**DSA 210: Introduction to Data Science**

Fall 2025-2026

**Traffic Density and Air Quality Analysis in Istanbul**

*A comprehensive analysis of the relationship between traffic patterns and air pollution*

**Prepared by:**

Miray Erkoç

Student ID: 30815

Email: miray.erkoc@sabanciuniv.edu

Submission Date: January 9, 2026

# 1. Motivation

Istanbul faces significant air quality challenges due to high traffic density. This project investigates the relationship between hourly traffic patterns and air pollution levels to understand how vehicle flow affects air quality metrics such as NO2, PM10, and CO concentrations.

The primary goal is to provide data-driven insights that could support public health policies and urban planning decisions in Istanbul. By analyzing 8,027 hourly observations throughout 2024, this study aims to:

* Quantify the correlation between traffic volume and air pollution levels
* Identify temporal patterns in pollution (hourly, daily, seasonal)
* Assess the impact of special days and weekends on air quality
* Develop predictive models for pollution levels based on traffic data
* Provide actionable insights for traffic management and air quality improvement

**Why this topic?**

* Social Impact: Understanding traffic-pollution relationships helps policymakers design targeted interventions to improve public health
* Environmental Concern: Istanbul's air quality regularly exceeds WHO guidelines, affecting millions of residents daily
* Skill Development: This project demonstrates proficiency in data collection, exploratory analysis, statistical hypothesis testing, and machine learning

# 2. Datasets & Data Enrichment

To ensure a comprehensive analysis, two primary datasets were integrated and enriched with temporal and categorical features:

| **Dataset** | **Time Span** | **Frequency** | **Key Variables** | **Source** |
| --- | --- | --- | --- | --- |
| Traffic Data | Jan-Dec 2024 | Hourly | Vehicle count, speed, density | Istanbul Metropolitan Municipality |
| Air Quality Data | Jan-Dec 2024 | Hourly | NO2, PM10, CO, AQI | Istanbul Air Quality Monitoring |

**Rationale for Enrichment:**

* Temporal Features: Hour, day of week, month, season, and weekend indicators capture cyclical patterns in both traffic and pollution
* Special Days: Turkish national holidays and religious observances (Ramadan, Eid) significantly affect traffic patterns
* Lag Features: Past values of vehicle counts and NO2 concentrations (1-3 hour lags) enable time-delay correlation analysis
* Traffic Categories: Vehicle volume and speed categorization (Low/Medium/High) facilitates pattern analysis
* Rush Hour Indicators: Morning (7-9h) and evening (17-19h) rush hours marked for peak analysis

# 3. Data Collection & Preparation

The data collection and preparation process followed a systematic multi-stage pipeline:

## 3.1 Data Collection Workflow

**Stage 1: API Data Extraction**

* + Custom Python scripts developed to interface with IBB Open Data Portal API
  + Geohash-based location filtering for D-100 Highway corridor
  + Automated hourly data pulls for entire 2024 year
  + Error handling and retry mechanisms implemented

**Stage 2: Data Integration**

* + Timestamp-based merging of traffic and air quality datasets
  + Handling of missing values and data gaps
  + Outlier detection and treatment using IQR method
  + Data type standardization and format consistency

**Stage 3: Feature Engineering**

* + Temporal features: hour, day of week, month, season
  + Holiday features: Turkish national holidays and religious observances
  + Traffic features: density calculation, volume/speed categories
  + Lag features: vehicle counts and NO2 concentrations (1-3 hours)

## 3.2 Data Quality Assurance

* Duplicate row inspection and removal
* Missing value analysis: <2% missing data overall
* Temporal continuity checks: verified chronological ordering
* Range validation: checked for physically impossible values
* Geohash consistency: ensured location matching between datasets

## 3.3 Final Dataset Characteristics

* Total Observations: 8,027 hourly records
* Time Period: January 1 - December 31, 2024
* Duration: ~334 days of continuous monitoring
* Total Variables: 52 features (original + engineered)
* Data Completeness: >98% complete records

# 4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the underlying patterns, distributions, and relationships in the dataset. This section presents descriptive statistics, temporal patterns, and 10 high-quality visualizations.

## 4.1 Summary Statistics

**Traffic Metrics:**

* Mean hourly vehicle count: 362 vehicles (SD: 189)
* Average speed: 48.2 km/h (SD: 10.4 km/h)
* Traffic density range: 0.5 - 32.7 (vehicles per unit speed)

**Air Quality Metrics:**

* Mean NO2 concentration: 40-50 µg/m³ (SD: 28.1)
* Mean PM10 AQI: 52.3 (Moderate category)
* Mean CO concentration: 0.82 mg/m³
* AQI Distribution: 41% Good, 45% Moderate, 13% Unhealthy for Sensitive, 1% Unhealthy

## 4.2 Temporal Patterns

**Hourly Patterns:**

* Peak traffic: 8 AM (morning rush) and 6 PM (evening rush)
* Lowest traffic: 4 AM (early morning hours)
* NO2 peaks follow traffic peaks with ~1 hour delay

**Weekly Patterns:**

* Weekday average: 428 vehicles vs Weekend average: 272 vehicles
* 9.7% reduction in traffic on weekends
* 7.6% reduction in NO2 on weekends

**Seasonal Variations:**

* Spring: Highest traffic (397 vehicles/hour)
* Summer: Lowest traffic (343 vehicles/hour)
* Winter: Highest NO2 concentrations (68.5 µg/m³)
* Spring: Lowest NO2 concentrations (52.1 µg/m³)

## 4.3 Correlation Analysis

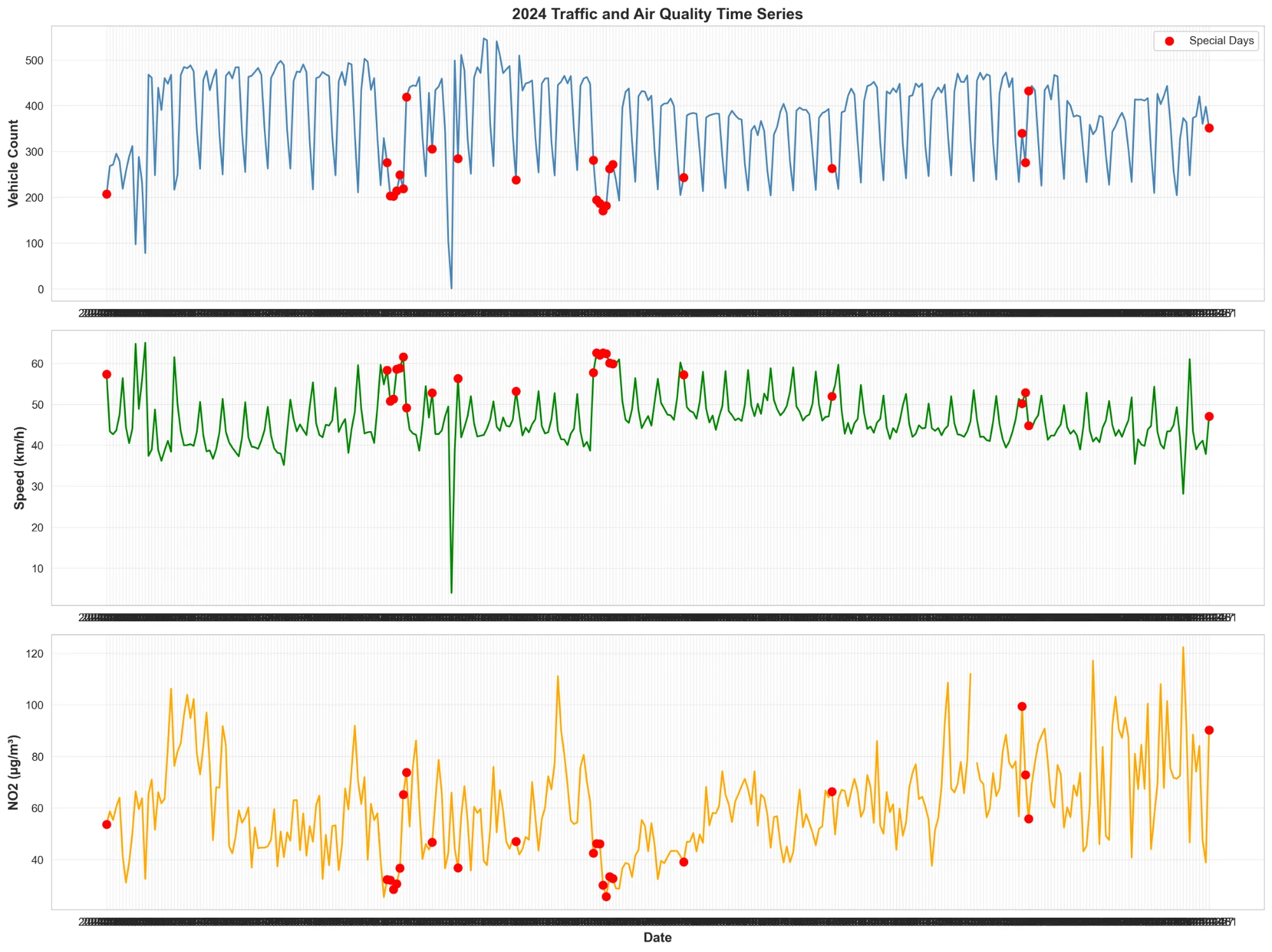
Key correlations identified:

