Transit Operations Monitoring Dashboard

by Krupali Shinde

Abstract

Urban transportation systems are fundamental to the movement and functioning of modern cities. As populations grow and congestion increases, there is a greater need for data-driven monitoring and optimization of public transit services. This project focuses on building a comprehensive Transit Operations Monitoring Dashboard using data from New York City's Metropolitan Transportation Authority (MTA). By leveraging historical bus operations data, the dashboard provides real-time insights into key metrics such as bus trip frequency, average travel time, average road speed, and equipment malfunction occurrences across different boroughs and routes.

The analysis begins with the cleaning and transformation of raw data to ensure quality and consistency. Using tools such as Python for preprocessing and Power BI for interactive visualization, the data is explored across multiple dimensions—temporal (by hour and day), geographical (borough-level analysis), and operational (equipment status and route behavior). By aggregating and visualizing this data, the project highlights patterns such as rush hour delays, high-traffic routes, borough-wise disparities in speed, and a notable percentage of trips impacted by equipment issues. The dashboard supports dynamic filtering, allowing stakeholders to drill down into specific boroughs, routes, incident types, and times of day for focused analysis.

The ultimate goal of the dashboard is to empower transit authorities, planners, and analysts with a tool that can facilitate quick decision-making and performance monitoring. The insights derived from this system can be used to prioritize maintenance schedules, optimize route planning, improve passenger experience, and reduce unplanned downtimes. By identifying inefficiencies in real time and over historical trends, this solution supports both strategic planning and tactical response, contributing to a more efficient and reliable public transportation network.

Table Of Contents

1. INTRODUCTION

1.1 Background and Context

Public transportation forms the backbone of urban infrastructure, especially in a densely populated metropolis like New York City. With millions of residents and tourists relying on MTA (Metropolitan Transportation Authority) buses daily, the efficiency and reliability of these transit services directly impact economic productivity, commuter satisfaction, and environmental sustainability. Buses offer a flexible, affordable, and relatively low-emission alternative to personal vehicles, but only when managed with precision. Delays, frequent breakdowns, and unpredictable schedules erode public trust in the transit system and can cause ripple effects across the broader transportation network.

In a city characterized by traffic congestion, road construction, weather fluctuations, and varied terrain across boroughs, monitoring the health of bus operations becomes not only a logistical requirement but also a policy imperative. Timely and data-driven decisions by transit authorities can lead to optimized routing, reduced service downtime, equitable transportation access, and enhanced commuter experience. Historically, such insights were drawn from manual surveys or incident logs; however, with the integration of GPS systems, IoT devices, and automated data collection in buses, it is now possible to obtain granular, near real-time data that can be harnessed for operational intelligence.

This project builds upon that foundation by analyzing historical MTA bus data with the aim of uncovering operational patterns, service bottlenecks, and incident distributions. In doing so, it helps stakeholders gain a clearer view of the current state of the MTA's transit services. It also equips them with the insights necessary to implement timely interventions—whether it's allocating more maintenance crews to malfunction-prone routes, rescheduling trips to avoid congestion, or identifying boroughs underserved by transit services. Through powerful visual analytics, the project aspires to elevate transit performance monitoring from a reactive exercise to a proactive, data-informed strategy.

1.2 Overview of the Dataset

The dataset used in this project originates from New York City's MTA Bus Time platform and Open Data NYC, capturing a wide range of operational metrics for bus routes across multiple boroughs. This includes detailed GPS-level logs of bus trips, aggregated travel speeds, average travel time between stops, timestamps, route identifiers, and trip counts. Importantly, the dataset also includes flags for incidents, such as equipment malfunctions or anomalies, offering a critical lens into downtime and maintenance challenges that impact service delivery.

Each record in the dataset reflects a composite of real-time and historical measurements captured at various intervals of the day and across multiple days or weeks. Attributes like route_id, borough, hour_of_day, and day_of_week allow for time-series and geospatial analysis, while quantitative fields such as average_road_speed, bus_trip_count, and average_travel_time enable performance benchmarking. Incident-related columns such as incident_flag help isolate problematic trips, adding a qualitative dimension to the dataset that can be cross-analyzed with operational variables.

Moreover, the dataset encapsulates not just operational metrics but behavioral insights—such as when and where delays are most likely to occur, how congestion patterns shift throughout the day, and what boroughs show consistently lower travel speeds. This multidimensionality enables the creation of rich visualizations that tell the full story behind bus operations. However, as with any real-world data, this dataset also presents challenges like missing values, inconsistent data types (e.g., travel time logged as text), and occasional GPS noise, all of which had to be addressed during preprocessing.

By thoroughly understanding the nature, scope, and limitations of this dataset, this project is able to extract meaningful insights and develop an actionable dashboard that supports data-informed decision-making for transit performance management.

2. OBJECTIVE

2.1 Goals of the Project

The primary goal of this project is to design a comprehensive transit monitoring solution that can track and evaluate the operational performance of MTA bus services in New York City. With a high volume of daily commuters relying on the city's public transportation system, there is an ongoing need to ensure that buses operate efficiently, adhere to schedules, and minimize delays or breakdowns. This project aims to provide a centralized dashboard that brings together critical operational metrics such as bus trip frequency, road speed, travel time distribution, and incident reporting by route and borough. By visualizing this data, stakeholders can easily identify service gaps, performance issues, and areas for strategic improvement.

Another important goal is to empower transit authorities and planners with the ability to pinpoint congestion zones and determine the impact of incidents, especially equipment malfunctions, on overall bus service performance. By identifying patterns based on route ID, time of day, and borough, decision-makers can proactively respond to recurring issues. For instance, by observing which hours see the highest delays or which routes frequently report equipment problems, targeted interventions can be implemented. The project bridges the gap between raw operational data and actionable intelligence.

Lastly, this project seeks to create a user-friendly and interactive Power BI dashboard that serves both high-level executives and operational managers. Users can filter and slice the data across different dimensions such as boroughs, route types, and incident flags, allowing for customized insights tailored to various operational concerns. Whether it's optimizing driver schedules or allocating maintenance crews, the dashboard supports data-driven decisions that improve transit reliability and commuter satisfaction.

