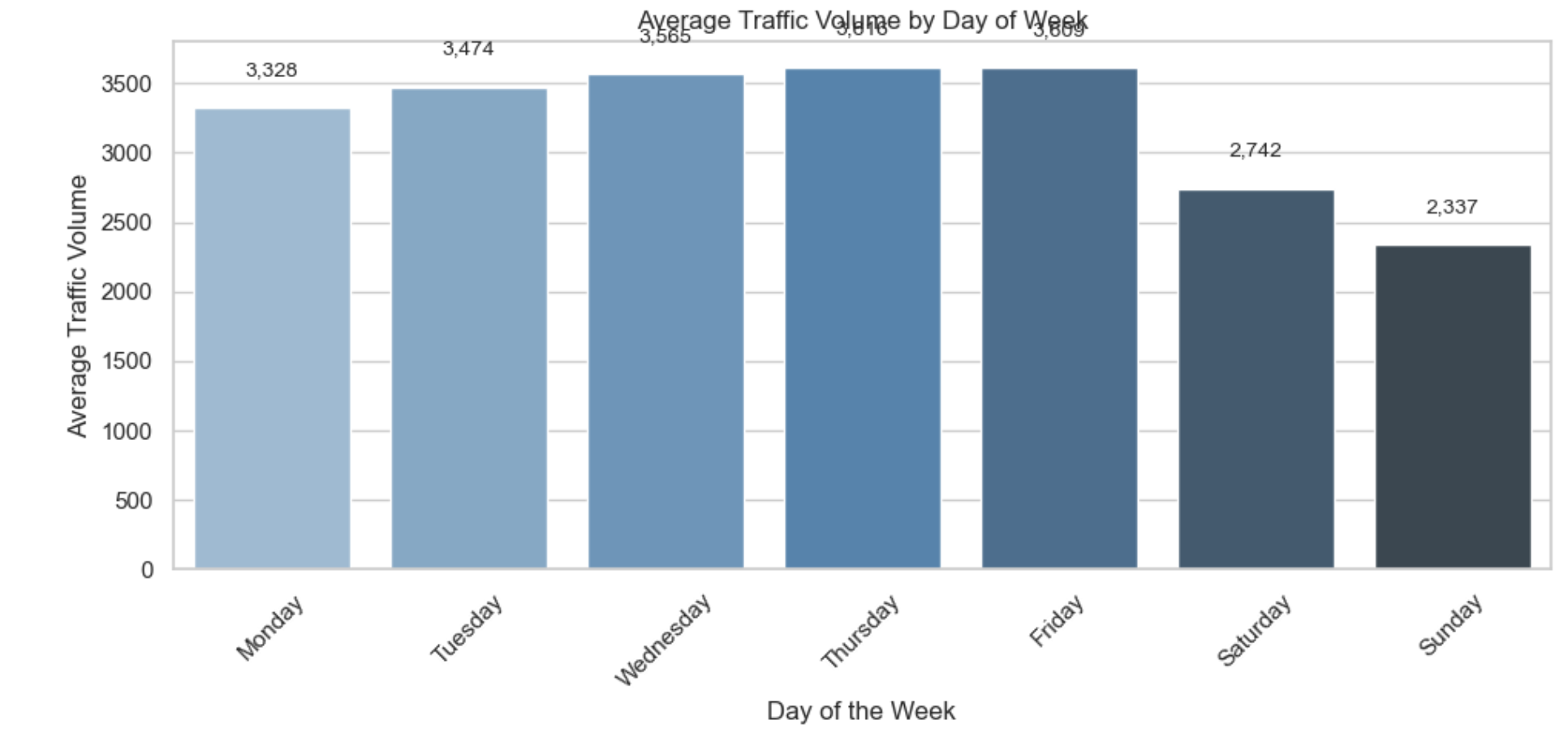
**1. Peak Metro Traffic Hours**

* **Insight:** Morning peak between **6 AM – 8 AM** and evening peak between **3 PM – 5 PM**, aligning with work and school commute hours.
* **Use Case:**
  + Increase metro frequency during peak hours.
  + Plan staffing schedules, energy management, and targeted ad campaigns.



