**Exploring Multimodal Transportation Planning: A Study of Blue Bikes and Public Transit in Boston**

*By Michael Brady, Justin Wells, and Daniel Heo*

Tufts University

DATA 200: Foundations of Data Analytics

December 9, 2024

1. **Introduction**

Efficient urban connectivity is crucial for creating livable and sustainable cities. Public transportation serves as a cornerstone of this connectivity—linking individuals to jobs, schools, and recreation while reducing traffic congestion and environmental damage caused by car dependency. In Boston, where the city’s compact layout and dense population create unique commuting challenges, the need for seamless transportation options is particularly important. A well-integrated transit system benefits not only individuals but also communities and businesses, supporting economic growth and enhancing the quality of life.

Boston’s Massachusetts Bay Transportation Authority (MBTA) forms the backbone of the city’s public transit system. With a network that includes subway, bus, commuter rail, and ferry services, the MBTA connects neighborhoods within Boston and extends to the broader metropolitan area. The system plays a key role in mitigating traffic, reducing the city’s carbon footprint, and ensuring equitable access to transportation. Complementing the MBTA is the Blue Bikes system, Boston’s public bike-share service. Blue Bikes offers residents and visitors an alternative to walking or driving, with stations distributed throughout the city and its surrounding areas. As an environmentally friendly solution for short trips, Blue Bikes fills gaps where MBTA services may not reach, making it a critical component of the city’s multimodal transit network.

There is a significant amount of current research that explores interactions between public transit systems and micro mobility solutions like bike-sharing. Much of this research focuses on specific elements such as seasonality, trip distance, or competition. Many of the studies we looked at, for example, highlight how shared mobility systems can serve as first-mile and last-mile solutions and complement traditional transit during service disruptions. However, these studies primarily analyze dockless bike-sharing systems, which differ significantly from Boston’s docked Blue Bikes model. Dockless bikes are not tied to a station, and interpreting their data does not include uncertainty about behavior prior to getting on the bike thus resulting in a gap in the current research surrounding the extent to which docked systems align with public transit and how they might be optimized to improve urban mobility.

We aim to address this gap by setting the overall research goal for our project as describing Blue Bike usage patterns in Boston and understanding how Blue Bikes interact with MBTA. More specifically, we investigate how the spatial proximity of Blue Bike stations to MBTA stops impacts bike-share usage, whether certain MBTA lines or stops are more frequently associated with Blue Bike activity, and how well Blue Bike usage aligns with peak MBTA commute times. Through spatial and temporal analyses, we seek to uncover patterns and trends that highlight the complementary roles of these two systems. To achieve this, we used geospatial analysis, descriptive statistics, and machine learning techniques to explore Blue Bike usage in relation to MBTA infrastructure. By answering questions about proximity, popularity, and temporal alignment, we aim to provide actionable insights for improving the integration of these systems and supporting a more connected and sustainable urban transit network.

1. **Literature**

Understanding how shared mobility systems interact with public transit networks has been a growing area of research, and the insights from previous studies were invaluable in helping to shape our analysis of the relationship between Boston’s Blue Bikes and MBTA transit systems. The studies we looked at not only provided frameworks and methodologies for our analysis, but also gave us some key examples of how bike-share systems function within broader transportation networks.

The first article, *Differences in First-Mile and Last-Mile Behavior in Candidate Multi-Modal Boston Bike-Share Micromobility Trips[[1]](#footnote-1)*, focused on the challenge of connecting transit systems with users' origins and destinations. This study emphasized how shared micro mobility systems like Blue Bikes can extend the reach of public transit through first-mile and last-mile connectivity. Using proximity-based categorization, it identified distinct patterns of bike-share usage, revealing that first-mile trips are more prevalent in Boston’s urban core, while last-mile trips dominate in outlying neighborhoods. This approach directly informed our spatial analysis by reinforcing the use of proximity buffers to define interaction zones around MBTA stops. On top of that, its analysis of spatiotemporal ridership patterns motivated our research into the ways in which timing and proximity affect Blue Bikes' integration with MBTA transit systems, especially during rush hour.

The second article, *Understanding the Competition and Cooperation Between Dockless Bike-Sharing and Metro Systems[[2]](#footnote-2)*, provided a detailed methodology for analyzing the interactions between bike-sharing and public transit. It introduced the concept of defining catchment areas around transit stations and categorizing trips into competitive, complementary, or feeder relationships. The use of proximity buffers such as the 150-meter zone adopted in this study, directly influenced our own analysis by helping us define the spatial connection between Blue Bikes and MBTA stops, allowing us to classify ‘close’ and ‘far’ proximity.

The final article, *Analysis of Spatial Interactions Among Shared E-Scooters, Shared Bikes and Public Transit[[3]](#footnote-3)*, emphasized the importance of spatial and temporal analysis in understanding the role of shared mobility systems. The study highlighted that micro mobility systems, like shared bikes and e-scooters, have become increasingly important in multimodal transportation networks, often complementing public transit by bridging gaps in accessibility. It also detailed how shared usage fluctuates throughout the day, with peaks typically occurring during traditional commuting times. By examining how trips align with transit stops and the times they occur, this study helped inform our investigation into Blue Bike trips during weekday commuting hours.

Each of these studies provided essential frameworks and methodologies that directly informed our project, enabling us to craft meaningful research questions and develop targeted objectives to approach for analyzing the relationship between Blue Bikes and the MBTA infrastructure. The emphasis on spatial distribution in the literature guided our use of proximity buffers to examine how closely Blue Bike stations align with MBTA stops, while insights into trip categorization and temporal usage patterns helped us explore broader themes of rider behavior. By incorporating these unique perspectives, we were able to analyze not only where and when these systems interact but also identify potential opportunities for improving their integration to create a more efficient multimodal transit network in the city of Boston.

1. **Methods**

We combined two datasets to most comprehensively perform our analysis. The first was a Blue Bikes Trip Data dataset that is released monthly by the Boston Blue Bikes organization and contains the following variables: Trip Duration (seconds), Start Time and Date, Stop Time and Date, Start Station Name, ID and Location, End Station Name, ID and Location, Bike ID, User Type (Member or Casual), and Rider Birth Year. It is preprocessed to remove any trips that were below 60 seconds in length as these might capture false starts or users trying to re-dock a bike to ensure it was secure.

