Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Anomaly Detection for Shared Mobility

Master Candidate

Elnur Isgandarov

Student ID 2014431

Supervisor

Prof. Gian Antonio Susto

University of Padova

Co-supervisor

Prof. Federico Chiariotti

University of Padova

Academic Year
2024/2025

To my parents
and friends

Abstract

Shared mobility systems, such as bike-sharing networks, have gained significant traction in urban transportation due to their environmental and economic benefits. Understanding anomalies within these systems is critical for improving operational efficiency, user satisfaction, and system resilience. Current state-of-the-art approaches to anomaly detection in shared mobility often focus on user behavior or system-level metrics. However, these methods frequently lack interpretability and fail to address the need for fast, unsupervised techniques, as labeled data is typically unavailable in this domain.

This thesis addresses these limitations by proposing a comprehensive approach to anomaly detection in shared mobility systems. The approach integrates diverse data sources, including bike-sharing data, weather data, and mass transit data, to provide a richer contextual understanding of anomalies. Machine learning techniques, such as Isolation Forest, are utilized alongside feature selection methods like Depth-based Isolation Forest Feature Importance (DIFFI) to identify atypical patterns in usage that may indicate operational inefficiencies or unusual user behavior.

Key contributions of this work include the development of spatiotemporal, environmental, and transit-related features, the application of interpretable anomaly detection models, and an in-depth analysis of the relationships between anomalies and external factors. These findings offer actionable insights for shared mobility operators, enabling enhancements in system performance and user experience.

Contents

List of Figures

1 Introduction

Shared mobility systems, such as bike-sharing networks, have become vital components of urban transportation due to their environmental benefits and cost-effectiveness. However, the complex dynamics of these systems often lead to operational inefficiencies, fluctuations in user demand, and challenges in maintaining service quality [3]. Anomalies within shared mobility systems, such as unexpected drops in ridership, unbalanced station usage, or irregular user behavior, can disrupt service reliability and user satisfaction [18]. The ability to detect and interpret these anomalies is essential for improving system resilience, optimizing operations, and enhancing user experience.

Current anomaly detection methods in shared mobility often lack interpretability. Existing approaches frequently focus on user behavior or system-level metrics but fail to provide actionable insights for operators [17]. Moreover, the absence of labeled data in shared mobility complicates the application of supervised learning techniques, necessitating the use of robust unsupervised methods that can still offer meaningful interpretations [6].

This thesis aims to develop an interpretable anomaly detection framework tailored to shared mobility systems. The objectives are to integrate multi-source data including bike-sharing trip data, weather data, and public transit information to enrich the context of anomaly detection; apply machine learning techniques, specifically Isolation Forest, for unsupervised anomaly detection; leverage feature importance methods like DIFFI (Depth-based Isolation Forest Feature Importance [2]) to enhance model interpretability; and analyze the spatial and temporal dynamics of anomalies while evaluating their correlation with external factors.

The primary contributions of this thesis include the development of a comprehensive anomaly detection framework that integrates diverse data sources for richer contextual analysis, the application of DIFFI to improve interpretability in Isolation Forest models, and a two-stage anomaly detection approach transitioning from trip-level to station-level analysis for improved reliability. Additionally, the thesis presents in-depth case studies exploring specific anomaly patterns, such as zero-duration anomalies and temporal spikes.

The thesis is structured as follows: Chapter 2 reviews existing literature on shared mobility systems, anomaly detection techniques, and feature engineering. Chapter 3 outlines the methodology, including data collection, feature engineering, and the anomaly detection framework. Chapter 4 presents the preliminary trip-focused analysis, highlighting initial findings and challenges. Chapter 5 details the final station-focused analysis, offering deeper insights into anomaly patterns. Finally, Chapter 6 concludes the thesis, summarizing contributions and proposing directions for future research.

2

Related Work (Literature Review)

Shared mobility systems, particularly bike-sharing networks, have become integral to urban transportation, offering sustainable and efficient mobility alternatives. However, ensuring their operational effectiveness requires the early detection and interpretation of anomalies that may indicate system inefficiencies, demand fluctuations, or atypical user behavior. While prior research has explored various machine learning-based methods for anomaly detection in mobility systems, challenges remain in achieving model interpretability and effectively integrating external factors such as weather conditions and public transit availability into these analyses. This chapter reviews the literature on anomaly detection in shared mobility, feature engineering approaches, and interpretability methods, highlighting key limitations and framing the contributions of this thesis.

2.1 Shared Mobility Systems and Their Challenges

Anomaly detection is a valuable tool for improving shared mobility systems, especially in the context of rebalancing bike-sharing networks. Rebalancing ensures that bicycles and docking spaces are evenly distributed across stations to meet fluctuating user demand. Unexpected usage patterns, such as sudden demand surges, low-activity stations, or irregular trip behaviors, can disrupt this balance and reduce system efficiency. Detecting these anomalies allows operators to make informed decisions that optimize resource allocation, minimize downtime, and enhance overall user satisfaction.

A key challenge in maintaining bike-sharing services is the management of rebalancing trucks, which ensure an adequate distribution of bicycles and open docking stalls across stations, despite ever-changing user demand. In *A Dynamic Approach to Rebalancing Bike-Sharing Systems*, the authors highlight how the dynamic nature of bike usage makes it difficult to continuously meet demand at every station [3]. Even a small misalignment between predicted and actual demand can lead to empty stations in high-demand areas or an oversupply of bikes in regions with low ridership, both of which create operational inefficiencies and negatively impact user satisfaction.

Beyond the logistical complexity of rebalancing, numerous studies emphasize the significant role of external factors particularly weather conditions in influencing ridership patterns. Although it is generally acknowledged that biking decreases in poor weather and increases in fair conditions, the intensity and direction of these effects can differ across individual stations. For example, stations near tourist attractions might experience reduced demand when it is raining, whereas stations located near business districts could still maintain moderate levels of usage. Hence, it is vital to analyze data at both the system level and the station level, as this nuanced view can better capture localized behaviors. In *Investigation on the Effects of Weather and Calendar Events on Bike-Sharing*, Kim et al. demonstrate how variables like precipitation, temperature, and public holidays uniquely shape daily ridership patterns [7]. They show that high temperatures exceeding 30°C can suppress biking activity, highlighting the need to incorporate both severe cold and extreme heat events into forecasting models. Additionally, their findings indicate that non-working days affect demand differently at various hours, suggesting that time-dependent analysis is essential for accurate modeling and service optimization.

Disruptive events, such as the COVID-19 pandemic, further underscore the resilience of bike-sharing systems in comparison to other public transit options. *The Link Between Bike Sharing and Subway Use During the COVID-19 Pandemic* reveals that bike-sharing ridership experienced a relatively smaller decline compared to the subway system 71% versus 90% and registered a notable increase in average trip duration (from 13 minutes to 19 minutes) [14]. The study also finds evidence for a modal shift, where some subway users opted for bike-sharing, demonstrating how shared mobility can bolster the robustness of urban transport infrastructures under crisis conditions. Such insights are pivotal for designing strategies that sustain mobility and public health in post-pandemic contexts, where flexible and adaptive transport modes may be particularly advantageous.

In addition to weather and disruptive events, the interplay between bike-sharing networks and conventional mass transit shapes overall demand patterns. *Spatial and Temporal Analysis of Bike-Sharing Use in Cologne* provides evidence that public transit disruptions can induce shifts in bike-sharing usage [12]. Although the short-term public transit changes caused by a construction site did not dramatically impact overall biking rates, proximity to universities and the presence of amenities such as restaurants and shops boosted ridership. Moreover, higher temperatures correlated positively with usage, whereas rainfall suppressed it-trends that align with previous findings. These results suggest that bike-sharing and mass transit can operate synergistically and highlight the importance of spatial planning, where stations are strategically placed near high-demand points of interest to enhance ridership and encourage intermodal journeys.

In summary, shared mobility systems offer substantial benefits for sustainable urban travel, but they face significant operational and analytical challenges. This section demonstrates the importance of investigating a variety of external influences ranging from environmental conditions to public health emergencies at both the station and system levels. The following subsections delve deeper into methods for detecting anomalies in these systems, strategies for crafting representative features, and approaches for enhancing the interpretability of predictive models.

2.2 Anomaly Detection in Mobility Systems

Anomalies in shared mobility systems refer to unusual patterns in ridership behavior, station occupancy, or bike flow that deviate from expected trends. Such anomalies may signal operational inefficiencies, demand surges, or even system failures that require intervention. Detecting these irregularities is crucial for improving service reliability, optimizing resource allocation, and enhancing user experience. Researchers have proposed a wide range of anomaly detection techniques, spanning traditional statistical methods to advanced machine learning approaches.

One approach is functional outlier detection, which targets deviations in temporal occupancy trends. In *Data Adaptive Functional Outlier Detection: Analysis of the Paris Bike Sharing System Data* [9], a novel two-stage method is introduced for identifying abnormal patterns in real-time station occupancy data. In the first stage, a clean reference dataset is generated using extreme statistics from random sampling to filter out potential outliers. The second stage applies a multiple hypothesis testing approach with an adaptive false discovery rate (FDR) control mechanism to refine the detection process. This method achieves higher accuracy than traditional outlier detection techniques and assists operators in pinpointing station-level anomalies that may necessitate policy adjustments or operational interventions.

Another method involves clustering-based event detection, which focuses on identifying significant deviations in ridership patterns linked to external events. *Detecting (Unusual) Events in Urban Areas Using Bike-Sharing Data* [8] proposes a spectral clustering approach to detect unusual spatiotemporal events by analyzing changes in bike-sharing network dynamics. By modeling bike stations and flows as an evolving graph, the method tracks fluctuations in edge and node values over time to determine whether observed ridership clusters align with expected commuter behavior or represent unexpected anomalies. The findings reveal that even localized events such as university activities or community gatherings can imprint measurable signatures on bike-sharing data, demonstrating the potential of bike-sharing systems as passive sensors for urban mobility disruptions.

Beyond temporal and event-driven anomalies, some studies concentrate on spatial anomaly detection to assess station-level imbalances. *Using Spatial Outliers Detection to Assess Balancing Mechanisms in Bike-Sharing Systems* [13] applies an improved version of Morans scatterplot using Gowers similarity metric to detect stations with usage patterns significantly different from their spatial neighbors. Analysis on the Paris Vélib dataset identifies several outlier stations characterized by persistently high or low occupancy levels. To mitigate these imbalances, the study introduces a user-driven redistribution approach whereby minor behavioral adjustments such as encouraging riders to return bikes to underutilized stations enhance overall system homogeneity. This work highlights how anomaly detection can serve not only as a monitoring tool but also as a basis for strategic intervention in shared mobility networks.

Despite the promise of these approaches, several challenges persist. One major limitation is the lack of labeled data, which complicates the evaluation and validation of model predictions. Many real-world bike-sharing datasets do not include explicit ground truth labels for anomalies, necessitating the use of unsupervised learning techniques. In addition, scalability remains a concern, as high-dimensional spatiotemporal datasets require algorithms that can efficiently manage large-scale urban networks in real time. Finally, distinguishing true anomalies from short-term fluctuations is challenging; while temporary disruptions (e.g., minor weather changes) may not warrant intervention, prolonged imbalances do.

In summary, the evolution of anomaly detection in mobility systems has progressed from rule-based and statistical methods to techniques such as functional outlier detection, clustering-based approaches, and spatial anomaly identification. Although these methods provide valuable insights, persistent issues such as unlabeled data, scalability, and the differentiation of genuine anomalies from transient variations must be addressed to develop more robust and interpretable models.

2.3 Feature Engineering and Model Interpretability

Feature engineering is fundamental to improving the performance of anomaly detection models, particularly in complex systems like shared mobility. Features derived from spatiotemporal data, weather conditions, and transit availability provide essential context for identifying anomalies and enhancing model accuracy.

Automatic Bike Sharing System Planning from Urban Environment Features highlights the significant impact of weather conditions on bike-sharing demand, demonstrating how factors such as precipitation and temperature influence ridership patterns [15]. Additionally, prior research emphasizes the dynamic interaction between bike-sharing and public transit systems, especially during periods of high demand or service disruptions [14]. Despite this, most studies focus on individual data sources, such as weather or transit data, without fully integrating them. This thesis adopts a multi-source feature engineering approach, incorporating bike-sharing, weather, and transit data to provide a comprehensive view of system anomalies.

While well-designed features are vital for accurate anomaly detection, model interpretability is equally crucial to ensure reliability and transparency. Techniques like Isolation Forest are commonly used in unsupervised anomaly detection due to their efficiency and scalability, but their black-box nature often limits insight into how specific features influence anomaly identification.

To improve interpretability, researchers have introduced methods that clarify feature contributions within anomaly detection models. Carletti et al. [2] proposed Depth-based Isolation Forest Feature Importance (DIFFI), a method tailored for Isolation Forest that quantifies the impact of individual features on anomaly detection outcomes. DIFFI enhances transparency, enabling operators to understand the drivers behind detected anomalies.

Although general interpretability tools like SHAP and LIME offer valuable explanations, they present challenges in unsupervised contexts. SHAP can be computationally intensive, and LIME relies on labeled data, limiting their effectiveness in unsupervised scenarios. DIFFI addresses these challenges by offering a more efficient and practical alternative, making it particularly suitable for anomaly detection in shared mobility systems.

The literature highlights the importance of both comprehensive feature engineering and model interpretability in effective anomaly detection. However, gaps remain in fully integrating multi-source data and applying interpretable unsupervised models within shared mobility contexts. Addressing these gaps could lead to more robust and actionable anomaly detection frameworks, ultimately supporting better operational decision-making in urban mobility systems.

2.4 Summary and Research Gaps

The literature reviewed in this chapter demonstrates that shared mobility systems, while offering significant benefits for sustainable urban transport, face numerous operational challenges. Research has shown that external factors such as weather conditions, public transit dynamics, and disruptive events profoundly affect bike-sharing demand at both the station and system levels. Various anomaly detection techniques, including functional outlier detection, clustering-based event detection, and spatial analysis methods, have been proposed to monitor these systems. However, several critical challenges remain:

- **Integration of Multi-Source Data:** While individual studies have examined the impacts of weather, transit availability, or local events, few have comprehensively integrated these heterogeneous data sources into a unified framework.
- **Interpretability of Anomaly Detection Models:** Many advanced detection methods, particularly those based on machine learning, operate as black boxes. The limited interpretability hinders the ability of operators to

understand and trust the underlying mechanisms driving anomaly identification.

- **Lack of Labeled Data and Scalability:** The absence of explicit ground truth labels in many bike-sharing datasets forces reliance on unsupervised approaches, which can struggle with scalability and accurately distinguishing true anomalies from short-term fluctuations.

Addressing these challenges is pivotal for the development of robust, real-time anomaly detection systems that can not only identify irregular patterns but also provide actionable insights for system optimization. This thesis contributes to the literature by proposing an integrated framework that combines multi-source feature engineering with interpretable, unsupervised anomaly detection techniques specifically leveraging methods such as DIFFI to enhance both detection speed and model transparency in shared mobility contexts.

3 Methodology

This chapter presents the methodology used to detect anomalies in bike-sharing systems by integrating multiple data sources and employing machine learning techniques. The approach consists of three key components: data collection and preprocessing, feature engineering, and anomaly detection modeling.

First, we describe the three main data sources used in this study: *bike-sharing trip data*, *weather data*, and *public transit data*. These datasets are combined to enrich anomaly detection with temporal, environmental, and spatial context.

Next, we outline the feature engineering process, where raw data is transformed into meaningful input variables. This step includes extracting time-based features, mapping weather conditions, computing transit accessibility, and encoding categorical attributes.

The core anomaly detection framework is based on the *Isolation Forest* algorithm, which is well-suited for identifying outliers in high-dimensional, unlabeled data. To enhance model interpretability, we apply *DIFFI* (Depth-based Isolation Forest Feature Importance), which provides insights into the most influential features driving anomaly detection.

3.1 Data Sources

The datasets used in this study come from five primary sources: bike-sharing trip records, weather data, public transit information, neighborhood boundaries, and holiday calendars. These datasets were combined to provide a richer context for anomaly detection in shared mobility, incorporating temporal, environmental, and spatial dimensions.

3.1.1 Bike-Sharing Data

The primary dataset is sourced from the BlueBikes system, specifically the file `202301-bluebikes-tripdata.zip`, which was downloaded from the BlueBikes system data portal [1]. This dataset contains detailed records of bike trips conducted in January 2023, including information on trip start and end times, trip duration, start and end station locations, and user type, indicating whether the rider was a registered member or a casual user.

During the initial exploration of the dataset, differences between the 2023 and 2024 BlueBikes data were observed. Notably, certain features, such as bike type distinguishing between electric and standard bicycles, were introduced in 2024 but were not present in the 2023 dataset. While initially considering a broader analysis using a full-year dataset, time constraints led to the focus on January 2023, ensuring consistency in feature availability while maintaining a manageable dataset size for anomaly detection. Additionally, data quality was assessed, revealing that the 2023 dataset contained no missing latitude or longitude values, whereas the 2024 data exhibited inconsistencies, potentially due to differences in data collection methods or increased system usage in the following year.

