Analyzing NYC CitiBike Usage Data: Trends, Patterns, and Insights

Abstract - This analysis focuses on usage patterns of NYC Citi Bikes in the year 2023 and aims to uncover insights into patterns, trends, and user behavior. With the rise of urbanization and urban migration, the need for sustainable means of transport has never been more important. Bike sharing services have become a popular option in major cities around the world such as London, Paris, Amsterdam, Japan, and so on. As such, this paper aims to conduct an analysis of user behavior in order to identify popularity, strengths and potential scope for improvement within the current bike share network.

I. BACKGROUND

Citi Bike, New York City's most popular bike sharing program, was launched in 2013 as an affordable, sustainable alternative to traditional urban transportation such as cars, buses, and the subway system. It initially began with 5000 bikes and 330 stations, offering annual memberships and short terms access passes for tourists. Over the years, Citi Bike gained massive success and expanded significantly, with over 2000 stations and 36,000 bikes as of November 2024, and has become the largest bike-share system in North America. Users can pick bikes from any station and drop them off at the nearest available dock, providing a hassle free option for commuters. The system was part of New York's broader strategy to reduce congestion, lower carbon emissions, and lighten traffic load on the streets (New York City Department of City Planning, 2008). However, managing such a large scale system comes with operational challenges as well such as fluctuating demand patterns, even distribution of bikes across the city, customer retention, and so on. Additionally, studies have shown that density and accessibility of bike stations are the key factors in the success of such systems (Sobolevsky et al., 2018; O'Mahony & Shmoys, 2015; Fritz, 2017). This analysis looks into some of the key issues that Citi Bike may face and aims to use usage and spatial to identify scope for improvements.

II. THE DATA

The full 2023 Citi Bike usage data set was sourced from Kaggle. The information in the dataset was sourced from the official Citi Bike website, which makes its monthly system data publicly available. The dataset provides a comprehensive view of trip data, allowing for insights into spatial patterns, user behavior, and temporal trends. Key variables within the dataset include timestamps indicating the start and date and time of each trip, names of the stations where trips started and ended, latitude and longitude of said stations to allow for spatial mapping, and user type to indicate whether the individual is a member or a casual user.

In order to prepare the dataset for analysis, some cleaning and preprocessing was required.

A. Handling Missing Values

Rows with missing or blank values in the start and end station column were identified from

the dataset and eliminated to ensure the validity of the analysis.

B. Date-Time Conversion

The ‘started at’ and ‘ended at’ columns were converted from chr to POSIXct format using R’s lubridate package. This allowed for easy extraction of time based features such as hour, day, and month for analysis.

C. Trip Duration and Distance Calculation

An additional column for trip duration was added by calculating the difference between ‘started at’ and ‘ended at’. Trip distance was calculated using the coordinates and the Haversine formula.

These preprocessing steps ensured that the dataset was clean, consistent, and enhanced its analytical potential.

III. TOOLS

R is an open-source programming language used extensively for data analysis, statistical computing, and creating visualizations.

IV. METHODOLOGY

This analysis followed a structured methodology designed to uncover patterns, trends, and relationships within the Citi Bike data. It comprised of five main steps: data preprocessing, exploratory data analysis, clustering, regression modelling, and result visualization.

A. Exploratory Data Analysis

An EDA was conducted to identify temporal trends and spatial patterns within the data. Results of this analysis highlighted peak usage hours and months, busiest stations, the bike routes most popular with users, and how trip distance varied with time of the day

B. Clustering

In order to group stations based on usage and geographic locations, k-means clustering was used. The input variables included the average latitude, longitude, and total trips originated from each station. Each station was assigned to a cluster based on location and the results were visualized on a to identify spatial groupings and usage patterns. The clusters promptly formed an outline of the boroughs of New York City.

C. Regression Modelling

A linear regression model was created to analyze factors influencing station popularity and usage and measured the total number of trips originating from each station. The model’s predictor variables included calculations of the average trip distance, average trip duration, and the latitude and longitude. The model was used to give insight into how spatial and operational factors influence station usage. The model was then evaluated using the R-squared values and then plotted onto a scatterplot of predicted vs actual total trips to visualize the accuracy.