Each month contained almost 500,000 rows of data, so to focus our analysis, we elected to only analyze data from September, 2024, as we felt this month conveyed peak Blue Bike usage with the weather still warm, and both schools and work in full session post summer holidays.

We decided to analyze these Blue Bike trips alongside an MBTA train infrastructure dataset in order to deduct the most relevant and widely significant conclusion from our analysis. The MBTA stop dataset contained MBTA Stop Name, Stop ID, Stop Location, Station Line, Terminal Stop (Y/N), Route, and District.

**Data Preprocessing**

Our first preprocessing and data cleaning step was to drop the rider birth year column from our Blue Bike Trip Data dataset and the Terminal Stop column from the MBTA dataset as they were both irrelevant to our analysis. We also dropped any rows that contained blank or missing data from the Blue Bikes Trip dataset which only consisted of a few hundred trips that did not have a registered start or end station.

We used ArcGIS Pro to perform the rest of our preprocessing and joining. This entailed creating a 150 foot buffer around each MBTA stop which we used to classify any Blue Bike station that fell within that buffer as ‘close to an MBTA stop’ as informed by a previous study.[[4]](#footnote-4) We also downloaded a third dataset containing Blue Bike station locations, and we did a spatial one-to-one join to get the closest MBTA stop and line for each Blue Bike station.

We finally joined the three datasets to get a final version of the Blue Bike Trip Data dataset that also included Close/Far from MBTA stop columns for every Blue Bike station, # of docks at every Blue Bike station, as well as the associated MBTA stop and line data for both the start and stop Blue Bike stations.

**Analytical Methods**

To break down our larger research goal of describing Blue Bike usage patterns in Boston and understanding how Blue Bikes interact with MBTA infrastructure, we split our analyses into five smaller questions and assigned an appropriate analytical method to each question.

We used GIS functions such as spatial intersect and point density to create heat maps that displayed the areas of boston and the MBTA lines/stops that are most popular with Blue Bike stations.

We used more general EDA and graphic techniques to sum up total trips for both member and casual users as well as for the Blue Bike stations that are close or far from MBTA stops. We then compared the output tables and created bar charts to highlight key trends and information.

We employed a logistic regression to test our hypothesis that the riders who use Blue Bikes in conjunction with MBTA for regular commuting were more likely to be members than casual riders. To take this analysis a step further, we ran two regressions to identify potential differences between morning and evening commute activity. We created two new variables in our dataset. The first was morning commute (Y/N) which = 1 if a trip ended at a ‘close’ station between 8am and 9am on a weekday. This variable captured—to the best of our capabilities— all Blue Bikes ride to a MBTA stop before work. Morning commute (Y/N) was the dependent variable for our morning regression. We then created another variable, evening commute (Y/N) which = 1 if a trip started at a ‘close’ station between 5pm and 6pm on a weekday—implying a ride from an MBTA stop after work—and set evening commute (Y/N) as the dependent variable for the evening regression. Our independent variables for this model were trip duration, member or casual rider (1 = member, 0 = casual), and electric or classic bike (1 = electric, 0 = classic).

Finally, we used a k-means clustering analysis to find the features that defined the most popular clusters of Blue Bike stations in Boston. We used the elbow method of plotting within-cluster sum of squares (WSS) vs different values of K to identify 3 as the optimal value for K, and used the following parameters as defining features for each cluster: Total Trips, Peak-Hour Usage Proportion (amount of trips between 8-9am and 5-6pm), Member-Casual Rider Type Proportion, Average Trip Duration, Station Docks, and the Proportion of Stations ‘close’ to MBTA stops.

**III. Results**

For this section, we will discuss the results for each of our smaller research questions, and then discuss how together, they address our overall research goal.

**Question 1: What MBTA lines and Boston Districts are most popular with Blue Bike stations?**

The analysis we performed in ArcGIS led us to conclude that the Red and Green lines are the most integrated with Blue Bike stations (figure 1) and that the majority of Blue Bike trips occur in Boston’s central districts (Downtown Boston, Cambridge, Brookline and Somerville). It is intuitive that Blue Bike trips follow Boston’s population density patterns. However, these results are valuable to our overall conclusion as they identify MBTA lines that might benefit from more Blue Bike stations such as the Orange and Silver lines which also have stops in some of these densely populated areas.