3.1.2 Weather Data

To incorporate environmental factors into the analysis, historical weather data was retrieved from the Meteostat service [11]. This dataset includes variables such as temperature, wind speed, and precipitation, which were selected due to their potential influence on ridership patterns. The weather data was processed and aligned with the bike-sharing dataset by matching weather conditions to trips based on the start hour of each ride. This integration ensures that each trip is associated with the corresponding weather conditions at the time of departure, providing additional contextual information for anomaly detection.

3.1.3 Public Transit Data

Public transit availability is another external factor that influences bike-sharing patterns, particularly in cases of transit disruptions or service gaps. To account for this, transit schedule data from the Massachusetts Bay Transportation Authority (MBTA) was incorporated into the dataset. The transit feed provides information on bus, subway, and commuter rail services, including stop locations, scheduled arrival and departure times, and service frequency.

The most recent available feed at the time of data extraction, published on January 3, 2024, covering the period from December 26, 2023, to April 6, 2024, was used [10]. A specific challenge encountered with the MBTA data was that the `stop_times` table lacked calendar dates in its `arrival_time` column. To resolve this, service IDs from the `calendar.txt` file, which defines service start and end dates along with operational days, were merged with `trips.txt` and `stop_times.txt` to reconstruct accurate date-time values for transit schedules.

To ensure accurate integration of transit data, a relational database approach was implemented to systematically link key tables, including `stops`, `stop_times`, `trips`, and `calendar`, using unique identifiers such as `stop_id`, `trip_id`, and `service_id`. By iterating through the calendar data and expanding service dates, transit schedules were properly aligned with bike-sharing data, enabling a structured integration for temporal analysis.

3.1.4 Neighborhood Data

Neighborhood boundaries were integrated into the dataset to provide spatial context for bike-sharing trips. Boston neighborhood boundaries were sourced from the city's open data portal [4], while Cambridge neighborhood boundaries were obtained from the Cambridge GIS repository [5]. These datasets, provided in GeoJSON and shapefile formats, respectively, were merged using a spatial join operation to assign a neighborhood to each bike station based on its geographic coordinates.

3.1.5 Holiday Data

To account for temporal variations in ridership, holiday data was incorporated into the dataset. The U.S. holiday calendar was retrieved using the `holidays` Python library, which provided official holidays within the study period. Since most bike trips occur within the same day, only the trips start date was used to determine whether it coincided with a holiday.

3.2 Feature Engineering

To enhance the anomaly detection process, feature engineering was applied to transform raw data into meaningful variables. This involved two stages: feature extraction at the trip level and aggregation at the station level.

3.2.1 Trip-Level Feature Engineering

Initially, features were extracted at the trip level to provide insights into individual bike trips. The dataset included fundamental attributes such as trip duration, distance, start and end timestamps, and user type. Additionally, a speed feature was calculated to better understand dependencies between trip duration and distance. Further features were derived to incorporate temporal, environmental, and spatial contexts.

Temporal features were extracted to capture usage patterns across different periods. These included the start and end hour, day, day of the week, and month of each trip. Since the dataset only covered January, the month feature was not used as a separate variable at either the trip or station level. Binary indicators for weekends and special days were also computed to analyze behavioral shifts on holidays and weekends. However, the special day feature was not included in model training but was retained for post-hoc analysis.

Environmental factors were incorporated by extracting weather-related features, including temperature, wind speed, precipitation, and the `coco` feature (an encoded categorical representation of general weather conditions such as clear, cloudy, rain, and snow). These attributes were mapped to each trip based on the starting hour, ensuring alignment with actual weather conditions at the time of departure.

Public transit accessibility was represented by computing the number of transit stops within a 500-meter radius of each trips starting and ending stations. Unlike the initially planned transit frequency feature, which aimed to capture transit arrivals within a specific time window, computational constraints led to its exclusion from the final dataset. To simplify interpretation, we categorized these features into three classes:

- **Badly Connected**: Fewer than 5 stops.
- **Moderately Connected**: Between 6 and 15 stops.
- **Well Connected**: More than 15 stops.

These thresholds were determined through a combination of statistical analysis and intuition, as illustrated in Figure 3.1.

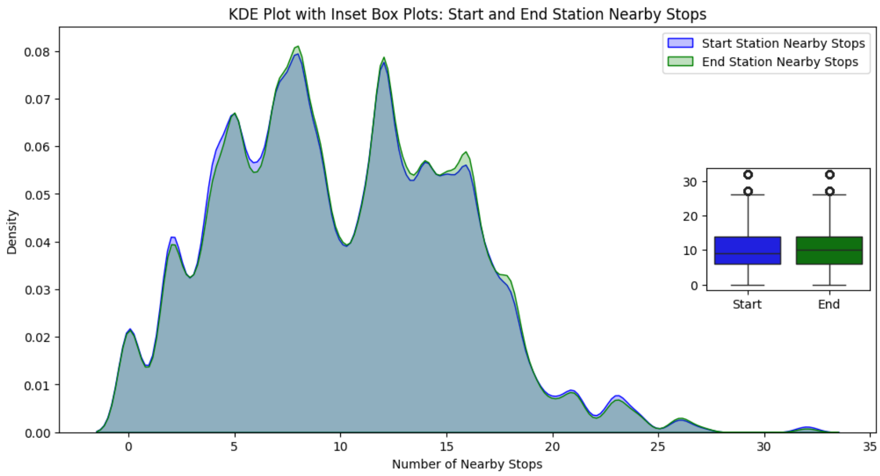


Figure 3.1: Distribution of transit stops within a 500-meter radius.

Spatial features were developed to analyze the geographical context of trips. Each station was assigned to a neighborhood using a spatial join with Boston and Cambridge neighborhood boundary datasets. While most stations were successfully mapped, approximately 10% remained unassigned due to their locations falling outside the defined boundaries. Rather than expanding the dataset with additional maps, these stations were labeled as *"Out of Area"*, as this designation itself provided meaningful insights into trip patterns.

To address missing neighborhood assignments, trips outside mapped regions were categorized into the following groups before neighborhood encoding for model training:

- **Out of Area Nearby:** Stations within 2.5 km of the mapped regions, a distance deemed appropriate based on station distribution.
- **Out of Area Far:** Stations beyond 2.5 km from the mapped regions.

Figure 3.2 illustrates the distribution of stations, showing their spatial relationships with neighborhood boundaries.

To further contextualize trips, a route-related feature was implemented, classifying trips based on their spatial coverage. This feature provided valuable

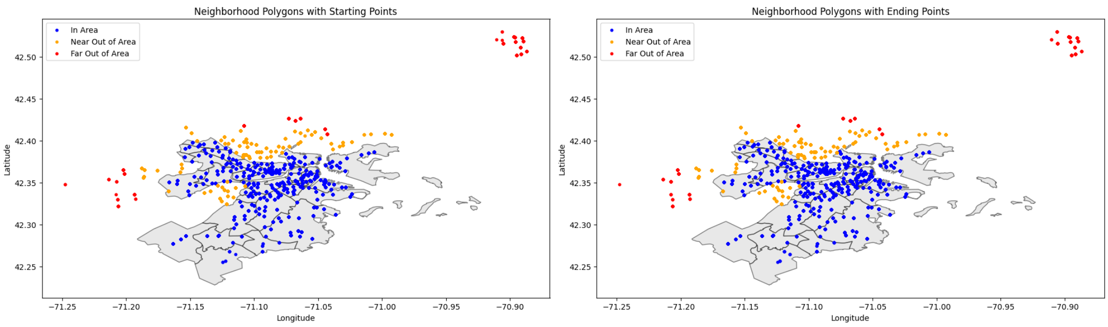


Figure 3.2: Spatial distribution of stations and their assigned neighborhoods.

insights into trip patterns and facilitated route visualization. Routes were categorized as follows:

- **In Area:** Both the starting and ending points are within the Boston/Cambridge area.
- **Near Out of Area:** At least one point (starting or ending) is classified as "Out of Area Nearby."
- **Far Out of Area:** At least one point is classified as "Out of Area Far."

This classification allows for a nuanced understanding of trip dynamics and provides additional information that can improve model performance and fairness. The spatial distribution of these route categories is illustrated in Figure 3.3.

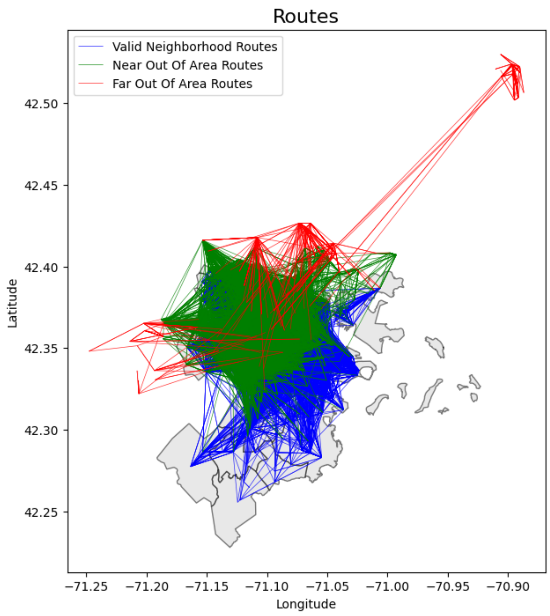


Figure 3.3: Visualization of categorized routes based on spatial coverage.

Following feature extraction, the dataset was saved in parquet format as `bike_trip_focused_data` for further processing.

3.2.2 Station-Level Feature Aggregation

To shift the focus from trip-level data to station-based analysis, trip features were aggregated at the station level, enabling the identification of station-specific anomalies. Aggregation was performed using the keys `station_id`, `hour`, `day`, `month`, and `is_start`. The resulting station-level dataset included aggregated statistics such as total trip count per station, as well as the mean and standard deviation of trip durations, distances, and speeds. For model training, mean values were primarily used, as some stations had an insufficient number of trips to compute meaningful standard deviations.

Temporal features, including station activity by hour, day, and day of the week, were retained to capture usage patterns. Similarly, the mean and standard deviation of weather attributes such as temperature, precipitation, and wind speed were computed for each station to assess environmental correlations at an aggregate level. However, only mean values were used for model training, as some stations had too few trips to yield meaningful standard deviation calculations. The presence of public transit was represented through the nearby transit stops feature, which remained the only transit-related variable included in model training.

Several features were maintained for contextual understanding, including station name, neighborhood, and whether the station served as a trip origin or destination. However, station name was excluded from model training. The final set of training features consisted of station trip count, mean distance, mean trip duration, mean speed, weather attributes, transit accessibility, temporal indicators such as hour, day and day of the week, mean user type, and an indicator specifying whether the station was a trip starting or ending location.

This two-stage feature engineering approach resulted in a structured dataset, allowing the model to detect anomalies at the station level while incorporating temporal, environmental, and transit-related factors.

3.3 Anomaly Detection Approach

Anomaly detection in shared mobility presents unique challenges due to the lack of labeled anomalies and the complex nature of urban transportation patterns. To address this, unsupervised machine learning methods were employed, allowing for the identification of atypical patterns in bike-sharing data without requiring predefined labels. The selected approach consists of two key components: *Isolation Forest* for anomaly detection and *DIFFI* for feature importance analysis.

3.3.1 Isolation Forest for Anomaly Detection

Isolation Forest (IF) is an unsupervised learning algorithm designed specifically for anomaly detection. Unlike traditional clustering or density-based methods, IF operates by constructing multiple isolation trees and recursively partitioning the dataset. Anomalies, which are rare instances with distinct characteristics, tend to be isolated more quickly than normal data points, resulting in shorter path lengths within the trees.

The primary advantages of IF include its scalability to large datasets, robustness to noise, and ability to handle high-dimensional data. Given the absence of labeled anomalies in bike-sharing data, IF was chosen for its effectiveness in identifying outliers in complex, dynamic systems. Across different analyses, the contamination parameter, which determines the proportion of data points classified as anomalies, was typically set to the default value or a similar range. This decision was based on domain knowledge and exploratory analysis.

While extensive hyperparameter tuning was not conducted due to the unsupervised nature of the task, the key parameters include:

- **`n_estimators`**: The number of trees in the forest. A higher number generally improves performance by capturing more diverse patterns.
- **`contamination`**: Defines the proportion of anomalies in the dataset. Since labeled data was unavailable, this parameter was adjusted iteratively based on exploratory analysis. Changes in this parameter significantly affect the results.
- **`max_samples`**: Specifies the fraction of the dataset used to build each tree. For large datasets, using only a portion can improve computational efficiency.

Tuning and Parameter Selection Initial tuning focused on trip duration and distance, as these were expected to be strong indicators of anomalies. While this provided a good starting point, incorporating additional features, such as weather attributes, improved model performance by enabling the detection of more nuanced anomalies beyond extreme trip lengths. Increasing the number of trees (`n_estimators`) also led to better results, particularly when more features were included, suggesting that the model benefited from capturing a wider range of patterns.

For computational efficiency, a reasonable set of parameters was determined. Generally, using `n_estimators` in the range of 100 to 500 and setting `max_samples` to 60% of the dataset yielded stable results. The contamination parameter was initially tuned experimentally; however, it was later decided to use the default value of 0.01, as it provided reasonable results aligned with domain knowledge and standard anomaly detection practices. In some exploratory analyses, a higher value (e.g., 0.3) was tested to observe its effect.

3.3.2 Feature Importance with DIFFI

While Isolation Forest effectively detects anomalies, it does not inherently provide explanations for why a particular instance is classified as an anomaly. To address this limitation, the Depth-based Isolation Forest Feature Importance (DIFFI) method was applied. DIFFI quantifies the contribution of each feature to the anomaly detection process by analyzing how feature values influence isolation depth [2].

Integrating DIFFI into the analysis not only enabled anomaly detection but also provided insights into the underlying factors contributing to each detected anomaly. The method supports both global and local interpretability, allowing for an assessment of whether anomalies were primarily influenced by extreme trip durations, unusual weather conditions, or transit-related factors. This interpretability component was crucial in evaluating whether the detected anomalies aligned with domain knowledge and expected mobility patterns.

4 Preliminary Work: Trip-Focused Analysis

4.1 Motivation for Trip-Focused Analysis

The initial direction of this study was to analyze anomalies at the individual trip level, as it was expected to provide valuable insights into unusual bike-sharing behaviors. The hypothesis was that anomalous trips – those with extreme durations, unrealistic distances, or abnormal speeds – could indicate operational inefficiencies, external disruptions, or unexpected user behavior. By identifying these irregularities, the goal was to determine whether anomalies were primarily driven by external conditions, such as weather and transit availability, or if they resulted from underlying system issues or natural user variations.

Trip-level analysis was chosen due to its fine-grained nature, allowing for a detailed examination of individual mobility patterns rather than aggregated trends. This approach was expected to reveal anomalies emerging from external disruptions, route choices, or behavioral inconsistencies. Additionally, trip-level anomaly detection could serve as a foundation for further analyses, including the development of new features for downstream models. Since trip-level data represents the most granular aspect of bike-sharing behavior, it provided a rich source for feature engineering, capturing variations that might be lost in aggregated data.

While this level of granularity provided valuable information, it also presented challenges in terms of interpretability and direct applicability, which were further explored in later stages of the study.

4.2 Exploratory Data Analysis (EDA)

Before applying Isolation Forest for anomaly detection, an initial exploratory data analysis (EDA) was conducted to understand the distribution of key variables and detect inherent patterns in the dataset. This analysis focused on trip characteristics, environmental influences, and spatial distributions to ensure a comprehensive understanding of potential anomalies.

Trip duration, distance, and speed were analyzed to identify natural outliers and unusual distributions. The majority of trips have relatively short durations and distances, while a considerable number of longer trips suggest potential anomalies. To gain further insight, Figure 4.1 presents boxplots of these key variables, highlighting extreme values.

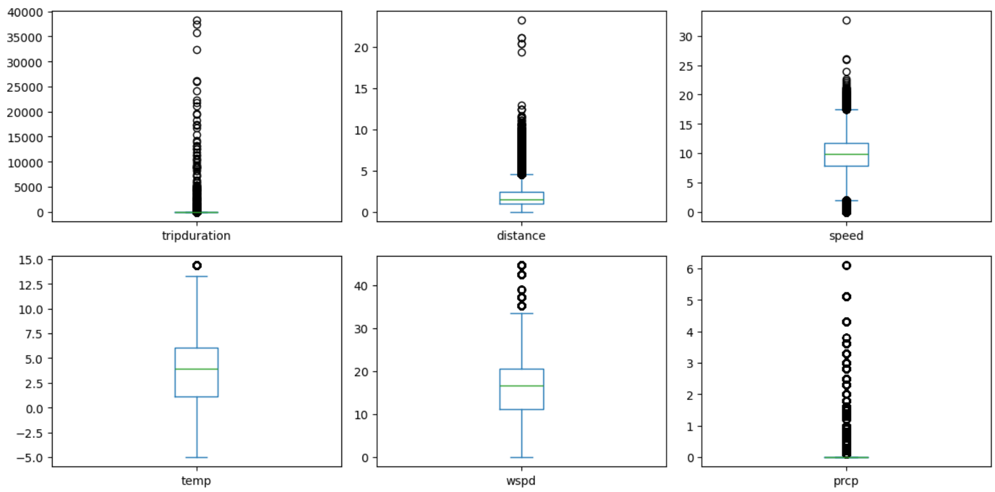


Figure 4.1: Boxplots of trip characteristics: trip duration (minutes), distance (km), speed (km/h), and weather-related variables.