2.2 Business and Technical Objectives

Business Objectives:

From a business standpoint, the key objective is to reduce transit system inefficiencies that contribute to rider dissatisfaction and operational losses. Equipment malfunctions, for example, can significantly impact the reliability of bus services. By identifying high-risk routes

and malfunction-prone time windows, this dashboard supports more proactive maintenance scheduling and resource allocation. The project also aims to highlight service disparities across boroughs, enabling equitable transit planning. Ultimately, the goal is to enhance commuter experience, reduce costs associated with delays and downtime, and improve public perception of MTA bus services.

Another critical business objective is improving transparency and accountability within public transportation management. The dashboard allows stakeholders to monitor trends over time and respond to performance dips with appropriate corrective measures. Whether it's increasing service frequency in underperforming boroughs or addressing operational bottlenecks during peak hours, the project serves as a foundation for strategic urban transit planning.

Technical Objectives:

On the technical side, the first objective is to clean, preprocess, and model the operational MTA dataset for meaningful insights. This includes handling missing values, standardizing data formats, and converting text-based features into numerical form where necessary (e.g., converting travel time from text to numeric). Another goal is to implement robust data modeling and DAX measures in Power BI to create aggregated summaries such as average road speed, total bus trips, and incident percentages.

Additionally, the project aims to deploy a well-structured Power BI dashboard with multiple visualizations including bar charts, line graphs, pie charts, and stacked columns. These visuals are enhanced with slicers for filtering by borough, route type, incident flag, and time dimensions (hour of day or day of week), offering a rich exploratory analysis experience. The dashboard is not only informative but also dynamic and customizable, ensuring adaptability for future data updates and scalability to other transit systems.

3. DATASET OVERVIEW

3.1 Description of the Dataset

The dataset used in this project, titled MTA_cleaned.csv, represents timepoint-level operational data collected from the Metropolitan Transportation Authority (MTA) bus system in New York City. Each row in the dataset corresponds to a timepoint (a GPS-tracked location at a stop) on a bus route, combined with multiple layers of information, including geolocation data, trip counts, average travel time between stops, and road speeds.

The data is timestamped, allowing it to be segmented by hour, day of the week, and month, making it well-suited for temporal trend analysis. In total, the cleaned dataset contains over 1000 records, covering multiple boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island) and dozens of bus routes, with both local and express service types. The dataset is designed to reflect not only scheduled service data but actual operational measurements, enabling performance analysis and real-world issue detection such as equipment malfunctions.

3.2 Key Features and Variables

Column Name	Description
timestamp	Full datetime stamp of when the data was recorded. Crucial for building time-series trends.
date	Extracted date component from the timestamp. Used for aggregation and daily trend visualization.
year, month, week	Temporal segmentations used to analyze monthly and weekly performance patterns.
day_of_week, hour_of_day, time_of_day	Temporal tags used to understand peak hours, weekend vs weekday trends, and operational load across times of day.
route_id	Unique identifier for each bus route (e.g., B1,

	Q44). Used to compare performance between routes.
route_type	Categorizes the bus route as Local, Express, etc. Important for understanding speed and trip dynamics.
borough	Indicates the borough in which the bus trip segment occurred (e.g., Queens, Manhattan). Enables spatial analysis.
stop_order	Denotes the sequence of stops within the route. Useful for analyzing travel time by segment.
timepoint_stop_id / next_timepoint_stop_id	Represents the current and next stop pair, forming the trip segment.
timepoint_stop_name / next_timepoint_stop_name	Human-readable names for the stops. Helps in visualizing location-specific bottlenecks.
timepoint_stop_latitude, timepoint_stop_longitude	Geo-coordinates of the current stop. Useful for mapping.
road_distance	Distance in miles between the current stop and the next. Important for calculating speed.
average_travel_time	The average time (in minutes) it takes buses to travel between two timepoints. It's a key indicator of delay.
average_road_speed	Computed speed based on road distance and average travel time. Lower speeds may indicate congestion.
bus_trip_count	Number of trips recorded during the time window. Used as a proxy for volume or demand.
IncidentFlag	A derived flag column created in preprocessing: If average_road_speed < 3 mph and bus_trip_count < 5, it's marked as "Equipment Malfunction"; else "Normal". Used to surface downtime events.

3.3 Challenges in the Dataset

During the preprocessing and analysis phase, several challenges were encountered and addressed:

1. Data Type Conversion: The original values in average_travel_time, average_road_speed, and bus_trip_count were stored as strings in the CSV. These had to be converted into numerical types after cleaning commas, empty values, or corrupted entries.

2. Missing Geolocation Data: While most records contained complete geolocation information, some stop pairs lacked latitude/longitude, which could hinder mapping. These were retained only if they didn't break downstream aggregations.

3. Duplicate Records: Several rows were exact duplicates, particularly due to multiple readings in short succession. These were removed based on a composite key of timestamp, route_id, and stop_order.

4. Temporal Gaps: The dataset had data only for specific days (e.g., March 1 and April 1), which may limit long-term trend analysis. However, these dates still offered enough granularity for hour-level performance assessment.

5. Borough-Level Imbalance: Some boroughs had denser coverage than others (e.g., Queens had significantly more trips than Manhattan), requiring normalization or per-borough filtering in dashboards.

6. Manual Feature Engineering: Columns like IncidentFlag had to be created using conditional logic in Python to reflect operational status, as the raw dataset did not explicitly indicate malfunctions or anomalies.

