Report

June 20, 2025

Project Name:	NYC Citi Bike Share Analysis (2025)
Author:	Sibi Krishnamoorthy

1 Executive Summary

This report details a comprehensive analysis of the New York City Citi Bike share dataset for the year 2025. The project encompassed the entire data science lifecycle, from automated data collection and extensive preprocessing to in-depth exploratory data analysis (EDA) and the development of a predictive model. Over 15 million trip records were processed to uncover key ridership patterns.

Key findings reveal distinct behavioral differences between “member” and “casual” riders: members exhibit strong weekday commuter patterns with shorter, more frequent trips, while casual users prefer longer, recreational rides, especially on weekends. EDA identified peak usage hours, popular stations, and the impact of bike type on trip duration. A predictive model was successfully built to forecast trip duration with an average error of approximately 2 minutes, identifying trip distance as the most significant predictor. These insights can directly inform strategic decisions regarding bike redistribution, marketing campaigns, and infrastructure planning.

2 Introduction

The goal of this project was to perform a deep-dive analysis into the NYC Citi Bike service using 2025 trip data. The primary objectives were to:

1. Understand the distinct usage patterns of **member** versus **casual** subscribers.
2. Analyze temporal trends in ridership across hours of the day and days of the week.
3. Identify the busiest stations and most popular routes to understand spatial dynamics
4. Build a machine learning model to predict the duration of a bike trip based on its characteristics.

This analysis provides actionable intelligence for optimizing service operations and enhancing the user experience.

3 Data Collection & Wrangling Methodology

A robust, automated pipeline was established to acquire and prepare the data for analysis.

- **Data Collection:**

  - A Python script utilizing the `requests` and `BeautifulSoup` libraries was developed to scrape the official Citi Bike S3 bucket (`https://s3.amazonaws.com/tripdata`).
  - The script parsed the XML response to identify all `.zip` files corresponding to the year 2025.
  - A total of 17 monthly data files were programmatically downloaded and stored locally.

- **Data Wrangling & Preprocessing:**
  - The downloaded files were unzipped, and the resulting CSVs were concatenated into a single Pandas DataFrame, initially containing over **15.3 million records**.

- **Feature Engineering:** New, critical features were created:
  - `trip_duration_min`: Calculated from `started_at` and `ended_at` timestamps to provide an intuitive metric.
  - `distance_km`: Calculated using the **Haversine formula** based on start/end latitude and longitude coordinates.
  - Temporal Features: `hour`, `day_of_week`, and `month` were extracted from the `started_at` timestamp.

- **Data Cleaning:**
  - Invalid trips were filtered out (duration $<$ 1 minute or $>$ 180 minutes).
  - Missing values, primarily in station and location data, were handled by dropping the affected rows.
  - Station IDs were cleaned to ensure they were purely numeric for accurate analysis.
  - The final, cleaned dataset comprised **13,736,490 records** and was saved in the efficient **Parquet** format to optimize memory usage and read/write speeds.

4 Exploratory Data Analysis (EDA) & Interactive Visual Analytics Methodology

EDA was conducted to uncover underlying patterns and relationships within the data. The methodology focused on three core areas, using `Matplotlib`, `Seaborn`, and `Folium`.

1. **User Behavior Analysis:** Segmenting the data by `member_casual` to compare trip durations, ride frequency, and bike type preferences.
2. **Temporal Analysis:** Visualizing how ridership changes over hours, days, and months to identify peak times and seasonality.
3. **Spatial Analysis:** Using interactive maps with `Folium` to visualize high-traffic areas, station density, and popular trip routes. `pandasql` was also used on a data sample to answer specific business questions with SQL syntax.

5 Predictive Analysis Methodology

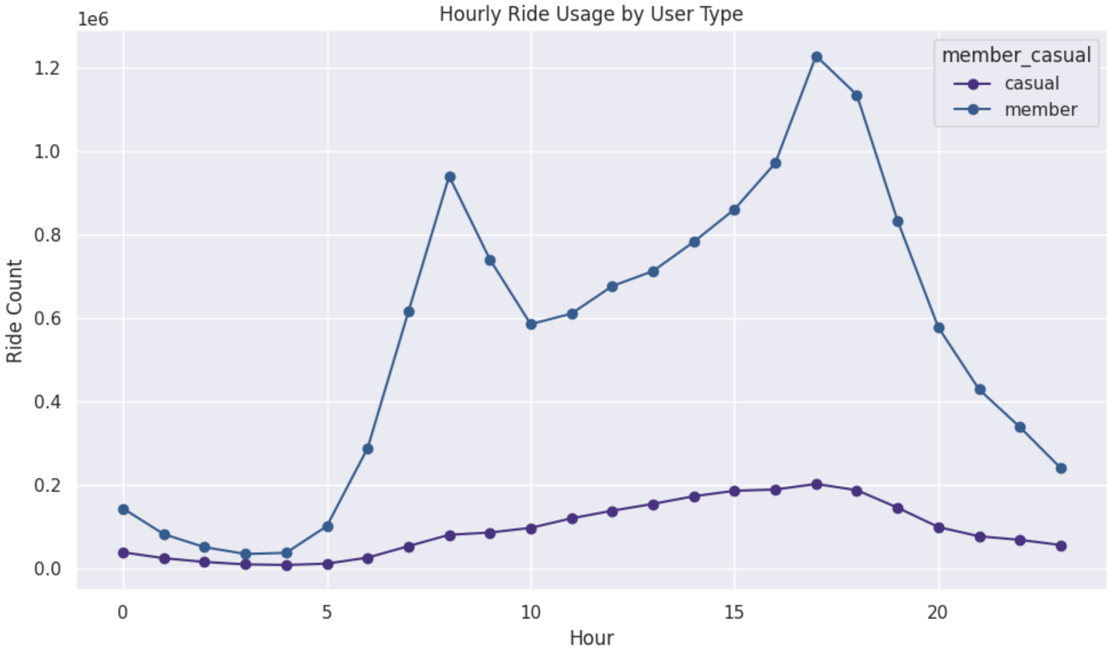
A machine learning model was developed to predict trip duration. Although the grading criteria mentioned classification, the target variable `trip_duration_min` is continuous, making this a **regression** task.