To categorize trip durations more systematically, trips were grouped into the following four categories based on their length:

- **Very Short Trips:** Less than 15 minutes.
- **Short Trips:** Between 15 minutes and 1 hour.
- **Long Trips:** Between 1 hour and 6 hours.
- **Very Long Trips:** More than 6 hours.

The frequency distribution of trips within these categories is presented in Figure 4.2. As expected, most trips fall into the very short and short categories, with significantly fewer long or very long trips. Understanding this distribution helps in assessing which trip lengths may be more prone to anomalies.

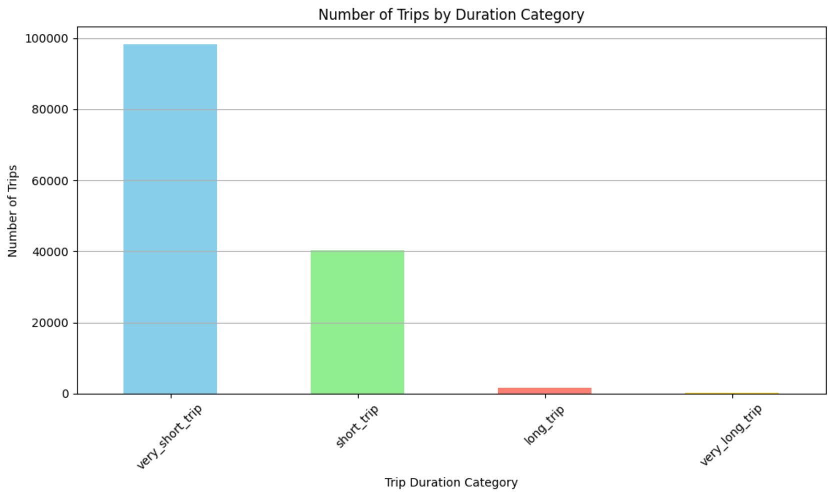


Figure 4.2: Distribution of trips by duration category.

Weather conditions are known to influence bike-sharing behavior, particularly during extreme conditions such as heavy rain or snow. To assess this relationship, Figure 4.3 illustrates how trip duration, distance, and speed vary across different weather types, represented by the categorical weather feature `coco`. The majority of trips occur under clear or partly cloudy conditions, but some cases involve adverse weather, which may correlate with anomalies.

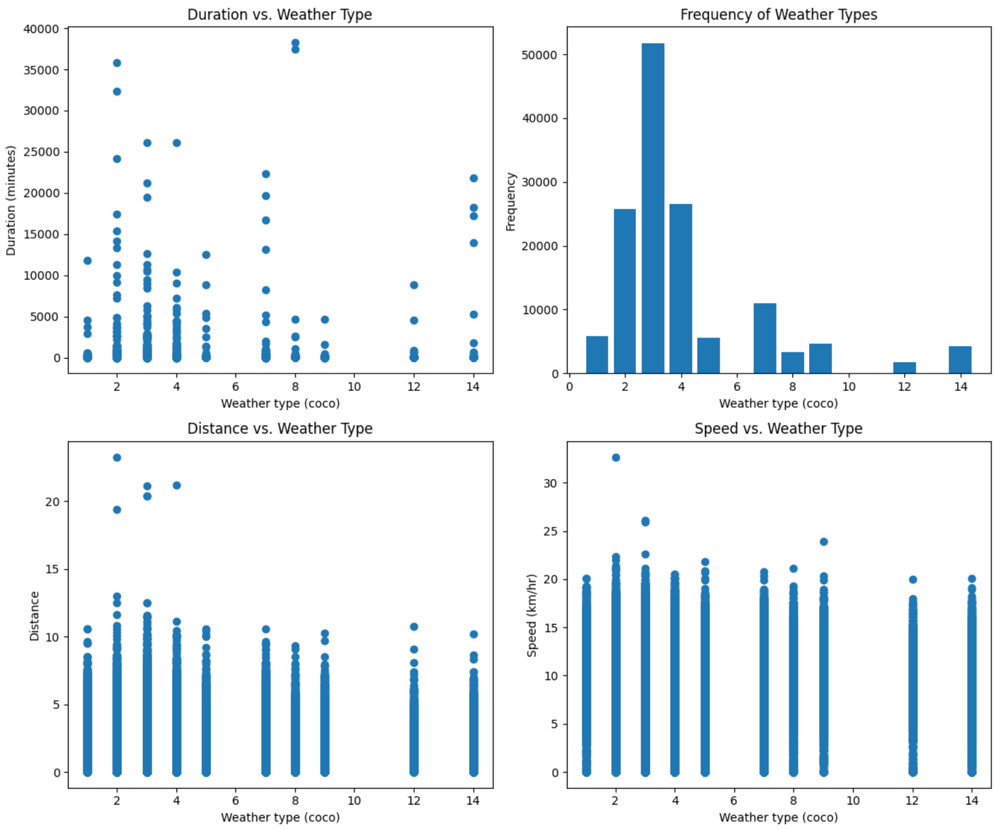


Figure 4.3: Analysis of trip characteristics across different weather conditions.

Beyond trip characteristics and weather effects, the spatial distribution of routes was also examined to identify potential anomalies. Figure 4.4 presents a segmented view of trip routes, categorized into three groups: trips occurring entirely within mapped areas, those slightly extending beyond the defined regions, and those significantly exceeding mapped boundaries.

The majority of trips (117,688) remained within the designated Boston/-Cambridge area, while 21,966 extended slightly beyond, categorized as Near Out-of-Area Routes. A smaller fraction (686 trips) traveled well beyond the expected service regions, classified as Far Out-of-Area Routes.

Among the Far Out-of-Area trips, an interesting pattern emerged involving the Salem service area. Although Salem is part of the broader BlueBikes network, it functions as a separate service zone, geographically distinct from the Boston/Cambridge area. This separation increases the irregularity of trips that span over 100 kilometers, as they cross between two independent service areas. Such long-distance trips are not typically expected within the operational scope of bike-sharing systems and raise questions about their validity or the specific user behaviors behind them.

These geographically distant trips were of particular interest, as they could align with other anomaly indicators such as extreme trip durations or unusually high speeds, further emphasizing their outlier status in the dataset.

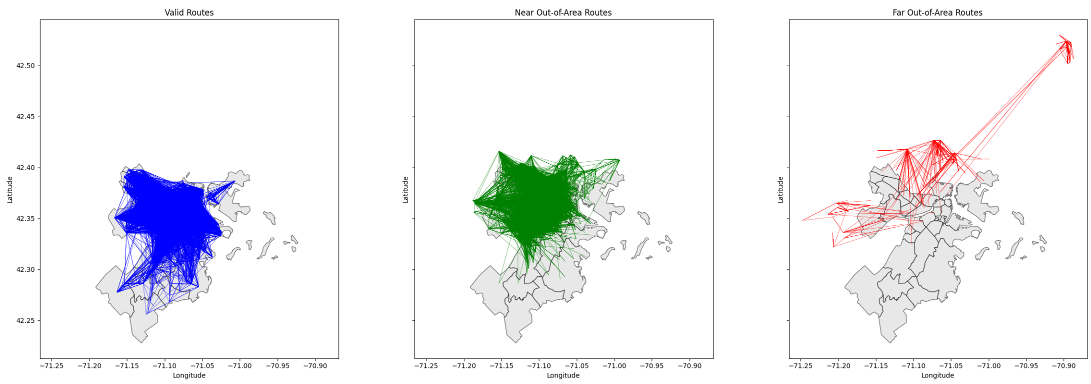


Figure 4.4: Visualization of routes categorized as In-Area (blue), Near Out-of-Area (green), and Far Out-of-Area (red).

This visualization provides a complementary perspective to the aggregated route map in Figure 3.3, where all trips are shown collectively. By examining routes separately, clearer distinctions emerge regarding potential spatial anomalies, helping to assess whether distant trips correlate with other unusual patterns in the dataset.

4.3 Initial Anomaly Detection Results

To validate the effectiveness of Isolation Forest (IF) for anomaly detection, the model was first applied using only basic features: trip related and weather related. Trip duration and distance were selected as dimensions to visualize them as they are strong indicators of anomalous trips. Figure 4.5 illustrates the identified anomalies in this two-dimensional space. As expected, extremely long durations and distances were flagged as anomalies. The contamination parameter was initially set to 0.005, and when increased further, the vertical cluster of normals (long-duration trips) became shorter, highlighting the sensitivity of the model to contamination adjustments.

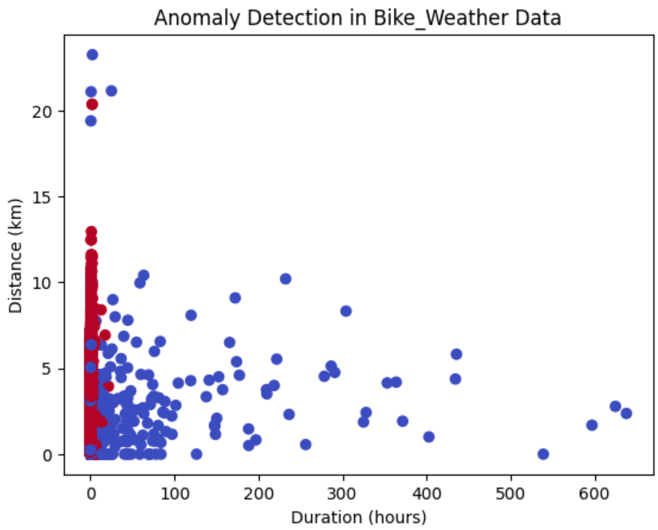


Figure 4.5: Anomaly detection results using only trip duration and distance. Anomalous points are highlighted in blue.

This initial analysis also revealed that the feature range affects the models behavior. Features with a wider range require more trees to be effectively isolated, reinforcing the importance of feature scaling and normalization. To further examine the relationship between anomalies and external conditions, we analyzed the `coco` weather feature, which categorizes weather conditions numerically. When using only bike-related and weather features, the mean `coco` values for anomalies and normal trips were 5.7 (freezing fog) and 4.0 (overcast), respectively. This indicates that trips classified as anomalies tended to occur under more severe weather conditions.

After incorporating all trip-level features and adjusting the contamination parameter to 0.03, the difference in weather conditions between normal and anomalous trips became more pronounced. The mean `coco` values for anomalies increased to 7.6 (indicating light to heavy rain), while normal trips remained at 3.9 (cloudy to overcast). This further supports the hypothesis that weather plays a role in defining anomalies in shared mobility.

The spatial distribution of anomalous trips was also analyzed to assess whether certain geographic regions were more prone to anomalies. Figure 4.6 presents the routes of detected anomalies. The similarity between this visualization and general trip routes suggests that many out-of-area trips were flagged as anomalies. Given that such trips represent a small proportion of total rides, their classification as anomalies is statistically reasonable.

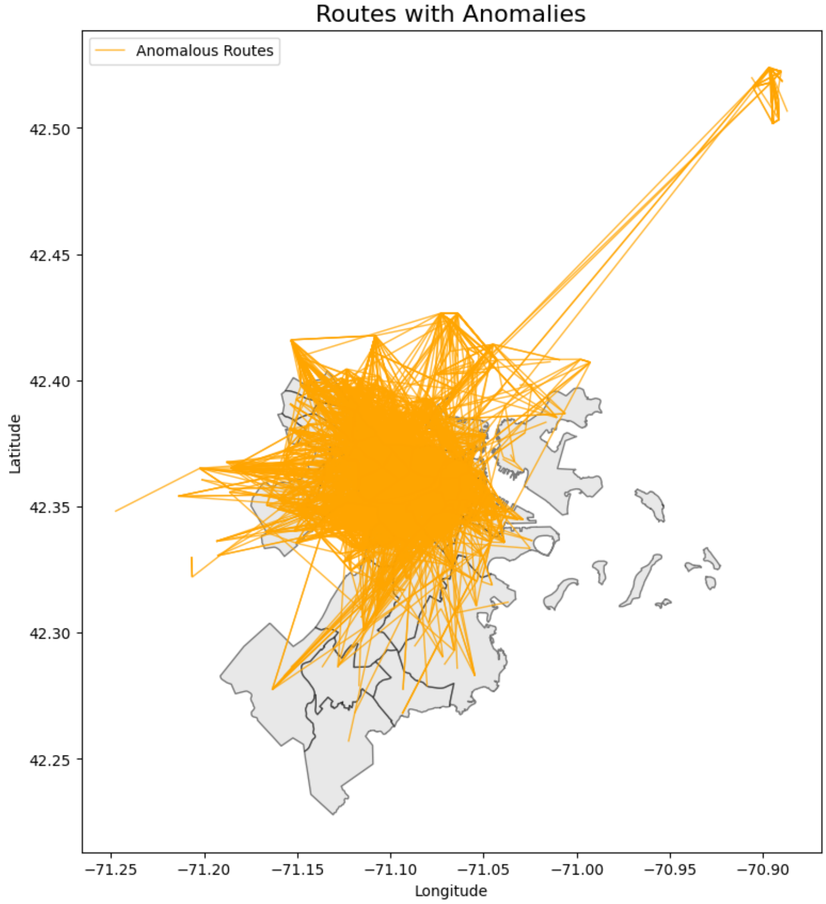


Figure 4.6: Visualization of detected anomalous routes.

To further explore station-level impacts, stations were classified based on the presence of anomalies. Figure 4.7 differentiates stations where only anomalous trips occurred, only normal trips occurred, or both were present. A particularly interesting observation emerged in the northeastern part of the map: stations that are geographically far from the central area were not necessarily associated with anomalies. By examining both the anomaly routes and station classifications, it became evident that some far-out stations consistently hosted normal trips, indicating that distance alone was not the sole determinant of anomalies.

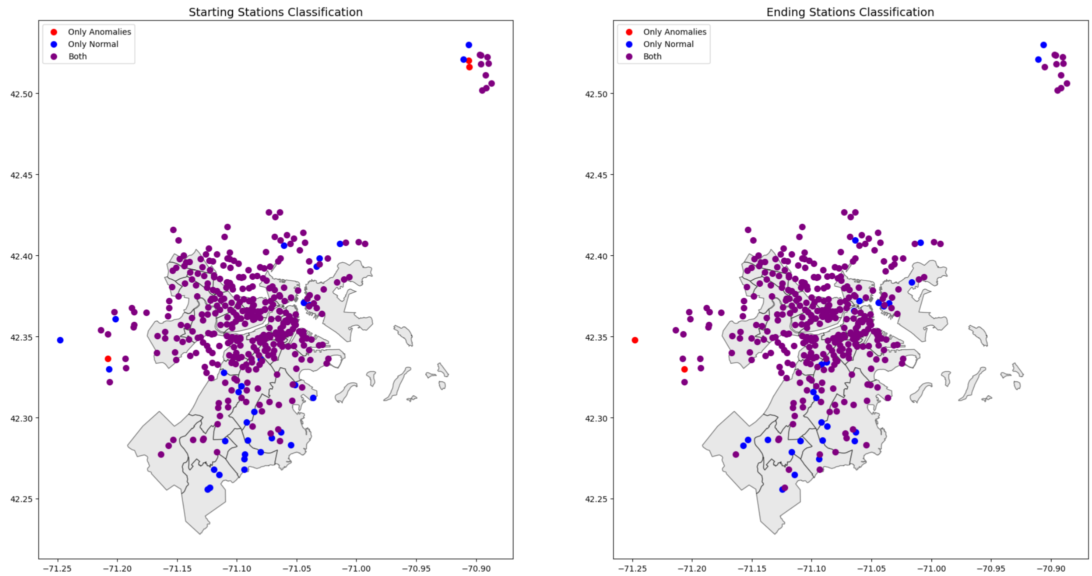


Figure 4.7: Station classification based on anomalies. Red points indicate stations with only anomalous trips, blue points represent stations with only normal trips, and purple points correspond to stations where both types occurred.

These findings provided essential insights into the nature of detected anomalies and their relationships with external conditions, trip characteristics, and geographic distribution. The results also underscored the need for further refinement in the anomaly detection approach, particularly in balancing sensitivity to false positives and false negatives.