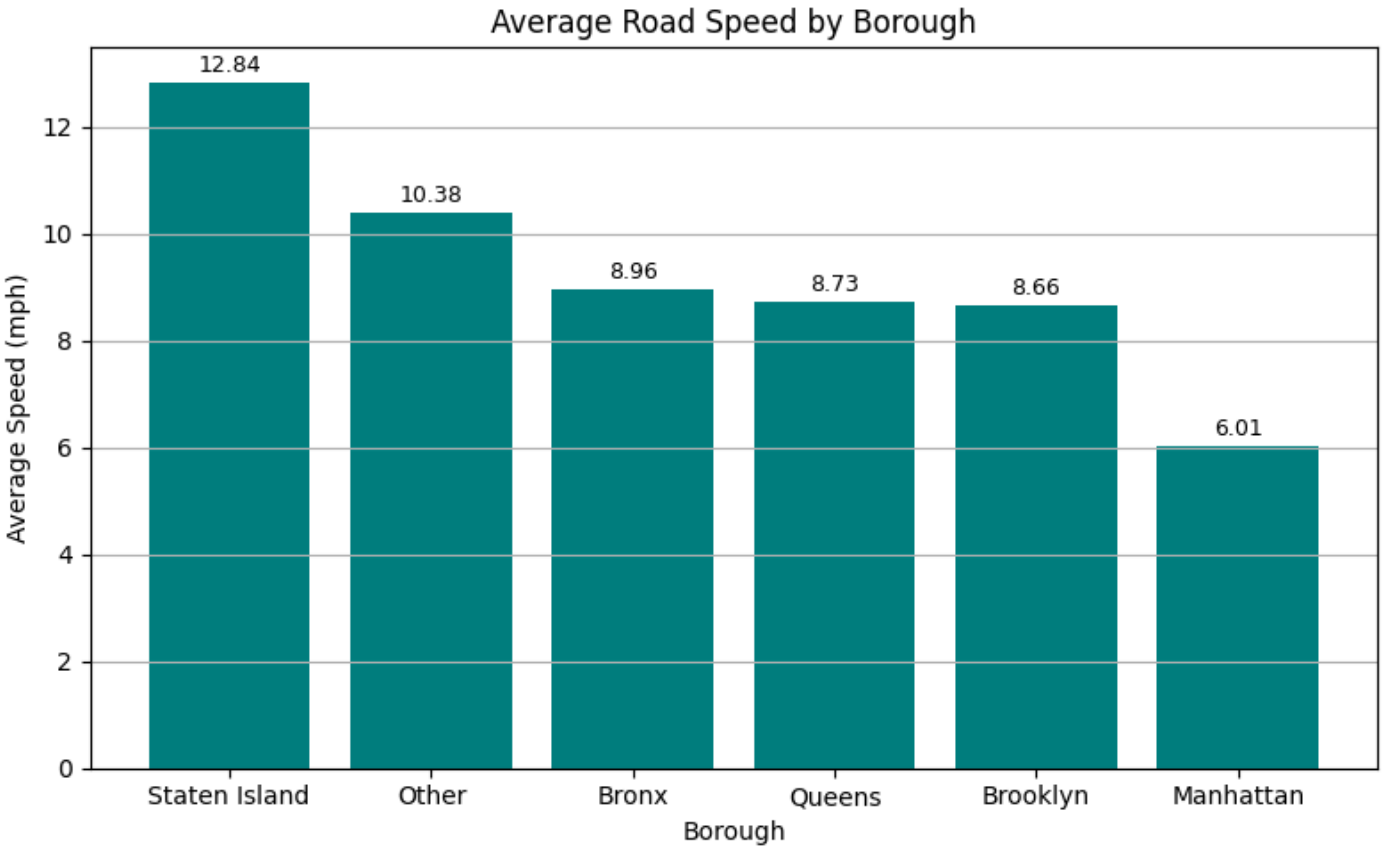
3.4 Summary of Dataset Characteristics

The final cleaned dataset is a structured and analysis-ready compilation of over 1000 rows, with each row representing a timepoint-to-timepoint segment of a bus trip. These records are rich in both temporal and spatial dimensions, with fields capturing dates, hours, day-of-week tags, route metadata, and precise geolocations. Through preprocessing, key performance indicators such as average travel time, average road speed, and bus trip count were successfully converted from string formats to numeric values, enabling accurate computation and visualization.