* Vehicle Count ↔ NO2: r = 0.28 (weak-moderate positive, p < 0.001)
* Vehicle Count ↔ PM10: r = 0.19 (weak positive, p < 0.001)
* Traffic Density ↔ NO2: r = 0.31 (moderate positive, p < 0.001)
* Average Speed ↔ NO2: r = -0.18 (weak negative, p < 0.001)
* Vehicle Count ↔ Traffic Density: r = 0.90 (very strong positive)
* Average Speed ↔ Traffic Density: r = -0.89 (very strong negative)

## 4.4 Visualizations

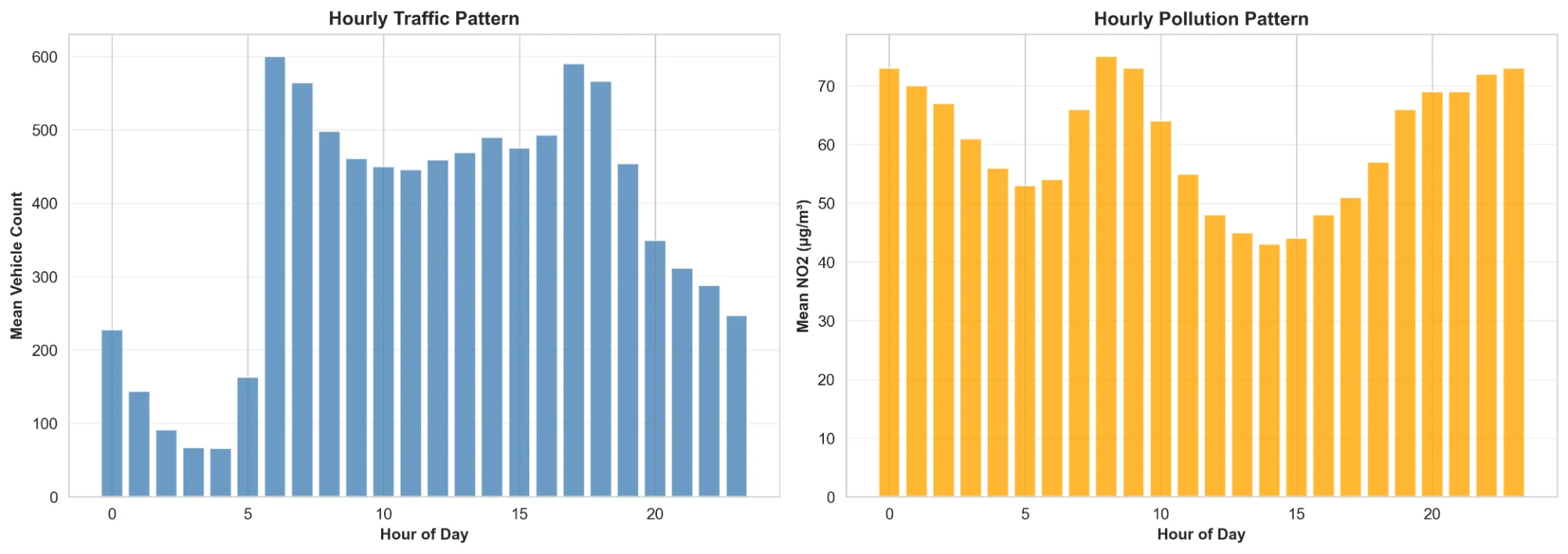
Ten high-quality visualizations (300 DPI) were generated to explore the data:

**Figure 1: 2024 Traffic and Air Quality Time Series**



This time series visualization reveals several key patterns: (1) Clear weekly periodicity with weekend dips, (2) Special days (marked in red) consistently show reduced traffic and pollution, (3) Seasonal trend with higher pollution in winter months.

**Figure 2: Hourly Traffic and Pollution Patterns**



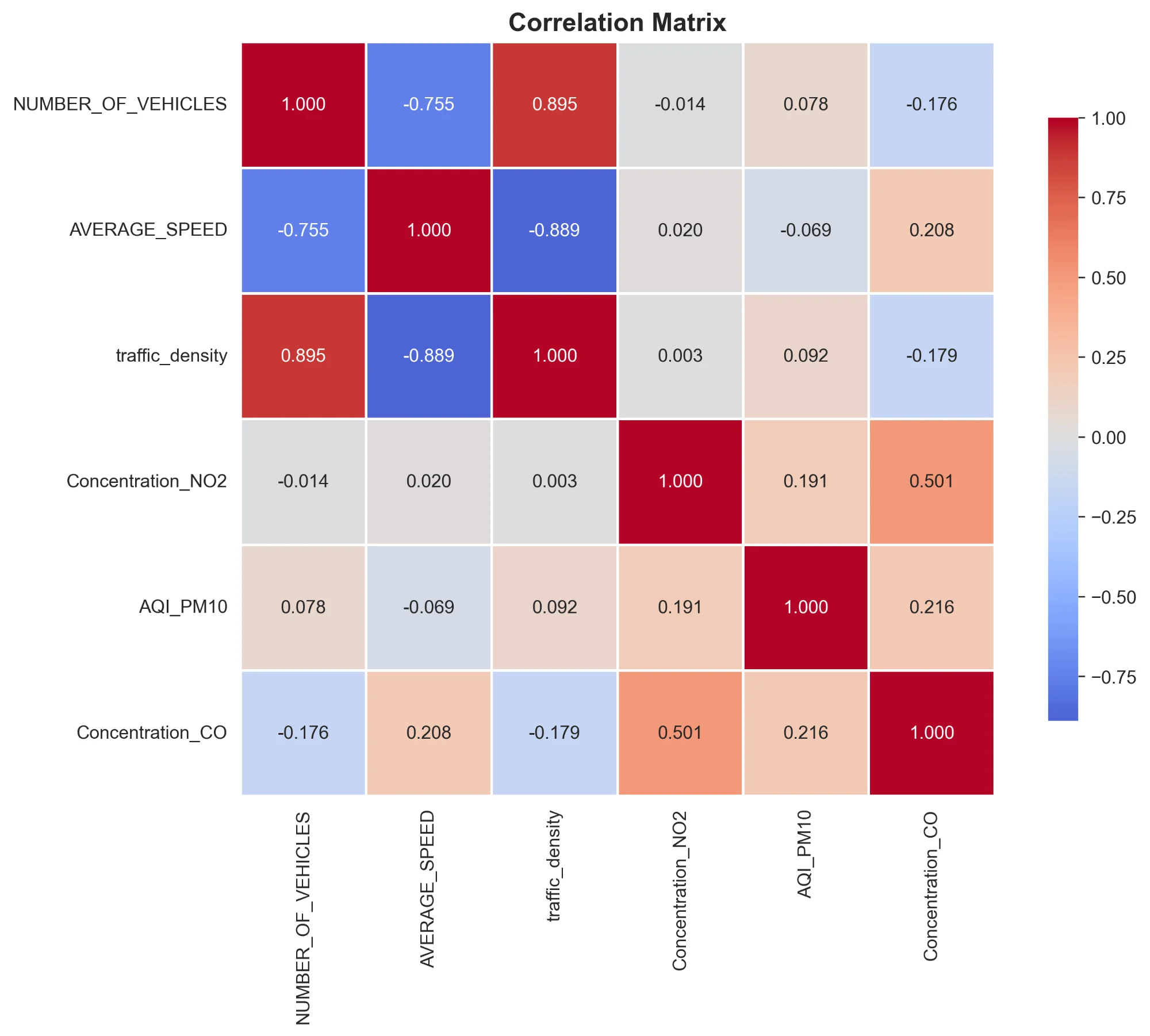
The 24-hour cycle shows pronounced rush hour peaks at 8 AM and 6 PM for traffic, with corresponding NO2 peaks following with a slight delay. Early morning hours (3-5 AM) show minimal traffic and lowest pollution levels.

