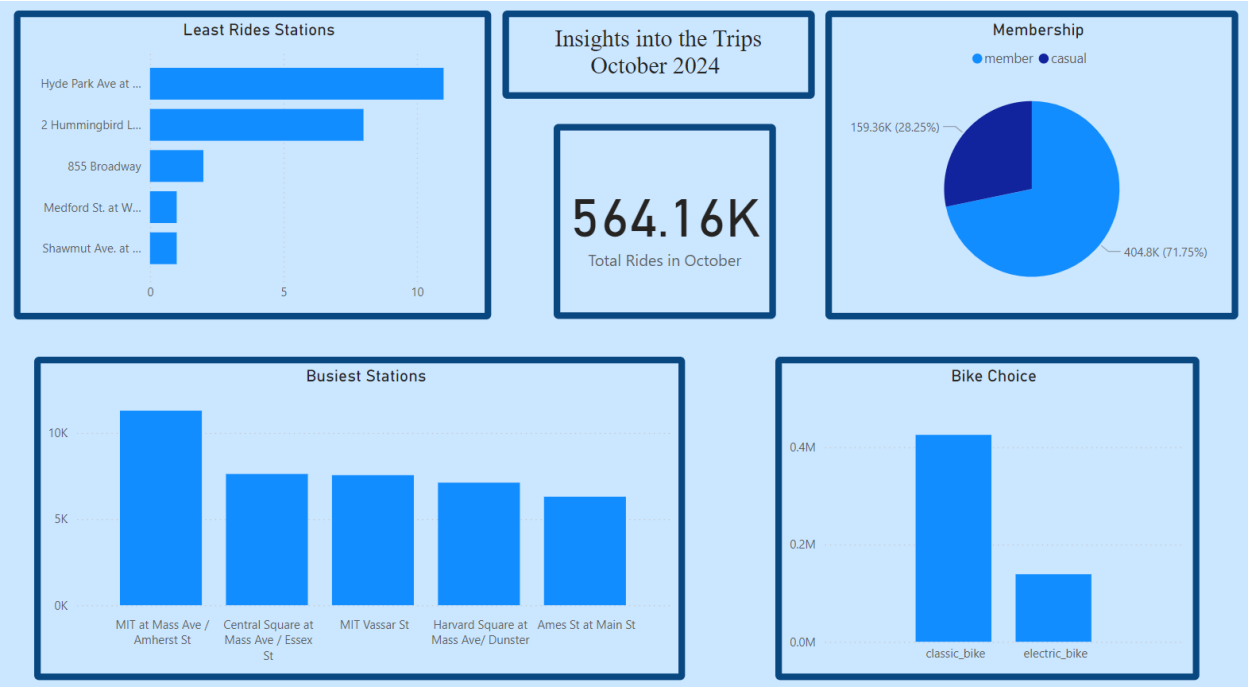
Boston Blue Bikes Rides Analysis – October 2024

Sreeja Kandukuri

11/19/2024



INTRODUCTION

The project's goal is to comprehend and forecast ride performance. This project's main objective is to create a model for forecasting future ride performance and analyze and interpret bike-sharing trip data to identify usage patterns and preferences among members and casual riders. An explanation of the goals is provided below.

- **Recognize the trip details:** This includes going into the raw data to extract crucial information about bike rides:

• Trip Duration: Examining the duration of rides to distinguish between members and casual users.
• Trip Distance: Calculating the distance covered on each ride in order to pinpoint outliers and average ride durations.
• Geographic Patterns: Use start and finish points to identify well-liked stations or routes.
The distance between the starting and finishing places to gauge riders' preferences for round-trip or one-way travel.
Members' and casual riders' preferred uses: By dividing the data into groups of members and casual riders, we hope to find: Average ride times, distances, and trip durations are examples of riding behavior. Members may commute by bicycle on weekdays for shorter distances.
Bikes can be used by casual riders for tourism or leisure (longer journeys, weekends, and holidays).

- **Estimating the Performance of a Ride**: Our goal is to develop a predictive algorithm to forecast future ride metrics using historical data. Time Series Analysis: To forecast general patterns and rider seasonality.

Regression models are used to forecast the length of a trip or its distance depending on variables such as rider type, day, time, and location.
Classification Models: To forecast the possibility that a member or casual rider will accept a ride.

- **Business Insights and Suggestions:** The results of the study and forecasts will offer practical insights like:

Streamlining Operations: Determining peak stations and hours to guarantee bike availability. Distributing bike supplies among stations according to travel trends.
Targeted Marketing: To promote membership, advertising should be tailored to casual riders.
Planning campaigns around areas and periods of high usage.
Improving the User Experience: o Providing benefits according to rider preferences, modifying prices, or improving routes.

- **Strategic Goals:** The ultimate goal is to use the insights and prediction models to increase the bike-sharing system's operational efficiency. By customizing services to rider preferences, you may increase client happiness. Boost income with improved resource management and marketing optimization.

In addition to meeting current analytical requirements, this project establishes the groundwork for a data-driven approach that can be adjusted to future opportunities and difficulties in bike-sharing systems.

DATA PREPROCESSING

Data Cleaning: With over 560,000 rows and 18 columns, the raw dataset offers extensive information regarding bike-sharing excursions. The ride ID, trip date, start and end station names, and other crucial trip details are all included in the key columns. The basis for identifying trends in bike usage and rider preferences is this extensive dataset.
Null value handling and data cleaning. It was crucial to confirm the accuracy and consistency of the dataset before beginning any sort of analysis. To guarantee the accuracy of the conclusions drawn from the data, a comprehensive data cleaning procedure was performed, identifying and eliminating null values from every column. To preserve data integrity, rows that were deemed invalid or incomplete were eliminated. Accurate analysis was made possible by strengthening the dataset by removing any missing or insufficient information.