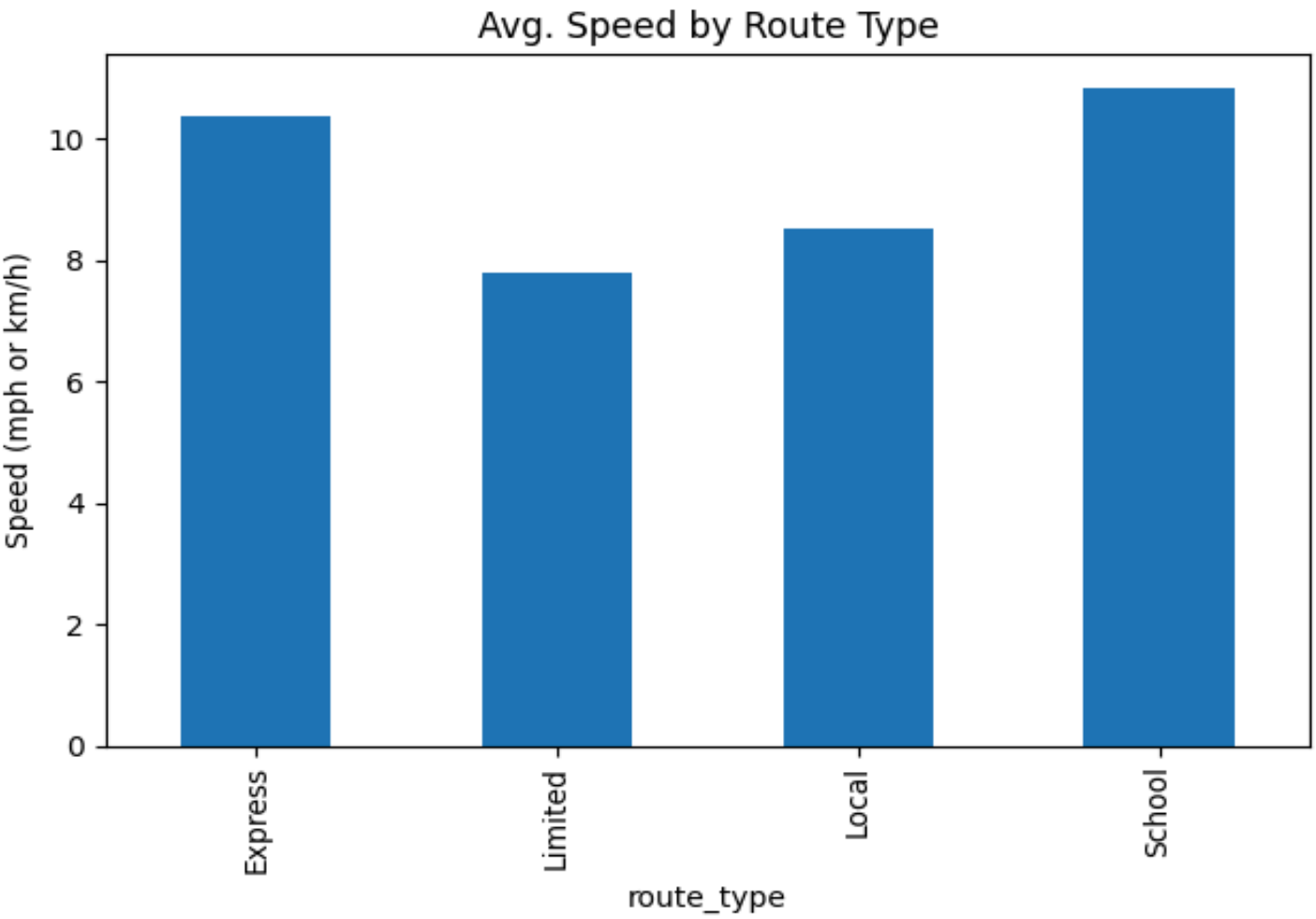
Derived columns such as IncidentFlag (used to tag low-speed/low-volume segments as "Equipment Malfunction") and hour_of_day, day_of_week, and time_of_day (for time-based analysis) were engineered to enhance the dataset's usability in both Python-based EDA and Power BI dashboards. No null values remain in the final dataset, and all key fields have been validated for consistency. Moreover, because the dataset includes both latitude and longitude coordinates for current and next stops, it is highly suitable for future use in geospatial dashboards or map overlays.

In summary, this dataset was transformed from a raw transit log into a streamlined analytical asset capable of powering operational monitoring tools. It supports aggregation across boroughs, comparison of route performance, incident flagging, and trend analysis by hour or day — all essential for real-time transit decision-making. This makes it ideal for both business intelligence applications and machine learning-based predictive systems.