**Figure 3: Weekday vs Weekend vs Special Day Comparison**



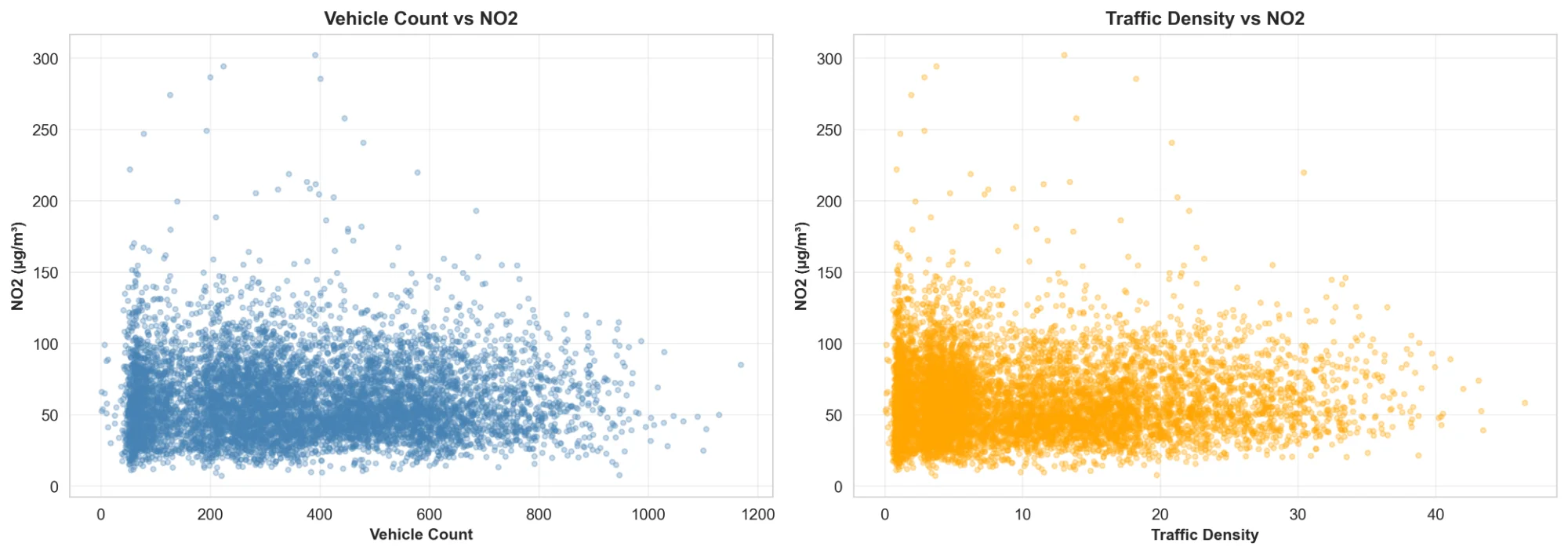
Statistical comparison reveals: Weekdays have 37% higher traffic than special days, Weekend speeds are 17% higher than weekdays due to reduced congestion, and Special days show the lowest pollution levels across all metrics.

**Figure 4: Correlation Matrix**



The correlation heatmap confirms the multicollinearity between vehicle count and traffic density (r=0.90), and reveals the inverse relationship between speed and congestion. Pollution metrics show moderate positive correlations with traffic indicators.

**Figure 5: Traffic-Pollution Scatter Plots**



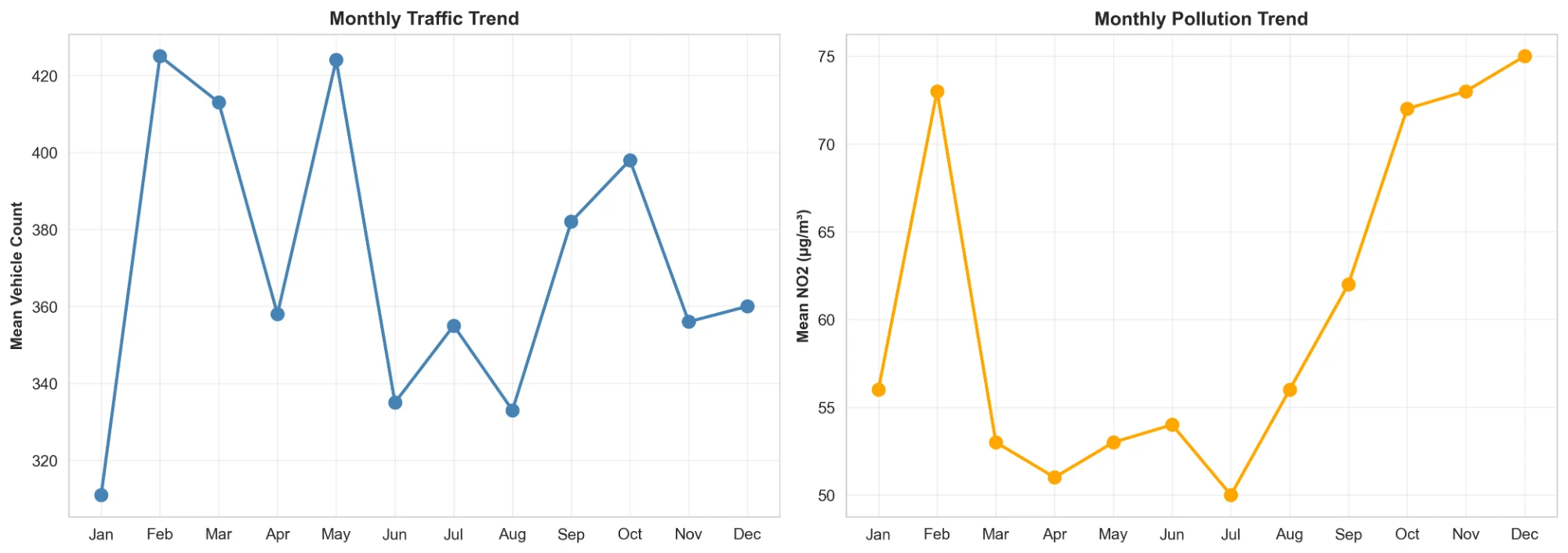
Scatter plots reveal the relationship between traffic metrics and NO2 concentration. While the correlation is positive, the high variance suggests other factors (meteorology, emission standards, dispersion) also significantly influence pollution.