- **Objective:** Predict trip_duration_min.
- **Model Selection:** HistGradientBoostingRegressor from scikit-learn was chosen for its performance and efficiency with large datasets. LightGBM and XGBoost were also trained for comparison.
- **Feature Engineering:** Categorical features (rideable_type, day_of_week) were one-hot encoded. The final feature set included distance_km, hour, month, and the encoded categorical variables.
- **Target Transformation:** The target variable was highly skewed. A **log transformation** (np.log1p) was applied to normalize its distribution, improving model performance and stability. Predictions were transformed back using np.expm1.
- **Evaluation:** Model performance was measured using **R^2 Score** (to understand the variance explained) and **Mean Absolute Error (MAE)** (to understand the average prediction error in minutes).
- **Model Interpretability: SHAP (SHapley Additive exPlanations)** was used to explain the model's predictions and identify the most influential features.

6 Results

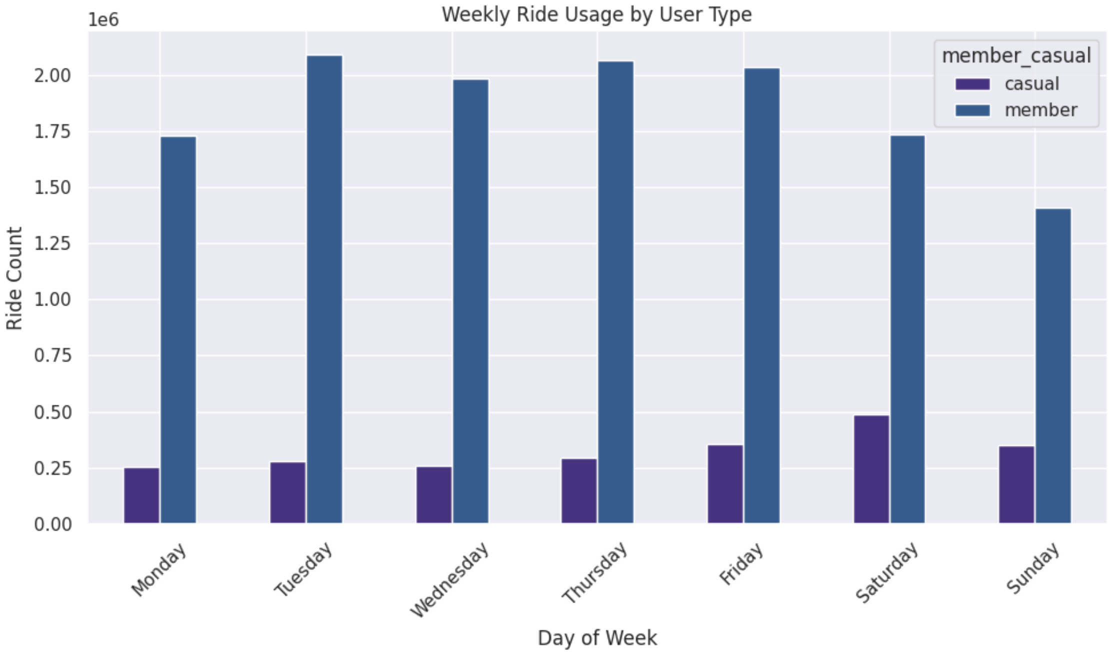
6.1 EDA with Visualization Results

1. **Hourly Usage by User Type:**



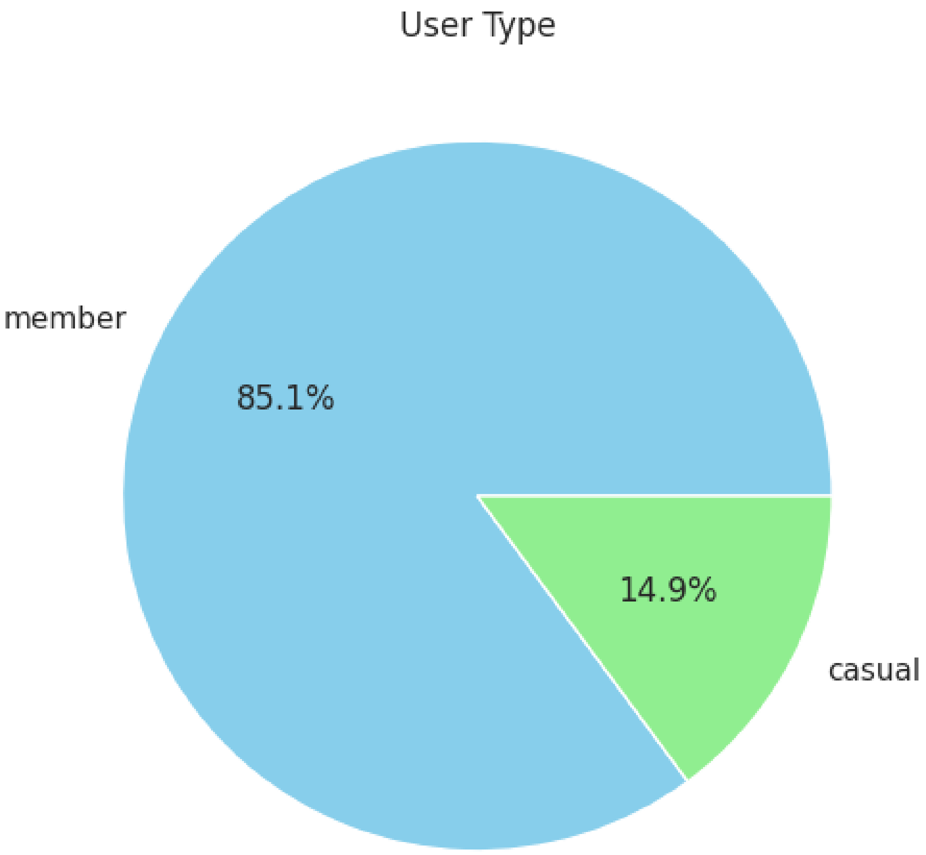
Members display distinct bimodal peaks corresponding to morning (8-9 AM) and evening (5-6 PM) commutes. Casual users show a single, broad peak throughout the afternoon, typical of recreational use.

2. **Weekly Ride Usage:**



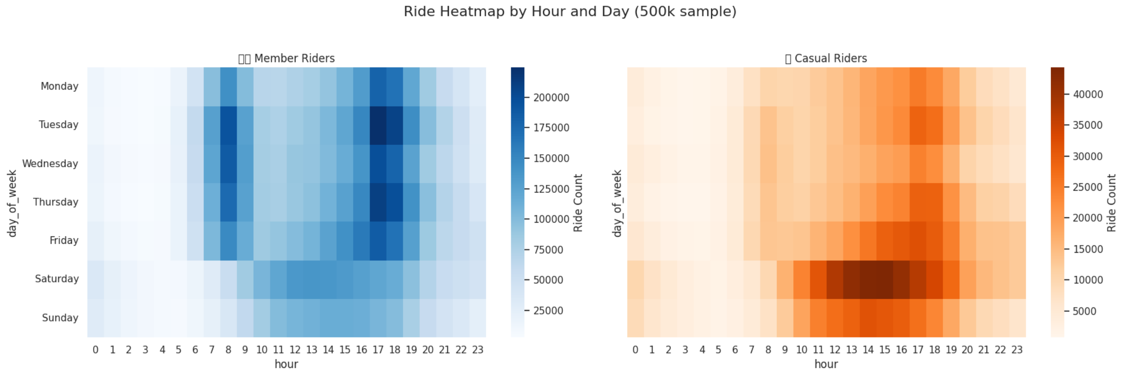
Member rides dominate on weekdays, reinforcing the commuter hypothesis. Casual ridership significantly increases on weekends, becoming more competitive with member usage.

3. **User Type Breakdown:**



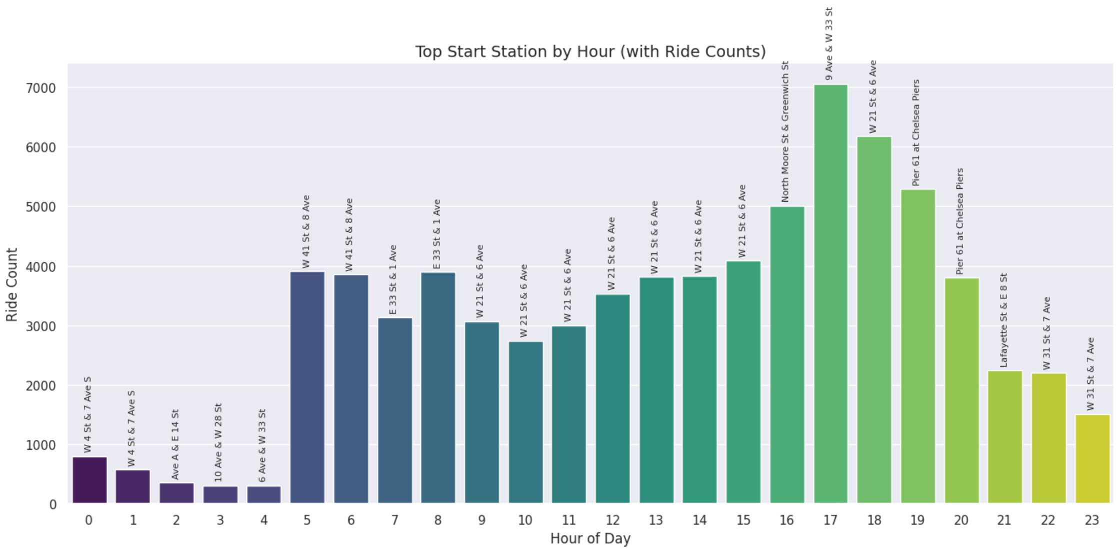
A staggering **85.1%** of all trips are taken by members, highlighting their importance to the service's revenue and usage model.