D. Result Visualization

To communicate the findings effectively, the results were visualized through various plots. Bar plots were used to visualize daily and

monthly usage patterns, cluster maps of the stations provided a visual for spatial trends.

V. RESULTS

A. *Temporal Trends: Usage by hour, day, and month*

The analysis of trip frequency across different times of the day shows a few clear and predictable peaks. The highest usage of the day happens between 4pm and 6pm, with the maximum usage at 5pm. This pattern reflects the rush hour during which people are likely commuting home after work. Rush hour traffic may also make bikes a more attractive option. 8am also sees a noticeable spike in usage, again likely due to people commuting to work. 4am records the lowest usage, indicating fewer riders in the early hours of the day. The daily variation in trip distance also lines up with these patterns: with fewer but longer rides during the late night and early morning hours, potentially for leisure or less congested travel. Shorter trips are more prevalent during morning and evening commutes, suggesting that Citi Bike might often be used for quick, short distance transport.

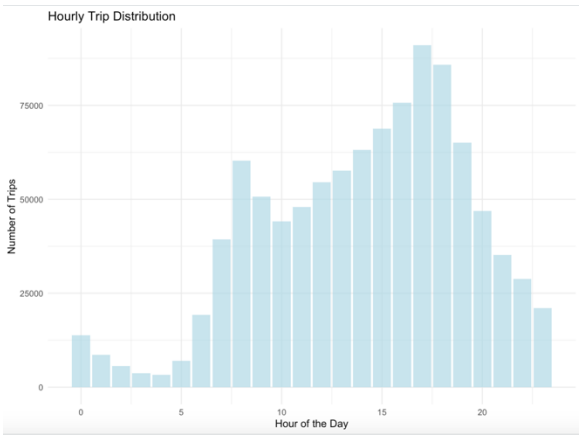

Fig. 1. “Plot Showing Temporal Trends of Citi Bike Usage by Hour.”

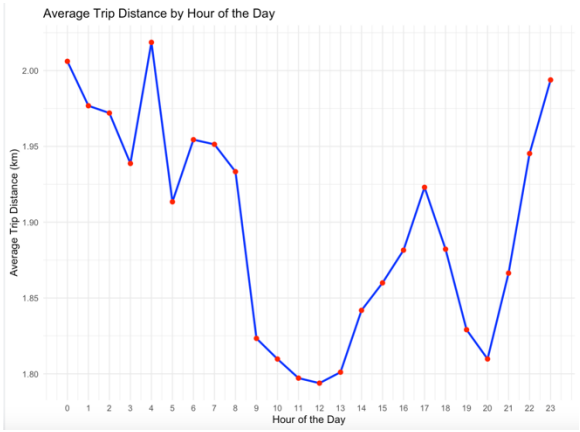

Fig. 2. “Line Graph of Average Trip Distance at Different Times of the Day.”

Analysis of trip frequency throughout the week shows that Wednesdays are the busiest days of the week, followed by Thursdays and Tuesdays. Sundays show the least trip frequency, reflecting a dip in activity, potentially indicating to greater reliance on other forms of transport on weekends or changes in commuting patterns to allow for more leisure. The high usage on Wednesdays and Thursdays may suggest that Citi Bike is used largely for work commute.

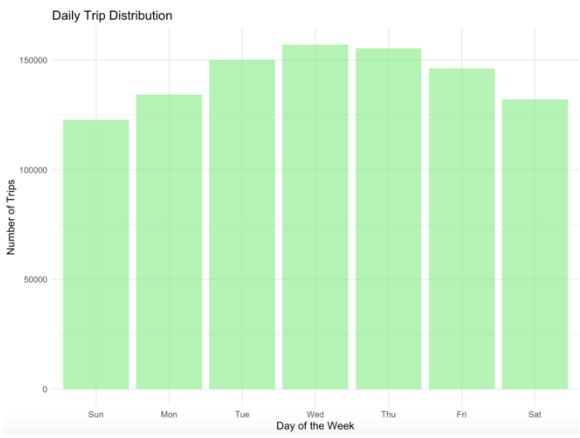

Fig. 3. “Plot Showing Temporal Trends of Citi Bike Usage by Day of the Week.”

