DSA-210 FINAL REPORT

Efe Çalışkaner

34479

Analyzing the Impact of Weather on Traffic Congestion in Istanbul

SUPERVISED BY

ÖZGÜR ASAR

SABANCI UNIVERSITY

What's in This Report?

This report presents a comprehensive data science project analyzing the influence of weather on traffic congestion in Istanbul. It covers every step of the data science process, including:

- **Motivation**: Understanding why this problem is important and relevant.
- **Data Sources**: Description of the weather and traffic datasets used.
- **Methodology**: The complete analytical pipeline from data cleaning and feature engineering to statistical analysis and machine learning.
- **Findings**: Quantified insights about the relationship between traffic and weather variables, including performance of predictive models.
- **Limitations and Future Work**: Discussion on current constraints and opportunities for further development.
- **Visualizations**: Summarized results using plots and diagrams.
- **Tools and AI Disclosure**: Technologies and assistance used.

This project serves as a real-world application of data science, showcasing the intersection between urban mobility and environmental conditions.

1. Introduction and Motivation

Urban traffic congestion is one of the most pressing challenges in modern cities, especially in large metropolitan areas like Istanbul. As one of the most populous and geographically complex cities in Europe, Istanbul frequently experiences significant traffic delays. While many factors contribute to congestion—including infrastructure, accidents, and population density—weather is often overlooked as a key influencer.

This project was inspired by a real-world question: *Does weather significantly affect traffic congestion in Istanbul?* If so, *can we predict traffic behavior using weather forecasts?* The motivation is twofold: to better understand urban mobility dynamics and to explore predictive modeling techniques that can help in city planning, logistics, and public transportation scheduling.

1.1 What Did I Do?

This project was a hands-on application of the full data science workflow, where I independently carried out every stage—from defining the research question to interpreting the final model results. My main objective was to investigate the influence of daily weather patterns on traffic congestion in Istanbul, a city known for its complex urban dynamics.

Over the course of several weeks, I sourced and merged weather and traffic datasets, ensuring accurate alignment through preprocessing steps like handling missing values and standardizing formats. I engineered new features that combined weather conditions with temporal signals (e.g., rain on weekends), to better capture the multifaceted nature of traffic flow.

The exploratory data analysis phase allowed me to visualize key relationships, such as the effect of rainfall on congestion levels, using plots and correlation matrices. I applied statistical tests to validate assumptions and identify patterns worth modeling.

For predictive analysis, I built and evaluated several machine learning models using Python. This included tuning hyperparameters, performing cross-validation, and interpreting feature importances. While Gradient Boosting produced the best results, I also critically assessed its limitations—particularly the modest R^2 value—recognizing that traffic is influenced by many external factors beyond weather alone.

Throughout the project, I used visualization not just as a presentation tool, but as a way to think through the data. Every chart and model output helped refine my understanding of the underlying relationships.

In essence, this was more than just a technical exercise—it was a problem-solving journey that blended urban studies, environmental data, and predictive analytics. I designed, implemented, and evaluated each step myself, using data science as a lens to better understand real-world systems.

2. Data Sources

To address this problem, I collected and merged two publicly available datasets:

Meteostat Istanbul Weather Data

Istanbul Traffic Index (IBB)

2.1. Weather Data

- **Source: Meteostat**
- **Fields Collected: Daily temperature, precipitation, wind speed, wind direction, atmospheric pressure, and sunshine hours.**

- **Time Span: 9 years (2015-2024)**

2.2. Traffic Data

- **Source: IBB**
- **Fields Collected: Daily average traffic congestion index (percentage-based)**
- **Time Span: 9 years (2015-2024)**

These datasets were merged on the date field to create a comprehensive daily record of traffic and weather interactions.