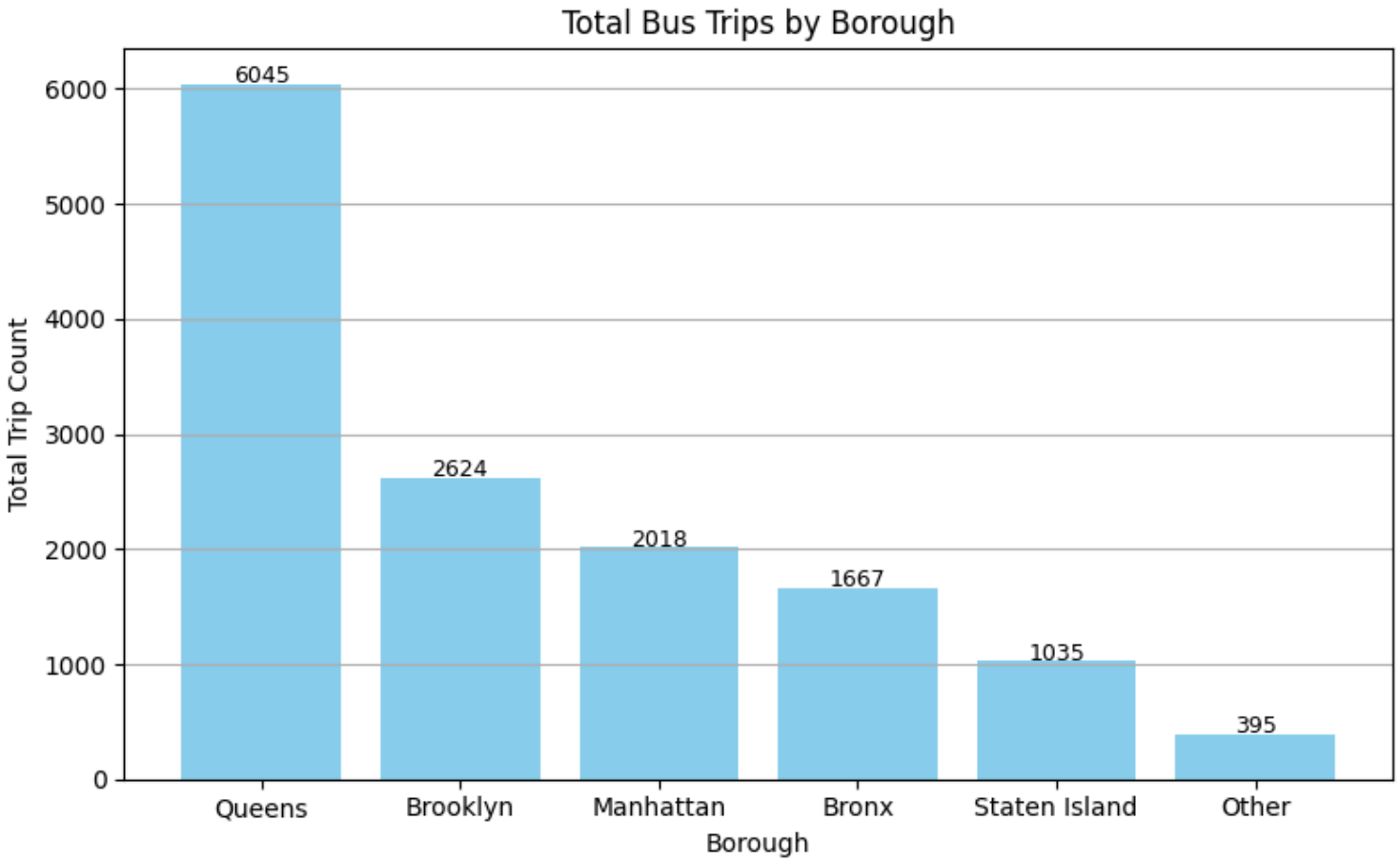
4. EXPLORATORY DATA ANALYSIS



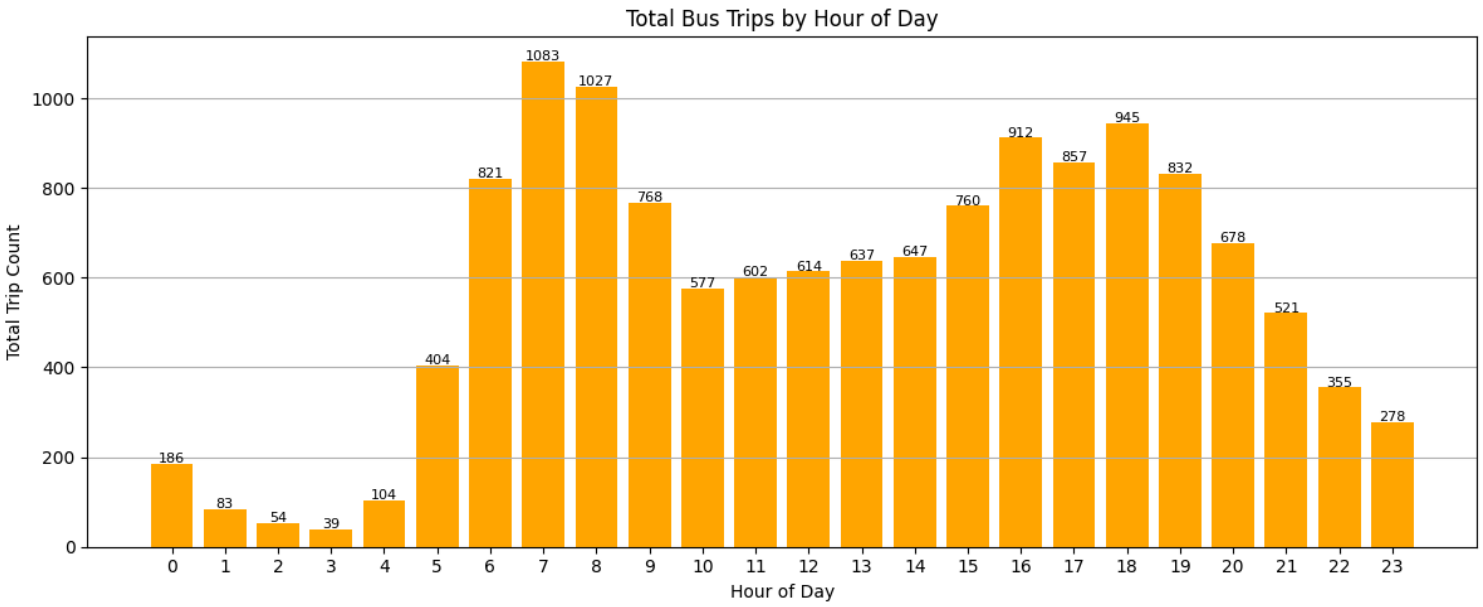
This bar chart presents the average speed (in mph) of buses across different boroughs. Staten Island leads with the highest average road speed of 12.84 mph, indicating smoother traffic flow or fewer congestion issues. On the other hand, Manhattan records the lowest average speed at 6.01 mph, highlighting significant delays likely due to high traffic density and frequent stops. Other boroughs like Bronx, Brooklyn, and Queens have moderately consistent average speeds ranging between 8.6 to 9 mph. These insights help in understanding performance variability due to geographic location and can guide infrastructure and scheduling improvement.



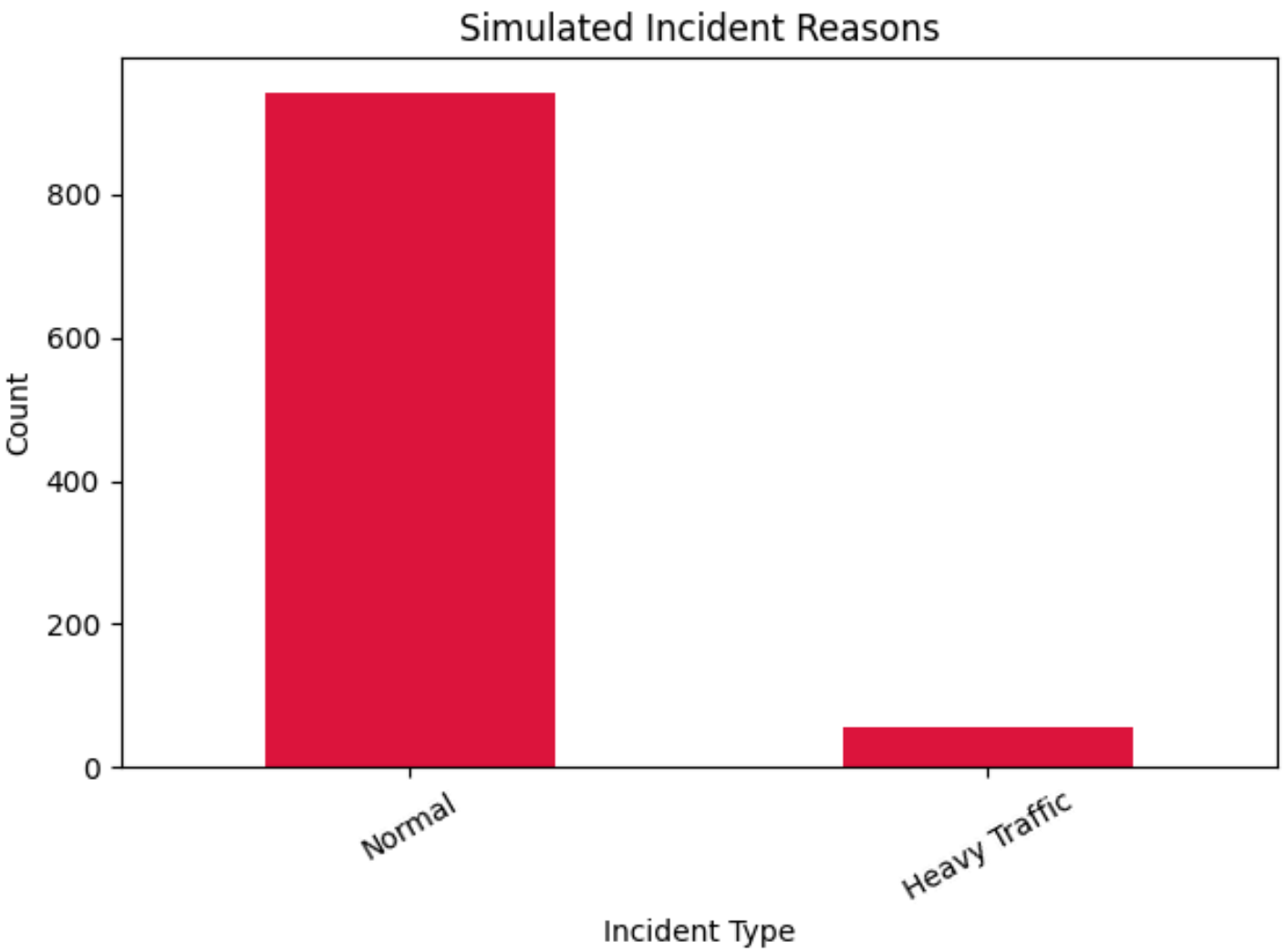
This chart illustrates the average speeds of buses grouped by route type—Express, Limited, Local, and School. Notably, School and Express buses operate at higher speeds (above 10 mph), reflecting their limited stops and dedicated routing strategies. Conversely, Limited and Local routes report lower average speeds, which can be attributed to more frequent stops and urban routing. This differentiation is critical for transit planning and suggests potential for optimizing route types based on performance benchmarks.