4.4 Feature Selection and Dimensionality Reduction

To refine the anomaly detection process, feature selection and dimensionality reduction techniques were explored. The motivation behind feature selection was to reduce the number of input variables while retaining the most relevant information. Initially, 12 features were considered, covering various trip-related and weather-related attributes. To systematically determine an optimal subset, the Depth-based Isolation Forest Feature Importance (DIFFI) method was applied. However, since the dataset lacked labeled anomalies, standard validation metrics such as the F1 score could not be used.

Instead, alternative evaluation metrics were necessary to approximate an optimal feature set. The feature selection process was first tested on the lympho dataset, a benchmark dataset with ground-truth labels, to compare the F1 score with alternative evaluation methods inspired by the feature selection experiment from the original DIFFI paper [2]. The evaluation focused on three key aspects: outlier scores, stability of predictions, and K-Means clustering inertia, aiming to determine a natural stopping point where additional features no longer provided significant benefits.

Before applying these evaluation techniques, we first outline how each method operates:

- **Outlier Scores**: Measures how anomalous a data point is based on a models decision function, such as that of an Isolation Forest. The median of these scores assesses how well a subset of features enhances the models ability to detect anomalies. As the number of features increases, the outlier score is expected to decrease, making anomaly detection more challenging.

- **Stability of Predictions**: Evaluates the variance of anomaly predictions across multiple runs of a model with different random seeds. Low variance suggests that a feature subset leads to consistent anomaly detection, while high stability indicates that the selected features generalize well. Increasing the number of features generally improves stability.

- **K-Means Clustering**: Assesses feature grouping quality using inertia, which represents the sum of squared distances between data points and their nearest cluster centroids. Lower inertia suggests better clustering. The elbow method is used to identify an optimal number of features where the inertia curve begins to flatten.

Figure 4.8 presents the evaluation results. In the original DIFFI experiment on the lympho dataset, the F1 score remained below 0.980 for feature sets with fewer than 12 features, reaching 0.986 at 12 features. While the difference was small, it provided a reasonable hypothesis for selecting an optimal number of features.

Applying our evaluation metrics (outlier scores, stability of predictions, and K-Means clustering) to the same dataset, we observed that the K-Means elbow point emerged around 11 to 12 features, indicating a natural grouping. Similarly, the stability metric exhibited reduced fluctuations at this point, supporting the idea that 12 features captured the data structure well. Additionally, while outlier scores generally declined as features increased, they remained relatively stable beyond 11-12 features, reinforcing this selection.

These findings provide further justification for the feature selection approach proposed in DIFFI. Even without labeled data, our alternative methods offer valuable insights into feature importance and selection.

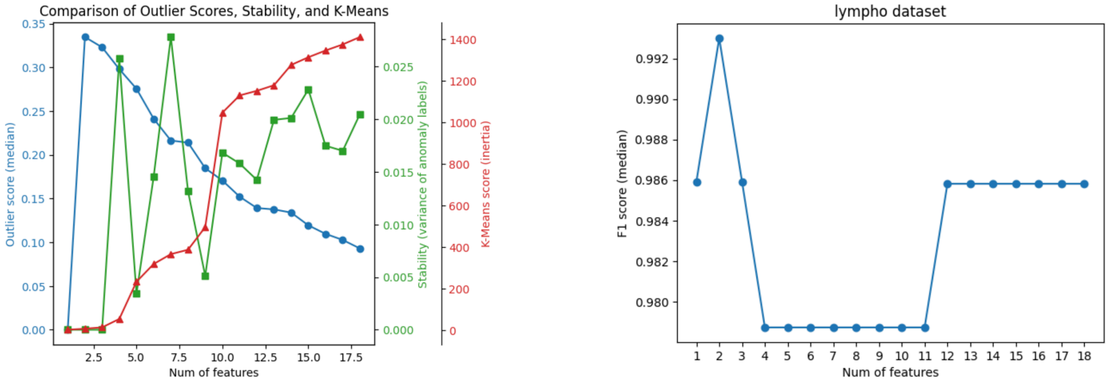


Figure 4.8: Comparison of evaluation methods and F1 score on the Lympho dataset.

For the bike_weather dataset, the same evaluation metrics were applied to determine a reasonable number of features for anomaly detection. Figure 4.9 illustrates that 9 features appeared to be the most suitable choice. The elbow point in the K-Means inertia curve around 9 features suggests that this number effectively captures the datasets structure. Additionally, the change in outlier scores from 8 to 9 features was smaller than the jump from 9 to 10, indicating that adding more features may not significantly improve anomaly detection. The stability metric also showed only minor fluctuations between 8 and 9 features, further reinforcing the selection of 9 features.

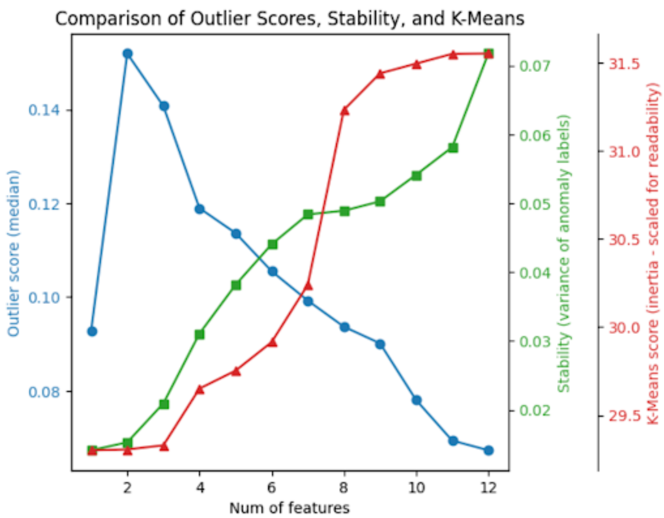


Figure 4.9: Isolation Forest results after DIFFI feature selection on the bike_weather dataset.

Once a subset of the most relevant features was determined, PCA was explored as an alternative dimensionality reduction technique. The goal was to assess whether transforming the dataset into a lower-dimensional space could yield similar or improved anomaly detection results. Three key approaches were compared: Isolation Forest on the given dataset, Isolation Forest on the DIFFI-selected features, and Isolation Forest on PCA-transformed data with the same number of components as DIFFI. Additionally, PCA with fewer components was tested to examine its impact.

Figure 4.10 compares the anomaly detection results. PCA transformation preserved the general anomaly structure but led to a more centralized clustering of normal points when nine components were used. With fewer components (five), anomalies appeared more dispersed, and the model classified points along the 45-degree line as normal rather than clearly separating anomalies. This suggests that while PCA effectively reduces dimensionality, it alters how anomalies are distributed.

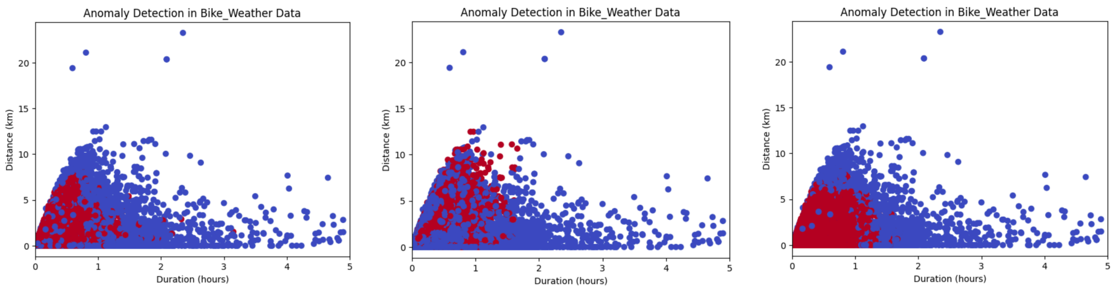


Figure 4.10: Anomaly detection results: Isolation Forest without PCA, with five PCA components, and with nine PCA components.

A direct comparison between DIFFI-based feature selection and PCA with nine components is shown in Figure 4.11. The results indicate that DIFFI-based selection produced results similar to the full-feature model, while PCA led to normal points clustering more tightly together. This suggests that DIFFI preserves feature relationships more effectively, while PCA may distort the anomaly distribution.

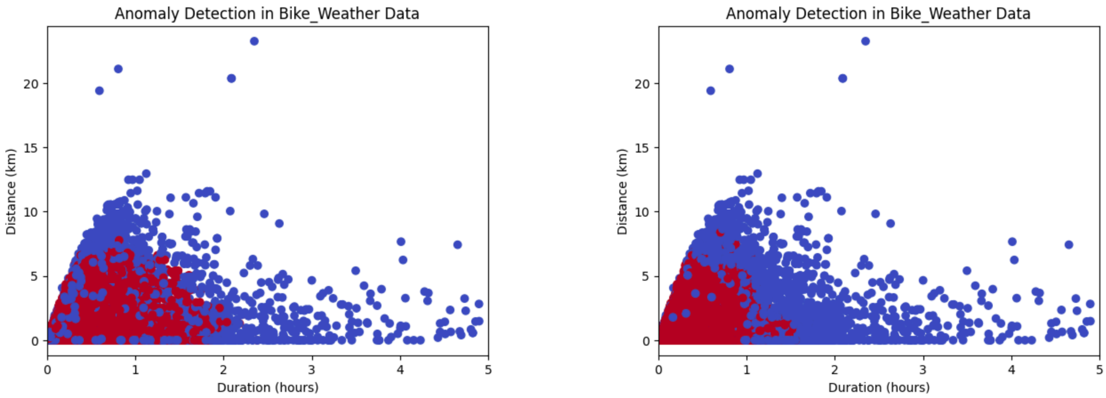


Figure 4.11: Comparison of anomaly detection: Isolation Forest trained with nine DIFFI-selected features (left) and PCA with nine components (right).

These findings suggest that while PCA effectively reduces dimensionality, it does not necessarily enhance anomaly detection. DIFFI-based feature selection preserved a meaningful subset of features while maintaining anomaly detection performance, making it a more interpretable and reliable approach for dimensionality reduction in this context.

4.5 SHAP vs. DIFFI: Feature Interpretability Comparison

To better understand the factors influencing anomaly detection, we compared two interpretability methods: SHAP (SHapley Additive Explanations) and DIFFI (Depth-based Isolation Forest Feature Importance). While both approaches explain the contributions of features to anomaly detection, they emphasize different aspects of the dataset, leading to slightly distinct interpretations.

Running Local DIFFI on the entire dataset required approximately 20 minutes, while SHAP is known from the literature to be even more computationally demanding. Given the high computational cost, a subset of 100 detected anomalies was chosen for interpretation to ensure a fair comparison while maintaining efficiency. This subset allowed for an interpretable yet computationally feasible comparison, with SHAP taking approximately 2 minutes to compute.

For the selected subset, Local DIFFI identified *distance* as the most important feature in anomaly detection, with *end_hour* ranking consistently high. The *coco* feature (weather condition encoding) frequently appeared as the second most significant feature, reflecting the role of environmental factors in anomalies.

Temperature (*temp*) followed a similar trend, confirming the correlation between extreme weather conditions and trip anomalies. Precipitation (*prcp*) was also consistently relevant, reinforcing the impact of weather on bike-sharing anomalies. Other features, such as *speed* and *tripduration*, while less dominant, still contributed meaningfully to the detection process. However, sudden changes in the rankings of dominant features between the first and second ranks in Figure 4.12 may indicate instability in interpretation, suggesting that increasing the number of trees could improve consistency.

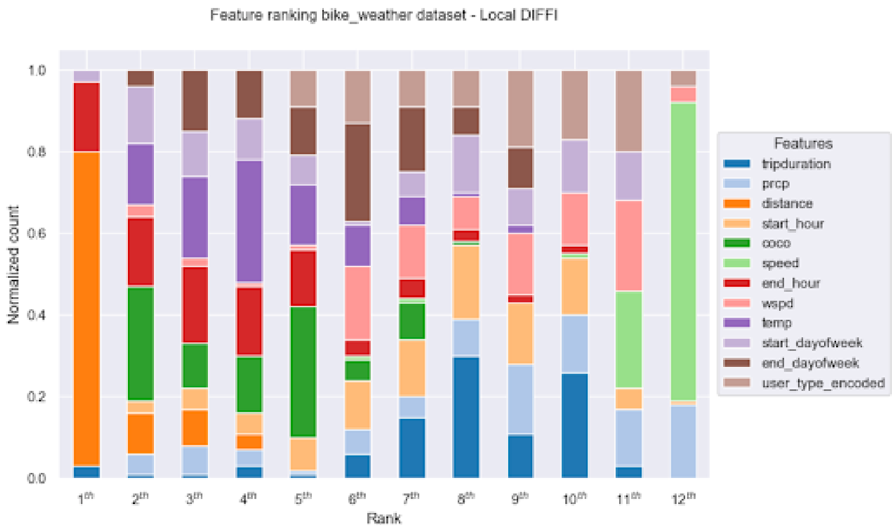


Figure 4.12: Feature ranking for a subset of 100 anomalies using Local DIFFI with 100 trees.

SHAP, when applied to the same subset, also highlighted *distance* as a critical factor, as shown in Figure 4.13. However, it placed significantly more emphasis on weather-related features, such as *temp* (temperature) and *wspd* (wind speed), while assigning less importance to temporal features like *end_hour* and *prcp* (precipitation). This contrast suggests that SHAP prioritizes weather-related anomalies, whereas Local DIFFI distributes importance more evenly across different types of anomalies.

To understand broader trends, Local DIFFI was also applied to the full dataset. The results showed that in the subset, some features played a more dominant role than others, particularly *distance* and *end_hour*, suggesting that anomalies in the subset were often linked to specific reasons. However, when applied to the full dataset, the importance of environmental factors such as *wspd* (wind speed) and *temp* (temperature) increased. This indicates that while

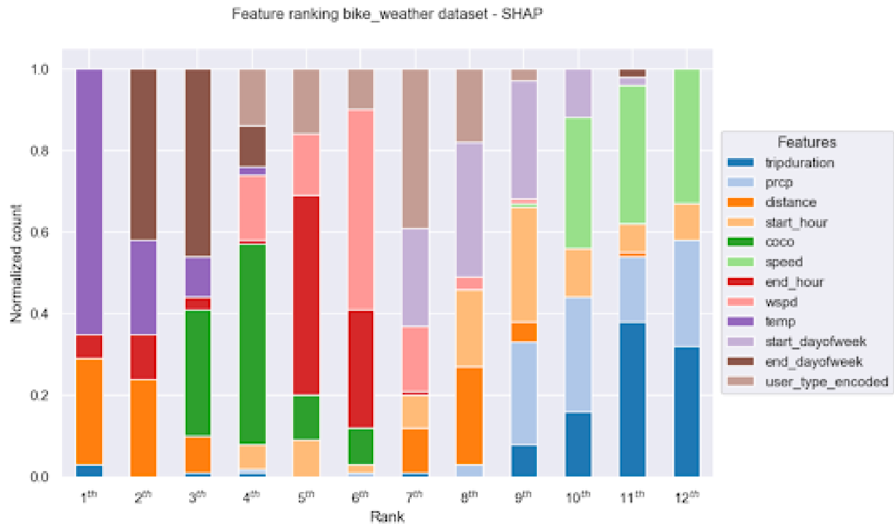


Figure 4.13: Feature ranking for a subset of 100 anomalies using SHAP.

anomalies in smaller subsets might be localized to specific conditions, large-scale anomalies in the full dataset are more influenced by overall weather patterns. Despite these differences, *tripduration* and *distance* consistently ranked among the most significant features across different scales, as seen in Figure 4.14.

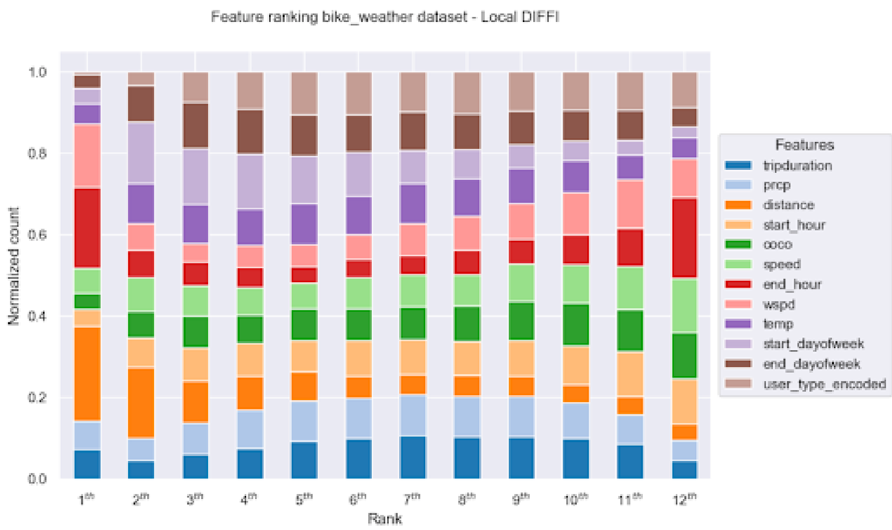


Figure 4.14: Feature ranking using the full dataset with Local DIFFI (100 trees).