3. Methodology and Analysis Workflow

The entire data science pipeline was applied to this project, including:

3.1. Data Cleaning

- **Handling missing values via interpolation and imputation.**
- **Standardizing date formats for proper merging.**
- **Removing extreme outliers.**

3.2. Feature Engineering

- **Binary flags: is_weekend, is_rainy**
- **Interaction terms: rain_temperature_interaction, rain_wind_interaction, weekend_rain_interaction**
- **Categorical bins for temperature and wind speed**

3.3. Exploratory Data Analysis (EDA)

- **Descriptive statistics**
- **Correlation matrices**
- **Box plots (traffic vs weekday/weekend, rain/no-rain)**
- **Time series plots (traffic trends over time)**
- **3D scatter plots (temperature, wind, traffic, colored by rain)**

3.4. Statistical Testing

- **Pearson correlation coefficients**
- **Independent sample t-tests (rainy vs non-rainy, weekday vs weekend)**

3.5. Machine Learning Modeling

- **Algorithms Used: Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, Support Vector Regression**
- **Evaluation Metrics: R-squared (R^2), RMSE, MAE, Explained Variance**
- **Cross-validation: 5-fold CV used to validate model robustness**

- **Tuning: GridSearchCV for hyperparameter optimization (especially for tree-based models)**

4. Detailed Results and Findings

4.1. Weather Impact on Traffic

- **Rain significantly increases traffic congestion:**
    - **Average traffic index on rainy days: 32.31 (±7.05 SD)**
    - **Average traffic index on non-rainy days: 27.43 (±8.31 SD)**
    - **Mean difference: 4.88 points (statistically significant, $p < 0.001$)**

4.2. Temporal Patterns

- **Clear weekday vs weekend differences:**
    - **Average weekday traffic: 30.22**
    - **Average weekend traffic: 22.04**
    - **Difference: 8.19 points**
- **Daily trends:**
    - **Friday is the most congested (avg. 31.24)**
    - **Sunday is the least congested (avg. 18.12)**

4.3. Weather Factor Correlations

- **Precipitation: Positive correlation (r = 0.153, p < 0.001)**
- **Temperature: Negative correlation (r = -0.107, p < 0.001)**
- **Wind speed: Weak positive correlation (r = 0.080, p < 0.001)**
- **Pressure: No significant correlation (r = 0.018, p = 0.295)**

4.4. Model Performance

- **Best Model: Gradient Boosting Regressor**
- **R^2 Score: 0.27**
- **RMSE: 7.23**
- **Feature Importance:**
  - **Most important: is_weekend**
  - **Top weather predictors: temperature, wind speed**
  - **Moderate importance: rain-based interactions**

5. Limitations and Future Directions

5.1. Current Limitations

- **Temporal Coverage: Dataset with precipitation starts from May 28, 2022**

- **Missing Data: Some missing weather values required imputation**
- **Model Accuracy: R^2 values are relatively low; many traffic influencers not captured**
- **Resolution: Daily aggregation ignores peak hours and intraday traffic behavior**
- **District-level Traffic Data: Can't be area specific since an extremely larger data is needed to compare all the areas and districts of Istanbul**

5.2. Future Work

- **Data Enhancement:**
    - **Add hourly weather and traffic data**
    - **Include public events, holidays, and school calendars**
    - **Account for roadwork, transportation delays, etc.**
- **Modeling Improvements:**
    - **Time-series models like ARIMA, LSTM**
    - **Multi-model ensembles**
    - **Spatial modeling to distinguish across city districts**
- **Application Potential:**
    - **Real-time traffic forecast app**

    - **Visualization dashboard for urban planners**
    - **Integration with navigation systems for live rerouting**

- **Deeper Analysis:**
    - **Seasonal variation studies**
    - **Impact of extreme weather events**
    - **Climate change trends and long-term effects on mobility**