Preprocessing and Feature Extraction: To convert raw data into a format suitable for analysis, a number of preprocessing procedures were carried out:
• Datetime Manipulation: For granular analysis, the trip_date column was divided into distinct date and time components. In order to examine weekly trends in rider behavior, the date data was also used to derive the day of the week.
• Trip Duration Calculation: The trip duration for every ride was determined using the start_time and end_time columns. This functionality is essential for comprehending user behavior and ride durations.
• Identification of Station Districts: While station IDs were present in the dataset, it lacked station districts. The Boston Blue Bikes Stations dataset was used in order to address issue. Missing station districts were mapped to their matching station IDs using Excel's Index-Match and VLOOKUP capabilities, enhancing the dataset for geographical analysis

Data Type Standardization: All columns were changed to their proper data types to enable seamless analysis. Datetime objects were created using the trip_date, start_time, and end_time datetime columns. Member_casual and other categorical columns had the proper labels. For uniformity, numerical fields like trip lengths were standardized.

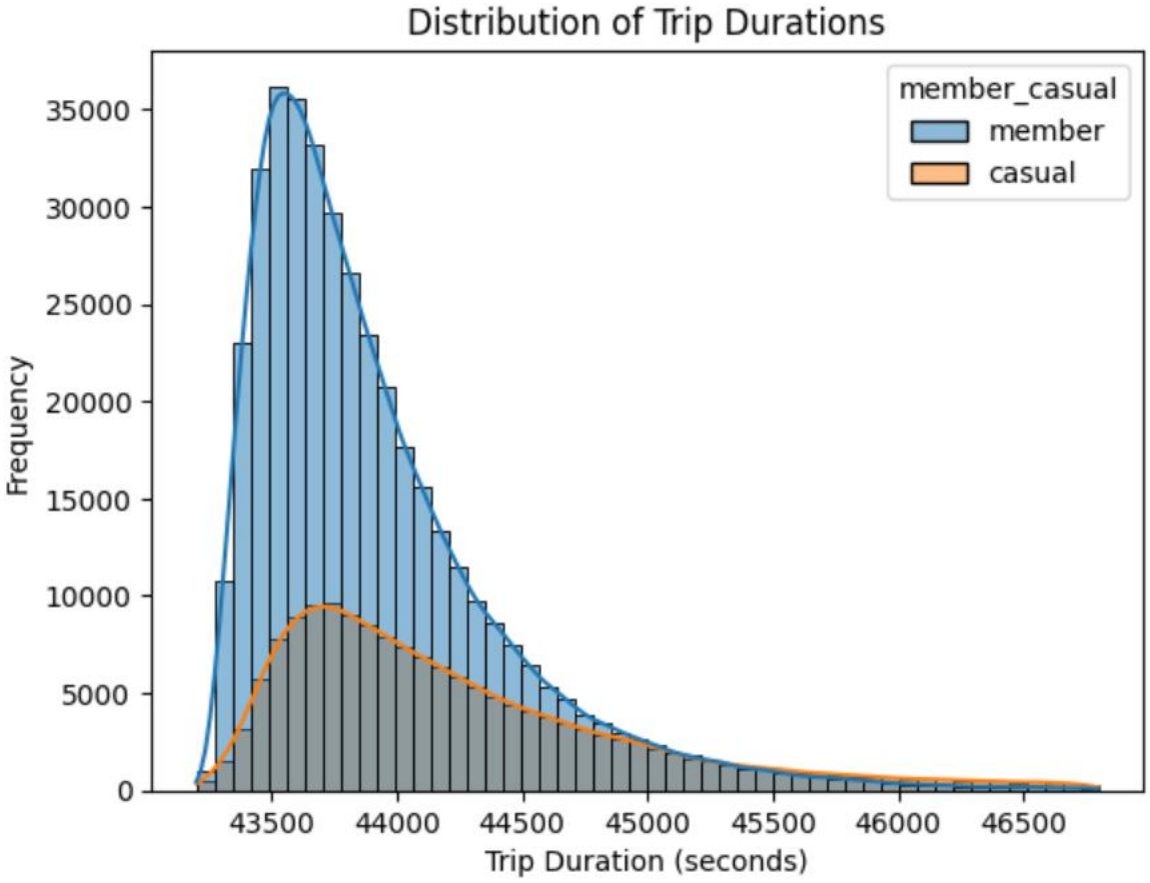
By ensuring that Python-based analysis could proceed without type-related problems, this standardization made it possible for data pretreatment and analysis to be seamlessly integrated.

Excel was used for the initial preprocessing and cleaning. The benefit of this option was that it provided an easy-to-use interface for managing big datasets, especially for manual chores like quickly validating data and mapping station IDs to names. The dataset was prepared for additional investigation and sophisticated analysis in Python by completing these procedures in Excel.

The raw data was transformed into a clean and organized format owing in large part to these preparation techniques. The dataset was transformed into a solid basis for gaining knowledge about ride utilization trends, preferences, and patterns by eliminating null values, extracting pertinent features, and standardizing data types. This methodical technique prepared the groundwork for in-depth investigation and predictive modeling by guaranteeing that the ensuing Python analysis was effective and error-free.

ANALYSIS

Average Trip Time: Members have a slightly shorter average trip time of about 43,913 seconds (about 12.2 hours), whereas casual riders' average trip time is about 44,241 seconds (about 12.3 hours). This suggests that casual riders typically travel a little farther than members. Although there is little variation between the two categories, the data indicates that subscribers, who probably use the service more frequently, could prefer shorter, more frequent trips, while casual riders might take longer trips, perhaps as a result of more impulsive or leisure-driven use.