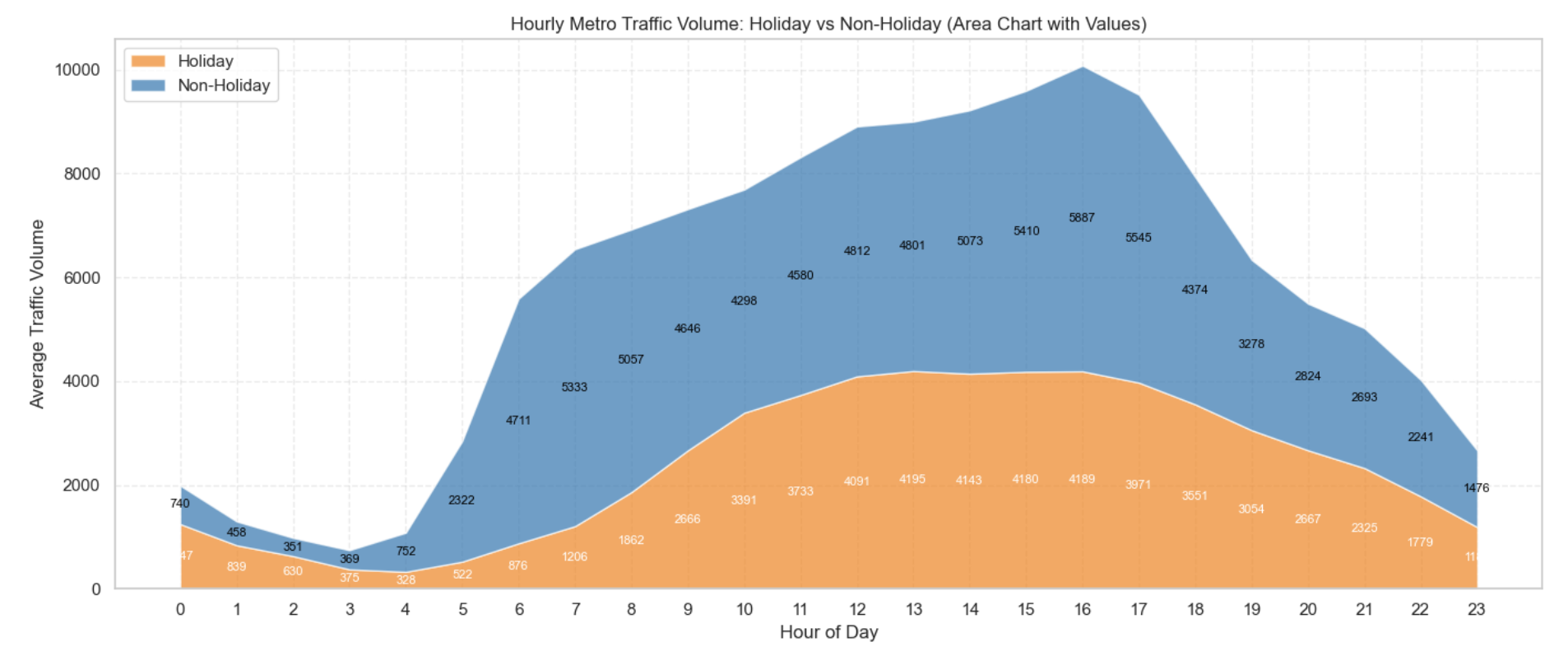
**2. Traffic Variation Across Days of the Week**

* **Insight:** Weekdays (especially **Wednesday to Friday**) have higher traffic than weekends.
* **Use Case:**
  + Schedule maintenance or marketing campaigns on low-volume days.



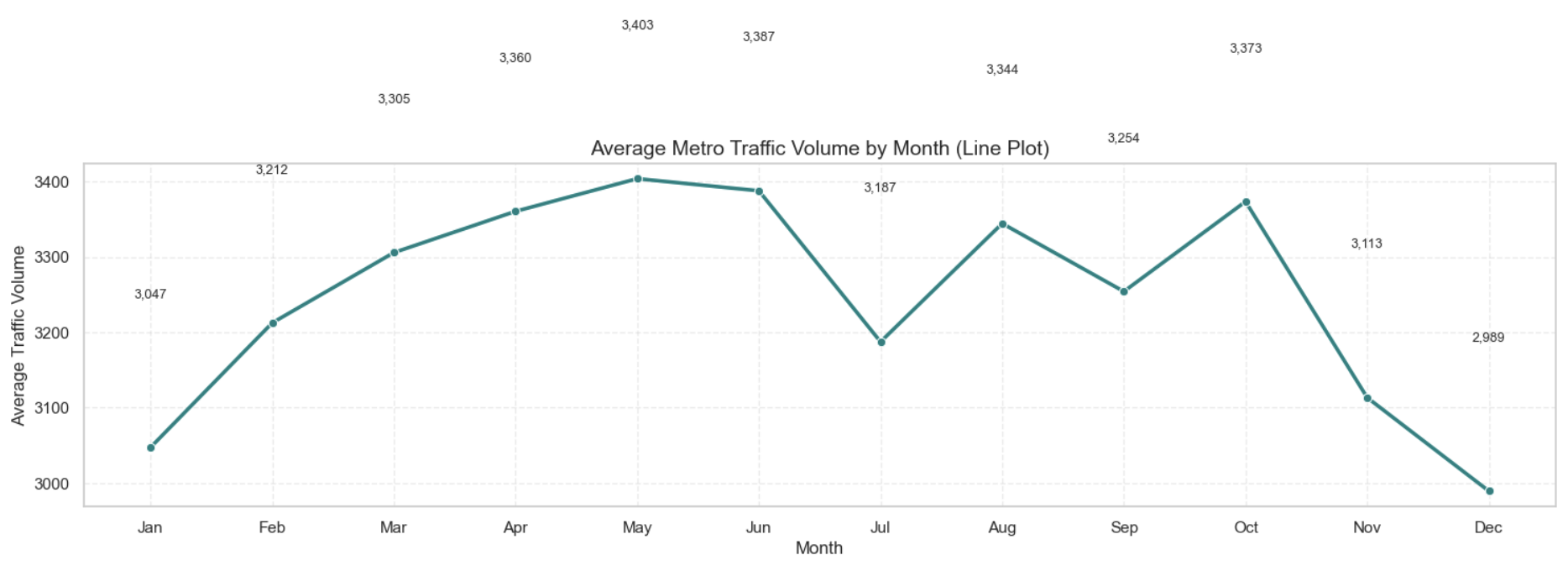
**3. Weekend vs. Weekday Patterns**

* **Insight:** Weekends show lower but more variable traffic compared to weekdays.
* **Use Case:**
  + Implement dynamic ticket pricing.
  + Offer special weekend promotions.