To further investigate the impact of model complexity on feature rankings, the number of trees in Local DIFFI was increased from 100 to 500. This resulted in more stable rankings, where weather-related features (*wspd* and *temp*) gained prominence while *distance* remained one of the dominant factors, as shown in Figure 4.15. Interestingly, despite an increase in tree count, SHAP continued to highlight similar feature rankings as before, reinforcing its emphasis on environmental factors. However, SHAP required over 20 times longer to compute, making it impractical for real-time or large-scale applications. This computational efficiency gives Local DIFFI a distinct advantage.

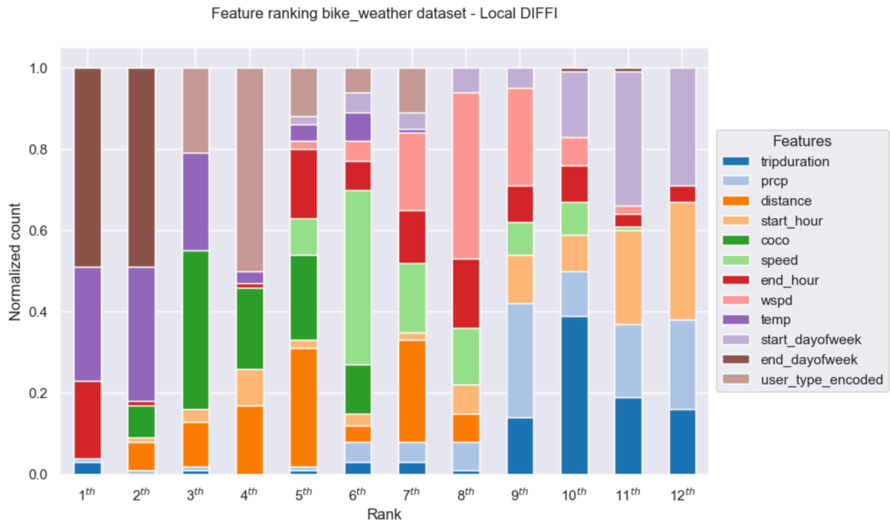


Figure 4.15: Feature ranking for a subset of 100 anomalies using Local DIFFI with 500 trees.

A specific example identified by Local DIFFI involved an anomalous trip taken in strong breeze conditions (*wspd* = 13 m/s). The user traveled a short distance at a relatively high speed during light rain, at midnight on a weekend, in moderately cold weather. Given the unusual combination of factors, this trip was correctly flagged as an anomaly.

This comparison highlights several key observations. When used with a higher number of trees, Local DIFFI produced more stable interpretations, indicating that hyperparameter tuning plays a role in enhancing its reliability. While SHAP showed promising results, its extreme computational inefficiency makes it unsuitable for practical use in Isolation Forest interpretability. Given its speed and well-balanced feature distribution, Local DIFFI emerges as the more practical choice for large-scale anomaly detection.

4.6 Hypothesis on Special Days

In the absence of labeled data, evaluating the results of anomaly detection becomes challenging. Given this limitation, the reasonable approach should prioritize minimizing false negatives over false positives to ensure that unusual patterns in city activities are captured, even at the cost of some false alerts. This is particularly useful for monitoring purposes, where missing anomalies could be more problematic than dealing with occasional false alarms.

To fine-tune the contamination parameter, a post-hoc evaluation method was introduced. The metric, called Special Day Precision, measures the proportion of anomalies detected on special days, such as holidays in our data. The hypothesis was that these days exhibit distinct bike usage patterns, making them more prone to anomalies. While this metric does not directly measure false negatives, it was assumed to provide insights into reducing them by examining deviations in bike usage.

Initially, it was expected that the Special Day Precision curve would show an elbow-like pattern, similar to the elbow method in K-Means clustering, which would help in identifying an optimal contamination rate through an inflection point. However, as shown in Figure 4.16, no clear elbow or plateau emerged. This suggests that while special days influence bike usage, many trips during these days still follow normal behavioral patterns, making it difficult to use this metric as a direct tuning reference.

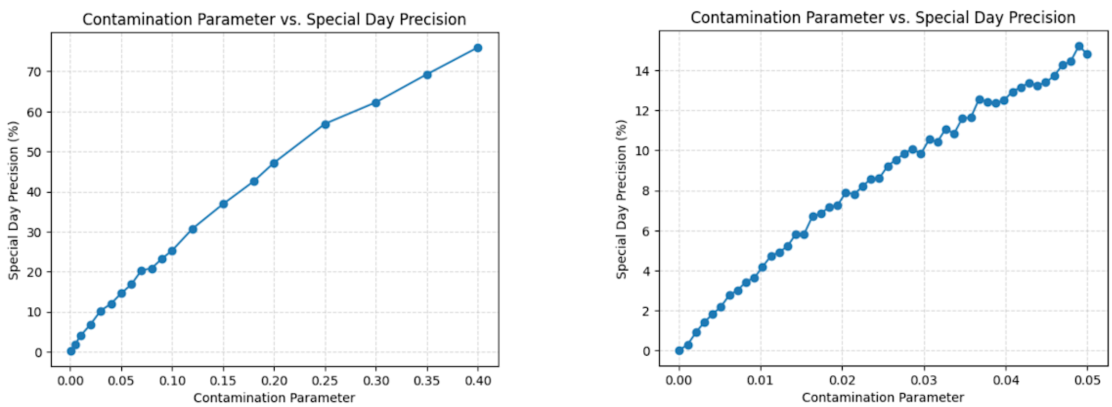


Figure 4.16: Relationship between the contamination parameter and Special Day Precision. The left plot shows the full range up to 40%, while the right plot zooms into lower contamination values (0-5%).

Since the results did not confirm the initial hypothesis, an alternative refinement would be to expand the definition of special days by incorporating additional contextual factors. Extreme weather conditions, such as storms and heavy rain, could influence bike usage in ways similar to holidays, making them relevant for anomaly detection. Future work could explore this extended definition to better capture anomalous mobility patterns in shared transportation systems.

4.7 Transition to Station-Level Analysis

While trip-level analysis provided valuable insights into individual bike trips, it also presented significant challenges in interpretability and reliability. A key issue was that even trips with similar anomaly scores or feature values could have entirely different underlying causes. Human mobility patterns are inherently unpredictable, making it difficult to conclusively determine whether a specific trip was an anomaly without deeper contextual understanding. Unlike aggregated analyses, where patterns emerge across multiple trips, trip-level anomalies were more isolated, making their interpretation less straightforward.

These limitations highlighted the difficulty of distinguishing genuine anomalies from normal behavioral variations. While trip-level anomaly detection successfully identified cases with unusual durations, distances, and external conditions, its sensitivity to certain parameters and its lack of spatial aggregation made it less effective for high-level system insights. Some of the key challenges included:

- High sensitivity to the contamination parameter, significantly impacting results.
- Individual trip variations sometimes led to misleading anomalies.
- The approach did not account for station-level usage patterns or network-wide trends.

To address these challenges, the focus shifted toward a station-level analysis, where anomalies could be detected based on aggregated station activity rather than isolated trips. Station-level aggregation provided more actionable insights by considering broader spatial and contextual factors influencing bike usage, such as station location, proximity to transit hubs, and surrounding neighborhood characteristics. This transition allowed for a more robust understanding of shared mobility patterns, reducing noise from individual trip variations while retaining key insights from trip-level analysis.

5 Final Results: Station-Focused Analysis

The focus of this chapter is to present the final anomaly detection results at the station level. After shifting from trip-level analysis, this approach allowed for a more structured understanding of anomalies by aggregating trip characteristics per station. The results provide insights into general anomaly patterns, temporal spikes, spatial clustering, and station-level interpretations.

5.1 Exploratory Data Analysis (EDA)

Before applying anomaly detection models, we conducted an exploratory data analysis (EDA) to gain a better understanding of the dataset at the station level. The goal was to observe differences between start and end stations, uncover patterns that may not be apparent at the trip level, and familiarize ourselves with the data before proceeding with anomaly detection.

One of the first aspects analyzed was the distribution of trip counts across different times of the day. Figure 5.1 illustrates the average trip counts at different hours for both start and end stations. While a similar analysis could be performed at the trip level, visualizing start and end stations separately at the station level provides additional insights. The graph reveals two distinct peaks: one at around 8 AM, corresponding to morning commuting hours, and another at approximately 6 PM, which aligns with typical return-home times.

This pattern is consistent with expectations based on urban mobility trends.

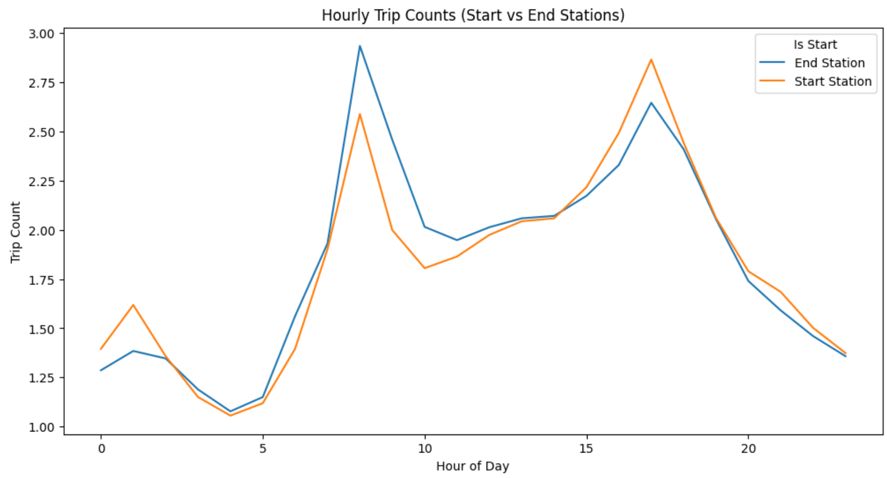


Figure 5.1: Hourly trip counts for start and end stations.

Next, we identified the top stations by total trip count. Figure 5.2 presents the 10 stations with the highest number of trips, which are predominantly located in central or high-traffic areas. This distribution aligns with expectations, as central locations typically have higher bike-sharing demand.

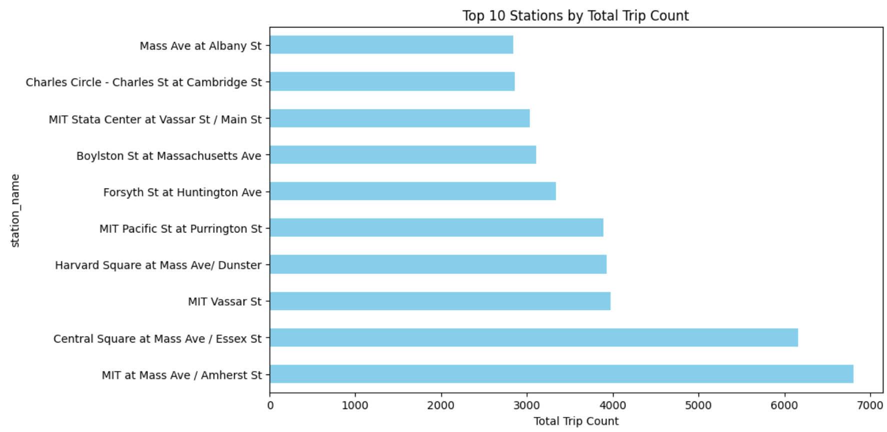


Figure 5.2: Top 10 stations by total trip count.

Weather conditions are known to impact bike usage significantly, so we examined the relationship between trip counts and weather variables such as precipitation and temperature. Figure 5.3 shows how trip counts vary with precipitation levels. As expected, there is a clear negative correlation: higher precipitation levels are associated with fewer trips, indicating that adverse weather conditions discourage bike usage.

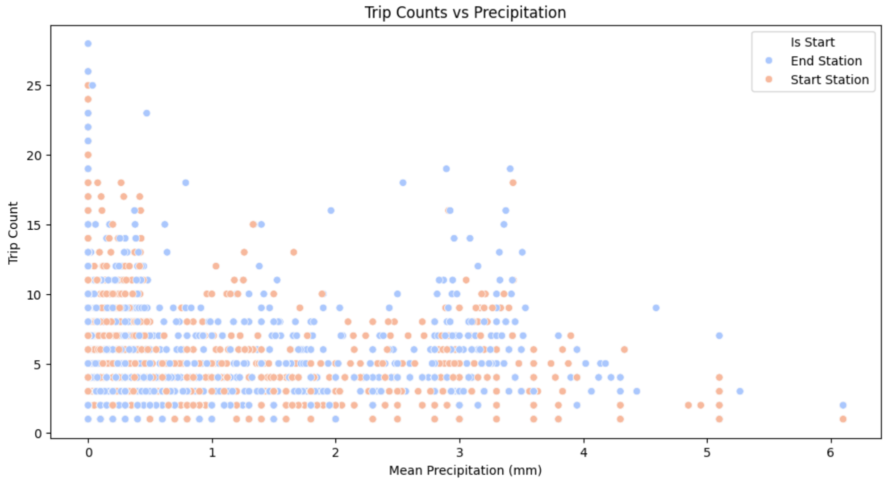


Figure 5.3: Trip counts vs. precipitation levels.

A similar pattern emerges when analyzing the impact of temperature on trip counts. Figure 5.4 illustrates the relationship between mean temperature and bike usage. According to [16], the average temperature in Boston during January 2023 was approximately 5řC. The data suggests that trip counts increase as temperatures rise to a level that is generally considered comfortable for cycling. However, the highest trip counts do not necessarily coincide with the warmest hours of the day, as peak usage still follows commuting patterns, occurring at 8 AM and 6 PM. This confirms the expected relationship between weather and bike-sharing behavior.

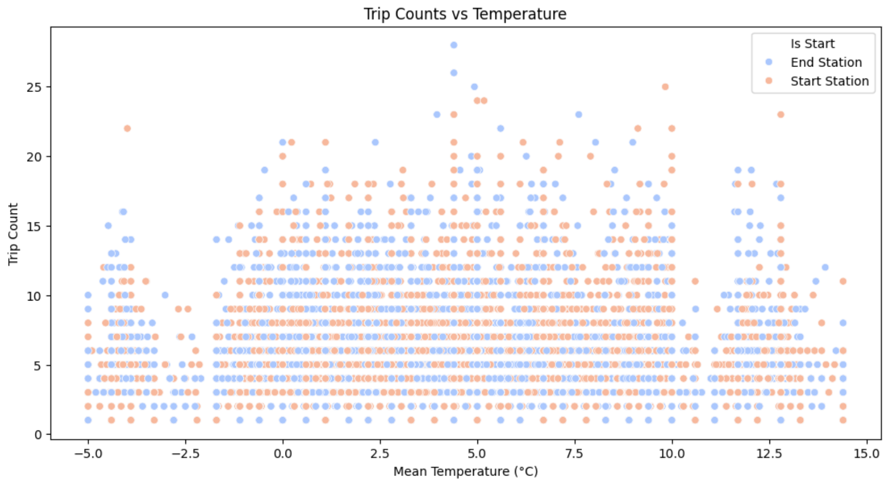


Figure 5.4: Trip counts vs. temperature.

These exploratory analyses provided additional insights into bike-sharing patterns from a station-level perspective, complementing the trip-level EDA conducted earlier. By examining temporal and environmental influences on bike usage, we gained a clearer understanding of the factors that may contribute to station-level anomalies. This analysis serves as a contextual foundation for applying the Isolation Forest model, allowing us to better interpret its results in the following sections.

5.2 Station-Level Anomaly Detection with Isolation Forest

After performing exploratory data analysis, we applied the Isolation Forest (IF) algorithm to detect station-level anomalies. The goal was to identify stations that exhibited unusual behavior in terms of trip patterns, weather conditions, and transit availability.

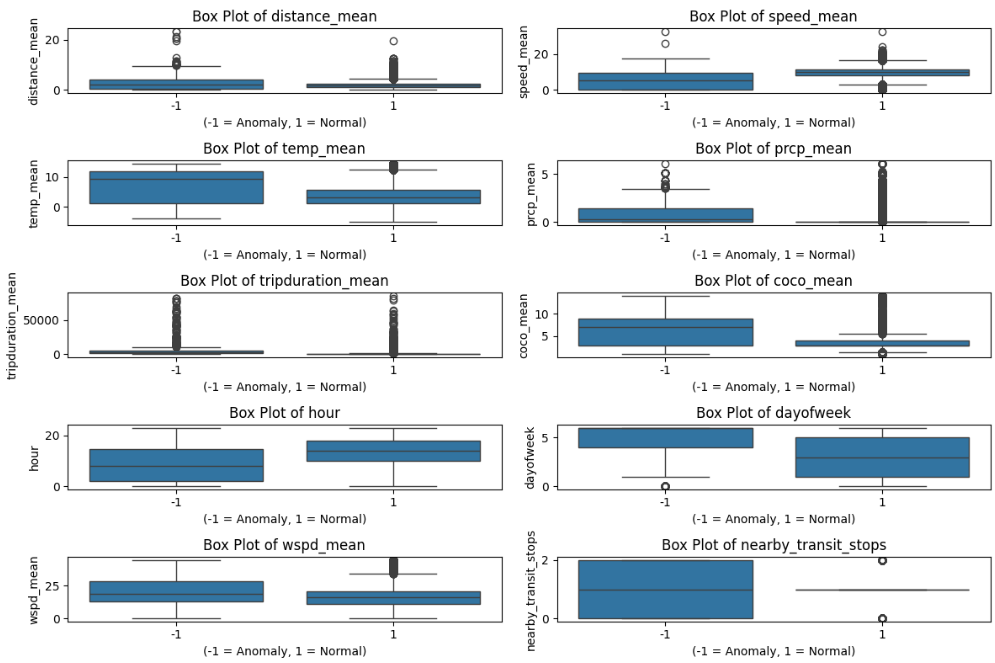


Figure 5.5: Boxplots of selected features for normal and anomalous stations.