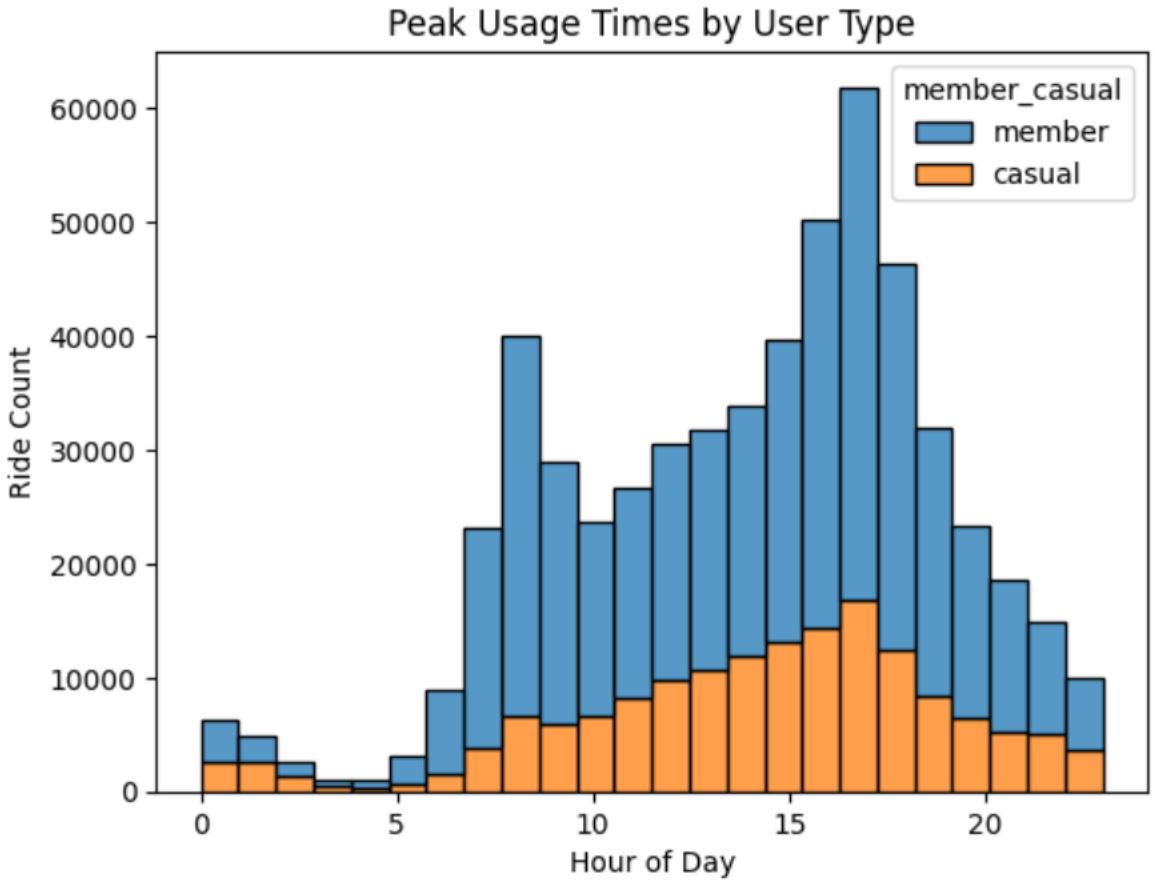


Bike Type Preferences: Members and casual riders differ significantly in the kind of bikes they use. With 123,425 rides overall, casual riders choose traditional cycles over electric bikes, which had 35,375 rides. However, with 301,602 rides on classic bikes and 102,795 rides on electric bikes, members exhibit a more evenly distributed preference between the two bike types. According to this trend, members have a more varied taste for bike types, including a large usage of electric bikes, which may provide additional convenience for longer or faster travels, whereas casual riders tend to choose traditional bikes, perhaps due to price or ease.

```
Bike Type Preferences:
 rideable_type  classic_bike  electric_bike
member_casual
casual                123425          35375
member                301602         102795
```

Usage Trends : The usage trends by time of day are revealed by the hourly distribution of trips. The morning and evening bike counts are the highest for both members and casual riders, which is consistent with normal commute schedules. For instance, both groups exhibit a significant rise in ride counts between 7:00 AM and 9:00 AM, with members hitting 33,328 trips at that time and casual riders reaching 6,710 trips at 8:00 AM. Similar to this, there are a lot of rides during the 5:00 PM to 7:00 PM evening commute, especially for users who consistently use the service at its highest levels during the day. Casual cyclists, on the other hand, exhibit less change throughout the day, riding less in the early morning and late evening.

These findings indicate that, while both casual riders and members utilize bikes for commuting, members are more likely to ride during peak periods, indicating consistent usage habits.



Average Trip Distance: Casual riders travel slightly further on average (2.13 km) than members (1.97 km). This distinction may indicate that casual riders use their bikes for more leisurely or lengthier trips, such as exploring neighborhoods or traveling longer distances. Members who use the bikes for specific objectives, such as commuting or running quick errands, may prefer shorter trips. This tendency shows that members are more likely

to utilize bikes for utilitarian, time-sensitive transport, whereas casual riders may be more willing to take longer trips for leisure purposes.

The data shows significant variations between casual and member riders in terms of trip duration, bike type preferences, peak usage periods, and trip distance. Casual riders take longer excursions, prefer older bikes, and go further on average than members. Members, on the other hand, have more structured usage patterns, with greater peak ride counts during rush hours and a more evenly split preference for classic and electric bikes. These analytics can help with operational choices like fleet management, bike distribution, and customized marketing campaigns for distinct rider segments.

```
Average Trip Distance (km):
 member_casual
casual    2.126892
member    1.970809
Name: distance_km, dtype: float64
```

Top Busy Stations: The top ten busiest stations exhibit a clear trend in which members (who are likely to use these stations more frequently) dominate trip counts. For example, the busiest stop is MIT at Mass Ave / Amherst St, which receives 8,836 trips from members and 2,460 from casual users. This shows that the station is popular with frequent users, most likely because to its proximity to significant locations like MIT. Similarly, Central Square at Mass Ave / Essex St sees a significant amount of journeys, with 6,087 by members and 1,531 by casual riders, demonstrating its popularity among both casual and habitual cyclists.

Stations including MIT Vassar St, Harvard Square at Mass Ave/Dunster, and Ames St at Main St follow a similar pattern, with members accounting for the vast majority of rides. The presence of educational institutions and commercial hubs in these places is expected to add to their popularity among members who commute frequently for job or study. In contrast, casual passengers continue to use these stations, albeit to a reduced proportion.

These stations are significant transit hubs in high-traffic locations, indicating that regular users (members) are more likely to utilize them because of their convenience, whereas casual users visit less frequently.

Least Busy Stations: In contrast, the Bottom 10 Least Busy Stations list stations that have little activity, with several showing only a few trips. For example, Medford St. at Warren St. and Shawmut Ave. at Herald St. each had only one trip recorded by members. This shows that some stations may be in less visited regions or do not serve a significant demand for rides. 855 Broadway and 2 Hummingbird Lane near Olmsted Green both have low trip counts, with 855 Broadway having only two casual trips and 2 Hummingbird Lane seeing balanced activity from both casual and member riders (four trips each).