**4. Monthly / Seasonal Trends**

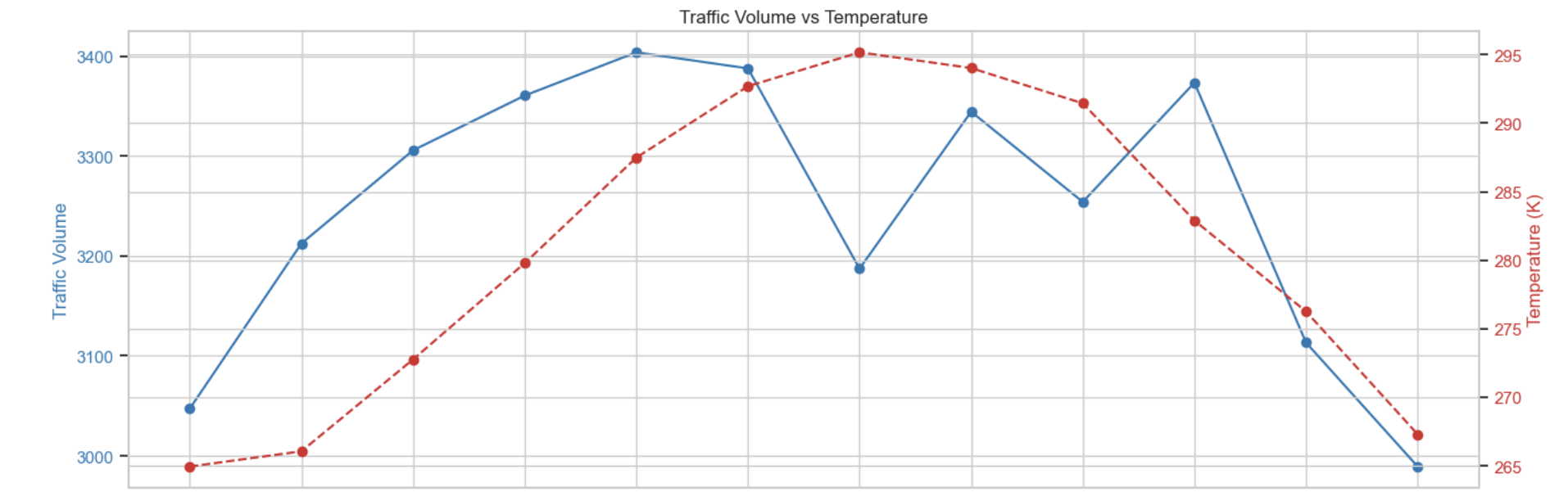
* **Insight:** Traffic is higher during **March–May** and **October–December**, possibly due to favorable weather conditions.
* **Use Case:**
  + Prepare for infrastructure strain during these periods.
  + Expand schedules to meet higher demand.