When it comes to annual frequency, August sees the highest activity, followed by October and then July. These months coincide with warmer and more pleasant weather, along with increased tourist activity, which may boost bike-share activity. On the other hand, December, January, and February have less frequent usage, with February seeing the least activity throughout the year. This coincides with colder winter months, where cycling is not as common due to weather conditions. These seasonal patterns highlight the impact of weather on bike-share usage.

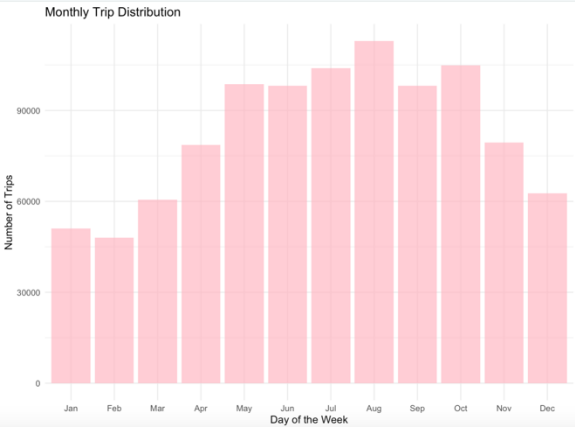

Fig. 4. “Plot Showing Temporal Trends of Citi Bike Usage by Month.”

B. Busiest Stations and Popular Routes

The busiest stations are predominantly located in areas that have high commercial and residential activity, such as W 21 St & 6th Ave, the busiest station in 2023. W 21 St & 6th Ave is located right in the heart of Manhattan and thus receives a high volume of footfall. Other busy stations include Broadway & W 58 St, and West St & Chambers St. All these stations are typically found near dense urban neighborhoods and transport hubs, allowing for both commute and leisure trips. Their strategic location near major landmarks such as Madison Square Garden, Alwyn Court, and Rockefeller Park, highlight Citi Bikes role in providing efficient transport options in high density areas.

```
# A tibble: 10 × 5
   start_station_name      total_trips avg_trip_duration latitude longitude
   <chr>                         <int>              <dbl>    <dbl>     <dbl>
 1 W 21 St & 6 Ave                4036               11.7     40.7     -74.0
 2 Broadway & W 58 St             3261               18.0     40.8     -74.0
 3 West St & Chambers St          3074               18.0     40.7     -74.0
 4 University Pl & E 14 St        3073               11.2     40.7     -74.0
 5 1 Ave & E 68 St                2969               13.1     40.8     -74.0
 6 11 Ave & W 41 St               2964               11.7     40.8     -74.0
 7 Broadway & W 25 St             2928               13.2     40.7     -74.0
 8 W 31 St & 7 Ave                2923               12.6     40.7     -74.0
 9 6 Ave & W 33 St                2842               13.0     40.7     -74.0
10 E 17 St & Broadway             2763               11.2     40.7     -74.0
```

Fig. 5. “Tibble Showing the 10 Busiest Stations in 2023.”

When it comes to the most popular/frequented routes, the analysis revealed that many involve short trips between nearby stations. For instance, the route between Forsyth & Broome St to Delancey St & Eldridge St, reflects a local demand for short distance commuting in the Lower East Side, a residential and commercial neighborhood where the stations are located. The most popular route was the loop at Central Park S & 6, suggesting the idea that in these areas Citi Bikes are mainly used for leisure and recreational purposes. These routes further highlight the importance of accessible bike-sharing infrastructure.

	start_station_name	end_station_name	total_trips	avg_trip_distance_km
	<chr>	<chr>	<int>	<dbl>
1	Central Park S & 6 Ave	Central Park S & 6 Ave	396	0.0209
2	7 Ave & Central Park South	7 Ave & Central Park South	350	0.0358
3	Forsyth St & Broome St	Delancey St & Eldridge St	343	0.111
4	Delancey St & Eldridge St	Forsyth St & Broome St	323	0.111
5	Grand Army Plaza & Central Park S	Grand Army Plaza & Central Park S	274	0.0656
6	Roosevelt Island Tramway	Roosevelt Island Tramway	243	0.0241
7	Broadway & W 58 St	Broadway & W 58 St	219	0.0321
8	Dock St & Front St	Old Fulton St	215	0.111
9	Greenwich St & Hubert St	North Moore St & Greenwich St	215	0.126
10	North Moore St & Greenwich St	Vesey St & Church St	201	0.888

Fig. 6. "Tibble Showing the 10 Most Popular Routes in 2023."