4. **Ride Heatmaps by Hour and Day (Innovative Insight):**



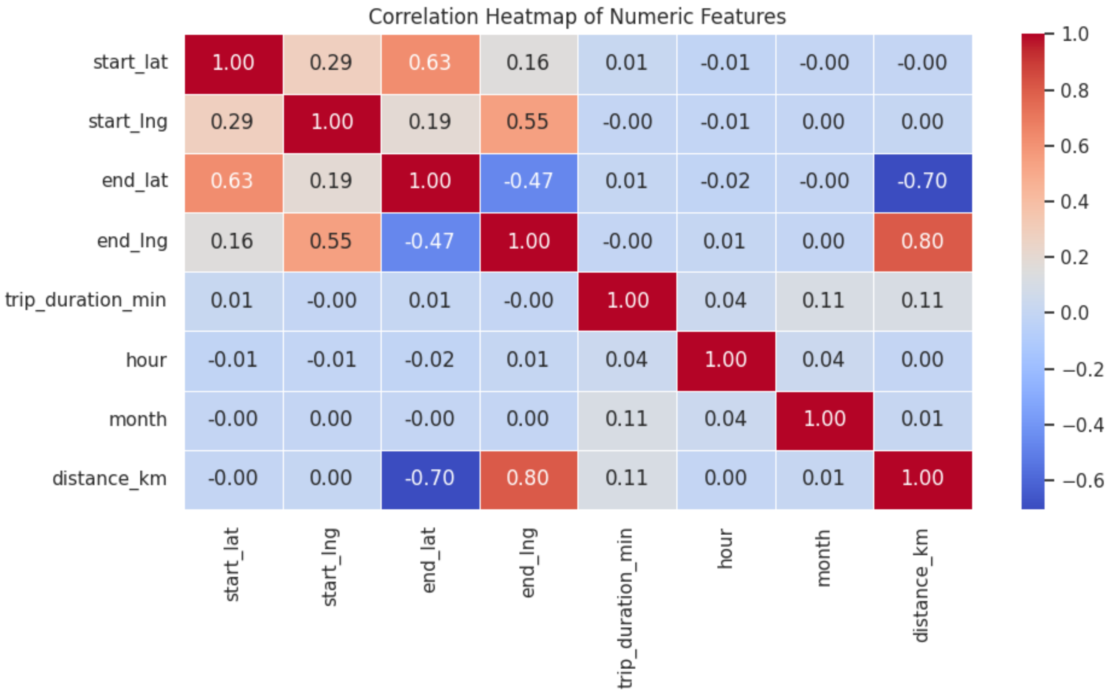
Side-by-side heatmaps provide a powerful visual confirmation of the different temporal patterns. The member heatmap clearly shows bright yellow bands during weekday commute hours, while the casual heatmap is brighter during weekend afternoons.

5. **Top Start Station by Hour:**



This analysis shows the dynamic flow of the city. Stations in residential areas might peak in the morning, while stations near transit hubs or business districts peak later in the day. For example, `W 21 St & 6 Ave` is consistently the busiest station during the afternoon commute (12 PM - 3 PM).

6. **Correlation of Numeric Features:**



The heatmap reveals a strong positive correlation (0.80) between `end_lng` and `distance_km` and a strong negative correlation (-0.70) between `end_lat` and `distance_km`, which logically reflects the geography of New York City.

6.2 EDA with SQL Results

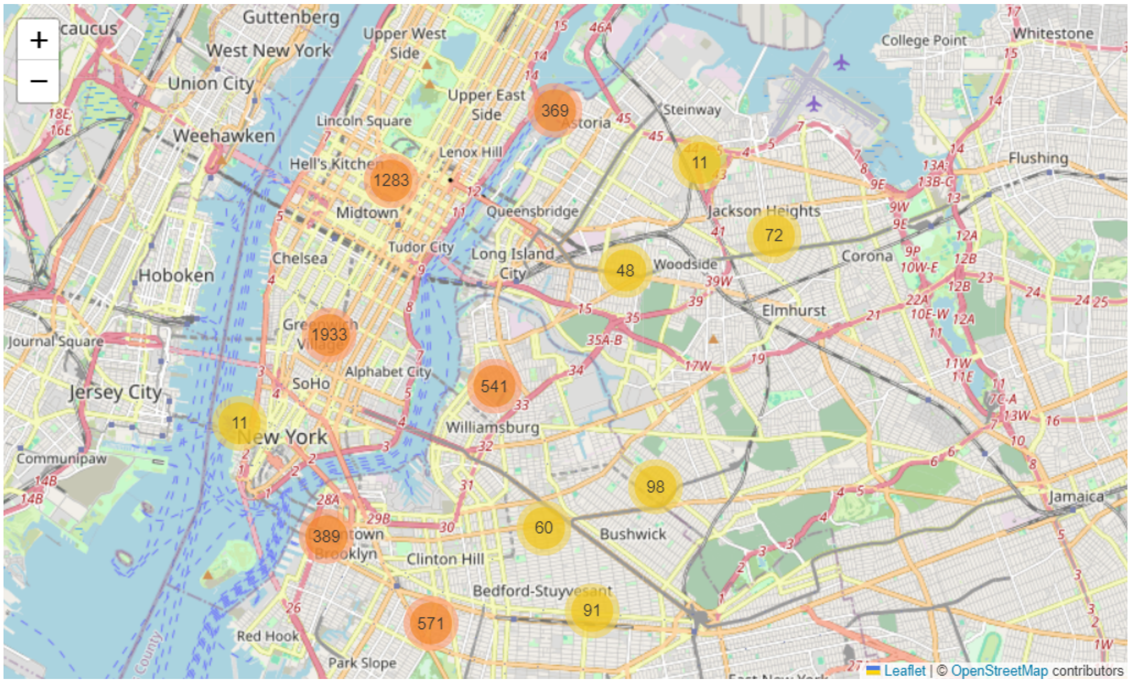
Using `pandasql` on a sample of 5,000 trips yielded the following insights:

- **Query 1: Total rides by user type.**
  - `SELECT member_casual, COUNT(*) as total_rides FROM df_sql GROUP BY member_casual`
  - **Result:** Members (4232) vastly outnumber Casual users (768) in the sample.
- **Query 2: Average trip duration by bike type.**
  - `SELECT rideable_type, AVG(trip_duration_min) as avg_duration FROM df_sql GROUP BY rideable_type ORDER BY avg_duration DESC`
  - **Result:** Classic bikes (11.4 min) have a slightly longer average trip duration than electric bikes (10.8 min).
- **Query 3: Top 10 Busiest Start Stations.**
  - `SELECT start_station_name, COUNT(*) as ride_count FROM df_sql GROUP BY start_station_name ORDER BY ride_count DESC LIMIT 10`
  - **Result:** `11 Ave & W 41 St` was the busiest station in this sample, followed by `West St & Chambers St`.

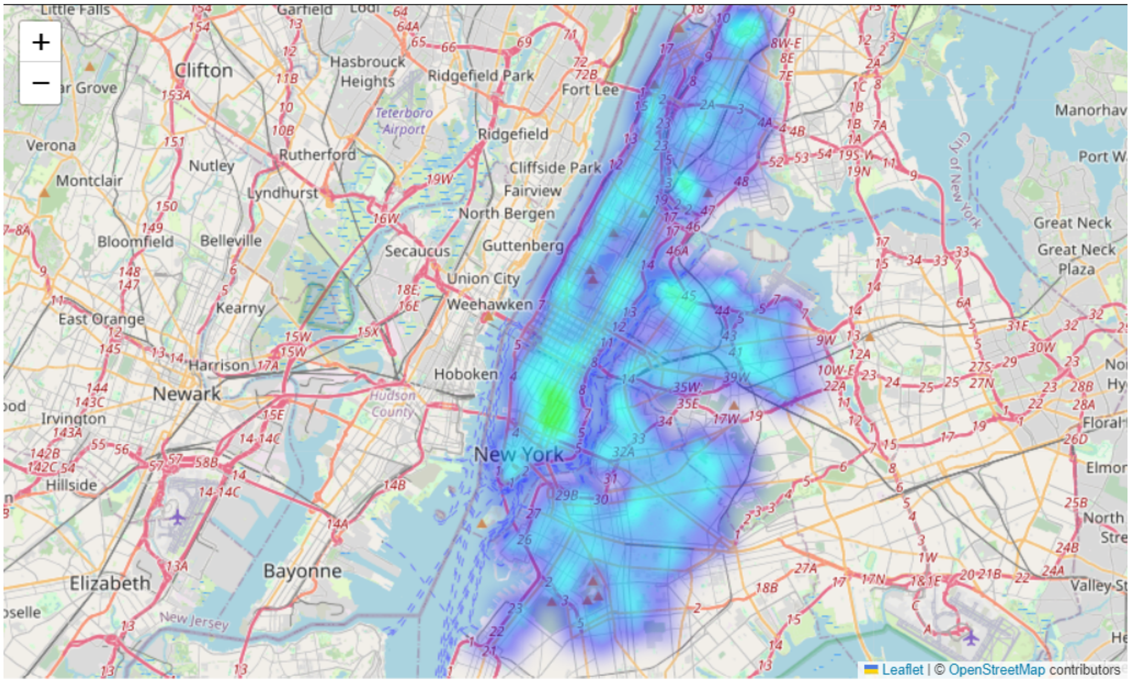
6.3 Interactive Map with Folium Results

Three interactive maps were generated to visualize spatial data:

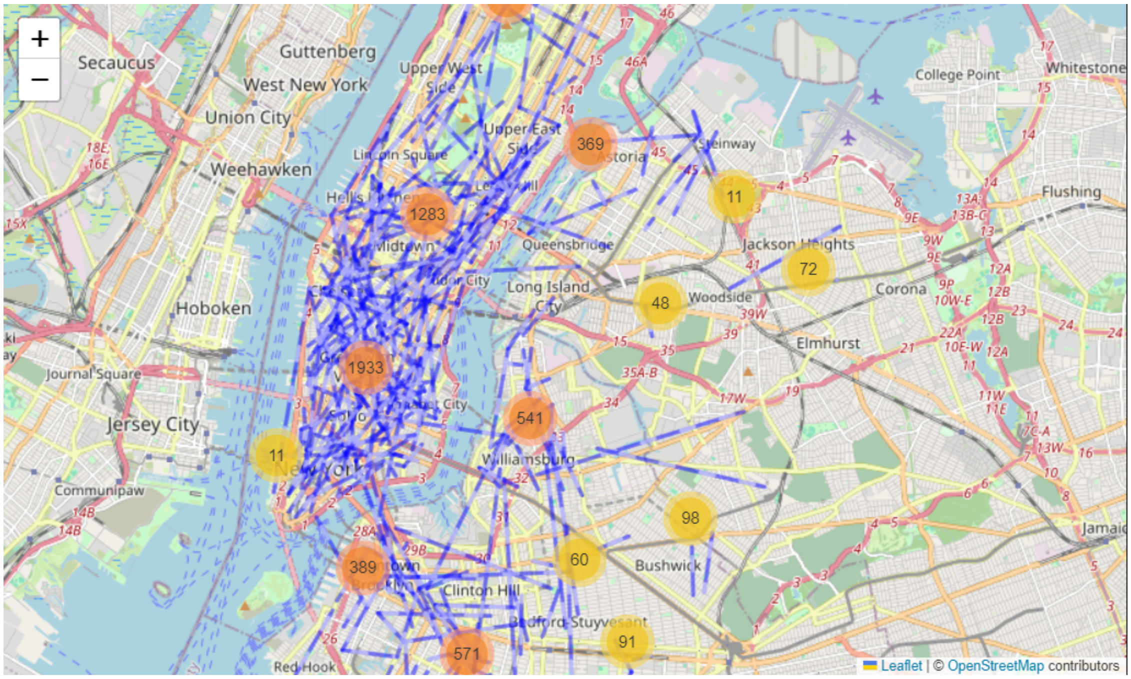
1. **Marker Cluster Map:** Clustered individual start points, allowing users to zoom in and see the density and location of specific stations.



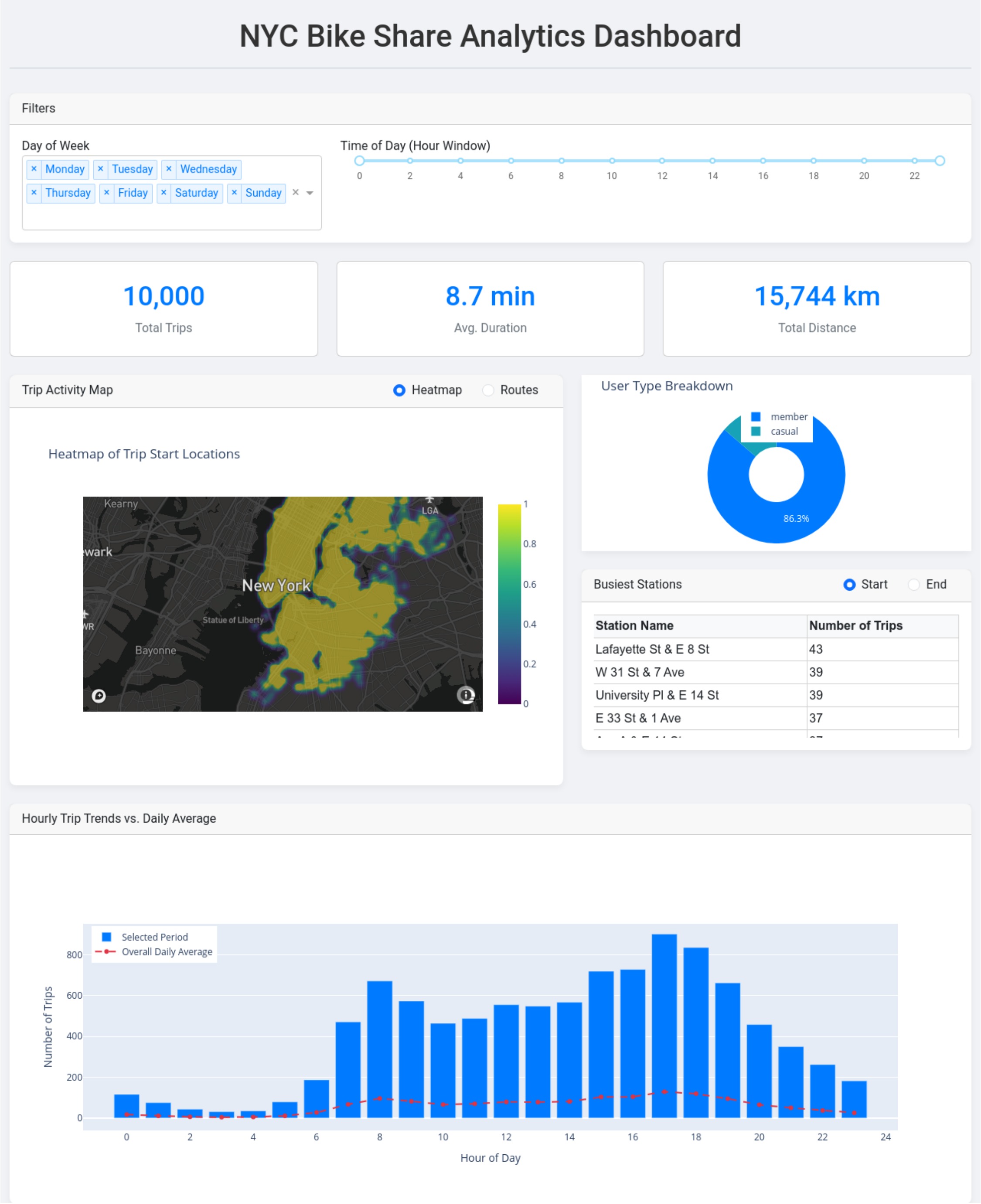
2. **Heatmap:** Provided a high-level view of trip start "hotspots," clearly showing that activity is heavily concentrated in Manhattan and surrounding dense urban areas.