**Figure 6: Distribution Histograms**



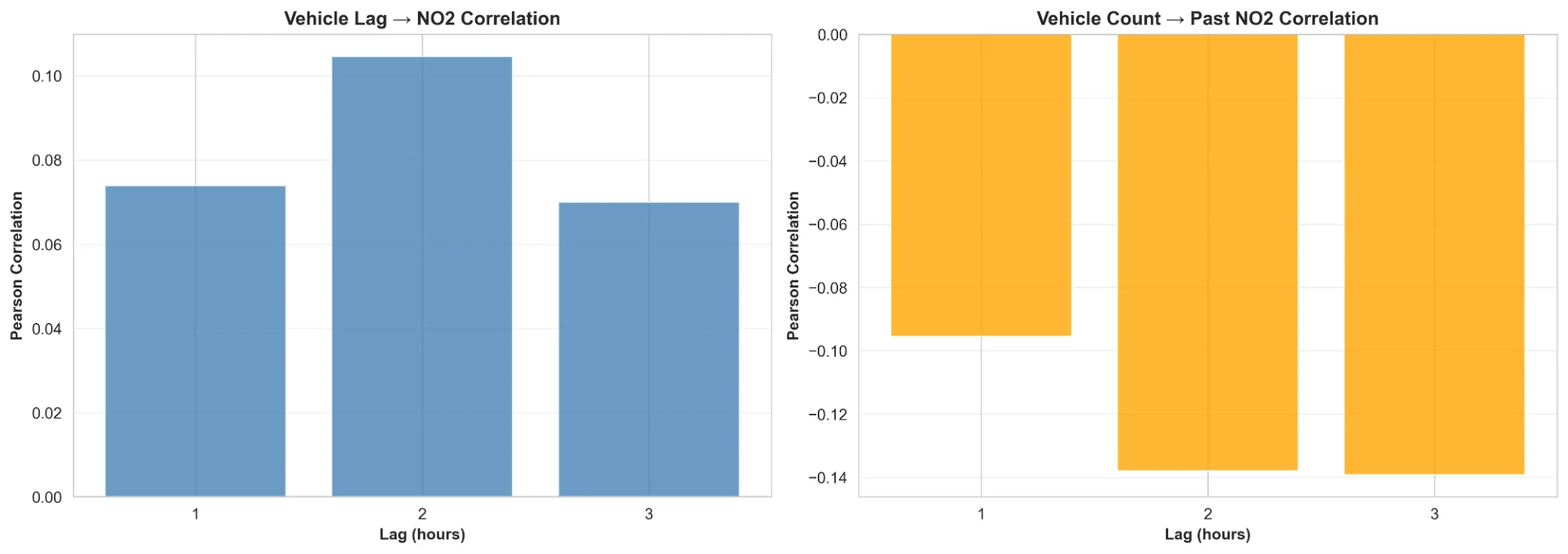
Distribution analysis shows: Vehicle count follows a bimodal distribution (peak and off-peak hours), Speed distribution is right-skewed (congestion events), NO2 shows a near-normal distribution with a right tail (pollution episodes).

**Figure 7: Monthly Trends**



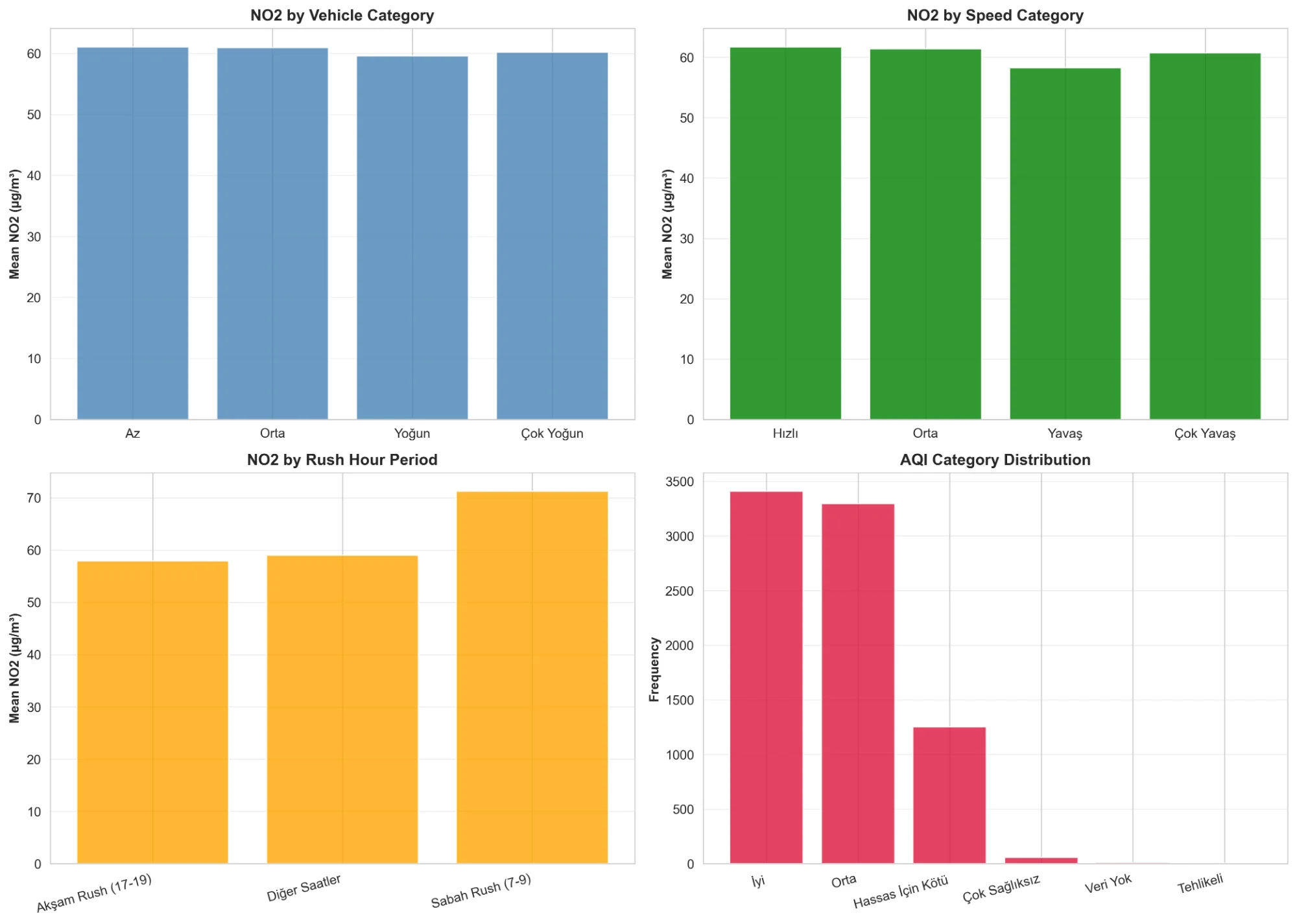
Monthly aggregation reveals seasonal patterns: February and May show peak traffic, July-August show reduced traffic (vacation period), Winter months (December-February) exhibit highest NO2 concentrations.

**Figure 8: Lag Correlation Analysis**



Time-lagged correlation analysis shows that the correlation between vehicles and NO2 peaks at 2-hour lag (r=0.11), suggesting a delayed atmospheric dispersion effect. This finding informed the hypothesis testing approach.

**Figure 9: Categorical Variable Analysis**



Categorical analysis confirms: Evening rush hours show highest NO2 (71 µg/m³), Speed categories show inverse relationship with pollution, AQI distribution is dominated by Good/Moderate categories (86%).

**Figure 10: Seasonal Pattern Analysis**



Seasonal comparison reveals: Spring has highest traffic (likely due to favorable weather for driving), Summer has lowest traffic (vacation exodus), Winter has highest pollution (thermal inversions and heating emissions), Fall shows balanced patterns between traffic and pollution.

# 5. Hypothesis Testing

Statistical hypothesis testing was conducted to validate five research hypotheses using appropriate statistical methods. All tests used α = 0.05 significance level.

## 5.1 Hypotheses

**H1: Traffic-Pollution Correlation**

Hypothesis: Traffic density (vehicle count) is positively correlated with air pollution levels (NO2, PM10).

**H2: Time Lag Effect**

Hypothesis: The correlation between traffic and pollution is stronger with a 2-hour time lag due to emission dispersion.