C. *Clustering Stations by Usage*

A Clustering analysis of the stations revealed distinct groupings of stations, and suggests that station demand is tied closely to location. Areas with denser clusters, such as Manhattan and Brooklyn (green and blue) show high bike usage, which likely correspond with greater footfall as these areas are highly commercial and residential and have more businesses and transport hubs in the surroundings. The clustering analysis further emphasized that location and proximity to major landmarks plays a key role in station popularity.

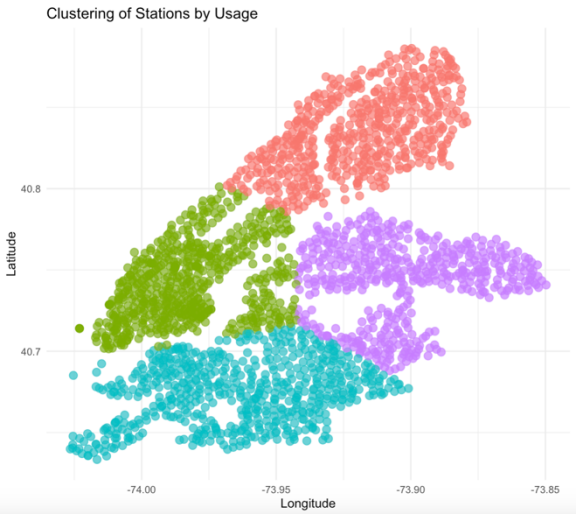

Fig. 7. "Plot Showing the Clustering of Stations by Usage and Location."

D. *Linear Regression*

The linear regression model was developed to examine the factors influencing station popularity. Average trip distance, average trip duration, and latitude and longitude were used as the predictors and total trips as the dependent variable. The model summary showed that the model explained approximately 36.3% of the variation in the total trips (R-squared value 0.363). The F-statistic of 315.8 and p-value $< 2.2e - 16$ confirms that the model is statistically significant. Getting into the key findings from the model, it was observed that:

- Average trip distance had a negative and significant effect (ß = -77.04 , $p < 0.001$), and indicates that as trip distance increases, station popularity decreases.
- Average trip duration was only marginally significant (ß = -3.79, $p = 0.0527$) and suggests a weak inverse relationship with station popularity.
- Latitude had a positive and highly significant effect (ß = 1834, $p < 0.001$), highlighting that stations located further north were associated with higher trip counts.
- Longitude had a significant negative relationship with popularity (ß = -9185, $p < 0.001$), thus indicating that stations towards the west were less popular.

Finally, the scatterplot of predicted vs total trips shows that the model exhibits heteroskedasticity, as the residuals are larger for higher trip counts. Most predictions cluster near the dashed line, however deviations become more obvious for stations with extreme popularity and points towards

limitations in the model's performance for outliers.

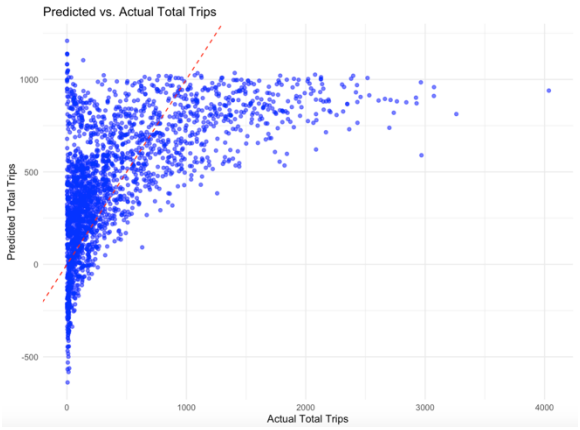

Fig. 8. "Scatterplot Showing the Regression Model of Predicted vs Actual Trips."