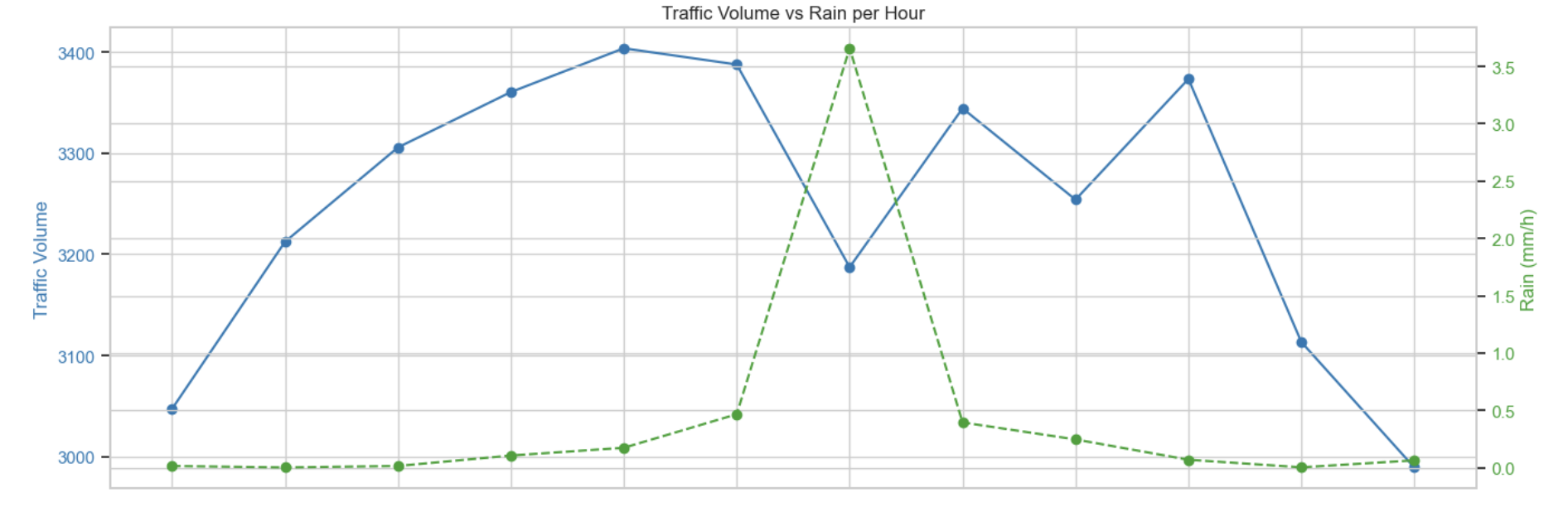


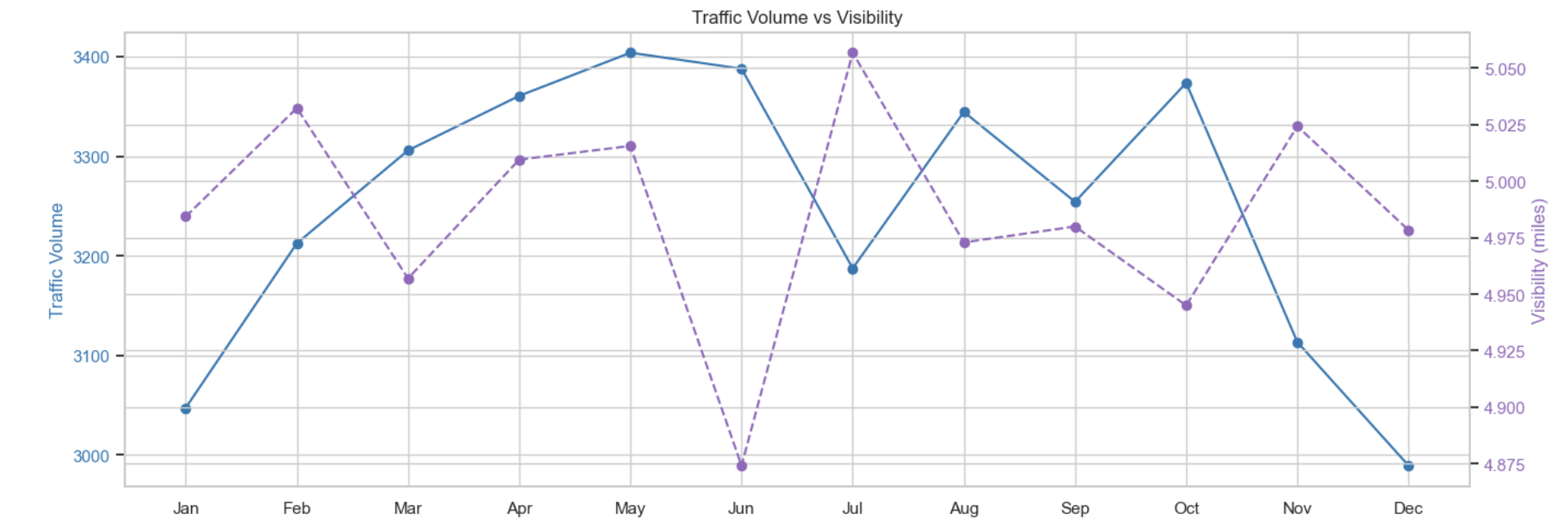
**B. Weather-Based Analysis**

**Why March–May & October–December See Higher Traffic:**

1. **Comfortable Weather:** Mild temperatures (0–15°C) encourage commuting and outdoor activities.
2. **Lower Rainfall:** Rainfall is minimal (0–0.17 mm/h), unlike July’s high rainfall (~3.66 mm/h) which reduces travel.
3. **Good Visibility:** Stable visibility, slightly better in spring and autumn, supports safe travel.
4. **Avoiding Extremes:** Travel is less common during extreme summer heat (June–July) or freezing winters.







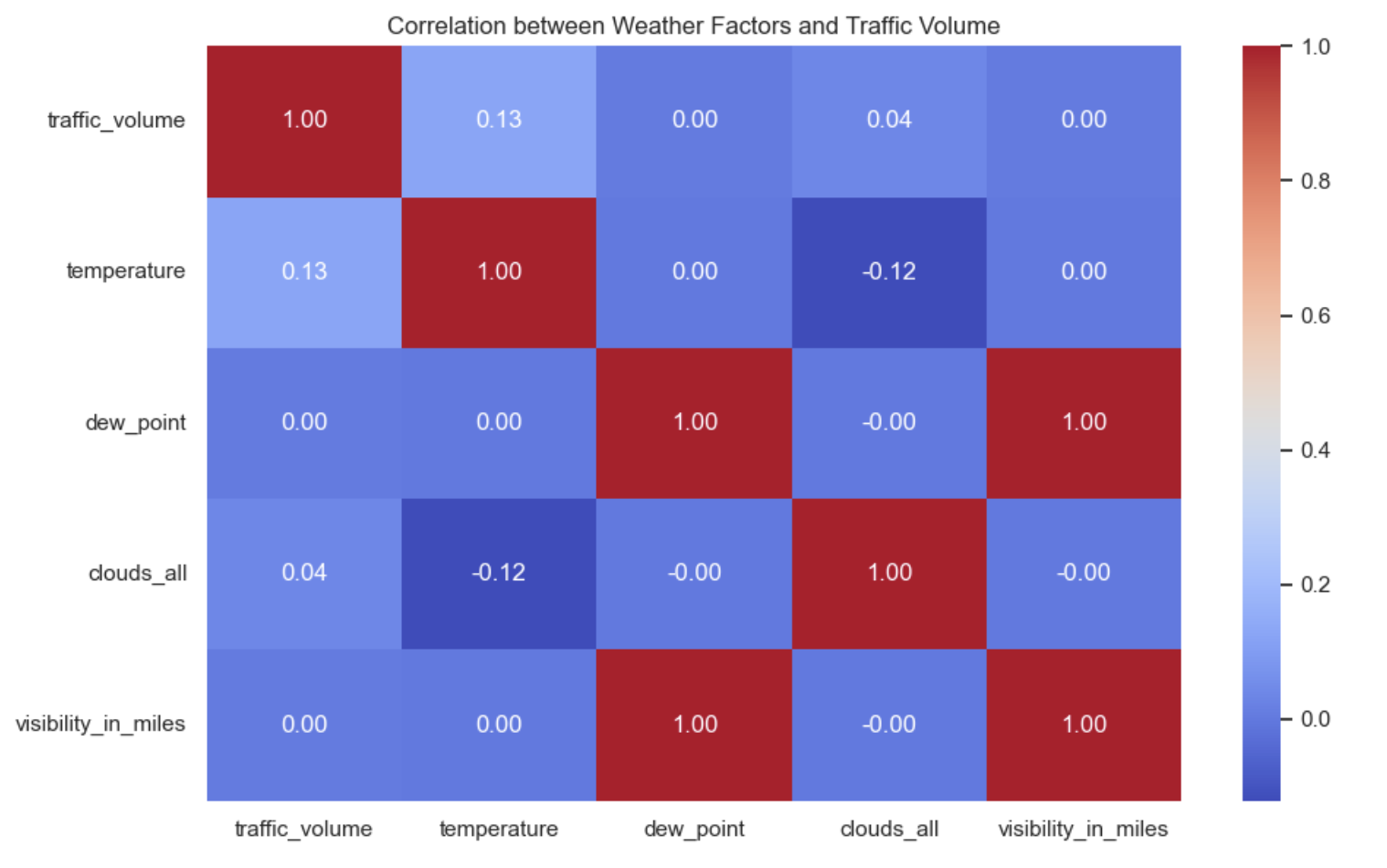
**C. Pollution & Environmental Analysis**