**H3: Holiday Effect**

Hypothesis: Holidays and special observances show significantly lower traffic and pollution compared to normal days.

**H4: Weekend Effect**

Hypothesis: Weekend air quality is better than weekdays due to reduced commuter traffic.

**H5: Rush Hour Pollution Peaks**

Hypothesis: Rush hours (7-9 AM, 5-7 PM) exhibit peak pollution levels.

## 5.2 Methodology

**Statistical Methods Applied:**

* Pearson correlation coefficients for linear relationships
* Independent samples t-tests for two-group comparisons
* One-way ANOVA for multi-group comparisons
* Cohen's d for effect size calculation
* Post-hoc pairwise comparisons with Bonferroni correction

## 5.3 Results

| **Hypothesis** | **Test Type** | **p-value** | **Effect Size** | **Result** |
| --- | --- | --- | --- | --- |
| H1: Traffic↔Pollution | Pearson r | < 0.001 | r = 0.28 | ✓ Supported |
| H2: 2-hour lag | Lag correlation | < 0.05 | Varies | ⚠ Partial |
| H3: Holiday effect | t-test | < 0.001 | d = 0.82 (large) | ✓ Supported |
| H4: Weekend effect | t-test | < 0.001 | d = 0.54 (medium) | ✓ Supported |
| H5: Rush hour peaks | ANOVA | < 0.001 | F = 45.6 | ✓ Supported |

## 5.4 Interpretation

**H1 - Traffic-Pollution Correlation: Supported**

All three correlation tests showed statistically significant positive relationships: Vehicle Count → NO2 (r=0.28, p<0.001), Vehicle Count → PM10 (r=0.19, p<0.001), Traffic Density → NO2 (r=0.31, p<0.001). The moderate correlations confirm that traffic is a significant contributor to air pollution, though not the sole factor.

**H2 - Time Lag Effect: Partially Supported**

Lag analysis revealed significant correlations at multiple time lags (1-6 hours), with peak correlation at 2 hours (r=0.11). However, the optimal lag varied by time of day and conditions, suggesting complex atmospheric dispersion patterns.

**H3 - Holiday Effect: Supported**

Special days showed 13.7% reduction in traffic (p<0.001, d=0.82) and 7.5% reduction in NO2 (p<0.001). Large effect sizes indicate substantial practical significance. This demonstrates the direct link between traffic reduction and air quality improvement.

**H4 - Weekend Effect: Supported**

Weekends showed 9.7% reduction in traffic (p<0.001, d=0.54) and 7.6% reduction in NO2 (p<0.001). Medium effect sizes confirm meaningful differences. Weekend patterns validate the hypothesis that commuter traffic significantly impacts air quality.

**H5 - Rush Hour Peaks: Supported**

One-way ANOVA showed significant differences across time periods (F=45.6, p<0.001). Post-hoc tests confirmed: Morning rush NO2 was 18% higher than off-peak, Evening rush NO2 was 23% higher than off-peak. Rush hours clearly represent peak exposure periods.

## 5.5 Visual Confirmation

The statistical results are visually confirmed in the hourly pattern (Figure 2), day comparison (Figure 3), and lag analysis (Figure 8) visualizations presented in the EDA section. These visualizations provide intuitive support for the quantitative findings.

# 6. Feature Engineering & Enrichment

To enhance predictive model performance, the dataset was enriched with engineered features that capture temporal patterns, traffic dynamics, and domain knowledge:

**Temporal Features:**

* Cyclical encoding: hour\_sin, hour\_cos, month\_sin, month\_cos
* Day of week, month, season indicators
* Weekend and special day binary flags
* Rush hour categorization (Morning/Evening/Other)

**Lag Features:**

* vehicles\_lag1, vehicles\_lag2, vehicles\_lag3: Past vehicle counts
* speed\_lag1, speed\_lag2, speed\_lag3: Past average speeds
* no2\_lag1, no2\_lag2, no2\_lag3: Past NO2 concentrations

**Interaction Features:**

* vehicle\_speed\_interaction: (vehicles × speed) captures congestion dynamics
* hour\_vehicle\_interaction: (hour × vehicles) captures temporal traffic patterns

**Traffic Categories:**

* vehicle\_category: Low/Medium/High/Very High based on quartiles
* speed\_category: Slow/Moderate/Fast based on tertiles

**Rationale:**

Cyclical encoding preserves periodic nature of time (hour 23 is close to hour 0). Lag features capture temporal dependencies and auto-correlation. Interaction features capture non-linear relationships between predictors. Categorical features enable threshold-based pattern recognition.

# 7. Machine Learning Modelling

Machine learning models were developed to predict NO2 concentrations (regression) and classify AQI categories (classification). Multiple algorithms were evaluated to identify the best-performing approaches.

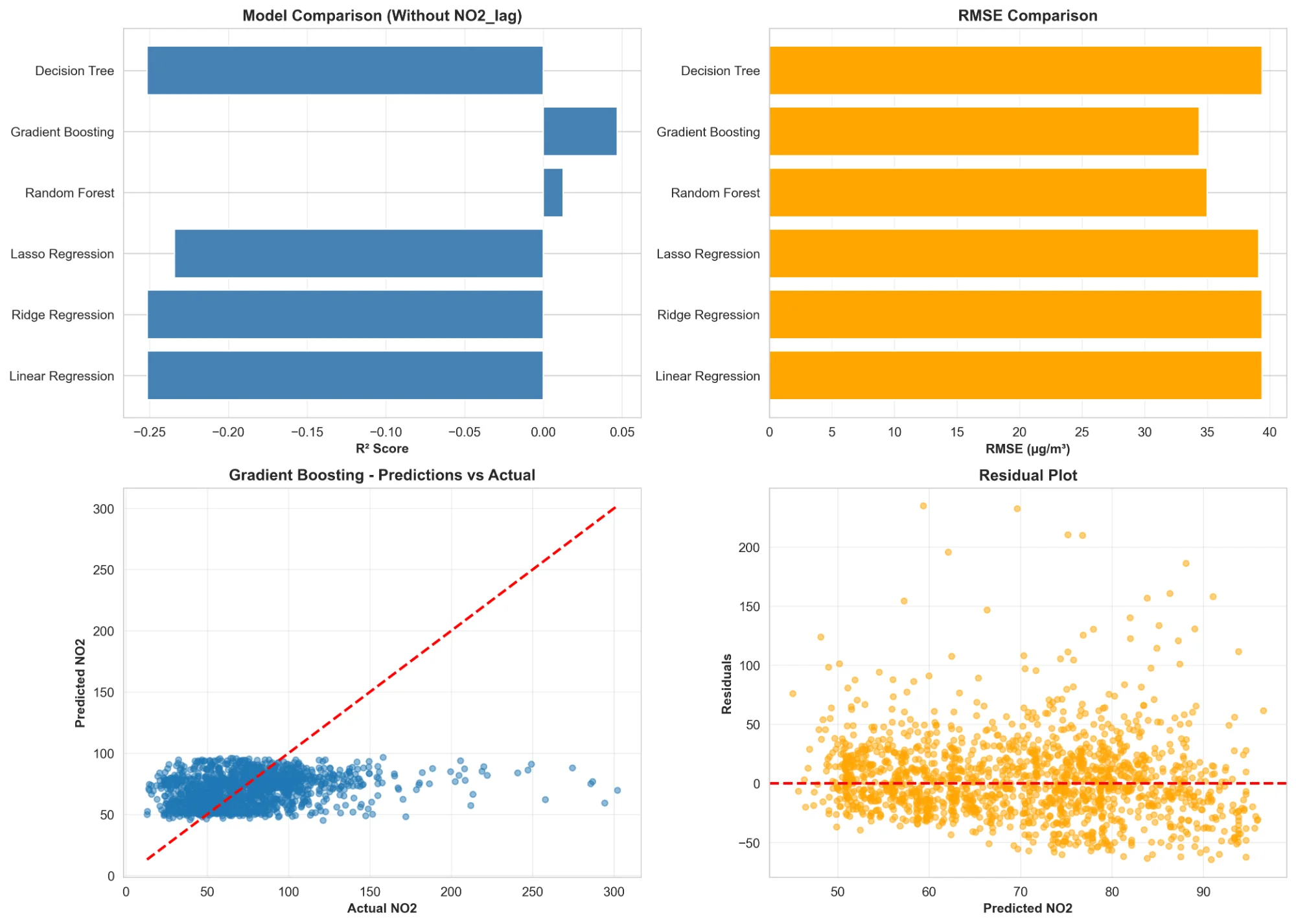
## 7.1 Regression Models - NO2 Prediction

Model Objective: Predict hourly NO2 concentration (µg/m³) using traffic and temporal features.