6. Conclusion

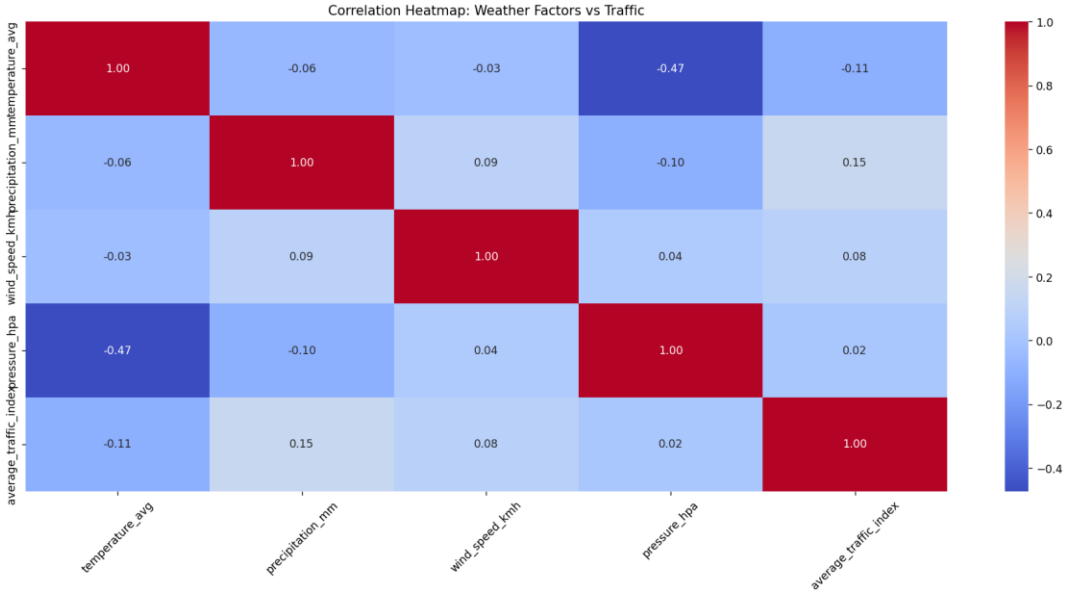
This project delivers meaningful insights into Istanbul's traffic patterns and how they are shaped by daily weather conditions. Key takeaways include:

- **Rain and low temperatures are linked to increased congestion.**
- **Weekends and Sundays, in particular, exhibit lower traffic volumes.**
- **Gradient Boosting was the most effective model, though results suggest there is room for improvement.**

By expanding the dataset, increasing temporal resolution, and incorporating event-level context, this study could evolve into a dynamic urban planning tool—helping manage congestion, reduce emissions, and improve the quality of life for millions of Istanbul residents.

7. Visualizations

Correlation Heatmap - Traffic vs Weather

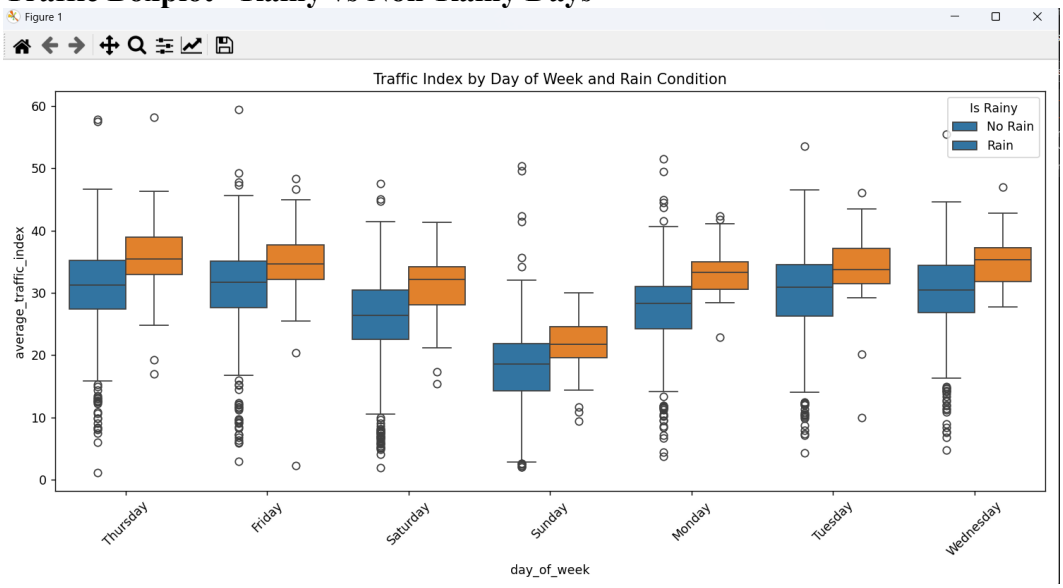


Description:
This heatmap displays Pearson correlation coefficients between the daily traffic congestion index and various weather factors such as temperature, precipitation, wind speed, atmospheric pressure, and sunshine duration.