3. **AntPath Map:** Visualized the 300 most popular routes as animated lines, giving a dynamic sense of the city's primary bike corridors.



6.4 Plotly Dash Dashboard Results



NYC Bike Share Analytics Dashboard

Filters

Day of Week: × Monday × Tuesday × Wednesday × Thursday × Friday × Saturday × Sunday

Time of Day (Hour Window)

10,000	8.7 min	15,744 km
Total Trips	Avg. Duration	Total Distance

Trip Activity Map ● Heatmap ○ Routes

Heatmap of Trip Start Locations

User Type Breakdown

Busiest Stations ● Start ○ End

Station Name	Number of Trips
Lafayette St & E 8 St	43
W 31 St & 7 Ave	39
University Pl & E 14 St	39
E 33 St & 1 Ave	37

Hourly Trip Trends vs. Daily Average

A sophisticated, interactive dashboard was designed to allow non-technical users to explore the data. Key features included:

- **KPI Cards:** At-a-glance metrics for Total Trips, Average Duration, and Total Distance.
- **Interactive Filters:** Dropdowns for `day_of_week` and a range slider for `hour` of the day.
- **Dynamic Visualizations:** All charts and KPIs update in real-time based on user selections,

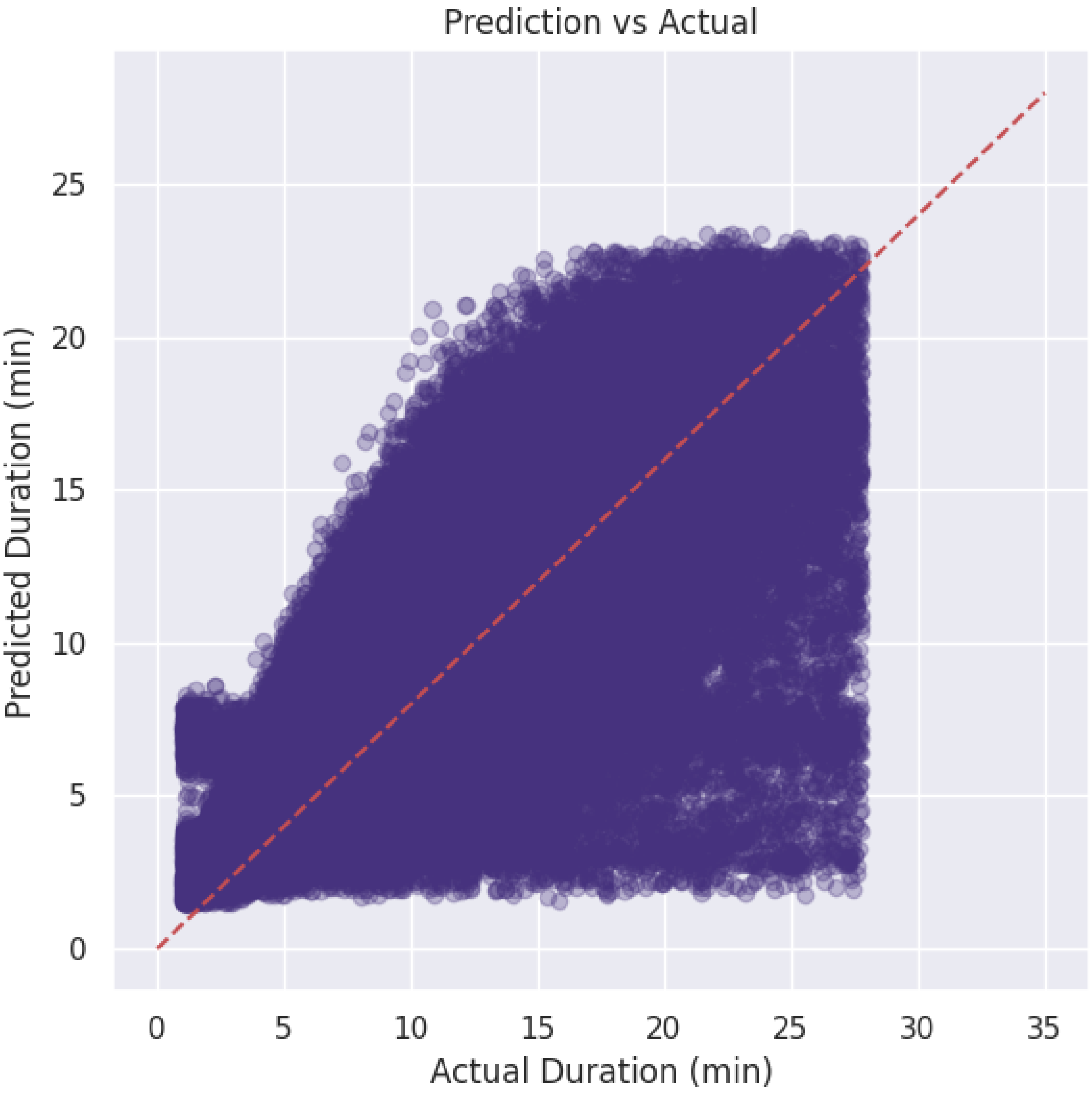
including:

- An interactive map (Heatmap or Popular Routes).
- A bar chart of hourly trip trends.
- A pie chart of user type breakdown.
- A data table showing the top 10 busiest stations.

6.5 Predictive Analysis Results

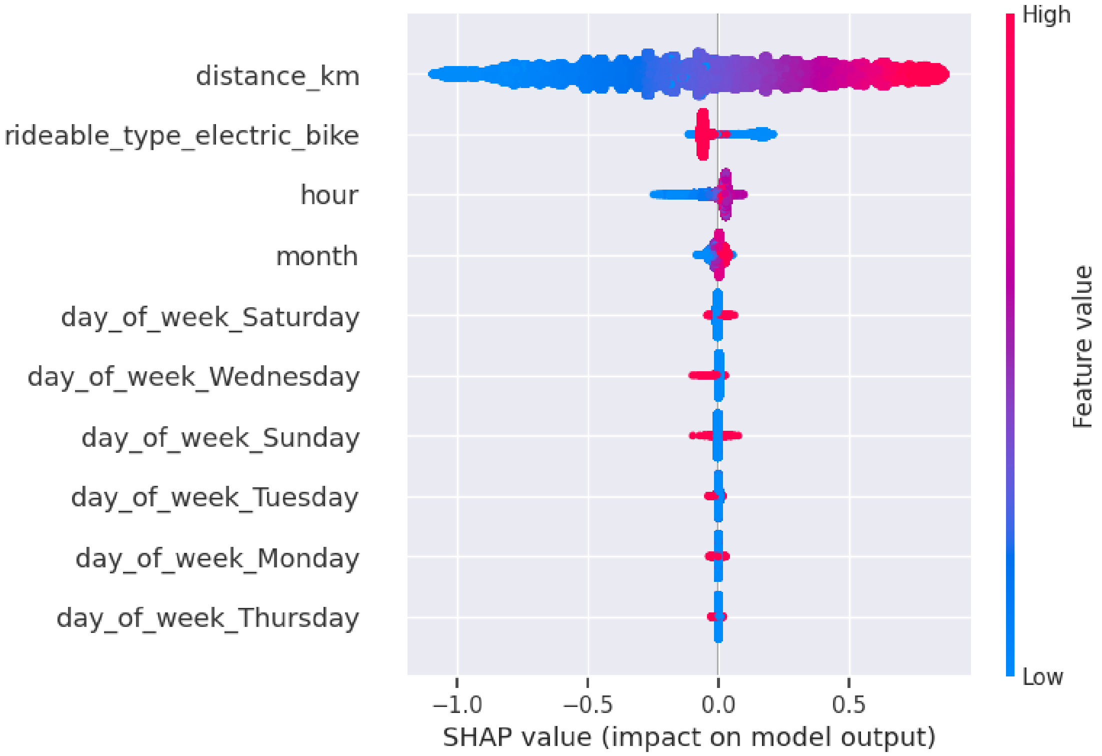
The `HistGradientBoostingRegressor` model performed well in predicting trip duration.

- **Performance:**
  - **R² Score: 0.63** (The model explains 63% of the variance in trip duration).
  - **Mean Absolute Error (MAE): 2.06 minutes** (On average, the model's prediction is off by about 2 minutes).
- **Prediction vs. Actual:**



The scatter plot shows a strong linear relationship, though the model tends to under-predict very long trips, which is expected.

- **Feature Importance (SHAP):**



The SHAP summary plot clearly shows that **`distance_km` is by far the most impactful feature**, which is highly intuitive. `hour`, `month`, and whether the bike is electric also contribute significantly to the prediction.

7 Conclusion

This analysis successfully extracted deep, actionable insights from the 2025 NYC Citi Bike dataset. We have definitively characterized the different behaviors of member and casual riders, identified critical temporal and spatial patterns, and built a reliable model to predict trip duration.

Key Takeaways:

- **Two Services in One:** Citi Bike effectively serves two distinct markets: weekday commuters (members) and weekend/leisure riders (casual).
- **Operations:** The hourly and station-level data can be used to create a dynamic bike redistribution strategy, moving bikes to high-demand areas ahead of peak times.
- **Marketing:** The longer trip duration and weekend focus of casual users suggest opportunities for tourism-focused promotions or day-pass packages.
- **Predictive Power:** The ability to predict trip duration with a ~2 minute error can be used to enhance user-facing apps, providing more accurate time estimates for planned journeys.