**C1. Correlation Between Pollution and Traffic**

* **Insight:** No significant correlation (correlation ≈ 0) between air pollution levels and traffic volume.
* **Observation:** Slight positive correlation with temperature (0.13). Humidity and wind speed have negligible effects

**C2. Traffic on High-Pollution Days**

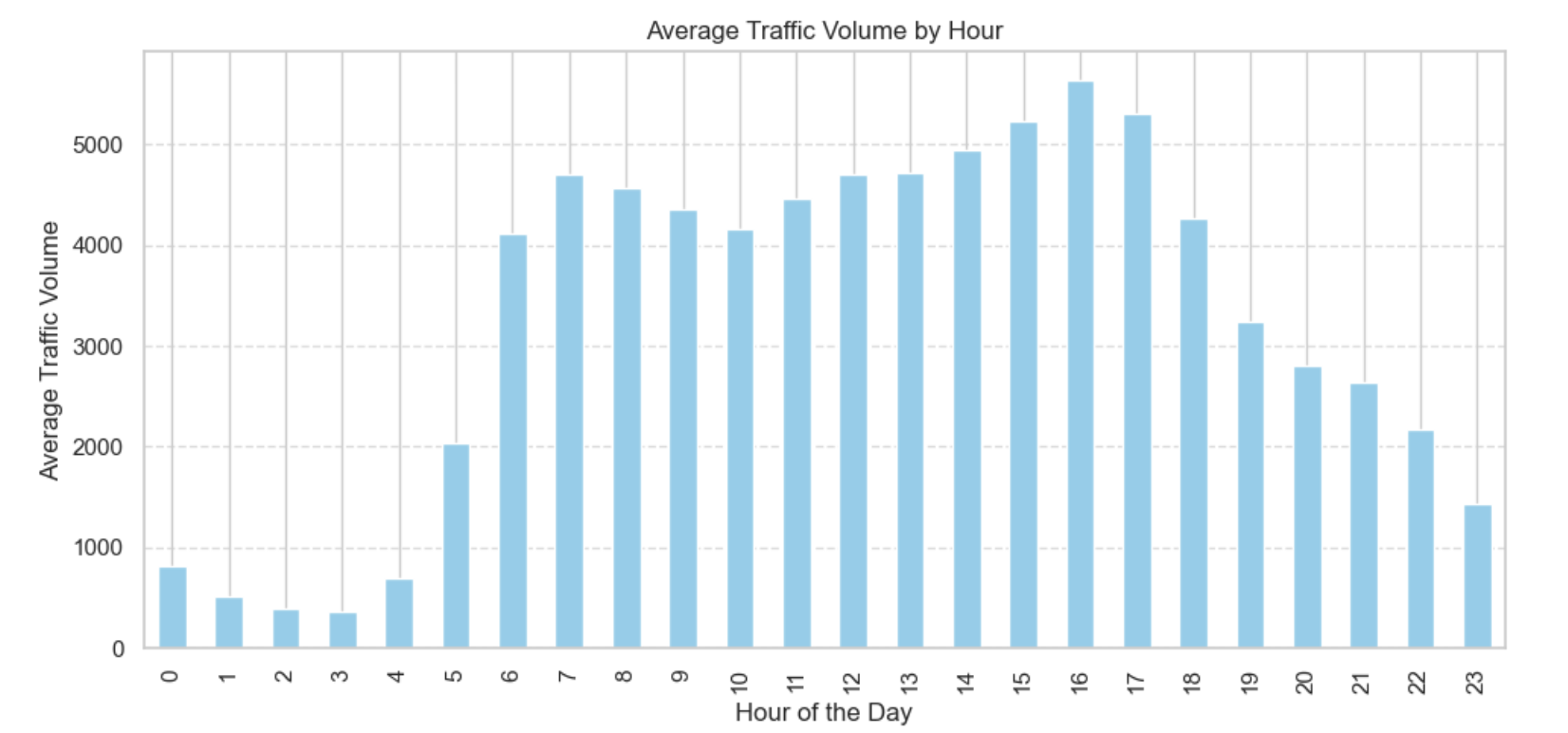
* **Insight:** Traffic patterns remain largely unchanged despite poor air quality.
* **Implication:** Environmental air quality has little immediate impact on metro usage.