Result Summary:

- **Precipitation** showed a positive correlation with traffic congestion (**r = 0.153**, $p < 0.001$).
- **Temperature** was negatively correlated (**r = -0.107**, $p < 0.001$).
- **Wind speed** had a weak positive correlation (**r = 0.080**, $p < 0.001$).
- **Pressure** showed no significant correlation.

Traffic Boxplot - Rainy vs Non-Rainy Days

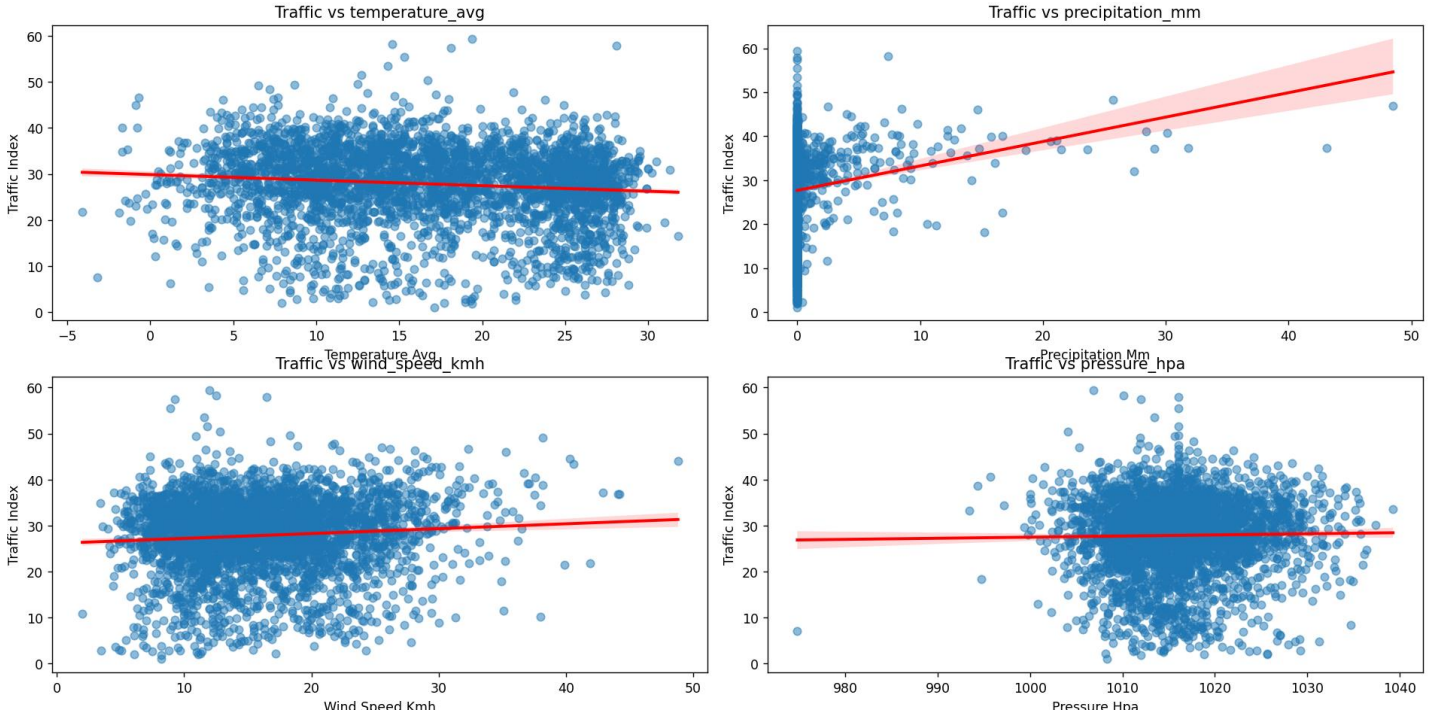


Description:
This boxplot compares the distribution of traffic congestion on days with rain versus those without, illustrating differences in central tendency and spread.