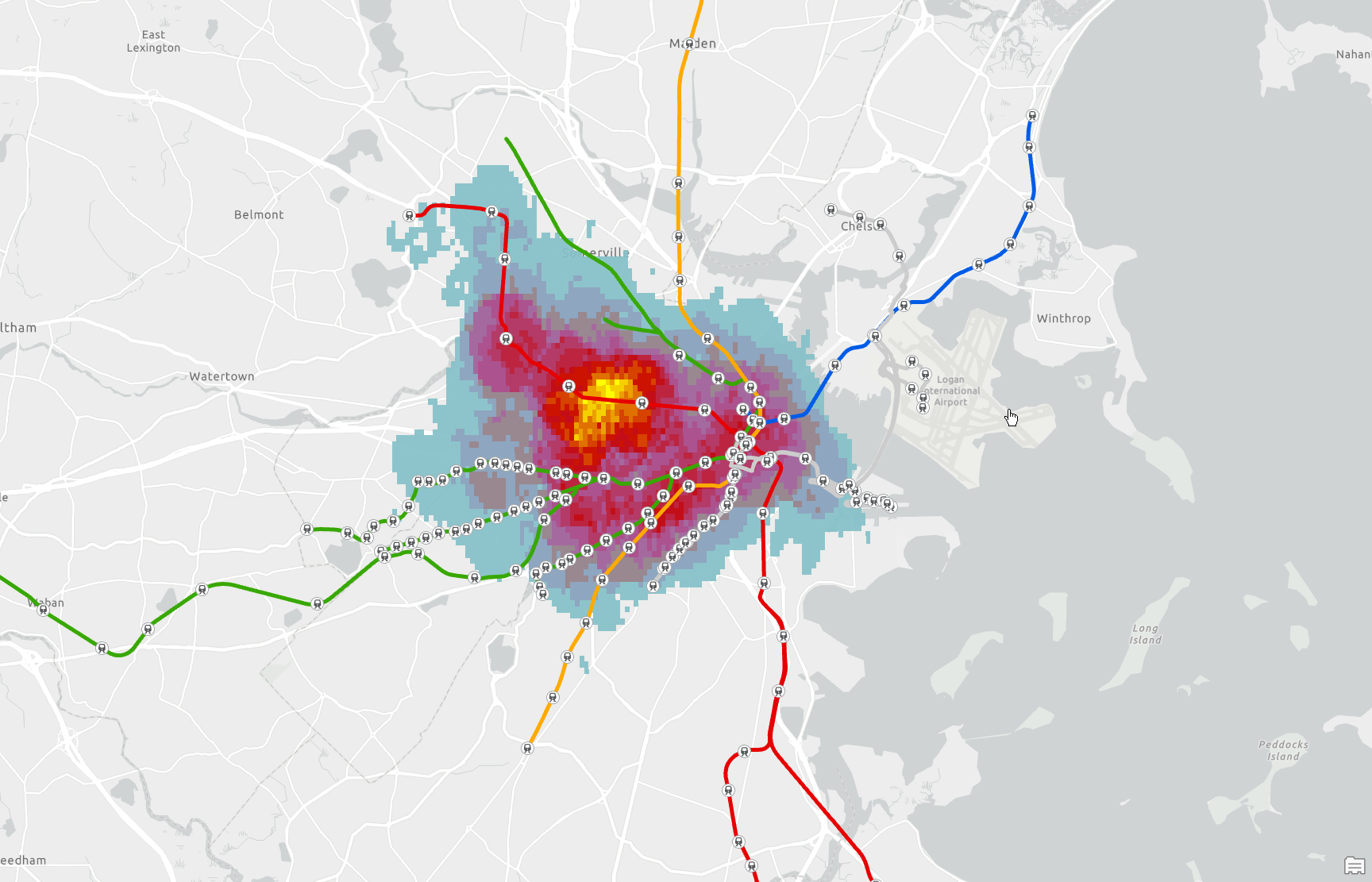


Figure 1

**Question 2: How does activity at Blue Bike stations near MBTA stops differ from those far away?**

We found there were over twice as many trips that started at Blue Bike stations that were ‘far’ from MBTA stops in September 2024. However, this is due to the fact that there are 319 ‘far’ stations and only ‘93’ close stations. When we normalized this and compared trips per station, we found that on average there were almost 660 more trips at stations close to MBTA stops than at ‘far’ stations. Given our analytical limitations, this result is the best way we could prove significant interactions exist between Blue Bikes and MBTA trains and that having bike stations ‘close’ to MBTA leads to higher usage.

**Question 3:How do member and casual usage differ?**

We found that 59% of all trips in September came from member riders indicating an effective Blue Bike membership program; We identified that trips taken by casual users are 11.3 minutes longer than member users pointing to more recreational and exploratory use for casual users; And finally, from this analysis we saw higher usage during the week than on the weekend for member riders and spikes during peak work commute times (8-9am and 5-6pm) for members, both of which highlight members as the primary users of Blue Bikes for commuting to work as well as the significance of the work commute to overall member usage (Figure 2).

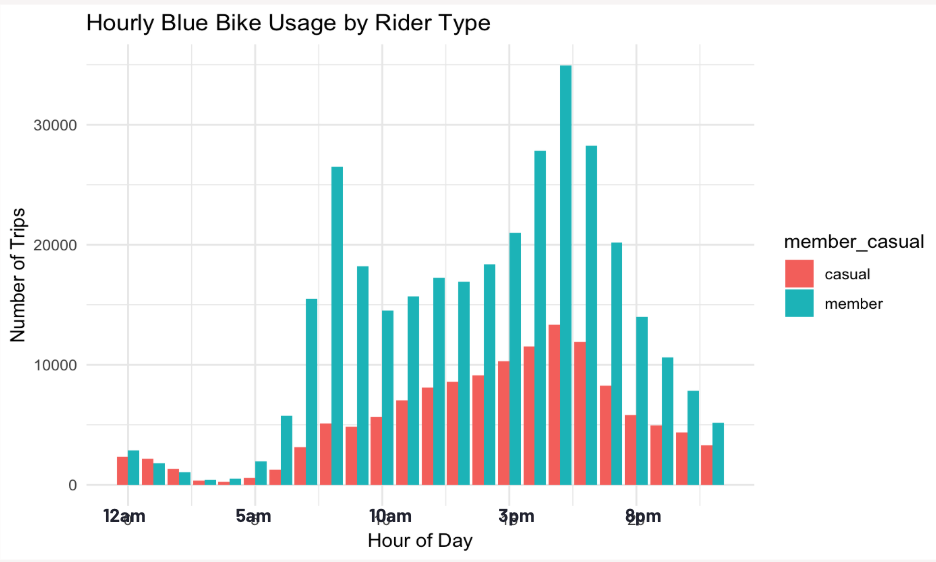


Figure 2

**Question 4: Are the people who use Blue Bikes for first/last mile public transit commutes primarily blue bike members?**

From our two logistic regression models, we were able to confirm our hypothesis that Blue Bike members are primarily the ones using Blue Bikes to travel to and from MBTA stops during peak commute times. For the morning regression, members were 93% more likely than casual users to end their trips at a ‘close’ station and 66% more likely to start their trip at a ‘close’ in the evening commute. The difference in these percentages makes sense as 5-6pm in the evening likely sees significantly more recreational usage than 8-9am. Although we were not able to prove this through analysis, we believe that by extension of these results, Boston Blue Bike has such a large membership because of its functionality in weekday work commuting.