Interestingly, several of these stations, such as Hyde Park Ave at Arlington St. and Cedar Grove T Stop, have slightly higher member counts (10-12), but still rank toward the bottom of the utilization spectrum. These sites may be more remote or less accessible to main destinations, making them less appealing for frequent cycling.While these stations are the least active, they nonetheless serve a small number of users, with a few, such as Community Life Center and Revere Public Library, displaying a balance of casual and member ridership. This balance shows that these stations could function as secondary hubs or community-based stations, with low but consistent use.

Overall, the analysis of the busiest and least crowded stations gives useful information about station consumption patterns. The busiest stations are actively used by members and are frequently placed near educational and commercial districts, whilst the least popular stations have less activity, potentially due to their location or lack of high demand. This data can be critical for improving station placement, understanding rider behavior, and maximizing the bike-sharing network's efficiency.

```
Top 10 Busiest Stations with Member and Casual Counts:
member_casual                                    casual  member
station_name
MIT at Mass Ave / Amherst St                       2460    8836
Central Square at Mass Ave / Essex St              1531    6087
MIT Vassar St                                      1135    6413
Harvard Square at Mass Ave/ Dunster                2415    4692
Ames St at Main St                                  740    5554
Beacon St at Massachusetts Ave                     1901    4088
MIT Pacific St at Purrington St                     398    5203
Christian Science Plaza - Massachusetts Ave at ...  1839    3419
Commonwealth Ave at Agganis Way                    1709    3349
Boylston St at Massachusetts Ave                   1662    3200

Bottom 10 Least Busy Stations with Member and Casual Counts:
member_casual                          casual  member
station_name
Medford St. at Warren St.                   0       1
Shawmut Ave. at Herald St.                  0       1
855 Broadway                                2       0
2 Hummingbird Lane at Olmsted Green         4       4
Hyde Park Ave at Arlington St               1      10
Cedar Grove T Stop                          8      12
Cummins Highway at Blue Hill Ave T Stop    10      10
Community Life Center                      15       6
Newton Library                             13       9
Revere Public Library                      14       9
```

District Count: The Start District Count indicates how many trips began in certain districts for both casual and member riders. The first notice is Boston's overwhelming domination in both categories. Boston accounts for 39,198 journeys by casual riders, while members make 105,812 trips. This is consistent with the concept that Boston is the key center for cycling activity, most likely due to its urban setting and high population density. Furthermore, Cambridge and Somerville appear to be major districts for both casual and member riders, with Cambridge receiving 11,867 trips from casual riders and 32,122 journeys from members, while Somerville receives 4,924 trips from casual riders and 12,567 trips from members.

These districts are anticipated to be popular destinations due to their proximity to universities, businesses, and residential neighborhoods. Other districts, such as Brookline, Chelsea, and Medford, exhibit noteworthy activity, but trip numbers are significantly lower than in Boston and Cambridge. Brookline, for example, has 1,927 trips from casual riders and 5,275 from members, demonstrating a consistent but less intense level of bike-sharing

utilization as compared to other major cities. Chelsea and Everett also contribute modestly to the overall count, with casual riders logging 2,312 and 2,306 journeys, respectively.

End Districts: The End District Count follows a similar trend to the start district statistics, although with minor differences in distribution. For casual riders, Boston remains the top finish district, accounting for 26,894 trips, which is consistent with the city's high number of start journeys. However, when comparing casual riders' end districts to members' end districts, we discover that members make much more visits in almost every category. Boston, for example, tops in terms of member travels, with 80,426. This high count matches the pattern seen in the start districts: members ride more frequently, and many of them end their trips in central cities like Boston and Cambridge.

Interestingly, while Cambridge is a popular starting place for both groups, the end district count for Cambridge shows that members are more likely to complete their travels there (30,437 trips) than casual riders (12,064). This could indicate that members use bikes to commute between neighboring locations (for example, from Boston or Somerville to Cambridge), whereas casual riders may prefer to begin and end their rides in Boston or nearby.

Several smaller districts, such as Arlington, Medford, Newton, and Revere, also exhibit excursion activity, but on a smaller scale. For example, Arlington had only 1,293 rides from casual riders and 1,222 from members, indicating that bike-sharing activity is minimal, maybe due to a lack of adequate infrastructure or reduced demand. Similarly, Medford has a low figure, with 1,083 trips from casual riders and 3,614 from members, which could indicate that the area is less central but still functions as a commuter hub for frequent riders.

In conclusion, the distribution of start and destination districts provides a few crucial insights into the bike-sharing usage patterns of both casual and member riders. Boston is the focal point for both groups, with a high proportion of journeys beginning and terminating in this district. Member users utilize the bike-sharing program more regularly in a broader range of districts, including Cambridge, Somerville, and other nearby cities. Casual riders, on the other hand, exhibit a more localized pattern of usage, with most trips concentrated in and around Boston. This data can be used to determine station placement, identify high-demand areas, and plan infrastructure improvements in less active districts.

Round Trip Details: The data shows that the majority of casual riders take non-round journeys, with a total of 150,059 rides that do not return to the starting station. This implies that casual riders, who are often occasional or recreational users, are more likely to utilize the bikes for one-way trips, such as exploring different parts of the city or running errands without having to return to their starting location. Only 8,741 rides are classified as round journeys, accounting for a modest percentage of their total utilization. This low round-trip count may imply that casual riders prefer to utilize the bike-sharing service for leisure, when they are less concerned with completing a round-trip excursion.

The data shows that member riders make a substantially higher number of non-round trips, with 397,697 one-way trips. However, unlike casual riders, members—who are likely regular commuters—have a significantly greater share of round journeys (6,700). Members' increased number of non-round journeys may represent their regular use of the service for commuting between different destinations, such as from home to work or school, rather than for recreational purposes, which are common among casual riders. Although members make fewer round trips than non-members, they are more likely to return to their starting station than casual riders.

This may indicate that members have more specific transit requirements, such as returning to the same spot after conducting errands or finishing their regular commute.

Weekly Ride Patterns: Thursdays and Wednesdays are consistently the busiest days for member rides, with over 70,000 rides each. These mid-week peaks could imply that members, who are likely regular commuters, are using the bike-sharing program as part of their daily work or study routine. The high volume of rides on these days may reflect a typical weekday commuting pattern, in which members prefer to use bikes for short excursions to and from work or university, leveraging the service's convenience and flexibility. In contrast, Sunday has the fewest member rides.