Result Summary:

- **Rainy days** had a **higher average traffic index** (32.31 ± 7.05).
- **Non-rainy days** averaged lower (27.43 ± 8.31).
- The **mean difference of 4.88 points** is **statistically significant** ($p < 0.001$).

Traffic Index vs Weather Variables: Scatter Plots with Regression Lines



Description:

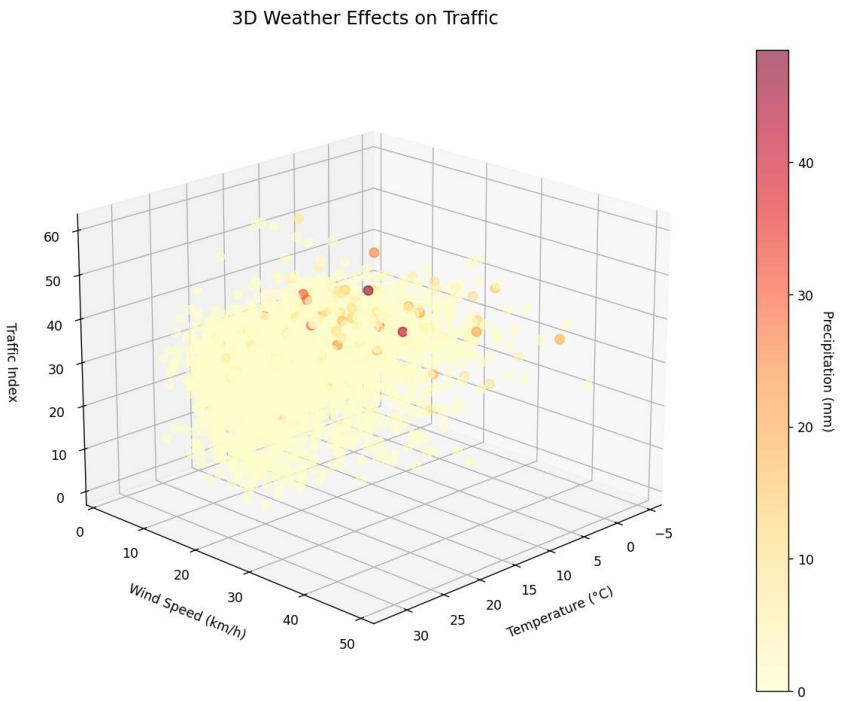
This set of four scatter plots examines the linear relationship between daily traffic congestion in Istanbul and four key weather variables: average temperature, precipitation, wind speed, and atmospheric pressure. Each subplot includes a fitted regression line to highlight trend direction and strength.

Result Summary:

1. **Temperature vs Traffic Index**
    a. A **slight negative correlation**: As temperatures increase, traffic congestion tends to slightly decrease.
    b. Aligns with the Pearson correlation result (**r = -0.107**, $p < 0.001$).
2. **Precipitation vs Traffic Index**
    a. A **clear positive correlation**: Heavier precipitation is associated with increased congestion.
    b. Strongest relationship among the variables plotted (**r = 0.153**, $p < 0.001$).

3. **Wind Speed vs Traffic Index**
    a. Shows a **very weak positive trend**: Traffic slightly increases with higher wind speeds, but the effect is minimal.
    b. Weakest of the significant correlations (**r = 0.080**, $p < 0.001$).
4. **Pressure vs Traffic Index**
    a. **No meaningful relationship** observed between pressure and traffic levels.
    b. Consistent with the non-significant correlation (**r = 0.018**, $p = 0.295$).