**Question 5: What features define the clusters of the most/least popular blue bike stations?**

The K-means Clustering Model split the Blue Bike stations into three clusters which were then classified as Low Usage, Moderate Usage, and High Usage. The results can be seen in Figure 3.

Figure 1


Figure 3

There are a few significant points to highlight from these results that align with our findings from the previous analyses. The first is that the Low Usage stations are more likely to be located in areas far from public transportation and are used more by casual riders for longer rides while the Moderate and High Usage stations are much more member reliant. Moderate stations see the most peak-commute time activity representing usage that comes primarily from commuting to work, while High Usage stations are the most likely to be close to MBTA stops, probably in high-demand areas that benefit from both regular and commuter-heavy usage.

**Overall Research Goal: Describe Blue Bike usage patterns in Boston and understand how Blue Bikes interact with MBTA infrastructure**

Together, the results from each of our research questions provide a comprehensive understanding of the relationship between Blue Bikes and MBTA public transit. The analysis of popular MBTA lines and Boston districts revealed that the majority of Blue Bike usage occurs near densely populated areas and along the Red and Green lines, highlighting opportunities for further station expansion in underserved areas like the Orange and Silver lines. Our comparison of activity at stations near and far from MBTA stops demonstrated that proximity significantly enhances station usage, which highlights the importance of strategic placement to maximize interactions between Blue Bikes and MBTA. The member vs. casual rider analysis emphasized the role of Blue Bike members as primary users for weekday commuting during peak hours. Our logistic regression model then confirmed that Blue Bike members predominantly use the system for commutes to and from MBTA stops, validating the system’s role in workday public transportation. Finally, our clustering analysis highlighted the features that determine station usage, reinforcing the importance of proximity to MBTA and member reliance for high-demand stations. Together, these results underscore the critical role Blue Bikes play in Boston’s transportation network and provide actionable insights for enhancing system integration and accessibility.

1. **Discussion**

Our analysis highlights the significant role of Blue Bikes in enhancing Boston's multimodal transit system, particularly as a first/last-mile solution that connects riders to MBTA stops. Our results show that Blue Bike stations located within 150 meters of MBTA stops experience higher trip density during peak hours, which we believe is due to members who rely on the bikes for commuting. This finding aligns with a prior study by Romm et al., which demonstrated that first-mile trips are more prevalent in urban centers, while last-mile trips are more common in outlying areas. These spatial trends show the importance of proximity-based categorization when examining interactions between bike-sharing systems and public transit, which we have done successfully. Additionally, our results align with Tang and Zhou’s findings, which emphasize the importance of integrating bike-sharing systems with public transit to enhance urban mobility.

A unique finding in our study is the preference for classic bikes over electric bikes among members commuting to and from MBTA stops. While electric bikes are popular in other cities, our findings in Boston that show an overwhelming preference for classic bikes could be due to factors such as pricing, availability, or user familiarity with classic models. Additionally, we found that there were distinct behavioral patterns between casual and member riders: casual riders typically take longer, recreational trips, while members predominantly use Blue Bikes for short, routine commutes during morning and evening peak hours. These temporal patterns align with findings by Lu et al., who noted that shared bike usage peaks during traditional commuting hours, which again shows its key role in multimodal transit.

While we were successful in obtaining these insights, our study is not without limitations. First, the dataset’s focus on September 2024 may be biased because it does not capture seasonal variations in bike usage. Second, our analysis did not incorporate socioeconomic variables or demographic data, which could provide a deeper understanding of accessibility and equity in Blue Bike usage. Third, while we examined proximity to MBTA stops, we did not fully consider the quality or safety of bike infrastructure, such as bike lane safety and connectivity, which could be key factors influencing station usage. Finally, without access to MBTA ridership data, we cannot say with full confidence that riders travelling to or from “close” Blue Bike stations actually got on or off the train as part of their commute. This means that although we were successful in developing an initial connection between Blue Bike usage and MBTA ridership based on proximity and peak commuting hours, we cannot state this relationship with absolute certainty. However, given the data available, we believe that our analysis provides the best possible approximation. Subsequent studies should incorporate MBTA ridership data to provide a more precise understanding of the interplay between these two modes of transportation.

Future research should address these gaps by analyzing year-round data to capture seasonal fluctuations and including demographic factors to better understand equity issues in bike-sharing access. Additionally, integrating MBTA ridership data will provide a more comprehensive view of how Blue Bikes complement public transit systems. Based on our analyses and limitations, policymakers should also consider promoting electric bike usage and expanding stations near underutilized MBTA lines to improve accessibility and efficiency. These proposals can support a more inclusive urban transit network in Boston.

1. **Conclusion**

This study examined the interaction between Boston’s Blue Bikes and MBTA infrastructure to better understand usage patterns and their implications for multimodal transportation planning. Using spatial analysis through GIS and logistic regression models, we analyzed trip data from September 2024 to identify trends across Blue Bike stations classified as either close to or far from MBTA stops.

Our results show that Blue Bike members are the primary users for first/last-mile connections to MBTA stops, with stations near transit hubs experiencing the most activity during peak commuting hours. Classic bikes were the preferred choice for these shorter trips, while casual riders favored longer, exploratory rides. These results highlight the important role Blue Bikes play in urban mobility and the opportunity for further integration with MBTA services in the future.

To enhance system efficiency and accessibility, planners should prioritize expanding Blue Bike stations near MBTA stops, promoting membership programs, and addressing barriers to electric bike adoption. Future research can build on our findings by incorporating broader datasets and examining equity implications of bike-sharing systems to support more inclusive and effective urban transit solutions. These insights contribute to ongoing efforts to create a more connected and accessible transportation network in Boston.