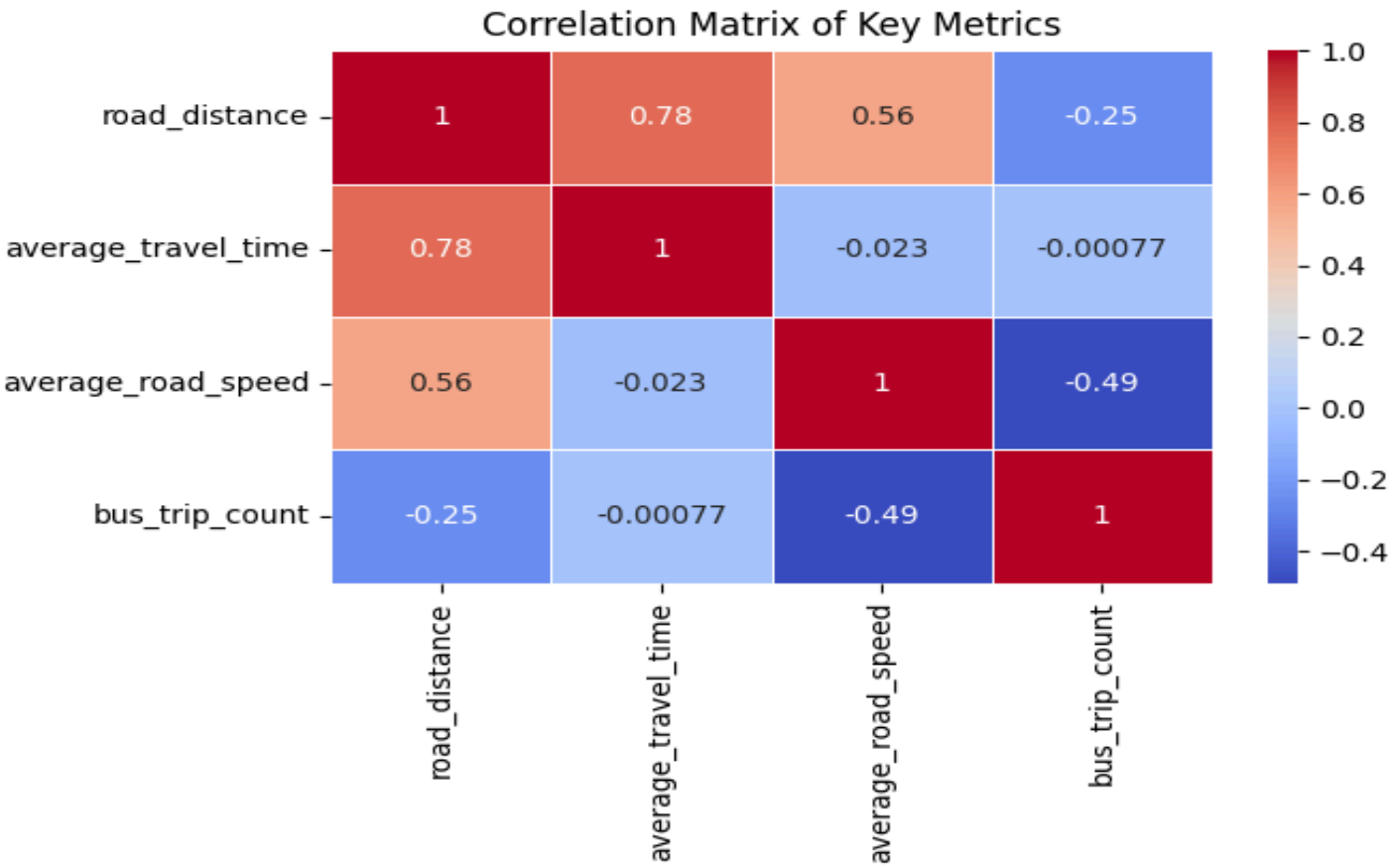
This bar chart displays the total number of bus trips recorded across boroughs. Queens has the highest trip count (6045), emphasizing its role as a central hub for bus activity. Following are Brooklyn and Manhattan with 2624 and 2018 trips, respectively. Staten Island and the 'Other' category have the least number of trips, potentially indicating less coverage or demand. This analysis highlights operational load by region and is useful for resource allocation and route planning.