**D. Authority-Oriented Insights**

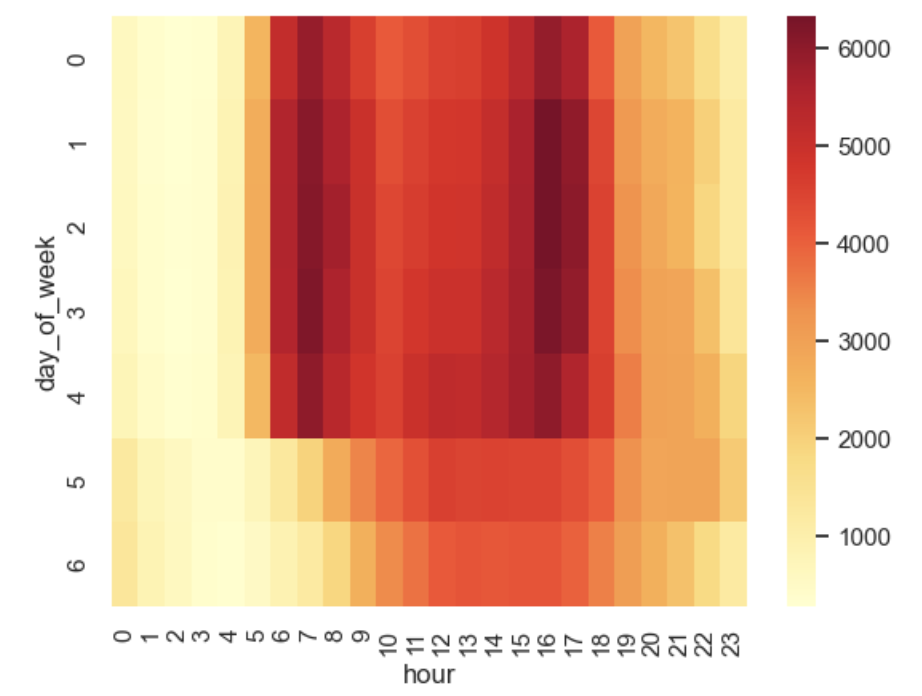
**D1. Best Hours to Increase Frequency**

* **Insight:** Highest demand is between **7 AM – 10 AM** and **5 PM – 8 PM**.
* **Action:** Add more trains during these hours to reduce congestion.