1. **References**
2. Romm, D., et al. (2022). Differences in first-mile and last-mile behaviour in candidate multi-modal Boston bike-share micromobility trips. *Journal of Transport Geography*. Pergamon. Retrieved from<https://www.sciencedirect.com/science/article/pii/S096669232200093X#s0010>
3. Lu, M., et al. "Analysis of Spatial Interactions Among Shared E-Scooters, Shared Bikes, and Public Transit." Journal of Intelligent Transportation Systems, vol. 28, no. 4, 2023, pp. 587–603, <https://doi.org/10.1080/15472450.2023.2174803>.
4. Tang, H., & Zhou, D. (2024). Understanding the competition and cooperation between dockless bike-sharing and metro systems in view of mobility. *Sustainability*, *16*(13), Article 5780. Retrieved from<https://www.mdpi.com/2071-1050/16/13/5780>
5. **Appendix**
6. **Appendix**

---

title: "FinalDataEDA"

output: html\_document

date: "2024-12-03"

---

# EDA

```

```{r}

# Load necessary libraries

library(dplyr)

library(readr)

library(ggplot2)

```

```{r}

# Define custom colors for MBTA lines

mbta\_colors <- c(

"BLUE" = "blue",

"BLUE/GREEN" = "cyan",

"BLUE/ORANGE" = "coral",

"GREEN" = "darkgreen",

"GREEN/ORANGE" = "brown",

"GREEN/RED" = "green",

"ORANGE" = "orange",

"RED" = "red",

"SILVER" = "gray"

)

```

```{r}

data <- read.csv("trip\_data\_with\_mbta.csv")

```

# Dataset Overview

```{r}

# Structure of dataset

str(data)

# Summarize the dataset

summary(data)

```

# Dataset View

```{r}

# view columns

colnames(data)

# view first few rows

head(data)

```

# Data Inaccuracies

```{r}

# Check for null values

cat("Null values in each column:\n")

print(colSums(is.na(data)))

# Remove rows with null values

data <- na.omit(data)

# Verify if null values are removed

cat("Are there any null values left? ", any(is.na(data)), "\n")

```

#Descriptive Statistics

```{r}

aggregate(trip\_duration ~ member\_casual, data, mean)

```

#Proximity Analysis

```{r}

# Start Station Proximity

table(data$member\_casual, data$start\_station\_proximity)

```

```{r}

# End Station Proximity

table(data$member\_casual, data$end\_station\_proximity)

```

# Member vs. Casual Trends → Trip Duration

\*Analyze and compare the average trip duration for members and casual riders\*

```{r}

# Summarize trip duration for member and casual riders

member\_casual\_summary <- data %>%

group\_by(member\_casual) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

median\_trip\_duration = median(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(member\_casual\_summary)

```

\*Explore trip duration distributions for each rider type.\*

```{r}

# Trip duration distribution by rider type

member\_casual\_distribution <- data %>%

group\_by(member\_casual) %>%

summarise(

min\_duration = min(trip\_duration, na.rm = TRUE),

max\_duration = max(trip\_duration, na.rm = TRUE),

sd\_duration = sd(trip\_duration, na.rm = TRUE)

)

print(member\_casual\_distribution)

```

# Time Trends → Trip Duration by Hour

\*Breakdown of hourly trends\*

```{r}

# Extract hour from the timestamp

data$hour <- format(as.POSIXct(data$started\_at), "%H")

# Summarize trip duration by hour

time\_trends <- data %>%

group\_by(hour) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

median\_trip\_duration = median(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(time\_trends)

```

\*Break down hourly trends by rider type (member vs. casual).\*

```{r}

# Hourly trends by rider type

hourly\_member\_casual <- data %>%

group\_by(hour, member\_casual) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(hourly\_member\_casual)

```

\*Analyze how trip duration changes throughout the week (weekday vs. weekend).\*

```{r}

# Weekday vs Weekend trends

data$day\_type <- ifelse(weekdays(as.Date(data$started\_at)) %in% c("Saturday", "Sunday"), "Weekend", "Weekday")

weekday\_weekend\_trends <- data %>%

group\_by(day\_type) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(weekday\_weekend\_trends)

# Plot: Weekday vs Weekend Trends in Trip Duration

ggplot(weekday\_weekend\_trends, aes(x = day\_type, y = mean\_trip\_duration, fill = day\_type)) +

geom\_bar(stat = "identity") +

ggtitle("Weekday vs Weekend Trends in Trip Duration") +

xlab("Day Type") +

ylab("Average Trip Duration (minutes)") +

theme\_minimal()

```

\*Analyze how trip duration changes throughout the week (weekday vs. weekend) by rider type.\*

```{r}

# Weekday vs Weekend trends

weekday\_weekend\_trends <- data %>%

group\_by(day\_type, member\_casual) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(weekday\_weekend\_trends)

# Plot: Weekday vs Weekend Trends in Trip Duration

ggplot(weekday\_weekend\_trends, aes(x = day\_type, y = mean\_trip\_duration, fill = member\_casual)) +

geom\_bar(stat = "identity", position = "dodge") +

ggtitle("Weekday vs Weekend Trends in Trip Duration") +

xlab("Day Type") +

ylab("Average Trip Duration (minutes)") +

theme\_minimal()

```

\*Compare trip durations during peak hours (morning/evening) vs. off-peak hours.\*

```{r}

# Peak hours (7-10 AM, 4-7 PM) vs off-peak

data$peak\_offpeak <- ifelse(data$hour %in% c("07", "08", "09", "10", "16", "17", "18", "19"), "Peak", "Off-Peak")

peak\_offpeak\_trends <- data %>%

group\_by(peak\_offpeak, member\_casual) %>%

summarise(

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

trip\_count = n()

)

print(peak\_offpeak\_trends)

# Plot: Peak vs Off-Peak Trip Trends by Rider Type

ggplot(peak\_offpeak\_trends, aes(x = peak\_offpeak, y = mean\_trip\_duration, fill = member\_casual)) +

geom\_bar(stat = "identity", position = "dodge") +

ggtitle("Peak vs Off-Peak Trends in Trip Duration by Rider Type") +

xlab("Time Period") +

ylab("Average Trip Duration (minutes)") +

theme\_minimal()

```

# Proximity Trends → Close vs. Far Trip Counts

\*Compare start station proximity for close vs far trip counts\*

```{r}

# Summarize trip counts by proximity

proximity\_trends <- data %>%

group\_by(start\_station\_proximity) %>%

summarise(trip\_count = n())

print(proximity\_trends)

```

# MBTA Line Trends → Trip Counts

\*Compare different MBTA Lines usage through trip count\*

```{r}

# Summarize trip counts by MBTA line

mbta\_line\_trends <- data %>%

group\_by(start\_mbta\_line) %>%

summarise(trip\_count = n())

print(mbta\_line\_trends)

```

## RESEARCH QUESTION BASED EDA (used for presentation)

# Question 1: How does the activity at Blue Bike stations far from MBTA stops differ from stations that are close?