This hourly distribution chart captures the fluctuation of bus trip counts throughout the day. Clear peaks are observed at 7 AM and 8 AM, with over 1000 trips, indicating morning rush hours. A secondary rise occurs around 4 PM to 6 PM, corresponding to evening commute periods. Late-night and early morning hours have minimal activity. Understanding these patterns helps in adjusting fleet deployment and driver scheduling to meet peak demand efficiently.



This visualization categorizes the reasons behind incidents. The majority are labeled as "Normal", indicating routine operations. A small portion is marked as "Heavy Traffic", which could affect average travel time and speed. While this is based on simulated labels (due to real incident details possibly being absent), it still highlights the importance of understanding and tracking disruptions to improve service reliability.



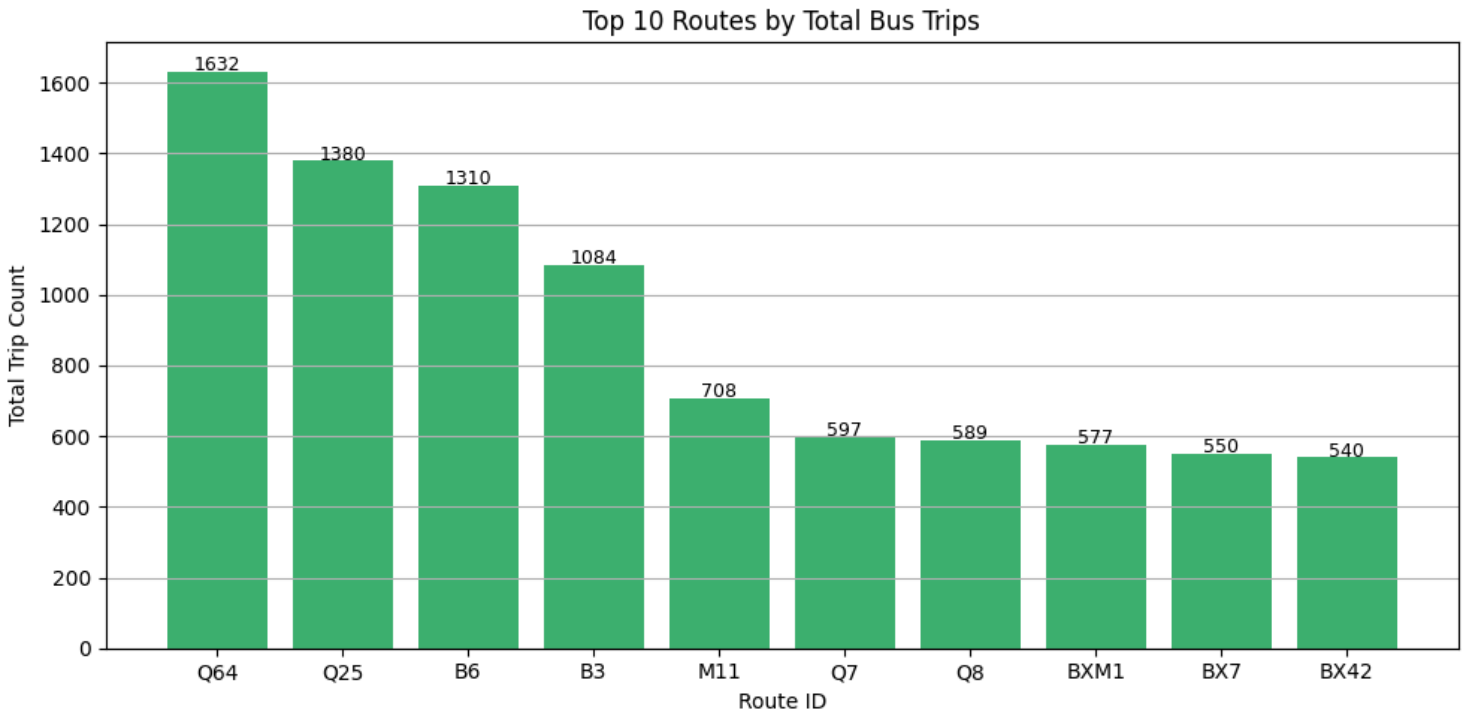
The heatmap displays Pearson correlation values between various continuous metrics:

- **Road distance and travel time** show a strong positive correlation (**0.78**), as expected.

- **Bus trip count** is **negatively correlated with average speed (-0.49)** and **distance (-0.25)**.