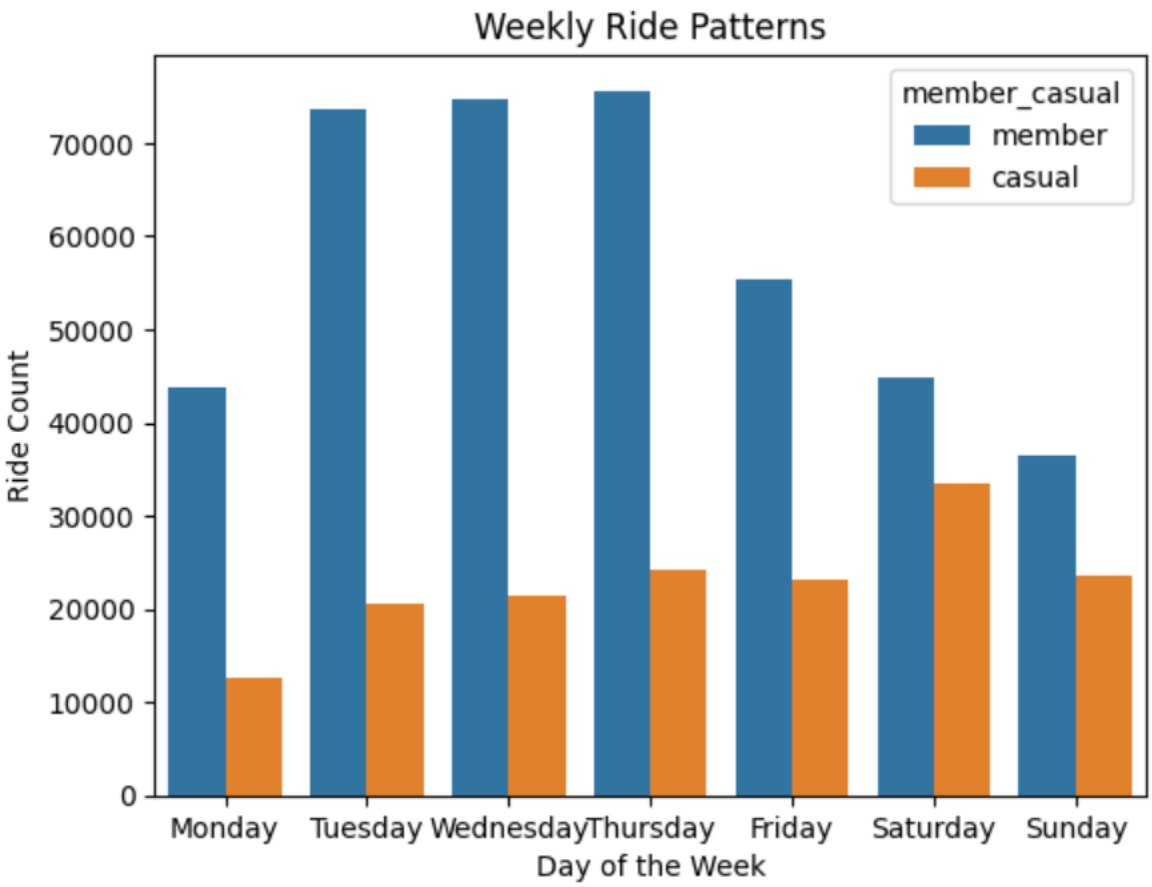
This could indicate that on Sundays, members may prefer to use other modes of transportation or engage in activities that do not necessitate frequent bike use, such as long-distance trips or leisure activities that do not include commuting. The drop in bike usage on Sundays could also be attributed to a general decrease in commuting activity as people take time off from their normal routines.

Casual Rides: Weekends are when casual riders see the most activity. Saturday is the most popular day for casual rides, with much more bikes being used than on other days of the week. This rise is most likely driven by leisure activities, with casual riders using the bike-sharing system to tour the city, participate in outdoor activities, or do errands on weekends. The increase in casual rides on Saturday indicates that people are more likely to utilize bikes for enjoyment when they have more spare time.

Although Sunday is the least popular day for member rides, casual ride activity increases when compared to weekdays. This may reflect the fact that many casual riders use bikes for leisurely rides or outdoor activities on weekends, yet utilization remains lower than on Saturdays. The reduced number of casual rides on Sundays as opposed to Saturdays could be attributed to weather conditions, personal schedules, or other reasons that reduce the desire to ride bikes on Sunday.