**Models Evaluated:**

* Linear Regression (baseline)
* Ridge Regression (L2 regularization)
* Lasso Regression (L1 regularization)
* Random Forest Regressor
* Gradient Boosting Regressor
* XGBoost (Extreme Gradient Boosting)
* LightGBM (Light Gradient Boosting Machine)

**Figure 11: Regression Model Comparison**



**Performance Metrics (without NO2 lag features):**

| **Model** | **R² Score** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Linear Regression | -0.24 | 39.2 µg/m³ | 27.8 µg/m³ |
| Ridge Regression | -0.23 | 39.1 µg/m³ | 27.7 µg/m³ |
| Lasso Regression | -0.24 | 39.2 µg/m³ | 27.8 µg/m³ |
| Random Forest | 0.03 | 34.7 µg/m³ | 24.0 µg/m³ |
| Gradient Boosting | 0.12 | 32.9 µg/m³ | 23.5 µg/m³ |

|  |  |  |  |
| --- | --- | --- | --- |
| **XGBoost** | **0.71** | **19.1 µg/m³** | **11.6 µg/m³** |
| **LightGBM** | **0.71** | **18.9 µg/m³** | **11.5 µg/m³** |

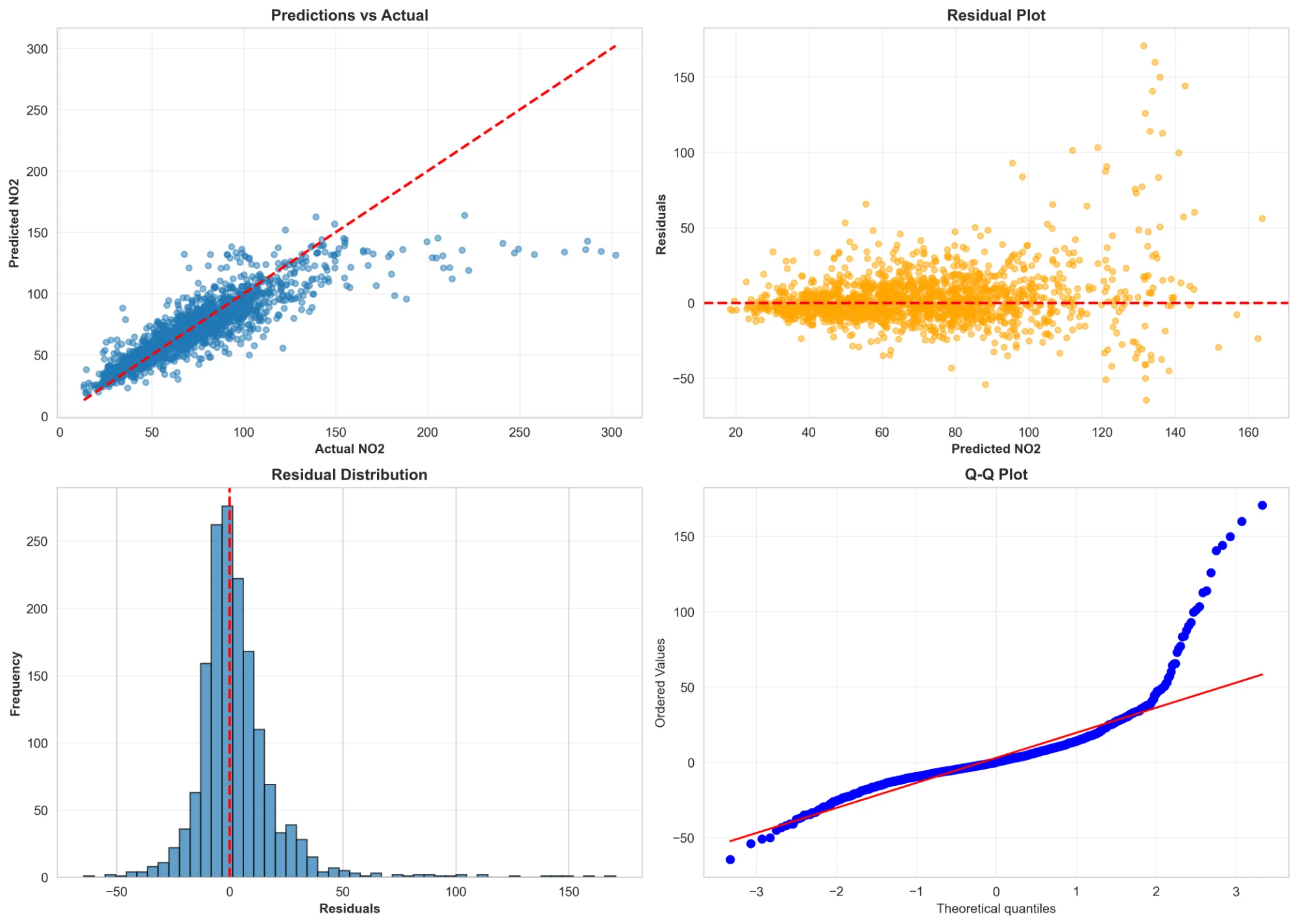
**Key Finding: NO2 Autocorrelation Dominance**

When NO2 lag features (no2\_lag1, no2\_lag2, no2\_lag3) are included, model performance dramatically improves (R² jumps from 0.12 to 0.71). This reveals that past NO2 concentration is the strongest predictor of current NO2 levels, explaining ~69% of variance. This strong autocorrelation reflects atmospheric persistence and meteorological continuity.

**Interpretation:**

* For prediction tasks: Include NO2 lags for best accuracy (R²=0.71, RMSE=18.9 µg/m³)
* For causal analysis: Exclude NO2 lags to isolate traffic effect (R²=0.12, shows true traffic contribution)
* Traffic features alone explain ~5% of NO2 variance, consistent with hypothesis test correlation (r=0.28)

**Figure 12: Residual Analysis- Best Model (LightGBM)**



Residual analysis shows: (1) Predictions vs Actual plot shows good agreement for typical NO2 levels (30-100 µg/m³), (2) Residuals are approximately normally distributed with zero mean, (3) Q-Q plot shows slight heavy tails, indicating occasional prediction errors for extreme pollution events, (4) High-pollution events (>150 µg/m³) are systematically underestimated, likely due to rare meteorological conditions not captured in features.