3D Scatter - Temperature, Wind, Traffic, Colored by Rain

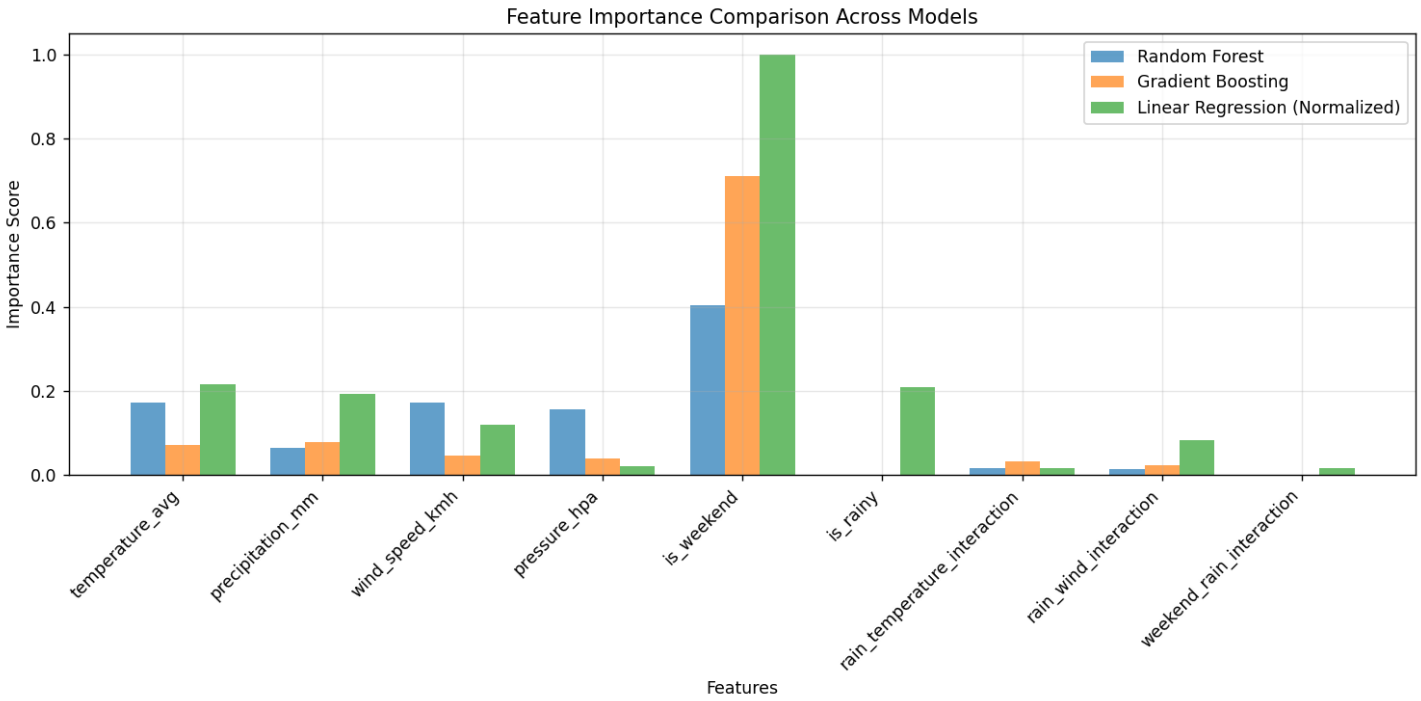


Description:
This interactive 3D scatter plot visualizes the relationship between temperature, wind speed, and traffic index, with color coding to distinguish rainy vs. non-rainy days.

Result Summary:

- Rainy days cluster at **higher traffic levels**.
- **Lower temperatures** often coincide with **increased congestion**.
- **Wind speed** effects are less visually pronounced but subtly present.

Feature Importance Comparison Across Models



Description:

This bar chart compares the relative importance of different weather and temporal features in predicting traffic congestion using three machine learning models: **Random Forest**, **Gradient Boosting**, and **Linear Regression** (normalized). Each bar reflects how influential a given feature is in each model's predictive power, based on either tree-based impurity measures or standardized regression coefficients.