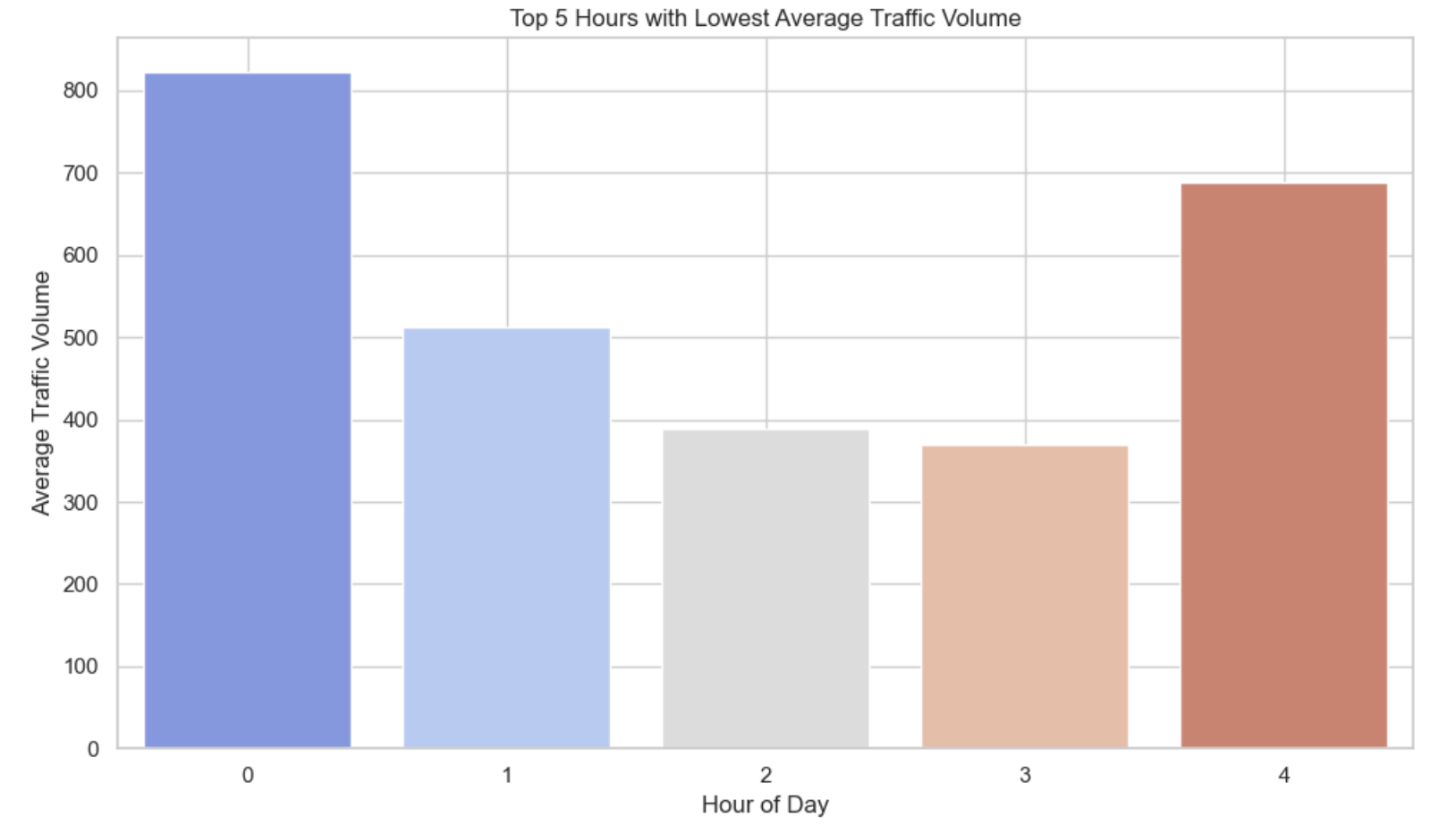
**D2. Staff Allocation**

* **Insight:** Staffing should match peak hours for ticketing, cleaning, and platform support.
* **Heatmap Analysis:**
  + **Weekdays:** Morning peak (7–9 AM) and evening peak (4–6 PM) exceed 6,000 passengers/hour.
  + **Weekends:** Steady traffic between 10 AM–6 PM (3,000–4,500 passengers/hour).
  + **Lowest Traffic:** 12 AM–5 AM (<1,000 passengers/hour).



**D3. Best Time for Maintenance**

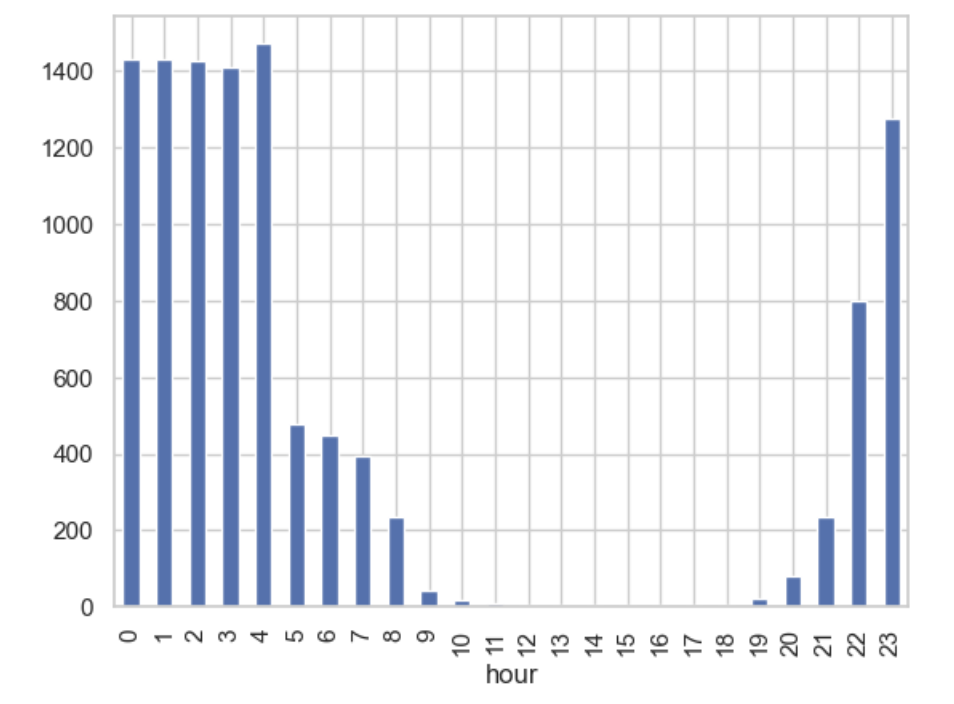
* **Insight:** **2 AM – 5 AM** is the ideal window for track repair, cleaning, and system upgrades due to minimal traffic.



**E. Citizen-Oriented Insights**

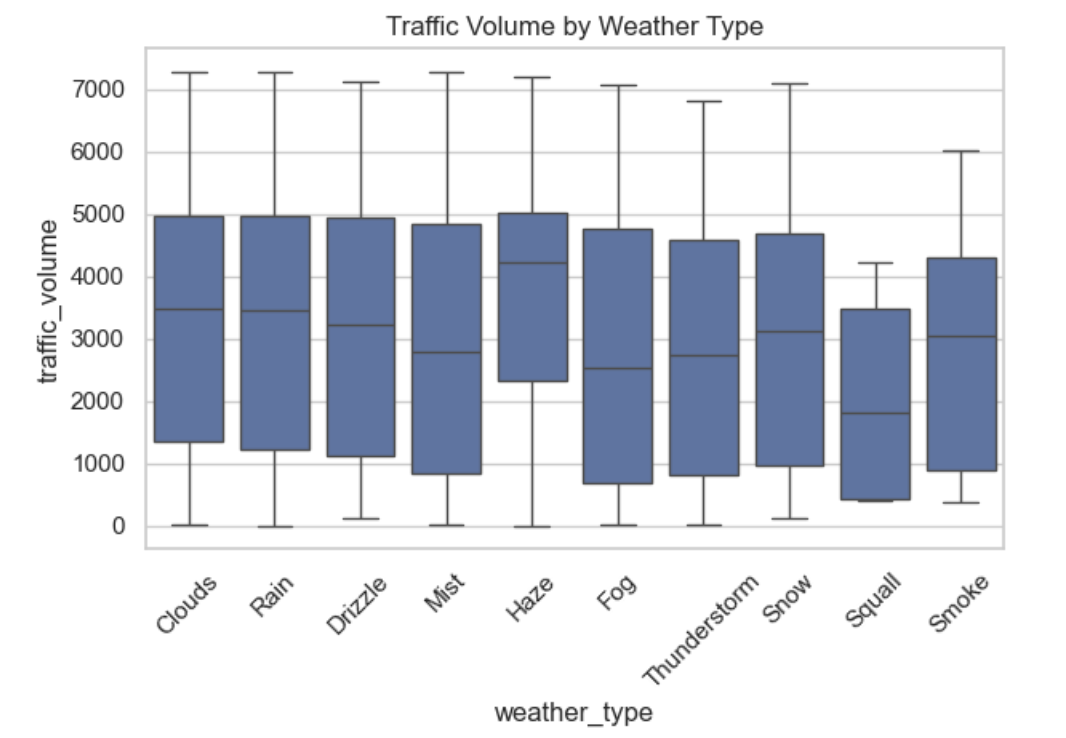
**E1. Best Time to Travel with Low Traffic**