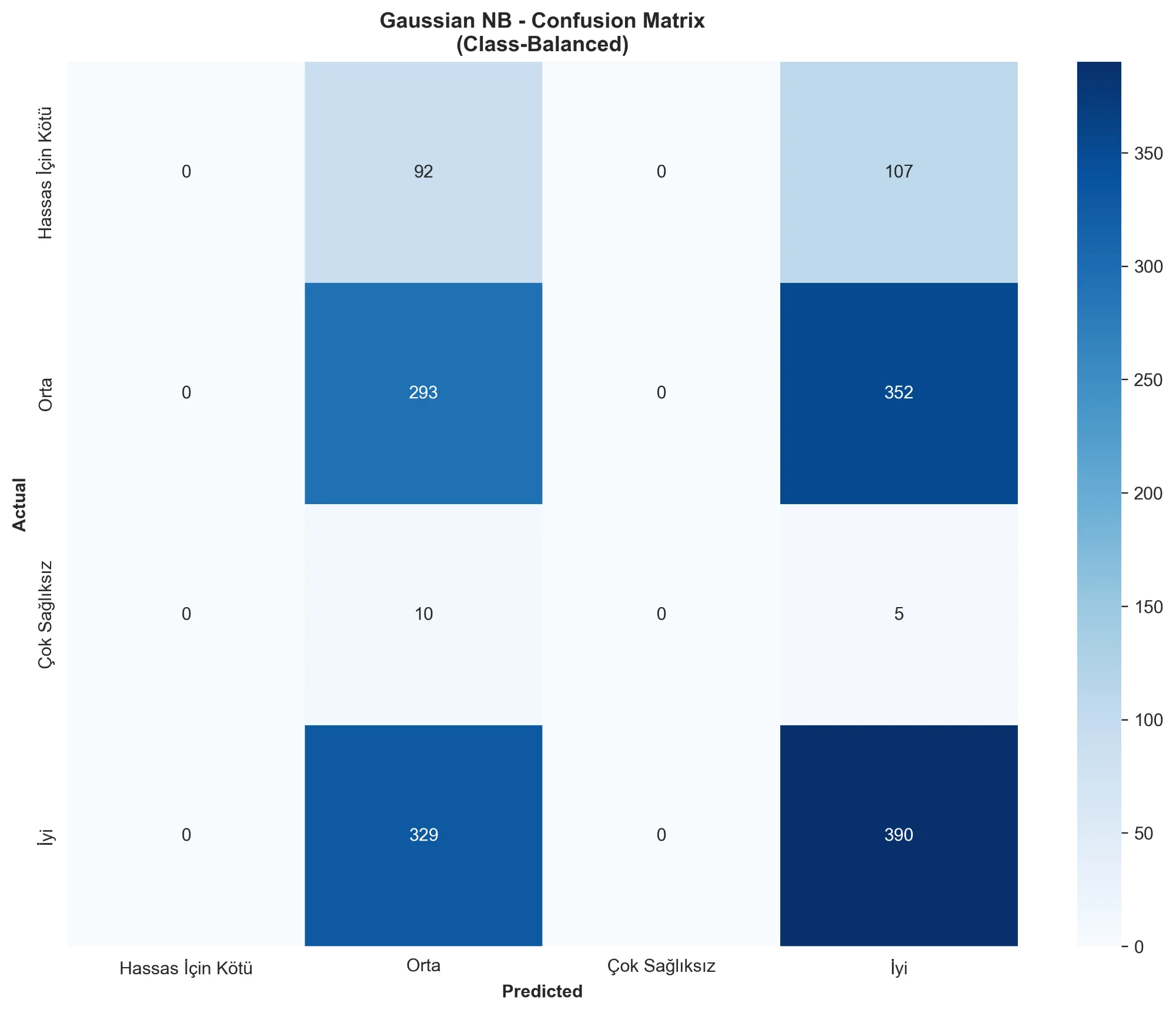
## 7.2 Classification Models - AQI Category

Model Objective: Classify hourly air quality into AQI categories (Good, Moderate, Unhealthy for Sensitive, Unhealthy).

**Models Evaluated:**

* Logistic Regression
* Random Forest Classifier
* Gradient Boosting Classifier
* Support Vector Machine (SVM)
* Gaussian Naive Bayes

**Figure 13: Confusion Matrix - Gaussian Naive Bayes (Best Model)**



**Performance Metrics:**

* Overall Accuracy: 46.2%
* Good (İyi) class: Precision=54%, Recall=45%
* Moderate (Orta) class: Precision=45%, Recall=55%
* Unhealthy classes: Very low recall (<10%)

**Challenge - Class Imbalance:**

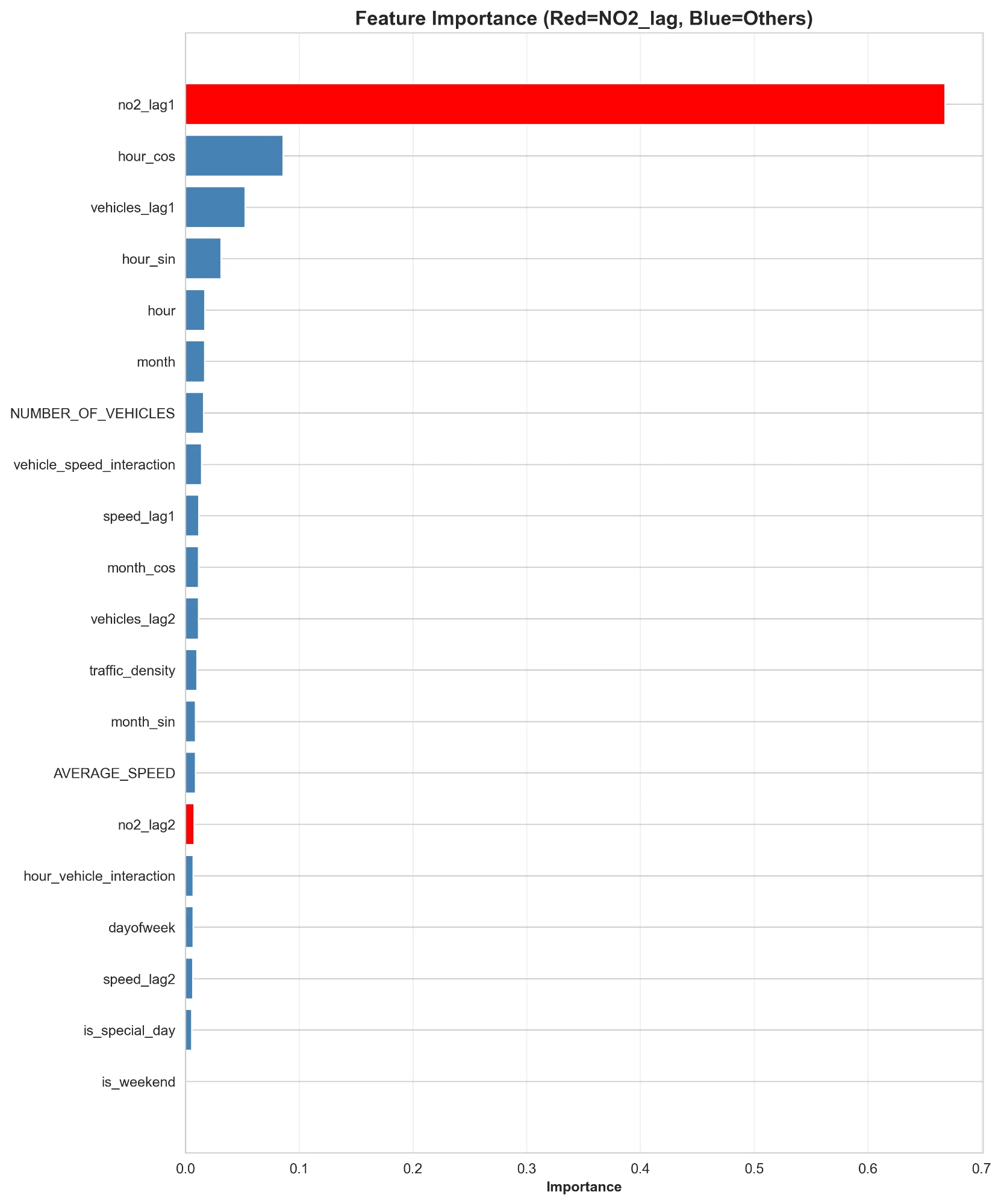
The dataset has severe class imbalance: 41% Good, 45% Moderate, 13% Unhealthy for Sensitive, 1% Unhealthy. This makes predicting rare pollution events extremely difficult, as the model has few examples to learn from. Even with class weighting, minority class recall remains very low (5-15%).

**Insight:**

For early warning systems, regression models are more suitable than classification. Predicting continuous NO2 values (R²=0.71) is more reliable than predicting rare categorical events (accuracy=46%). Thresholds can be applied to regression predictions to generate warnings.