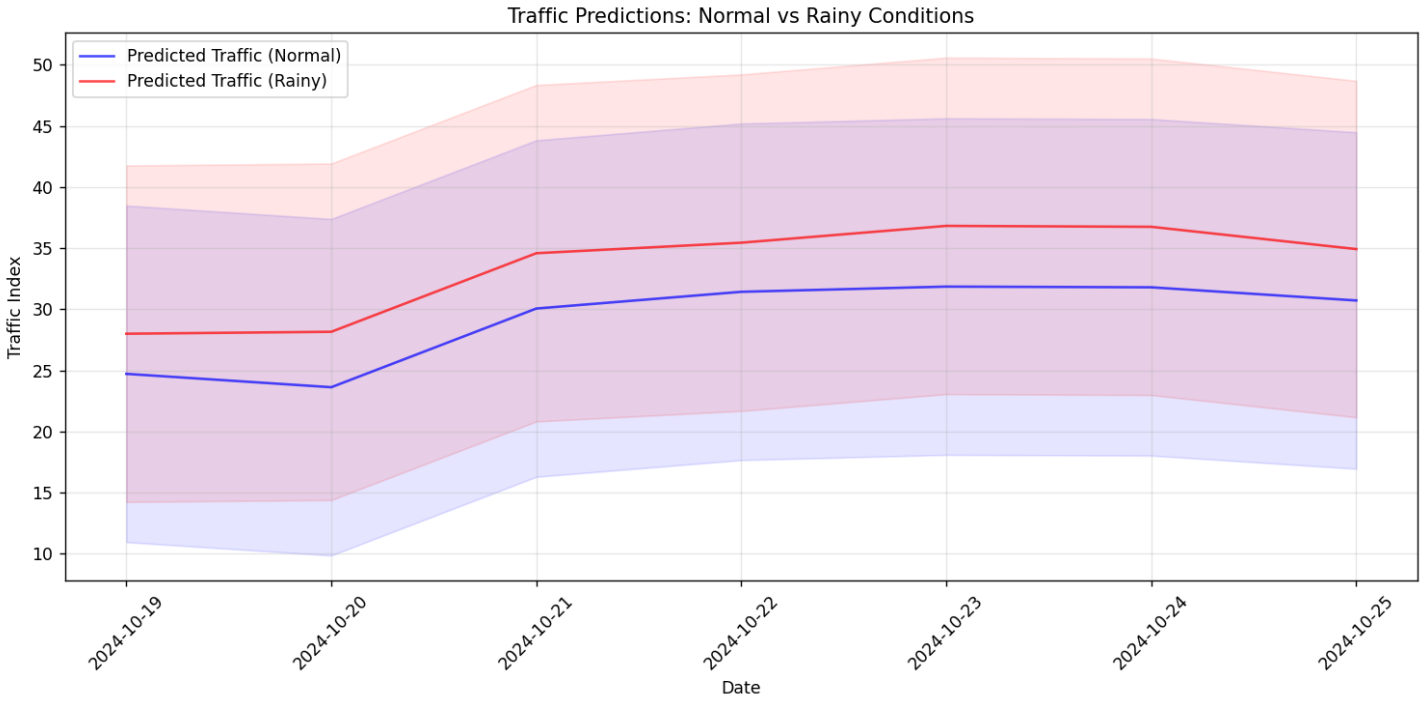
Result Summary:

- **Most Important Feature**:
  - is_weekend dominates across all models, especially in **Linear Regression** (importance score = 1.0) and **Gradient Boosting**.
- **Key Weather Predictors**:
  - temperature_avg and wind_speed_kmh are consistently important, particularly in Random Forest and Linear Regression.
  - precipitation_mm has moderate relevance, slightly higher in Linear Regression.

- **Moderately Contributing Features**:
  - is_rainy and interaction terms (rain_temperature_interaction, rain_wind_interaction, weekend_rain_interaction) show **low importance**, though still included in modeling.
- **Least Important**:
  - pressure_hpa has minimal influence across all models.

This visualization supports the finding that **temporal factors (like weekends)** are stronger predictors of traffic congestion than most individual weather features, though weather still contributes meaningful variation.