```{r}

# Summarize trip count by proximity

proximity\_activity <- data %>%

group\_by(start\_station\_proximity) %>%

summarise(

trip\_count = n(),

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

median\_trip\_duration = median(trip\_duration, na.rm = TRUE)

)

print(proximity\_activity)

# Plot: Trip Count and Duration by Proximity

ggplot(proximity\_activity, aes(x = start\_station\_proximity)) +

geom\_bar(aes(y = trip\_count, fill = start\_station\_proximity), stat = "identity") +

ggtitle("Activity at Blue Bike Stations by Proximity to MBTA Stops") +

xlab("Proximity to MBTA") +

ylab("Trip Count") +

theme\_minimal()

```

```{r}

# Count the number of Blue Bike stations by proximity

stations\_proximity <- data %>%

group\_by(start\_station\_proximity) %>%

summarise(station\_count = n\_distinct(start\_station\_name))

print(stations\_proximity)

# Plot: Blue Bike Stations by Proximity to MBTA Stops

ggplot(stations\_proximity, aes(x = start\_station\_proximity, y = station\_count, fill = start\_station\_proximity)) +

geom\_bar(stat = "identity") +

ggtitle("Blue Bike Stations by Proximity to MBTA Stops") +

xlab("Proximity to MBTA") +

ylab("Number of Blue Bike Stations") +

theme\_minimal() +

theme(legend.position = "none")

```

# Question 2: What MBTA lines/stops are most popular with Blue Bike stations?

```{r}

# Count trips by MBTA line

mbta\_line\_popularity <- data %>%

group\_by(start\_mbta\_line) %>%

summarise(

trip\_count = n()

) %>%

arrange(desc(trip\_count))

print(mbta\_line\_popularity)

# Plot: Popularity of MBTA Lines

ggplot(mbta\_line\_popularity, aes(x = reorder(start\_mbta\_line, -trip\_count), y = trip\_count, fill = start\_mbta\_line)) +

geom\_bar(stat = "identity") +

ggtitle("Popularity of MBTA Lines with Blue Bike Stations") +

xlab("MBTA Line") +

ylab("Trip Count") +

scale\_fill\_manual(values = mbta\_colors) + # Apply custom colors

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

```

# Question 3: Does Blue Bike usage align with the expected travel commute times of MBTA users (7-9am and 5-7pm during the week?

```{r}

# Extract day of the week and hour from the dataset

data$day\_of\_week <- weekdays(as.Date(data$started\_at))

data$hour <- as.numeric(format(as.POSIXct(data$started\_at), "%H"))

# Categorize time into peak commute and non-peak

data$commute\_period <- ifelse(data$hour %in% c(7, 8, 9, 17, 18, 19), "Commute", "Non-Commute")

# Analyze usage during commute vs. non-commute times

commute\_analysis <- data %>%

group\_by(commute\_period, member\_casual) %>%

summarise(

trip\_count = n(),

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE)

)

print(commute\_analysis)

# Plot: Commute vs Non-Commute Usage Rider Type Comparison (Bar Plot)

ggplot(data, aes(x = commute\_period, fill = member\_casual)) +

geom\_bar(position = "dodge") +

ggtitle("Blue Bike Usage During Commute vs Non-Commute Times") +

xlab("Time Period") +

ylab("Trip Count") +

theme\_minimal()

# Ensure day\_of\_week is in chronological order

data$day\_of\_week <- factor(data$day\_of\_week,

levels = c("Monday", "Tuesday", "Wednesday", "Thursday",

"Friday", "Saturday", "Sunday"))

# Plot: Daily Usage Rider Type Comparison (Bar Plot)

ggplot(data, aes(x = day\_of\_week, fill = member\_casual)) +

geom\_bar(position = "dodge", stat = "count") +

ggtitle("Daily Blue Bike Usage by Rider Type") +

xlab("Day of Week") +

ylab("Number of Trips") +

theme\_minimal()

```

# Additionally, does member usage of Blue Bikes decrease on the weekend?

```{r}

# Categorize days into weekday vs weekend

data$day\_type <- ifelse(data$day\_of\_week %in% c("Saturday", "Sunday"), "Weekend", "Weekday")

# Analyze member and casual usage by day type

member\_usage\_weekend <- data %>%

group\_by(day\_type, member\_casual) %>%

summarise(

trip\_count = n(),

mean\_trip\_duration = mean(trip\_duration, na.rm = TRUE)

)

print(member\_usage\_weekend)

# Plot: Member and Casual Usage by Day Type

ggplot(member\_usage\_weekend, aes(x = day\_type, y = trip\_count, fill = member\_casual)) +

geom\_bar(stat = "identity", position = "dodge") +

ggtitle("Blue Bike Usage by Day Type (Weekday vs Weekend)") +

xlab("Day Type") +

ylab("Trip Count") +

theme\_minimal()

# Plot: Blue Bike Usage by Hour and Day

usage\_heatmap <- data %>%

group\_by(day\_of\_week, hour) %>%

summarise(trip\_count = n())

ggplot(usage\_heatmap, aes(x = hour, y = day\_of\_week, fill = trip\_count)) +

geom\_tile() +

ggtitle("Heatmap of Blue Bike Usage by Hour and Day") +

xlab("Hour of Day") +

ylab("Day of Week") +

scale\_fill\_gradient(low = "lightblue", high = "darkblue") +

theme\_minimal()

# Analysis

```{r}

# Mean trip duration by proximity

trip\_durations <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_proximity, end\_station\_proximity) %>%

summarise(mean\_duration = mean(trip\_duration, na.rm = TRUE))

# View results

print(trip\_durations)