VI. CONCLUSIONS AND IMPLICATIONS

The results of the analysis highlight several important trends and insights:

- Peak Commute and Leisure Times: The peaks in usage during the evening and early-morning hours suggest that Citi Bike is primarily used for commuting during these times, and more recreationally during later hours of the day.
- Geographic Influence on Popularity: the clustering analysis as well as the linear regression model showed that geographic factors play a significant role in station popularity. Areas near transit hubs, businesses, parks, and tourist destinations see more frequent use, highlighting the importance of optimizing station placement and availability of bikes in these locations.
- Demand for Shorter Rides: the regression analysis also pointed out a preference for shorter, and more accessible trips, which may be driven by convenience during main commuting hours. This suggests that stations within close proximity of others that offer quick, local trips may be likelier to be more popular.
- Opportunities for Expansion: Areas with sparse clustering of stations or stations with fewer trips may be indicative of underserved neighborhoods. Targeted expansions in these areas could improve accessibility and help reduce inequalities in bike share availability.

VII. APPLICATIONS AND FUTURE SCOPE

Findings of the analysis offer several real world applications for urban planning, transport optimization, and sustainability initiatives:

- Optimizing Station Placement: insights from station clustering and usage patterns can guide the placement of new stations.
- Improving Customer Experience: analysis of peak usage times and routes can inform inventory allocation during rush hours. This includes increasing bike availability, lowering rates, offering promotions, and expanding docking capacities at busy stations and busy hours.
- Enhancing Equity and Accessibility: identifying underserved neighborhoods with fewer stations can highlight areas requiring increased investment. Targeted outreach and

infrastructure development in these areas can improve access for all socioeconomic groups.

- Developing Sustainable Tourism Initiatives: popular routes like those around Central Park and other tourist hotspots indicate opportunities to promote eco friendly tourism. In this case, customized packages or guided cycling tours could be introduced to encourage sustainable travel.
- Incentive Design: findings on trip distances and durations can provide data for designing policies such as incentives for off peak usage or discounts for longer trips, thus balancing demand throughout the day.

VIII. LIMITATIONS

While the analysis offers several insights into bike share usage and their real world applications, it does not come without it's limitations. For example, there is a lack of socioeconomic and demographic variables which limits the ability to fully explore equity and accessibility issues. Another limitation may be that while the clustering analysis provided valuable insights, the approach may have been too simple to fully capture the broader scope of what the data can tell us. Using a basic k-means method may not fully capture the complexity of spatial and usage dynamics. The linear regression model also explains only a moderate proportion of variability in station popularity, meaning that there are other unmeasured factors such as weather patterns or local events that significantly influence trip patterns. Lastly, the study only focuses on data from a single year and a single city, thus limiting the generalizability of the findings to other urban areas or timeframes. Extending the analysis by including longitudinal data from multiple cities could strengthen both the generalizability and the ecological validity of the conclusions.

References

An, R., Zahnow, R., Pojani, D., & Corcoran, J. (2019). Weather and cycling in New York: The case of Citibike. *Journal of Transport Geography*, *77*, 97–112. https://doi.org/10.1016/j.jtrangeo.2019.04.016

Chen, Y., Liu, Z., & Huang, D. (2018). Unlock a bike, unlock New York: A study of the New York Citi Bike system. In *Smart innovation, systems and technologies* (pp. 356–366). https://doi.org/10.1007/978-3-319-92231-7_37

Fritz, F. R. (2017). *BIKE SHARING IN NEW YORK CITY: HOW THE CITI BIKE SYSTEM SERVES POINTS OF INTEREST*[Bachelor of Science Thesis, University of Twente]. https://essay.utwente.nl/73652/1/Fritz_BA_TNW.pdf

New York City Department of City Planning. (2008). *Sustainable Streets Strategic Plan for the New York City Department of Transportation 2008 and beyond*. http://www.nyc.gov/html/dot/downloads/pdf/stratplan_compplan.pdf

O'Mahony, E., & Shmoys, D. (2015). Data analysis and optimization for (CITI)Bike Sharing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1). https://doi.org/10.1609/aaai.v29i1.9245

Schulze, J. (2024, February 12). *What is R Programming? Use cases and FAQ*. Coursera. https://www.coursera.org/articles/what-is-r-programming

Sobolevsky, S., Levitskaya, E., Chan, H. C. B., Postle, M., & Kontokosta, C. E. (2018). Impact of bike sharing in New York City. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1808.06606