To understand how anomalous stations differ from normal ones, we analyzed key feature distributions. Figure 5.5 presents boxplots comparing selected features between normal and anomalous stations. As observed, anomalies exhibit greater variability, particularly in features such as *tripduration_mean*, *distance_mean*, and environmental factors like *prcp_mean* (precipitation). Additionally, *public_transit_stops* shows more variability among anomalies, whereas normal stations are often classified as 1, indicating good level connectivity to public transit. Examining the *hour* feature, anomalies appear across all times of the day but are more frequent in the early morning and daytime. Similarly, looking at the *dayofweek* feature, anomalies show a tendency to occur more often on weekends compared to normal stations. These findings indicate that stations flagged as anomalies often operate under unique conditions or experience trip patterns that deviate significantly from the norm.

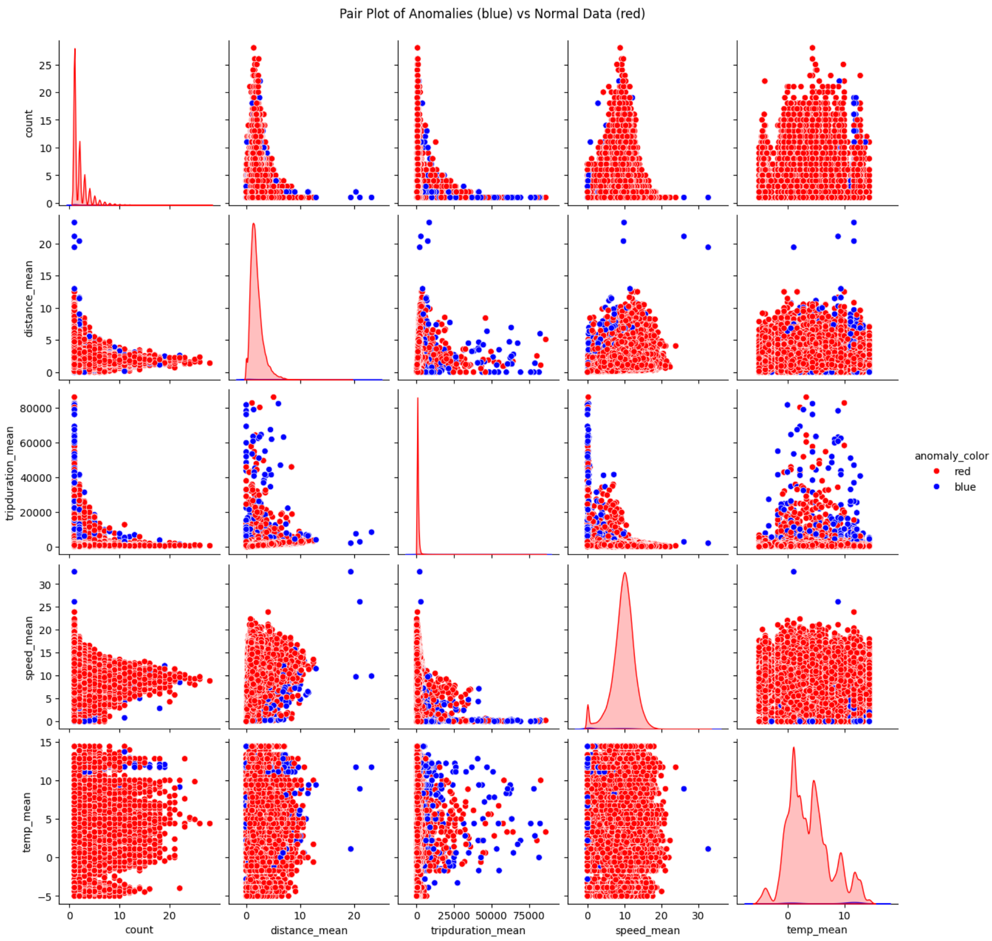


Figure 5.6: Pair plot comparing normal and anomalous stations based on key features.

Additionally, to visualize interactions between different features, we examined pairwise relationships using a pair plot, shown in Figure 5.6. This visualization highlights distinct patterns in anomalous stations (blue), which tend to cluster in regions with extreme values of trip duration, distance, and weather-related attributes.

Next, temporal analysis was performed to examine how anomalies are distributed across different time periods. Figure 5.7 presents the trip count distribution over different hours of the day and across days of the week, with anomalies highlighted separately. A notable spike in anomalies occurs around 8 AM, aligning with the morning commute period, while another peak is observed on Thursdays. These observations indicate that unusual activity tends to cluster around specific temporal patterns, suggesting potential external influences that merit further investigation.

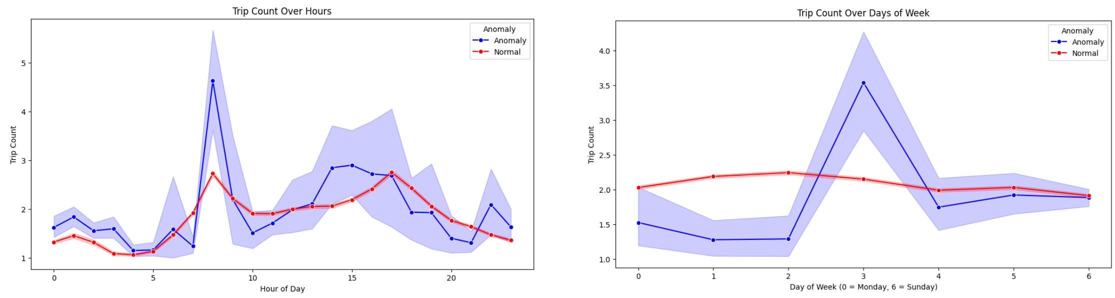


Figure 5.7: Temporal analysis of anomalies: hourly (left, over hours of the day) and daily (right, over days of the week) trip distributions.

The results of Isolation Forest provide a deeper understanding of anomalies at the station level. While temporal and environmental factors play a major role, further spatial analysis and interpretation are required to determine whether localized disruptions, such as public transport failures or external events, contribute to these anomalies.

5.3 Investigating the Anomaly Spikes

After detecting anomalies, we observed two prominent spikes in the anomaly distribution: one around 8 AM and another on Thursdays, as shown in Figure 5.7. This temporal overlap raised the question of whether an event occurring specifically on Thursdays at 8 AM could have driven these anomalies. To investigate this, we conducted targeted analyses and, ultimately, evaluated whether our DIFFI interpretation method could have surfaced these insights more efficiently.

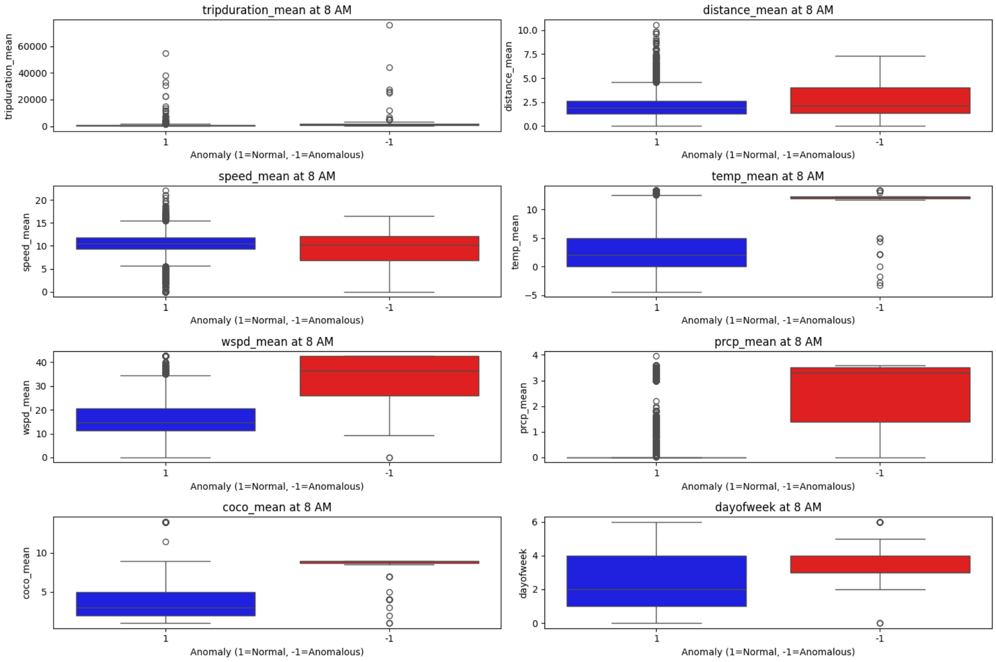


Figure 5.8: Boxplot analysis of anomalies occurred at 8 AM.

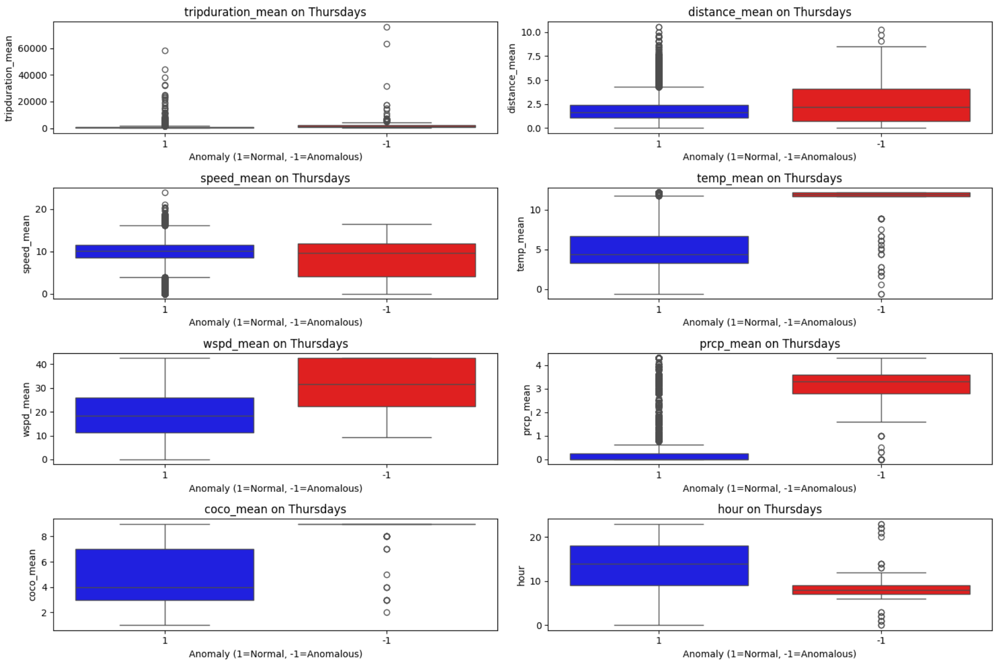


Figure 5.9: Boxplot analysis of anomalies occurred on Thursdays.

We began by analyzing anomalies that occurred at 8 AM to understand their distinct characteristics and potential causes. In Figure 5.8, Thursdays stand out among weekdays for exhibiting the highest concentration of anomalies, suggesting that specific patterns or external factors may be influencing these occurrences. Conversely, Figure 5.9, which focuses exclusively on Thursdays, highlights 8 AM as the peak anomaly time. This overlap between day and time strengthens the hypothesis that recurring events or systemic issues may be contributing to the observed spikes, prompting a deeper investigation into these irregularities.

Delving further, Figure 5.10 presents boxplots comparing various features for normal and anomalous stations specifically during 8 AM on Thursdays. This analysis reveals that anomalies tend to coincide with adverse weather conditions, marked by elevated wind speeds and higher precipitation levels. Such conditions can directly impact biking behavior, leading to irregular usage. Additionally, the *public_transit_stops* feature shows a notable reduction in nearby transit availability for anomalies, suggesting that limited access to alternative transportation may have further exacerbated the situation. These findings indicate that the anomaly spike was likely driven by a combination of environmental challenges and reduced public transit options.

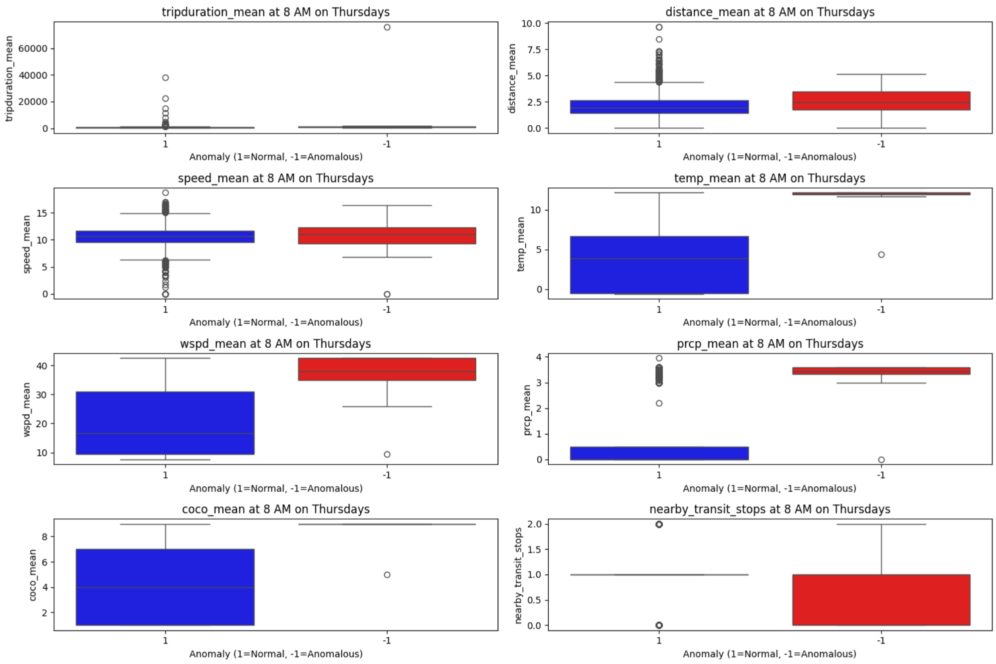


Figure 5.10: Boxplot analysis of feature distributions for anomalies at 8 AM on Thursdays.

To validate these findings and evaluate whether similar conclusions could be drawn more efficiently, we used the DIFFI interpretation method. As shown in

Figure 5.11, DIFFI identified public transit availability and environmental conditions as key contributors to the anomalies, reinforcing our earlier conclusions derived from manual exploratory analysis. One of DIFFI's key strengths lies in its ability to highlight influential features without requiring labor-intensive data exploration. In this case, DIFFI quickly emphasized the importance of both public transit accessibility and adverse weather conditions, streamlining a process that would have otherwise required multiple iterative analyses. The consistency between DIFFIs output and our prior findings also reinforces the methods reliability, demonstrating its capability to uncover complex relationships in the data while significantly reducing the time and effort needed for interpretation. This efficiency is particularly valuable in large-scale datasets, where manual feature analysis can become time-consuming and error-prone.

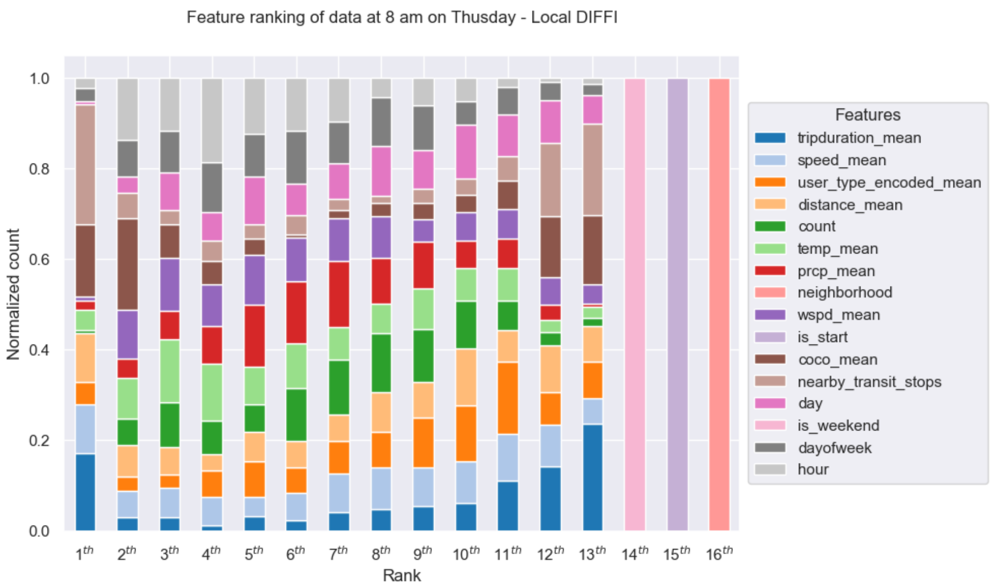


Figure 5.11: Feature ranking interpretation using DIFFI for anomalies at 8 AM on Thursdays.