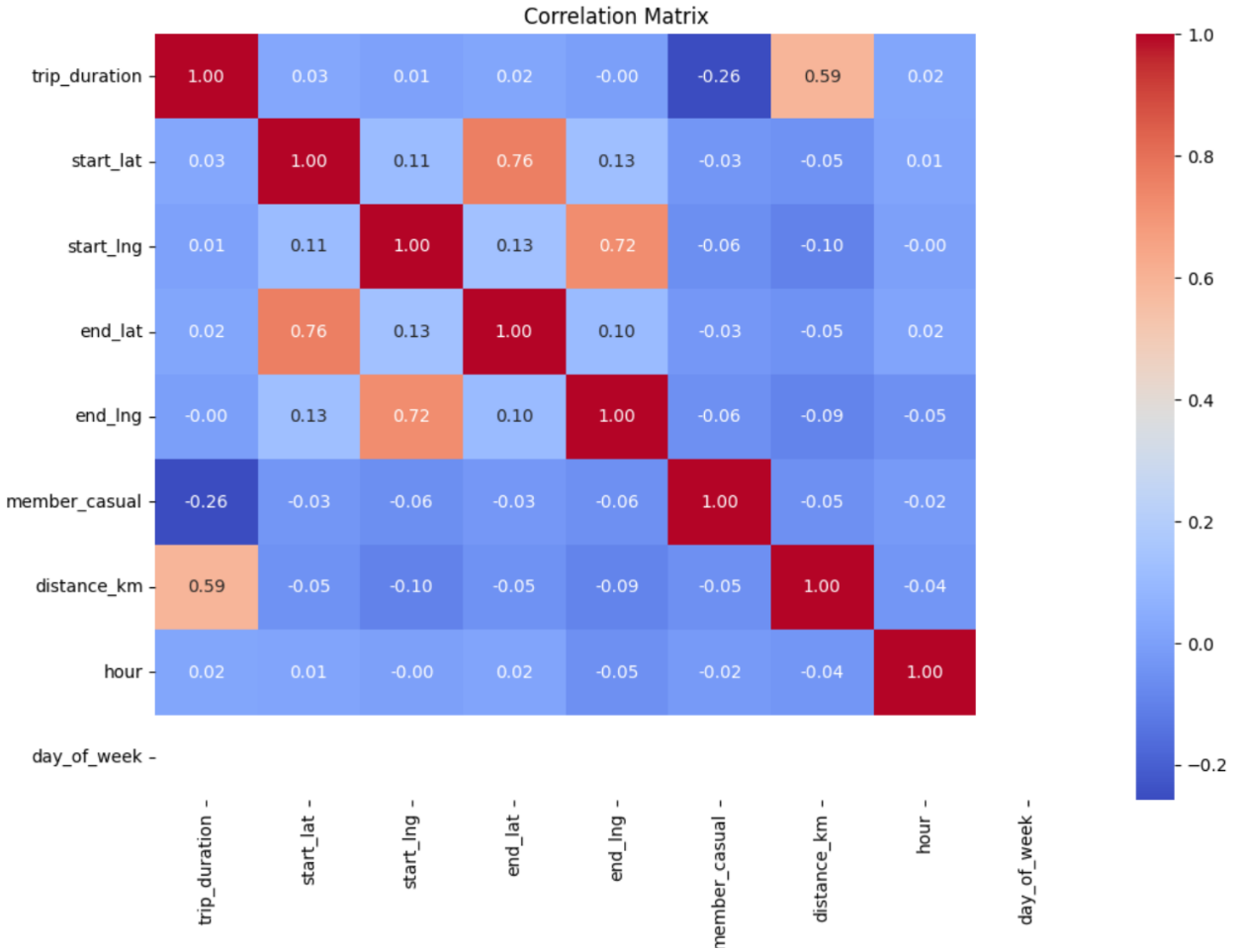
Monday: Interestingly, Monday has the fewest casual rides, at slightly more than 10,000. This could be attributed to normal start-of-week patterns, in which people may not prioritize leisure or recreational biking when they return to work or school following the weekend. People may opt to use other modes of transportation or focus on more time-sensitive activities at the beginning of the week, resulting in fewer casual users on Mondays.

In conclusion, the weekly ride patterns show clear differences between casual and member rider habits. Wednesday and Thursday are peak days for member rides, reflecting usual commute habits. In contrast, weekends—particularly Saturday—are dominated by casual riders utilizing bikes for recreational purposes, with Sunday showing a more moderate increase in informal usage. Monday's casual ride count is the lowest, in contrast to the mid-week peaks for members and the weekend surge. These trends offer useful insights into the temporal dynamics of bike-sharing demand, which can be used to influence operational strategies such as station placement, fleet management, and promotional efforts aimed at specific days or rider groups.



Correlation Matrix: The correlation matrix sheds light on the relationships between the dataset's various variables, allowing us to better understand how they influence trip time. The strongest positive correlation is found between trip_duration and distance_km, with a value of 0.5879. This demonstrates that journey duration grows dramatically with trip distance, a natural conclusion that is consistent with expectations. Interestingly, there is a weak negative connection between trip_duration and member_casual (-0.2579), indicating that casual riders may have longer trip durations than member riders.

Geospatial parameters such as start_lat, start_lng, end_lat, and end_lng have modest correlations with trip_duration, showing that while the start and end locations influence trip characteristics, they are not direct predictors of duration. Furthermore, temporal characteristics such as hour show negligible correlation (0.0216), indicating that the time of day has little impact on journey duration in this sample. Overall, these data highlight distance as the primary driver of journey duration, with other factors having more subtle and indirect effects.