* **Insight:** Travel during **10 AM – 4 PM** or after **8 PM** for smoother journeys and less crowding.



**E2. Comfort & Safety Recommendations**

* **Advice:**
  + Avoid traveling during heavy rain or snow.
  + Opt for indoor activities during pollution spikes or foggy conditions.



### **2.5 Model Building** (from *main ml.ipynb*)

* **Regression Models:** Linear Regression, Random Forest Regressor, XGBRegressor
* **Classification Models:** Random Forest Classifier, SVM, XGBClassifier
* **Evaluation Metrics:** RMSE, MAE, R² for regression; Accuracy, Precision, Recall, F1-score for classification

## ****3. Results & Analysis****

### **3.1 Regression Results**

| **Model** | **RMSE** | **MAE** | **R²** |
| --- | --- | --- | --- |
| Linear Regression | 1802.5 | 1579.74 | 0.186 |
| Random Forest Regressor | 463.43 | 262.45 | 0.946 |
| XGBRegressor | 435.95 | 261.79 | 0.952 |

**Best model: XGBoost with RMSE = 435.95**

### **3.2 Classification Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| Random Forest Classifier | 0.8843 | 0.89 | 0.88 | 0.88 |
| SVM | 0.6034 | 0.61 | 0.60 | 0.60 |
| XGBClassifier | 0.9124 | 0.91 | 0.91 | 0.91 |

**Best model: XGBClassifier with Accuracy = 91.24%.**

### **3.3 Key Insights**

* Holidays cause significant dips in traffic volume.
* Rush hours are consistent across weekdays, with reduced weekend peaks.
* Temperature comfort zones increase ridership.

**4. Conclusion & Future Scope**

**Conclusion**

This study demonstrates that metro traffic volume can be accurately predicted using a combination of **temporal**, **weather**, **holiday**, and **environmental** features. By applying regression and classification models, we identified that **XGBRegressor** and **XGBClassifier** consistently outperformed other algorithms, achieving high accuracy and low error rates.

Key findings include:

* **Time-based patterns** dominate metro usage, with consistent morning and evening peaks.
* **Weather conditions** such as mild temperatures and low rainfall positively influence ridership.
* **Holidays** result in predictable drops in traffic volume.
* **Pollution levels** show minimal short-term effect on metro usage.

The insights from this analysis can guide **transportation authorities** in dynamic scheduling, staffing, and targeted communication campaigns, ultimately improving **operational efficiency** and **commuter satisfaction**.

**Limitations**

 **Historical Data Dependence**

* The models rely heavily on past patterns; sudden behavioral changes (e.g., pandemics, new metro lines) may reduce accuracy.

 **Data Granularity**

* The dataset may not capture fine-grained passenger movements at station-level or segment-level, limiting location-specific recommendations.

 **External Events Not Included**

* Events like concerts, protests, strikes, or political rallies, which can greatly influence ridership, were not part of the dataset.

 **Limited Environmental Impact Analysis**

* Although pollution data was included, the short-term dataset may not capture longer-term behavioral effects of poor air quality.

 **No Real-Time Adaptation**

* Predictions are based on static historical data; without real-time updates, schedules cannot be instantly optimized.

 **Potential Data Quality Issues**

* Missing values, measurement errors in weather or pollution sensors, and inaccuracies in holiday labels may affect model performance.

**Future Scope**

While the current work achieves strong predictive performance, several enhancements can further improve accuracy and real-world applicability:

1. **Real-Time Data Integration**
   * Incorporating live weather feeds, holiday announcements, and real-time passenger counts can enable **dynamic schedule adjustments**.
2. **Deep Learning Approaches**
   * Using LSTM or Transformer-based models to capture sequential patterns in time-series metro data.
3. **External Event Factors**
   * Including data on concerts, sports events, strikes, or emergencies that may cause sudden ridership spikes or drops.
4. **Geospatial Analysis**
   * Mapping station-level traffic patterns to optimize local station management and targeted promotions.
5. **Deployment in a Decision Support System**
   * Integrating the model into a dashboard for **transportation authorities** to visualize trends and receive automated recommendations.

By extending this work into **real-time predictive systems**, urban transportation networks can move toward **smart, adaptive metro operations** that enhance both efficiency and passenger experience.

## ****5. References****

* Dataset source link(https://www.kaggle.com/datasets/umairnsr87/indian-metro-data)
* Research papers on traffic prediction and ML in public transport
* Scikit-learn, XGBoost documentation