Spatial analysis offered further insights into the distribution of anomalies across the city. Figure 5.12 maps the total number of anomalies at 8 AM on Thursdays. While anomalies appear to cluster around specific locations, these areas typically experience high traffic volumes, making absolute counts less informative for understanding the true anomaly density.

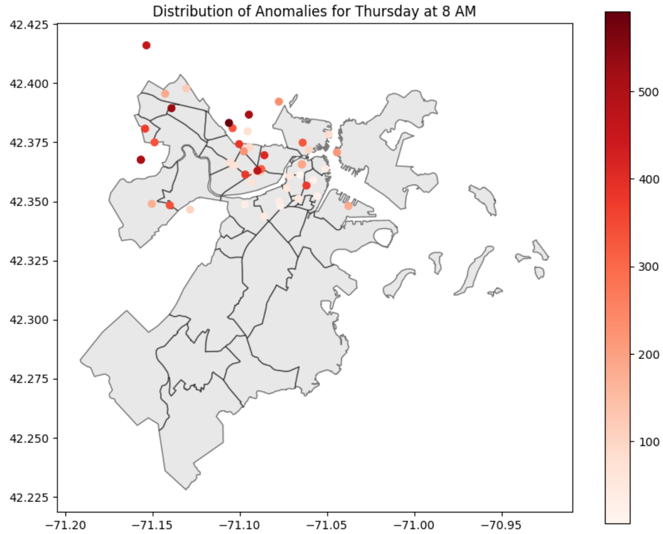
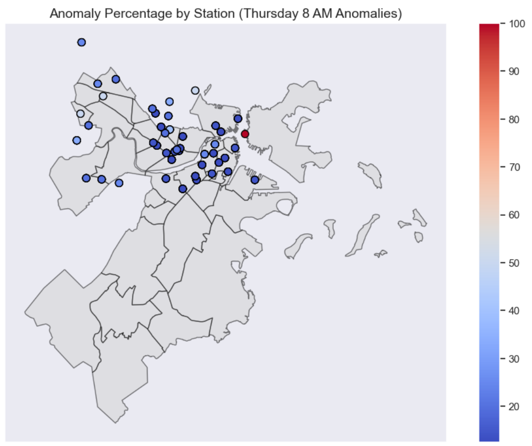
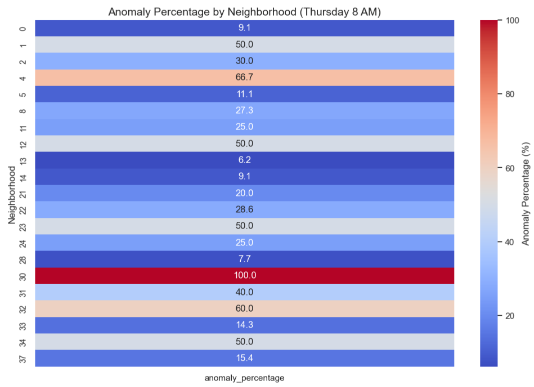


Figure 5.12: Spatial distribution of anomalies (absolute count) at 8 AM on Thursdays.

To provide a more balanced view, Figure 5.13 combines station-level and neighborhood-level analyses. The left panel shows the percentage of anomalies relative to total trips per station, normalizing for station activity. The right panel aggregates anomalies at the neighborhood level, indicating which areas were disproportionately affected. This dual perspective highlights specific neighborhoods and stations that experienced higher anomaly rates, reinforcing the likelihood of localized disruptions or events.



(a) Station-level anomaly percentage.



(b) Neighborhood-level anomaly percentage.

Figure 5.13: Spatial distribution of anomalies at 8 AM on Thursdays. The left panel displays station-level anomaly percentages, while the right panel highlights neighborhood-level distributions.

These findings confirm that the anomaly spikes at 8 AM on Thursdays were driven by a combination of adverse weather, reduced public transit availability, and localized disruptions. By integrating spatial and temporal analyses, alongside interpretability methods like DIFFI, we achieved a deeper understanding of the factors influencing these anomalies while also identifying efficient paths for future anomaly investigations.

5.4 Neighborhood and Station-Level Spatial Analysis

To further understand the spatial dynamics of anomalies, we conducted a neighborhood and station-level analysis. This exploration aimed to uncover patterns that could reveal localized disruptions, high-risk areas, and the spatial consistency of anomalies over time.

We first examined the overall anomaly distribution across neighborhoods to identify broader spatial trends. Figure 5.14 presents the total anomaly counts by neighborhood and day. This heatmap highlights areas consistently affected by anomalies, with certain neighborhoods exhibiting sustained high anomaly counts throughout the month.

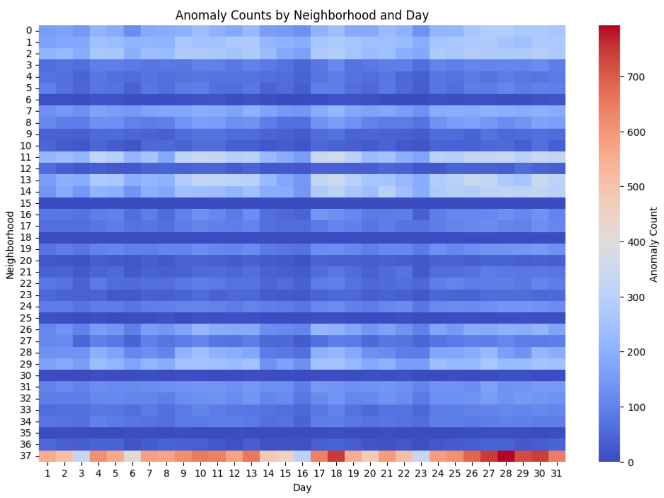


Figure 5.14: Anomaly counts by neighborhood and day.

While raw counts provide valuable insights, they can be biased toward neighborhoods with higher overall activity. To mitigate this bias, Figure 5.15 displays the percentage of anomalous stations relative to the total number of stations in each neighborhood. This normalized view highlights which days are more active within each neighborhood, allowing for a fairer comparison across areas. For instance, January 1st, which is widely recognized as a busy day due to New Year's celebrations, exhibited high anomalous activity across nearly all neighborhoods, reflecting the impact of large-scale events on anomaly rates.

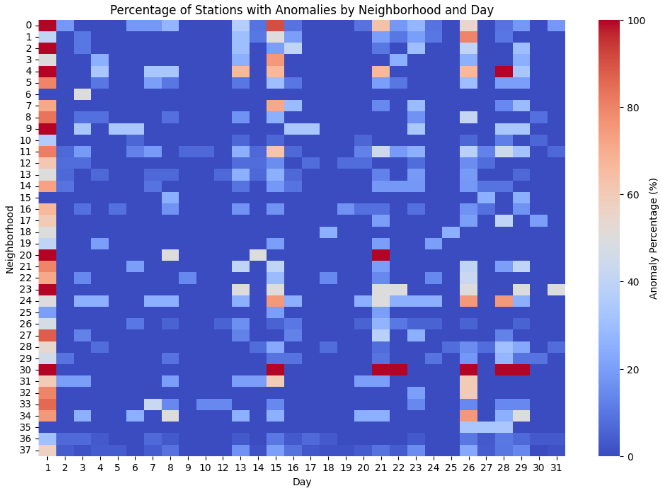


Figure 5.15: Anomaly percentage of stations by neighborhood and day.

To delve deeper into the relationship between anomaly frequency and trip activity, Figure 5.16 presents anomaly percentages relative to total trip counts in each neighborhood. This visualization helps identify neighborhoods where high anomaly rates occur on days with significant activity, pinpointing critical areas that require further investigation. From this heatmap, it is clear that January 1st remains distinct due to its elevated anomaly levels, though fewer neighborhoods show high percentages compared to the previous heatmap. This suggests that while January 1st was a busy day across the city, only specific neighborhoods experienced a high concentration of anomalies. The significance of anomalies highlighted in this heatmap is greater, as it indicates that on certain days and in specific neighborhoods, the anomalies were particularly concentrated. Depending on the investigation's focus, both heatmaps offer valuable perspectives, with this version emphasizing more severe anomaly cases.

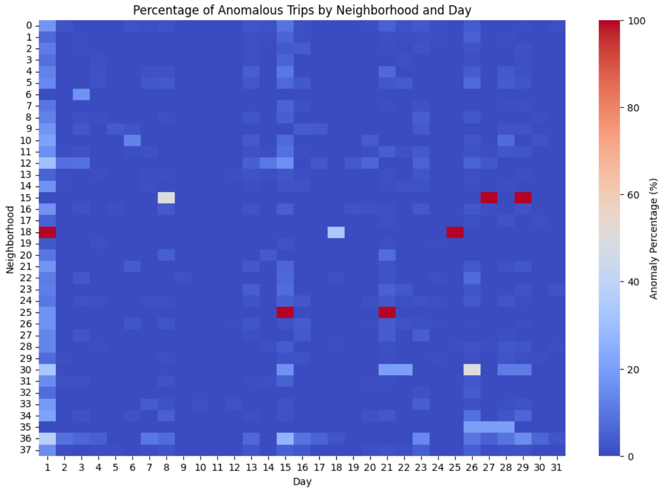


Figure 5.16: Anomaly percentage of total trip counts by neighborhood and day.

Shifting focus to the station level, we identified the top 20 stations with the highest anomaly counts for a more granular analysis. Figure 5.17 highlights the temporal distribution of anomalies at these stations. This heatmap reveals specific stations that consistently experience irregular patterns, suggesting localized factors or recurring events that may be driving these anomalies.

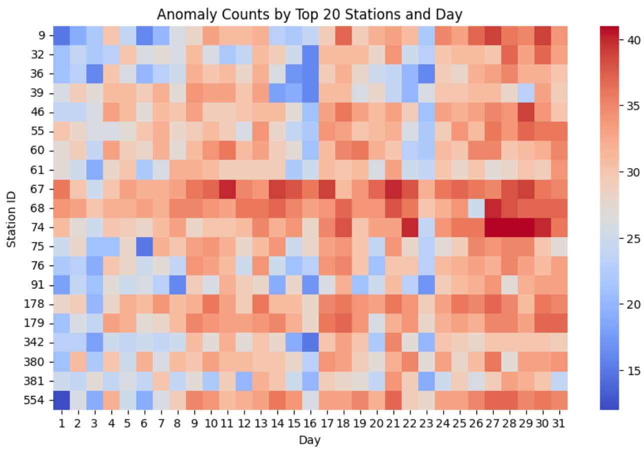


Figure 5.17: Anomaly counts by top 20 stations and day.

To better understand the spatial context of these high-anomaly stations, Figure 5.18 maps their locations across the city. This spatial visualization highlights clustering patterns and reveals neighborhoods that experience concentrated anomaly activity.

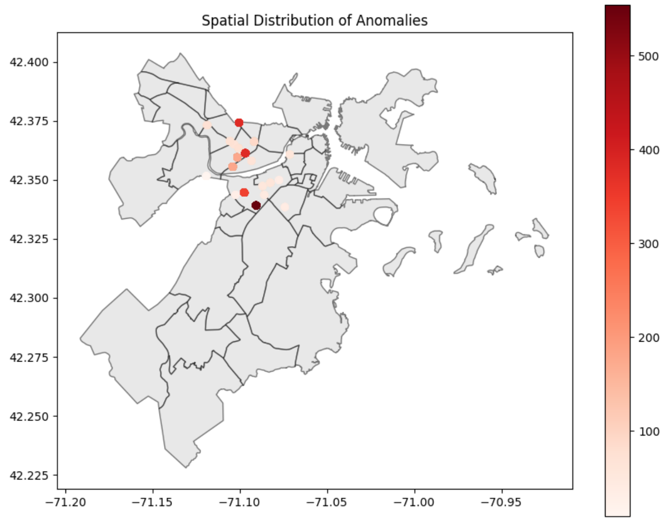


Figure 5.18: Spatial distribution of anomalies at top 20 stations.

In Figure 5.14, we observed that Neighborhood 37 frequently experiences a high number of anomalies on almost every day. To further investigate this neighborhood and simulate how such cases could be explored in real-world scenarios, we conducted a focused analysis of its internal dynamics. Figure 5.19 presents two perspectives: the left panel shows total anomaly counts by station and day, while the right panel normalizes these counts by total trips, providing insight into stations with disproportionately high anomaly rates despite lower overall activity.

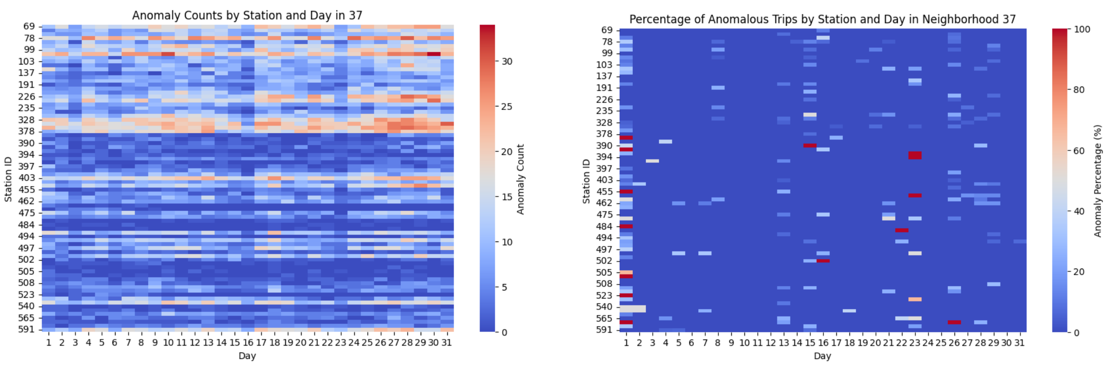


Figure 5.19: Anomaly analysis for Neighborhood 37: (left) total anomaly counts, (right) anomaly percentages.

To further contextualize these findings, Figure 5.20 maps the spatial distribution of anomalies within Neighborhood 37. This visualization highlights specific stations that serve as anomaly hotspots, enabling targeted investigations or potential interventions.

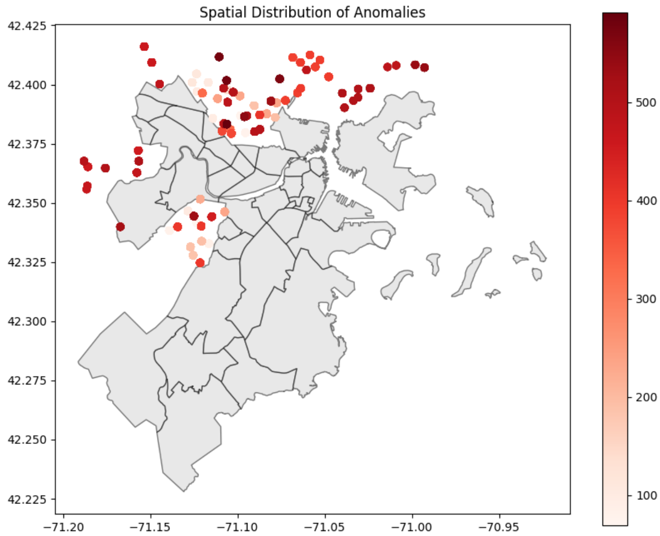


Figure 5.20: Spatial distribution of anomalies in Neighborhood 37.

Through this neighborhood and station-level spatial analysis, we identified areas of heightened anomaly activity, potential localized disruptions, and broader spatial dynamics that influence irregular patterns. These insights lay the groundwork for targeted interventions and future anomaly mitigation strategies, supporting more efficient and data-driven urban mobility management.

5.5 Case Study: Zero-Duration Anomalies

During the anomaly investigation, an interesting pattern emerged regarding trips with a duration of zero but with recorded distances. This finding raised questions since trips with zero distance had already been filtered out during the data cleaning phase. The presence of trips with zero duration but non-zero distance required a deeper examination to understand their nature and potential causes.

Figure 5.21 highlights the distribution of key features, showcasing a significant cluster of anomalies with zero duration. Despite the zero duration, these trips often registered considerable distances and non-zero speeds, indicating that these were not data entry errors but reflected some form of actual movement.