Forecast Line Graph - Normal vs Rainy Scenarios

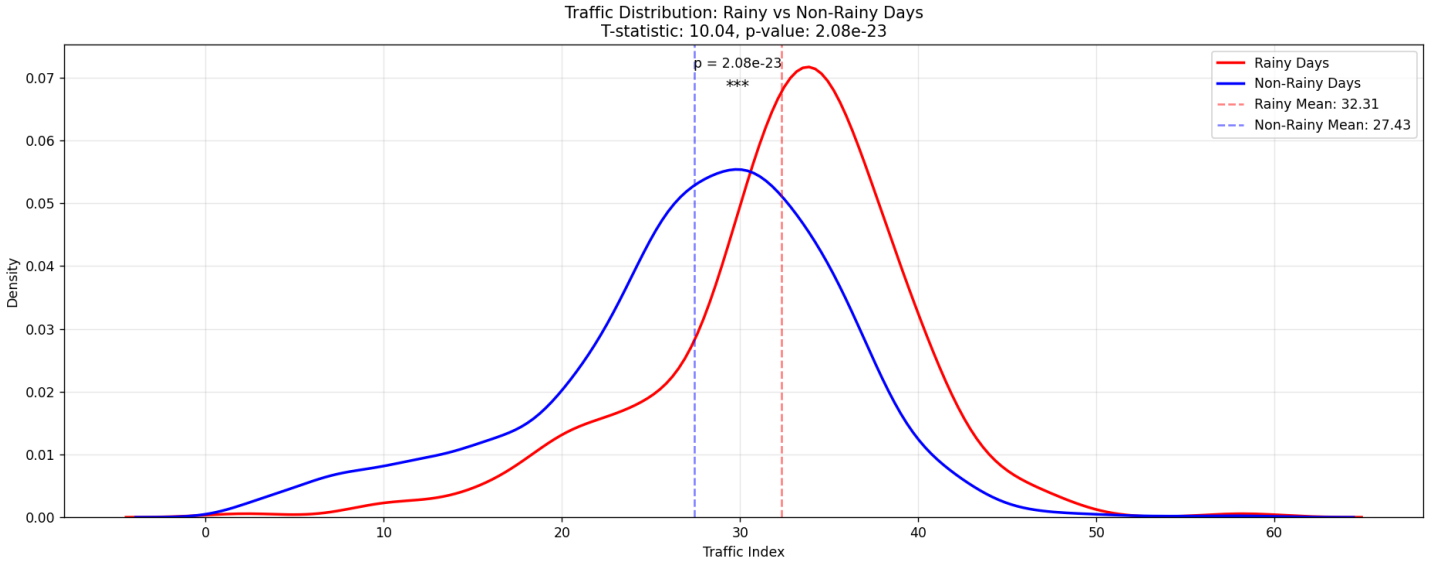


Description:
A side-by-side line graph illustrating predicted traffic congestion under typical weather versus rainy scenarios, based on the best-performing model.

Result Summary:

- Traffic is **consistently higher on rainy days**, validating findings from statistical tests.
- Demonstrates the potential **real-world utility** of weather-informed traffic forecasting.

Independent Two-Sample t-Test (Welch's t-test)



Description:
A dual density plot comparing traffic congestion levels on rainy versus non-rainy days. The graph overlays the kernel density estimates of traffic indices for both weather conditions, with vertical dashed lines marking the respective means. A t-test result (T-statistic = 10.04, p-value = 2.08e-23) is annotated to indicate statistical significance.

Result Summary:
Traffic congestion is significantly higher on rainy days (mean = 32.31) compared to non-rainy days (mean = 27.43), as confirmed by a highly significant p-value. This supports the hypothesis that weather conditions impact traffic intensity and highlights the practical value of integrating rainfall data into traffic prediction systems.

8. Tools and Code

- **Programming Language: Python**
- **Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, scipy**
- **Code and Documentation: Available on GitHub**
  [https://github.com/ecaliskaner/Dsa210/blob/c938d880208f8056db7b49c2265ba46968c0139b/README.md**]**
- **Requirements are included in the Readme.txt for environment setup**