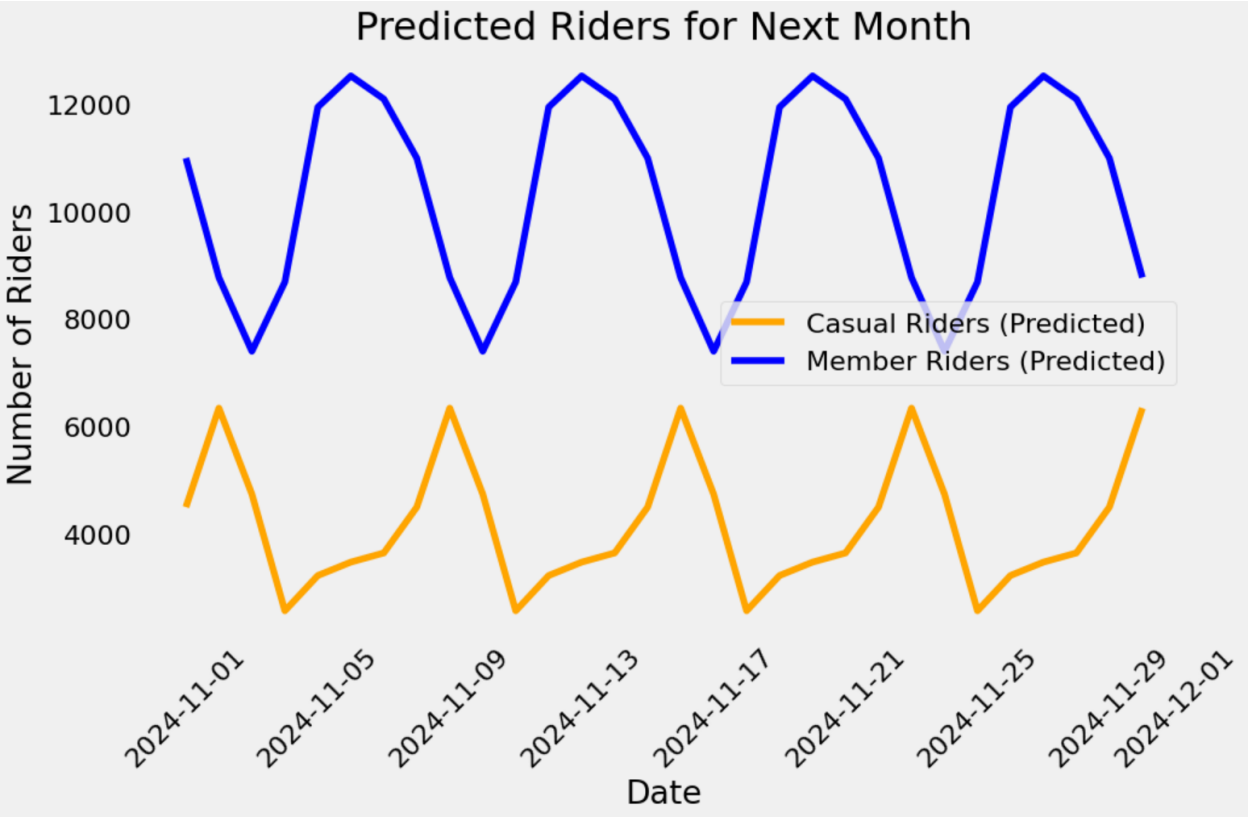


Random Forest Regressors: One of the first models used to forecast the duration of trips for both casual and member riders was the Random Forest Regressor. Several decision trees are used in this ensemble learning technique to provide reliable predictions, taking advantage of the data's intricate, non-linear interactions. Root Mean Squared Error (RMSE) values of 872.82 for member riders and 321.86 for casual riders were obtained using the Random Forest Regressor, indicating significant variations in model performance between the two rider groups.

The model performed poorly for member riders, most likely as a result of intrinsic complexity or fluctuation in their trip data, but it was reasonably accurate for casual riders. The disparity may have been caused by elements like an imbalance in the characteristics, variations in trip patterns, or a larger volume of data for member riders. Furthermore, Random Forests work well with numerical and categorical data, although their performance may have been constrained by the absence of explicit hyperparameter adjustment.

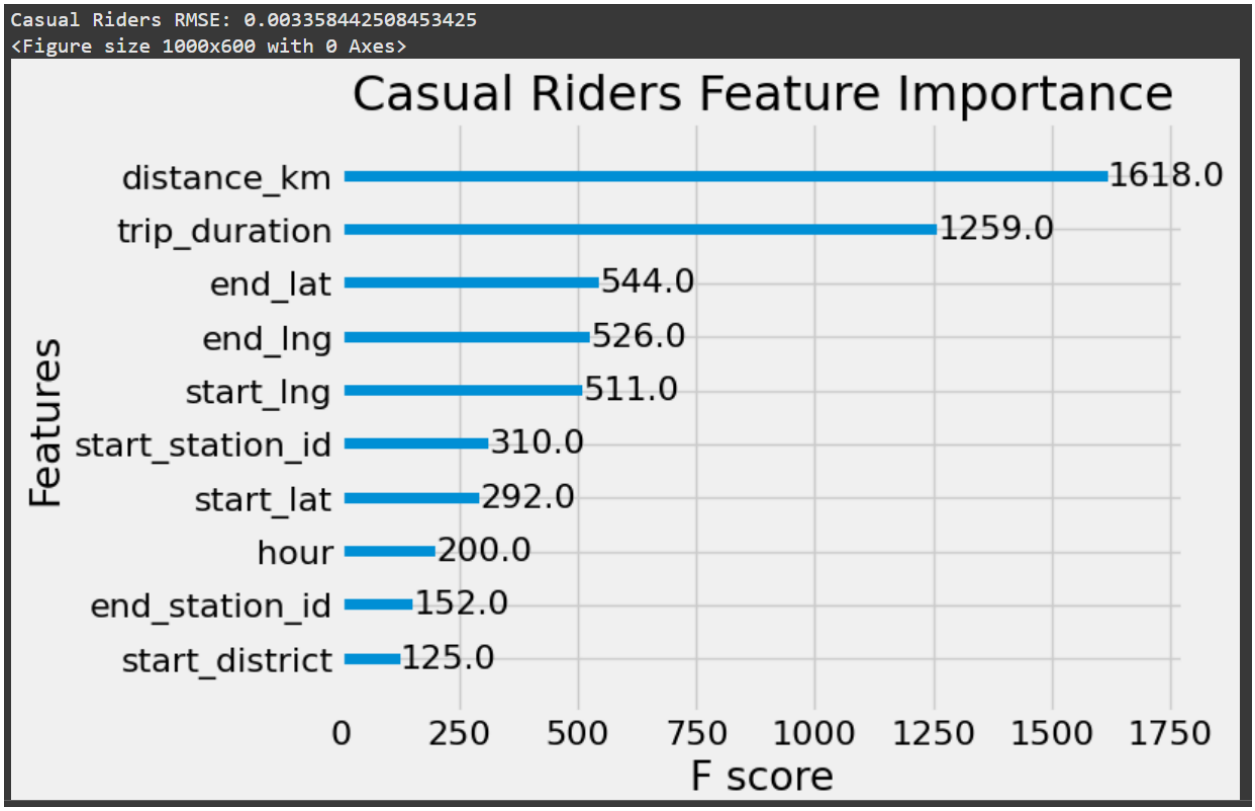
Not with standing these drawbacks, the Random Forest Regressor offered insightful baseline data that showed how parameters like distance, journey duration, and day of the week affected the target variable. Nonetheless,

the comparatively elevated RMSE for member riders indicated potential for enhancement, prompting the investigation of sophisticated machine learning models.