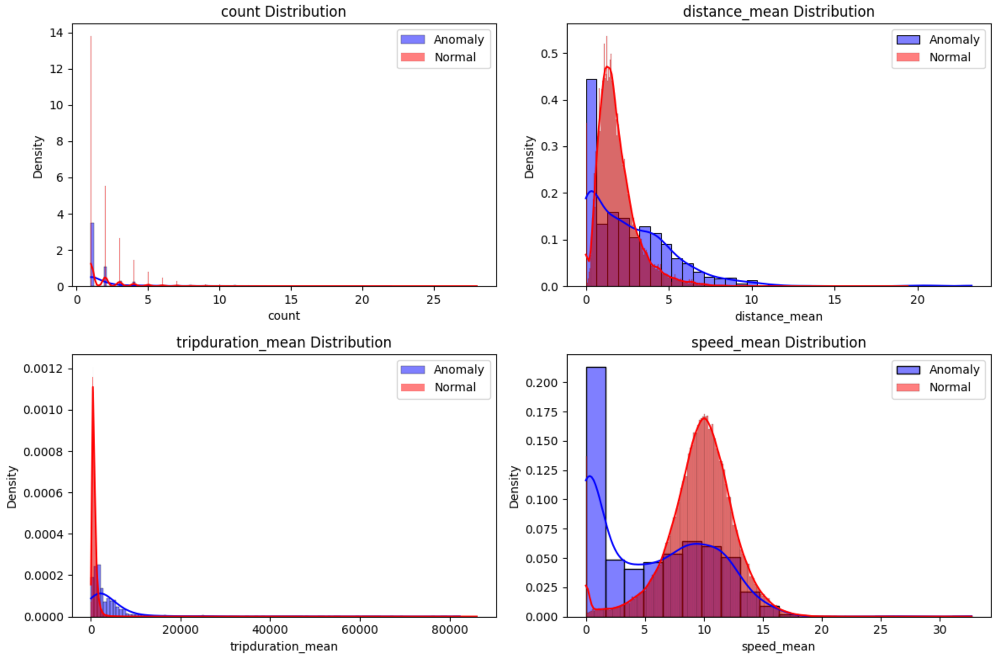


Figure 5.21: Density distribution of key features showing a concentration of zero-duration anomalies.

This observation led to the hypothesis that such trips could be self-looping rides, often associated with touristic routes, leisure activities, or park-based circuits. To further validate this assumption, we analyzed the temporal distribution of these zero-duration trips. Figure 5.22 illustrates the frequency of these trips across the days of the week, revealing a clear spike during weekends. This pattern aligns with the nature of leisure trips, which typically occur more frequently during weekends when users have more free time.

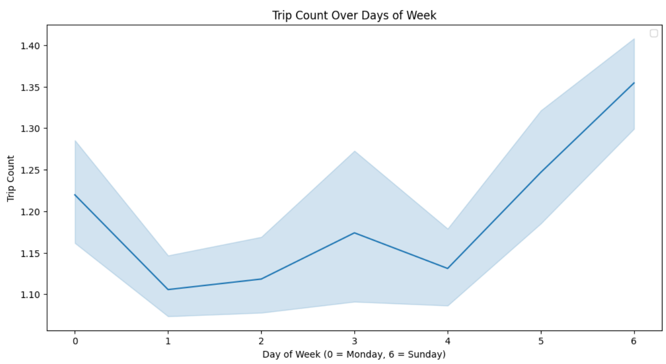


Figure 5.22: Distribution of zero-duration trips across days of the week, highlighting weekend peaks.

Further supporting this hypothesis, an analysis of station names associated with these anomalies revealed that many trips originated from or ended at locations near parks, recreational spaces, or touristic areas. Stations like *Discovery Park - 30 Acorn Park Drive*, *Rindge Avenue - O'Neill Library*, and *Murphy Skating Rink - 1880 Day Blvd* frequently appeared among these anomalies. The common presence of words such as "Park," "Skating Rink," and "Bikeway" in these station names reinforces the idea that these trips are often recreational or leisure-oriented.

To further validate these findings, we applied the DIFFI interpretation method to assess feature importance for these anomalies. As shown in Figure 5.23, the analysis highlights the strong influence of temporal features, particularly the *day*, *day of the week* and the *is_weekend* indicator, which consistently rank among the top contributing factors. This reinforces the observation that these anomalies are closely tied to weekends, aligning with the hypothesis of leisure-oriented trips. This interpretability analysis confirms that the observed anomalies are not random but are driven by specific temporal and spatial patterns.

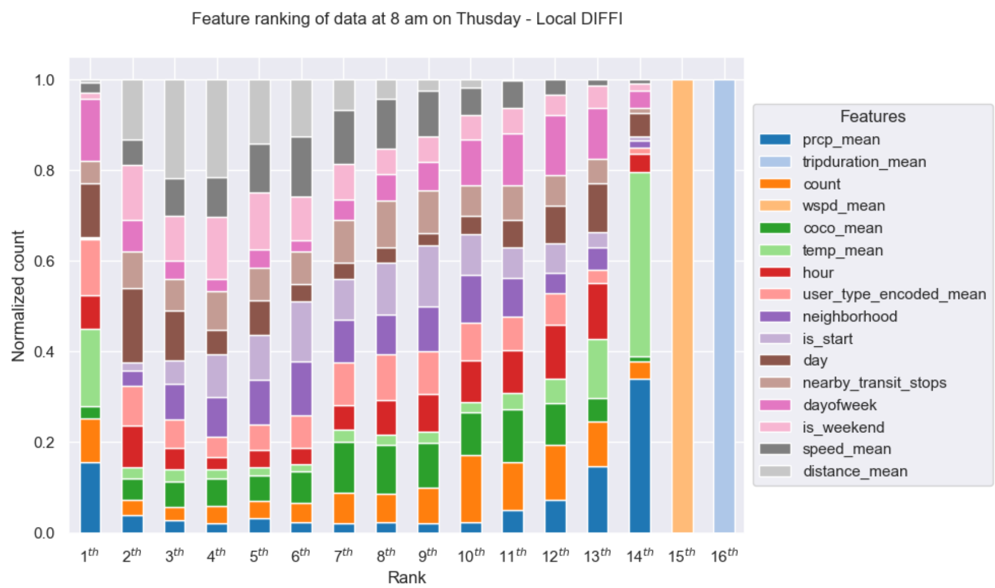


Figure 5.23: Feature importance analysis using DIFFI, emphasizing weekend-related features.

The combination of temporal trends, spatial patterns, and interpretability analysis confirms that the zero-duration anomalies largely represent recreational or self-looping trips, rather than errors or random outliers. These findings highlight the value of anomaly detection not only for identifying irregularities but also for uncovering behavioral insights that can inform system improvements and urban planning strategies. This case study demonstrates how interpretability methods can validate hypotheses and strengthen data-driven conclusions, providing a comprehensive understanding of seemingly anomalous behaviors.

5.6 General Insights and Final Interpretation

This chapter has explored the spatial and temporal dynamics of anomalies within the bike-sharing system, revealing key patterns, underlying causes, and context-specific insights. To consolidate these findings, we present a general interpretation of the entire dataset and compare station-level and trip-level analyses to evaluate their effectiveness in identifying and explaining anomalies.

The DIFFI interpretation method was applied to the full dataset to assess overall feature importance. Figure 5.24 visualizes the ranked contributions of different features in detecting anomalies. Notably, the station-level interpretation highlights the complex interplay between spatial, temporal, and contextual factors. Features such as *dayofweek*, *is_weekend*, *prcp_mean* (precipitation), and *nearby_transit_stops* consistently appear among the top ranks, underscoring the influence of weather conditions, transit accessibility, and broader temporal patterns on anomaly occurrences.

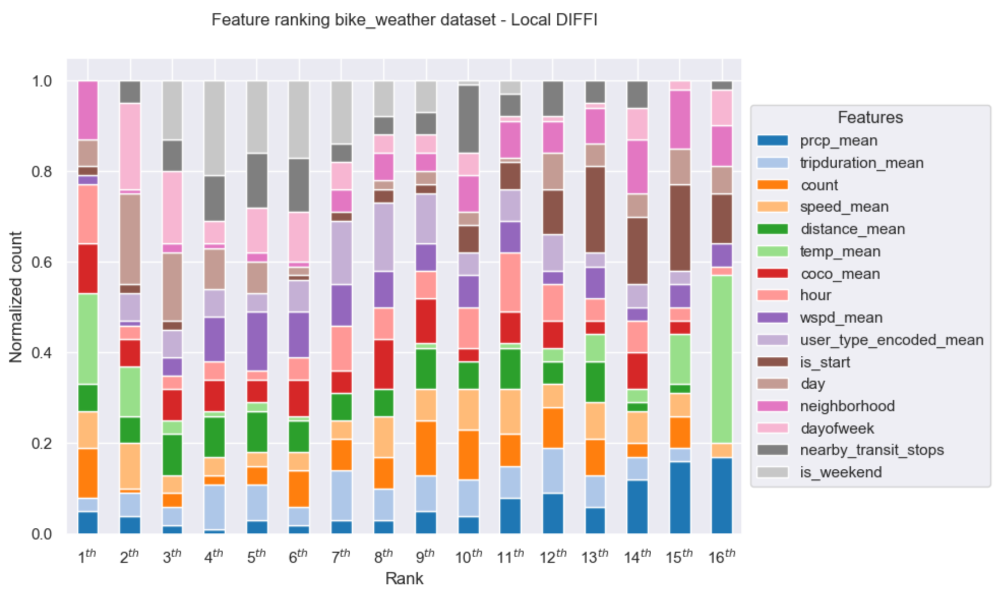


Figure 5.24: Feature ranking of the entire station-level dataset using Local DIFFI.

To contextualize these results, we compared them with the trip-level DIFFI analysis shown in Figure 4.14. The trip-level interpretation focuses predominantly on trip-specific attributes, with features such as *tripduration*, *distance*, *start_hour*, and *speed* consistently dominating the rankings. This approach is effective in capturing micro-level anomalies, such as irregular trip durations or atypical speeds, but it offers limited insight into external factors that may influence anomaly patterns.

The contrast between the two analyses reveals important insights. The station-level interpretation integrates broader contextual variables, such as neighborhood characteristics and public transit availability, leading to a more comprehensive understanding of anomalies. In this approach, feature diversity across ranks is more pronounced, with no single feature overwhelmingly dominating, indicating that anomalies often result from complex interactions among multiple factors.

Conversely, the trip-level analysis is narrower in scope, with a concentration of importance on a handful of trip-specific features. While this approach excels at identifying individual outlier trips, it lacks the depth to uncover systemic issues or localized disruptions that may manifest across multiple trips or stations.

This comparison highlights the added value of station-level analysis, especially when aiming to understand anomalies in a broader urban mobility context. By incorporating environmental, spatial, and temporal factors, the station-level approach provides richer, more actionable insights, making it a preferable choice for city planners and policymakers seeking to implement targeted interventions.

In conclusion, while trip-level analysis remains a valuable tool for detecting anomalies at the micro scale, station-level interpretation offers a more holistic perspective that captures the broader dynamics of shared mobility systems. The integration of both approaches allows for a multi-layered understanding of anomalies, enhancing both anomaly detection and the development of effective mitigation strategies.

6 Conclusions and Future Works

This thesis presented an interpretable anomaly detection framework for shared mobility systems, with a specific focus on bike-sharing networks. By integrating multi-source data and employing unsupervised learning techniques, the study addressed key challenges in anomaly detection, particularly the absence of labeled data and the need for model interpretability. The application of DIFFI enhanced the transparency of the Isolation Forest model, providing a clearer understanding of the factors driving detected anomalies.

The findings of this thesis offer valuable insights for shared mobility operators and urban planners. By identifying atypical patterns in bike usage, system operators can proactively address operational inefficiencies, optimize resource allocation, and improve user experience. Furthermore, the integration of external factors, such as weather conditions and public transit availability, allows for a more comprehensive understanding of mobility patterns, aiding strategic planning and infrastructure development.

Despite its contributions, this study faced several limitations. The reliance on unsupervised methods without ground-truth labels introduced challenges in evaluating model accuracy and fine-tuning hyperparameters. Additionally, the analysis was confined to a single city's dataset over a specific time frame, which may limit the generalizability of the findings to other urban environments. The exclusion of real-time data streams and potential external disruptions further constrained the scope of the study.

Future research can expand upon this work by extending the analysis to multiple cities and larger datasets to validate the framework's generalizability.

Investigating real-time anomaly detection systems could enable dynamic operational responses, enhancing system adaptability. Incorporating additional data sources, such as demographic information or real-time traffic data, could further enrich contextual analysis and improve anomaly detection accuracy. Finally, exploring advanced interpretability methods with a focus on minimizing the false negative rate could strengthen trust and transparency in anomaly detection models, ensuring that critical anomalies are not overlooked.

References

[1] BlueBikes. *BlueBikes System Data*. 2023. URL: https://www.bluebikes.com/system-data.

[2] Mattia Carletti, Matteo Terzi, and Gian Antonio Susto. "Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest". In: *Engineering Applications of Artificial Intelligence* 119 (2023), p. 105730. DOI: 10.1016/j.engappai.2022.105730.

[3] Federico Chiariotti et al. "A Dynamic Approach to Rebalancing Bike-Sharing Systems". In: *Sensors* 18.2 (2018), p. 512. DOI: 10.3390/s18020512.

[4] City of Boston. *Boston Neighborhood Boundaries (2020 Census Block Groups)*. Accessed: 2024-02-05. 2023. URL: https://data.boston.gov/dataset/boston-neighborhood-boundaries-approximated-by-2020-census-block-groups1.

[5] City of Cambridge GIS. *Cambridge Neighborhood Boundaries*. Accessed: 2024-02-05. 2023. URL: https://www.cambridgema.gov/GIS/gisdatadictionary/Boundary/BOUNDARY_CDDNeighborhoods.

[6] Giacomo D'Amicantonio, Egor Bondarau, and Peter HN de With. "uTRAND: Unsupervised Anomaly Detection in Traffic Trajectories". In: *arXiv preprint arXiv:2404.12712* (2024).

[7] Kyoungok Kim. "Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations". In: *Journal of Transport Geography* 66 (2018), pp. 309–320. DOI: 10.1016/j.jtrangeo.2018.01.001.

[8] Alex Lam, Matthew Schofield, and Shen-Shyang Ho. "Detecting (Unusual) Events in Urban Areas using Bike-Sharing Data". In: *3rd ACM SIGSPATIAL International Workshop on Analytics for Local Events and News (LENS'19)*. Chicago, IL, USA: ACM, 2019, p. 7. DOI: 10.1145/3356473.3365190.

[9] Chao Liu, Xiao Gao, and Xiaokang Wang. "Data Adaptive Functional Outlier Detection: Analysis of the Paris Bike Sharing System Data". In: *Information Sciences* 602 (2022), pp. 13–42. DOI: `10.1016/j.ins.2022.04.029`.

[10] Massachusetts Bay Transportation Authority (MBTA). *MBTA General Transit Feed Specification (GTFS) Data*. Data for the period 26 December 2023 - 6 April 2024. Accessed: 2024-02-05. 2023. URL: `https://transitfeeds.com/p/mbta/64/latest`.

[11] Meteostat. *Meteostat Weather Data*. 2023. URL: `https://meteostat.net`.

[12] Katja Schimohr and Joachim Scheiner. "Spatial and Temporal Analysis of Bike-Sharing Use in Cologne Taking into Account Public Transit Disruption". In: *Journal of Transport Geography* 92 (2021), p. 103017. DOI: `10.1016/j.jtrangeo.2021.103017`.

[13] Rayane El Sibai, Yousra Chabchoub, and Christine Fricker. "Using Spatial Outliers Detection to Assess Balancing Mechanisms in Bike Sharing Systems". In: *32nd IEEE International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2018, pp. 988–995. DOI: `10.1109/AINA.2018.00144`.

[14] João Filipe Teixeira and Miguel Lopes. "The Link Between Bike Sharing and Subway Use During the COVID-19 Pandemic: The Case Study of New York's Citi Bike". In: *Transportation Research Interdisciplinary Perspectives* 6 (2020), p. 100166. DOI: `10.1016/j.trip.2020.100166`.

[15] Nicolai A. Weinreich et al. "Automatic Bike Sharing System Planning from Urban Environment Features". In: *Preprint for submission* (2023). Draft for submission, April 13, 2023.

[16] World Weather. *Boston Weather Forecast - January 2023*. Accessed: February 18, 2025. 2023. URL: `https://world-weather.info/forecast/usa/boston/january-2023/`.

[17] Chen Xu et al. "A Survey on Explainable Anomaly Detection". In: *ACM Computing Surveys (CSUR)* 55.12 (2023), pp. 1–37. DOI: `10.1145/3609333`.

[18] Wei Zhang, Juan Li, and Rui Wang. "A Quantitative Study on Factors Influencing User Satisfaction of Shared Micro-Mobility". In: *Sustainability* 16.4 (2023), p. 1637. DOI: `10.3390/su16041637`.