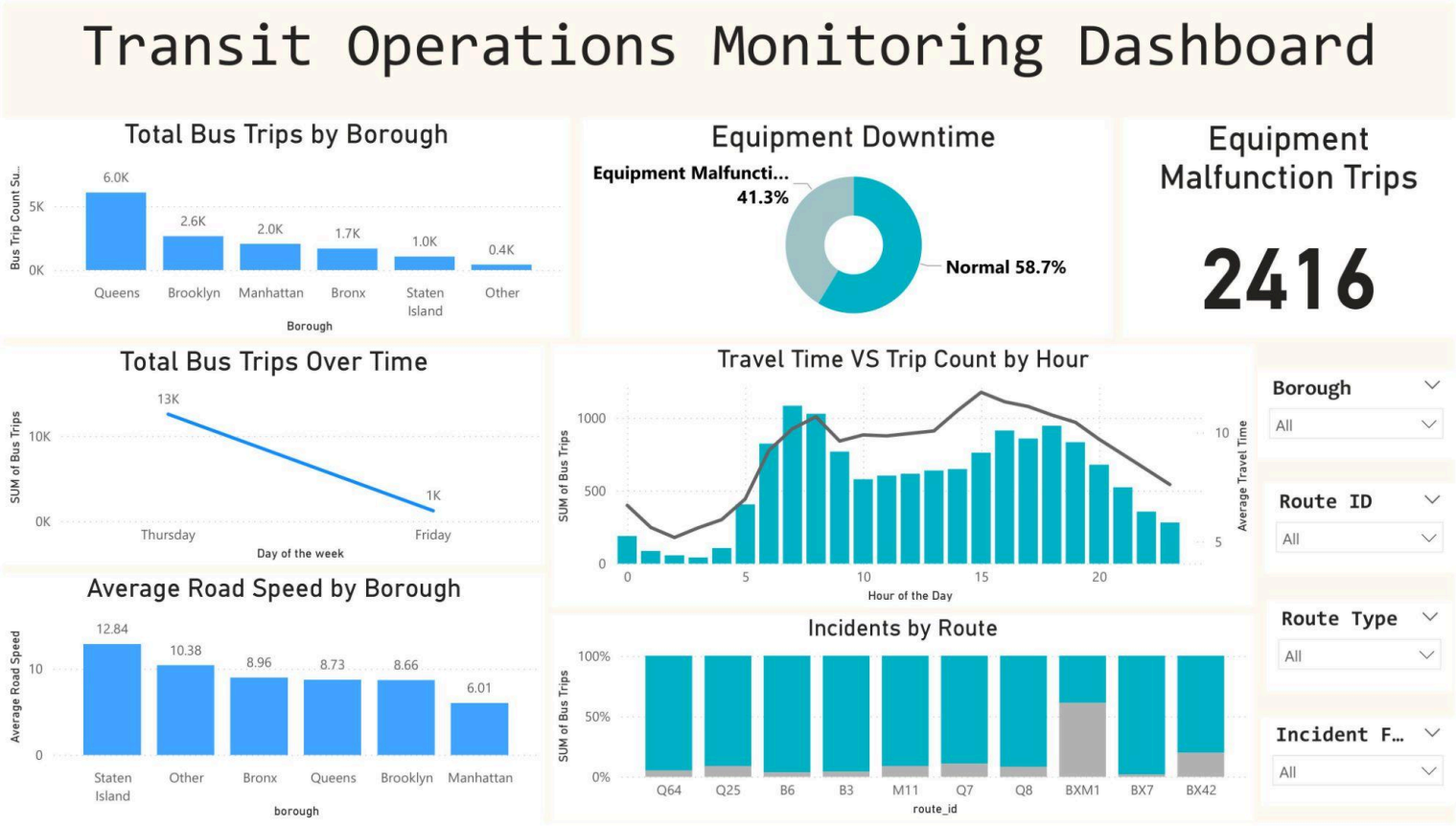
- **Average travel time and average speed** are weakly correlated, which may suggest delays are not solely distance-driven.

This matrix provides insight into how operational variables interact and which features may influence performance outcomes.



The chart shows the most utilized routes in terms of trip frequency. Route Q64 tops the list with 1632 trips, followed by Q25, B6, and B3, each having well over 1000 trips. This ranking allows transit managers to identify high-traffic routes requiring more frequent monitoring, maintenance, or optimization efforts. It also serves as a basis for identifying which routes may benefit from additional buses or driver shifts.

5. VISUALIZATIONS / DASHBOARD



The Power BI dashboard consists of the following key visualizations:

- **Bar Chart**: Total Bus Trips by Borough – highlights geographic trip distribution.
- **Pie Chart**: Equipment Downtime – shows proportion of trips affected by malfunctions.
- **Line + Bar Combo**: Travel Time vs Trip Count by Hour – dual-axis chart comparing volume and efficiency.
- **Line Chart**: Total Trips Over Time – temporal trend across days.

- **Box Plot**: Travel Time Distribution – highlights variance and outliers.
- **Stacked Bar**: Incidents by Route – shows breakdown of trip types by route and incident status.
- **Dropdown Slicers**: Enable filtering by Borough, Route Type, IncidentFlag, and Date.

The dashboard allows MTA staff to perform in-depth monitoring in real time and drill down into specific issues or time periods.

6. CONCLUSIONS

The Transit Operations Monitoring Dashboard offers a data-driven lens into the real-world performance of New York City's MTA bus system, providing comprehensive insights that can significantly enhance operational efficiency, maintenance scheduling, and commuter satisfaction. Through rigorous preprocessing and exploratory data analysis of timepoint-level data, this project uncovered key temporal and geographical patterns. Notably, borough-level disparities such as Staten Island's highest average road speed and Manhattan's lowest average speed revealed the uneven impact of traffic congestion and stop frequency on transit flow. Similarly, the spike in bus trip volumes during morning (7–8 AM) and evening (4–6 PM) hours reflected the pronounced influence of commuter rush periods on trip density and resource allocation.

The integration of multiple metrics—such as average travel time, equipment malfunction rate, and incident frequency—into a dynamic Power BI dashboard allowed for multi-dimensional filtering and real-time decision-making. Stakeholders can now isolate high-priority issues like malfunction-prone routes, analyze travel patterns by hour and day of the week, and examine road speed variations across boroughs. The dual-axis analysis of travel time versus trip volume revealed critical insights: higher trip volumes are often associated with dips in travel speed and surges in travel time, suggesting congestion-related delays. Moreover, the correlation matrix provided statistical validation of operational relationships, showing that factors like road distance and travel time are positively correlated, while trip count and average speed are inversely related—pointing to systemic slowdowns during high-demand periods.

From a business perspective, the findings empower the MTA to make strategic decisions regarding route optimization, driver scheduling, and maintenance prioritization. The identification of key bottlenecks, such as low-performing routes or boroughs with frequent breakdowns, enables targeted interventions that can improve reliability and reduce rider dissatisfaction. On the technical front, the success of the dashboard illustrates the power of combining structured data modeling, Python-based preprocessing, and intuitive visual analytics to tackle complex public infrastructure challenges. This end-to-end solution transforms raw transit logs into actionable intelligence, laying the groundwork for future scalability into predictive analytics, geospatial optimization, and broader smart city applications.