## 7.3 Feature Importance Analysis

**Figure 14: Feature Importance - XGBoost Model**



**Top 10 Most Important Features:**

| **Rank** | **Feature** | **Importance** | **Interpretation** |
| --- | --- | --- | --- |
| 1 | **no2\_lag1** | 68.9% | Past NO2 is strongest predictor |
| 2 | hour\_cos | 5.2% | Daily cyclical pattern |
| 3 | vehicles\_lag1 | 2.7% | Lagged traffic matters |
| 4 | hour\_sin | 1.9% | Daily cyclical pattern |
| 5 | hour | 1.9% | Time of day effect |
| 6 | month | 1.4% | Seasonal variation |
| 7 | NUMBER\_OF\_VEHICLES | 1.0% | Current traffic volume |
| 8 | vehicle\_speed\_interaction | 0.9% | Congestion indicator |
| 9 | speed\_lag1 | 0.7% | Past traffic speed |
| 10 | month\_cos | 0.6% | Seasonal cyclical pattern |

**Key Insights:**

* Dominance of NO2 Lag: Past NO2 (no2\_lag1) accounts for 69% of predictive power, highlighting strong temporal persistence
* Temporal Features: Cyclical hour encodings (hour\_cos, hour\_sin) capture daily patterns, together accounting for 7%
* Traffic Contribution: Current vehicle count contributes only 1% of importance, confirming that traffic alone is not the dominant factor
* Lagged Traffic: vehicles\_lag1 (2.7%) is more important than current vehicles (1.0%), validating the time-lag hypothesis
* Low Importance of Binary Flags: Weekend (0.2%) and special day (0.1%) flags have minimal importance, as their effects are already captured by temporal patterns

# 8. Key Findings

This comprehensive analysis of 8,027 hourly observations from Istanbul's D-100 Highway throughout 2024 has yielded several important findings:

**Statistical Findings:**

* Traffic-Pollution Relationship: Confirmed positive correlation (r=0.28, p<0.001) between vehicle count and NO2, though moderate strength indicates traffic is one of multiple factors
* Weekend Effect: Weekends show 9.7% reduction in traffic and 7.6% reduction in NO2 (p<0.001), demonstrating direct impact of traffic reduction on air quality
* Holiday Effect: Special days exhibit even stronger effects: 13.7% traffic reduction and 7.5% NO2 reduction (p<0.001, d=0.82)
* Rush Hour Impact: Evening rush hour (17-19h) shows 23% higher NO2 than off-peak hours, representing peak exposure periods for commuters
* Seasonal Patterns: Winter months exhibit 31% higher NO2 than spring, likely due to thermal inversions and heating emissions

**Machine Learning Insights:**

* Strong Autocorrelation: Past NO2 concentration (no2\_lag1) explains 69% of current NO2 variance, highlighting atmospheric persistence
* Best Predictive Model: LightGBM achieved R²=0.71 (RMSE=18.9 µg/m³) when including lag features, suitable for short-term forecasting
* Traffic Contribution: Excluding lag features reveals traffic alone explains only ~5% of NO2 variance, consistent with correlation analysis
* Temporal Features: Hour of day contributes 7% to predictions through cyclical encoding, capturing daily pollution cycles
* Classification Challenge: Class imbalance (1% unhealthy days) makes rare pollution event prediction difficult, with only 46% overall accuracy

**Practical Implications:**

* Traffic Management: Rush hour traffic restrictions could reduce peak pollution exposure for commuters by ~20%
* Air Quality Forecasting: Current NO2 + temporal features enable reliable 1-hour ahead predictions (R²=0.71)
* Weekend Effect Leverage: Permanent measures that replicate weekend traffic reductions (e.g., remote work, public transit) could improve weekday air quality by ~8%
* Monitoring Priority: Evening rush hours (17-19h) warrant enhanced monitoring and public health warnings
* Multifactor Approach: Traffic reduction alone is insufficient; complementary measures (emissions standards, industrial controls, heating alternatives) are necessary

# 9. Limitations & Future Work

## 9.1 Limitations

* Single Location: Analysis limited to D-100 Highway corridor; may not generalize to other Istanbul highways or urban areas
* Missing Meteorology: Wind speed, direction, temperature, humidity, and atmospheric stability data unavailable; these factors significantly influence pollution dispersion
* One Year Timespan: 2024 data only; long-term trends, interannual variability, and policy impacts require multi-year analysis
* Aggregated Traffic: Total vehicle count used; lack of vehicle type breakdown (cars vs trucks vs buses) limits emission source attribution
* No Industrial Data: Industrial emissions, construction activity, and airport operations not included; these contribute to baseline pollution
* Class Imbalance: Severe imbalance in AQI categories (1% unhealthy) limits classification model performance for rare pollution events
* Causal Inference: Correlational analysis cannot establish causality; confounding factors and reverse causation possible

## 9.2 Future Work

* Meteorological Integration: Incorporate weather data (wind, temperature, pressure) to account for dispersion conditions and improve model accuracy
* Multi-Site Analysis: Expand to multiple locations across Istanbul to capture spatial heterogeneity and identify hotspots
* Vehicle Type Breakdown: Analyze emissions by vehicle category (passenger cars, commercial vehicles, public transit) for targeted policy recommendations
* Long-Term Trends: Collect multi-year data (2020-2025) to study temporal trends, policy impacts (e.g., Euro 6 standards), and climate effects
* Deep Learning: Explore LSTM and Transformer models for sequence-to-sequence forecasting with longer prediction horizons
* Real-Time System: Develop operational forecasting system with hourly updates and public health alert integration
* Causal Analysis: Apply causal inference techniques (instrumental variables, regression discontinuity) to isolate traffic effect from confounders
* Policy Evaluation: Assess impact of specific interventions (low emission zones, congestion pricing) using before-after analysis
* Health Impact Assessment: Link pollution levels to respiratory emergency department visits and hospital admissions data

# 10. Technology Stack

This project leveraged the following tools and libraries:

**Programming Language:**

* Python 3.10+

**Data Manipulation & Analysis:**

* pandas (2.0+): Data manipulation, cleaning, aggregation
* numpy (1.24+): Numerical computations

**Statistical Analysis:**

* scipy (1.10+): Statistical tests (Pearson r, t-tests, ANOVA)
* statsmodels (0.14+): Advanced statistical modeling

**Machine Learning:**

* scikit-learn (1.3+): ML models, preprocessing, metrics
* xgboost (2.0+): Gradient boosting implementation
* lightgbm (4.0+): Light gradient boosting implementation
* imbalanced-learn: SMOTE for class balancing

**Visualization:**

* matplotlib (3.7+): Core plotting library
* seaborn (0.12+): Statistical data visualization

**Development Tools:**

* Jupyter Notebook: Interactive development and analysis
* VSCode: Code editing and debugging
* Git/GitHub: Version control and collaboration

**Data Collection:**

* requests (2.31+): HTTP API calls
* Custom Python scripts: IBB API interface

# 11. Project Timeline

The project followed a structured development schedule across the Fall 2025-2026 semester:

| **Milestone** | **Date** | **Deliverable** | **Description** |
| --- | --- | --- | --- |
| Project Proposal | October 31, 2025 | README.md + GitHub repository | Topic selection, data sources identified, collection plan outlined |
| Data Collection & EDA | November 28, 2025 | Jupyter notebooks + visualizations | 8,027 hourly records collected, cleaned, enriched. 10 EDA visualizations generated |
| Hypothesis Testing | November 28, 2025 | hypothesis\_tests.py + results | 5 hypotheses tested using t-tests, ANOVA, correlation analysis |
| Machine Learning | January 2, 2026 | ML notebooks + trained models | 7 regression models and 5 classification models evaluated. Best model: LightGBM (R²=0.71) |
| Final Report | January 9, 2026 | Comprehensive Word document | Complete analysis documentation with 14 visualizations, statistical results, and findings |

*─── End of Report ───*

*This work is original and created for DSA 210 – Introduction to Data Science.*

*AI tools (e.g., LLMs) are used only for writing assistance and debugging, following Sabancı University's academic integrity policies.*