```

```{r}

library(lubridate)

# Extract hour from start time

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(hour = hour(as.POSIXct(started\_at)))

# Trip counts by hour

hourly\_trends <- trip\_data\_with\_mbta %>%

group\_by(hour) %>%

summarise(total\_trips = n())

# View results

print(hourly\_trends)

```

```{r}

# Extract day of week

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(day\_of\_week = wday(as.POSIXct(started\_at), label = TRUE))

# Trip counts by day of week

daily\_trends <- trip\_data\_with\_mbta %>%

group\_by(day\_of\_week) %>%

summarise(total\_trips = n())

# View results

print(daily\_trends)

```

```{r}

# Trips by MBTA line

trips\_by\_line <- trip\_data\_with\_mbta %>%

group\_by(start\_mbta\_line, end\_mbta\_line) %>%

summarise(total\_trips = n())

# View results

print(trips\_by\_line)

```

```{r}

# Top MBTA stations

top\_mbta\_stations <- trip\_data\_with\_mbta %>%

group\_by(start\_mbta\_station, end\_mbta\_station) %>%

summarise(total\_trips = n()) %>%

arrange(desc(total\_trips))

# View results

print(top\_mbta\_stations)

```

```{r}

# User type analysis

user\_type\_analysis <- trip\_data\_with\_mbta %>%

group\_by(member\_casual, start\_station\_proximity, end\_station\_proximity) %>%

summarise(total\_trips = n())

# View results

print(user\_type\_analysis)

```

```{r}

duration\_by\_user <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_proximity, end\_station\_proximity, member\_casual) %>%

summarise(mean\_duration = mean(trip\_duration, na.rm = TRUE))

print(duration\_by\_user)

```

```{r}

# Group by start/end stations and user type

popular\_routes\_by\_user <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_name, end\_station\_name, member\_casual) %>%

summarise(total\_trips = n(), .groups = "drop") %>%

arrange(desc(total\_trips))

# View top 10 routes by user type

head(popular\_routes\_by\_user, 10)

```

```{r}

# Filter for routes involving Close proximity stations

close\_routes <- trip\_data\_with\_mbta %>%

filter(start\_station\_proximity == "Close" | end\_station\_proximity == "Close") %>%

group\_by(start\_station\_name, end\_station\_name) %>%

summarise(total\_trips = n(), .groups = "drop") %>%

arrange(desc(total\_trips))

# View top 10 Close proximity routes

head(close\_routes, 10)

```

```{r}

library(readxl)

blue\_bike\_stations <- read\_excel("Blue\_Bike\_Stations.xlsx") # Replace with the correct file path

# Clean Blue Bike Stations data

blue\_bike\_stations <- blue\_bike\_stations %>%

select(Name, District) %>% # Select relevant columns

rename(station\_name = Name) %>% # Rename column for consistency

mutate(station\_name = tolower(trimws(station\_name))) # Standardize names

# Clean the main dataset

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

start\_station\_name = tolower(trimws(start\_station\_name)),

end\_station\_name = tolower(trimws(end\_station\_name))

)

# Join district information for start stations

trip\_data\_with\_districts <- trip\_data\_with\_mbta %>%

left\_join(blue\_bike\_stations, by = c("start\_station\_name" = "station\_name")) %>%

rename(start\_station\_district = District)

# Join district information for end stations

trip\_data\_with\_districts <- trip\_data\_with\_districts %>%

left\_join(blue\_bike\_stations, by = c("end\_station\_name" = "station\_name")) %>%

rename(end\_station\_district = District)

# View the updated dataset

head(trip\_data\_with\_districts)

```

```{r}

# Count trips by start district

start\_district\_counts <- trip\_data\_with\_districts %>%

group\_by(start\_station\_district) %>%

summarise(total\_trips = n(), .groups = "drop") %>%

arrange(desc(total\_trips))

# Count trips by end district

end\_district\_counts <- trip\_data\_with\_districts %>%

group\_by(end\_station\_district) %>%

summarise(total\_trips = n(), .groups = "drop") %>%

arrange(desc(total\_trips))

# View results

print("Start District Trip Counts:")

print(start\_district\_counts)

print("End District Trip Counts:")

print(end\_district\_counts)

```

```{r}

# Group by start district, end district, and user type

district\_user\_counts <- trip\_data\_with\_districts %>%

group\_by(start\_station\_district, end\_station\_district, member\_casual) %>%

summarise(total\_trips = n(), .groups = "drop") %>%

arrange(desc(total\_trips))

# View the results

print(district\_user\_counts)

```

```{r}

library(dplyr)

library(lubridate)

# Add an hour column

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(hour = hour(as.POSIXct(started\_at)),

peak\_commute = ifelse(

(hour >= 8 & hour < 9 | hour >= 17 & hour < 18) &

(start\_station\_proximity == "Close" | end\_station\_proximity == "Close"),

1,

0 # Binary dependent variable

))

# Convert categorical variables to factors

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

member\_casual = as.factor(member\_casual),

rideable\_type = as.factor(rideable\_type)

)

```

```{r}

# Fit the logistic regression model

logit\_model <- glm(

peak\_commute ~ member\_casual + rideable\_type + trip\_duration,

data = trip\_data\_with\_mbta,

family = binomial()

)

# Summarize the model

summary(logit\_model)

```

```{r}

library(dplyr)

library(lubridate)

# Add weekday column

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

hour = hour(as.POSIXct(started\_at)), # Extract hour

day\_of\_week = wday(as.POSIXct(started\_at), label = TRUE, week\_start = 1), # Extract weekday

is\_weekday = ifelse(day\_of\_week %in% c("Mon", "Tue", "Wed", "Thu", "Fri"), 1, 0), # Binary for weekdays

# Define new binary variable for Close station trips during peak hours on weekdays

peak\_commute\_weekday = ifelse(

is\_weekday == 1 &

(hour >= 8 & hour < 9 | hour >= 17 & hour < 18) &

(start\_station\_proximity == "Close" | end\_station\_proximity == "Close"),

1,

0

)

)

# Convert categorical variables to factors

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

member\_casual = as.factor(member\_casual),

rideable\_type = as.factor(rideable\_type)

)

# Re-run the logistic regression model

logit\_model\_weekday <- glm(

peak\_commute\_weekday ~ member\_casual + rideable\_type + trip\_duration,

data = trip\_data\_with\_mbta,

family = binomial()

)

# Summarize the model

summary(logit\_model\_weekday)