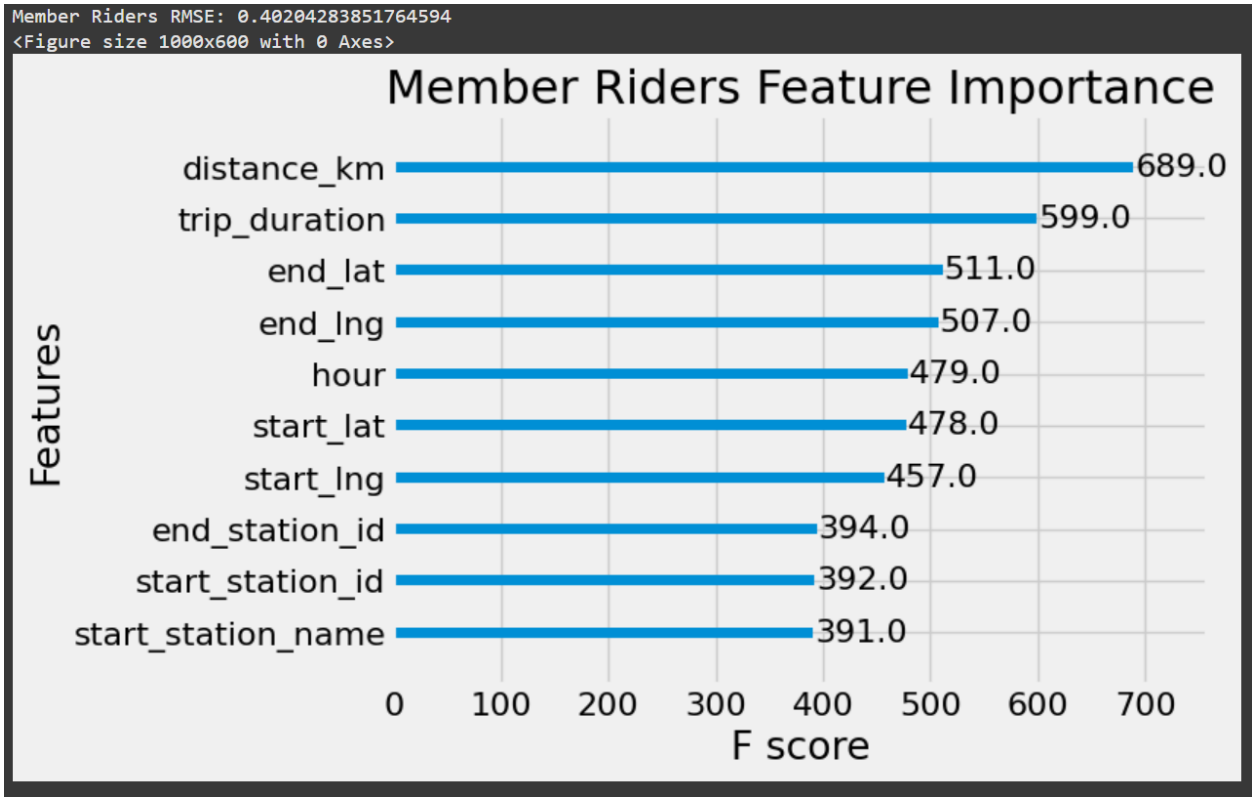
XGBoost: Based on a number of features, including distance_km, rideable_type, day_of_week, hour, and more, the model's output shows how well it predicts trip time for both casual and member riders.

Given the little discrepancy between the expected and actual values, the model's RMSE of 0.0034 indicates that it can forecast the trip duration for casual riders with very high accuracy. Distance_km and trip_duration are the most important parameters in forecasting the trip duration, according to the feature importance study. Distance_km has the greatest importance score (1618), followed by trip_duration (1259). This implies that the model's ability to produce precise forecasts is mostly dependent on the distance traveled during the ride and the length of the previous journey.



The model's predictions are less accurate for member riders, as evidenced by the Member Riders' much higher RMSE of 0.4020. Greater data variability or even more intricate patterns that are more difficult for the model to detect could be the cause of this larger inaccuracy. Despite this, the feature importance indicates that trip_duration and distance_km continue to be important factors in forecasting trip duration, with trip_duration coming in second (599), and distance_km being the most significant (689).

In conclusion, the model uses a variety of features to estimate the length of trips for both casual and member riders, with distance acting as the most reliable predictor. While predictions for member riders exhibit greater variability, as indicated by a larger RMSE value, the model performs remarkably well for casual riders.



BUSINESS RECOMMENDATIONS

Concentrate on Trip Optimization: Take into account improving trip recommendations and route ideas for both casual and member riders, as distance_km is a crucial element in determining trip duration. To improve user experience, highlight effective routes and give projected travel times and pricing estimation.

Improve Member Personalization: Given that member riders' RMSEs are greater, there may be a chance to customize services for them. To better serve members' requirements, examine other data, such as ride preferences, usage patterns by time of day, and station popularity.

Enhance Data Collection: Gather more information on weather, traffic patterns, and real-time ride usage trends to further lower prediction mistakes. These elements could aid in reducing variability, particularly for riders who are members.

Dynamic Pricing Strategies: To encourage off-peak trips and boost revenue during peak demand, use insights from the estimated trip durations to introduce dynamic pricing for casual passengers.

Tailored Marketing Campaigns: By demonstrating constant, data-backed ride duration reliability and route efficiency, marketing campaigns may be created to entice casual riders to become members

Focus on High-Impact elements: Since trip_duration and distance_km were shown to be the most influential elements, give priority to enhancing GPS precision, station connectivity, and distance measuring equipment in order to guarantee high-quality data for more accurate forecasts in the future.

CONCLUSION

Our analysis investigated the use of cutting-edge machine learning algorithms to forecast journey durations for both casual and member passengers. We were able to attain different degrees of accuracy for both rider groups by using models like Random Forest and XGBoost, which finally highlighted the differences between member and casual rider behavior. The most successful predictive tool was the XGBoost model, which performed exceptionally well for casual riders and had an astonishingly low RMSE of 0.0033, showing a high degree of accuracy in trip duration prediction. Even while the RMSE for member riders was marginally higher at 0.4020, the findings nevertheless offered insightful information about their riding habits.

One of the main conclusions is that, for both casual and member riders, distance_km is the most important element, followed by trip_duration. This emphasizes how crucial journey distance and past trip data are in figuring out how long a ride will last. The need for enhanced datasets to better capture the complexity of ride behavior, especially for members, is highlighted by the fact that other characteristics like time of day, start and end locations, and rider type contributed less significantly.

The analysis's findings highlight how useful machine learning models are for comprehending and forecasting rider behavior, empowering companies to make informed decisions. Although the models did well, particularly for casual riders, the marginal underperformance for member riders points to areas where data collecting and modeling techniques should be improved. If weather, traffic conditions, and user preferences are included, predicted accuracy may improve even more.

This study has established a solid foundation for analyzing trip dynamics, allowing the company to successfully optimize services and meet rider needs.