```

```{r}

library(dplyr)

# Morning commute: Trip ends at a Close station

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

morning\_commute = ifelse(

is\_weekday == 1 & hour >= 8 & hour < 9 & end\_station\_proximity == "Close",

1,

0

),

# Evening commute: Trip starts at a Close station

evening\_commute = ifelse(

is\_weekday == 1 & hour >= 17 & hour < 18 & start\_station\_proximity == "Close",

1,

0

)

)

```

```{r}

# Logistic regression for morning commutes

logit\_model\_morning <- glm(

morning\_commute ~ member\_casual + rideable\_type + trip\_duration,

data = trip\_data\_with\_mbta,

family = binomial()

)

summary(logit\_model\_morning)

# Generate predicted probabilities from the logistic regression model

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

predicted\_prob = predict(logit\_model\_morning, type = "response"), # Predicted probabilities

predicted\_class = ifelse(predicted\_prob >= 0.5, 1, 0) # Convert probabilities to binary class

)

# Calculate accuracy

accuracy <- mean(trip\_data\_with\_mbta$predicted\_class == trip\_data\_with\_mbta$morning\_commute)

print(paste("Model Accuracy:", round(accuracy \* 100, 2), "%"))

```

```{r}

# Logistic regression for evening commutes

logit\_model\_evening <- glm(

evening\_commute ~ member\_casual + rideable\_type + trip\_duration,

data = trip\_data\_with\_mbta,

family = binomial()

)

summary(logit\_model\_evening)

# Generate predicted probabilities from the evening commute logistic regression model

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

predicted\_prob\_evening = predict(logit\_model\_evening, type = "response"), # Predicted probabilities

predicted\_class\_evening = ifelse(predicted\_prob\_evening >= 0.5, 1, 0) # Convert probabilities to binary class

)

# Calculate accuracy for evening commute

accuracy\_evening <- mean(trip\_data\_with\_mbta$predicted\_class\_evening == trip\_data\_with\_mbta$evening\_commute)

print(paste("Evening Commute Model Accuracy:", round(accuracy\_evening \* 100, 2), "%"))

```

# CLUSTERING ANALYSIS

```{r}

# Join to get Total docks per start station

library(readxl)

bluebike\_mbta <- read\_excel("BlueBikeStations\_MBTAStops.xlsx")

# Rename columns for clarity

bluebike\_mbta <- bluebike\_mbta %>%

rename(

station\_name = BlueBikeStation,

latitude = Latitude,

longitude = Longitude,

total\_docks = Total\_dock,

mbta\_station = `MBTA Station`,

mbta\_line = `MBTA Line`

)

# Standardize station names in bluebike\_mbta

bluebike\_mbta <- bluebike\_mbta %>%

mutate(station\_name = tolower(trimws(station\_name)))

# Join Total\_dock and other features to trip\_data\_with\_mbta for start and end stations

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

left\_join(bluebike\_mbta %>% select(station\_name, total\_docks),

by = c("start\_station\_name" = "station\_name")) %>%

rename(start\_station\_docks = total\_docks)

# View result

head(trip\_data\_with\_mbta)

```

```{r}

library(dplyr)

# Aggregate features by station

station\_features <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_name) %>%

summarise(

total\_trips = n(), # Total trips starting at the station

peak\_trip\_proportion = mean(hour %in% c(8, 9, 17, 18), na.rm = TRUE), # Proportion of trips during peak hours

member\_proportion = mean(member\_casual == "member", na.rm = TRUE), # Proportion of trips by members

avg\_trip\_duration = mean(trip\_duration, na.rm = TRUE), # Average trip duration

total\_docks = mean(start\_station\_docks, na.rm = TRUE), # Total docks at the station

close\_to\_mbta = ifelse(mean(start\_station\_proximity == "Close", na.rm = TRUE) > 0.5, 1, 0) # Binary for proximity

)

# View the aggregated data

head(station\_features)

```

```{r}

# Normalize features

station\_features <- station\_features %>%

mutate(

total\_trips\_scaled = scale(total\_trips),

peak\_trip\_proportion\_scaled = scale(peak\_trip\_proportion),

member\_proportion\_scaled = scale(member\_proportion),

avg\_trip\_duration\_scaled = scale(avg\_trip\_duration),

total\_docks\_scaled = scale(total\_docks)

)

clustering\_data <- station\_features %>%

select(

total\_trips\_scaled,

peak\_trip\_proportion\_scaled,

member\_proportion\_scaled,

avg\_trip\_duration\_scaled,

total\_docks\_scaled,

close\_to\_mbta # Include as is (binary feature)

)

```

```{r}

set.seed(123)

# Compute total within-cluster sum of squares for different k values

wss <- sapply(1:10, function(k) {

kmeans(clustering\_data, centers = k, nstart = 25)$tot.withinss

})

# Plot the Elbow Method

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,

xlab = "Number of Clusters (k)",

ylab = "Total Within-Cluster Sum of Squares",

main = "Elbow Method for Optimal k")

```

```{r}

set.seed(123)

# Run k-means clustering with 3 clusters

kmeans\_model <- kmeans(clustering\_data, centers = 3, nstart = 25)

# Add cluster assignments to the dataset

station\_features$cluster <- as.factor(kmeans\_model$cluster)

library(ggplot2)

ggplot(station\_features, aes(x = total\_trips\_scaled, y = member\_proportion\_scaled, color = cluster)) +

geom\_point(size = 3) +

theme\_minimal() +

labs(

title = "Clusters of Bluebike Stations",

x = "Total Trips (Scaled)",

y = "Member Proportion (Scaled)",

color = "Cluster"

)

cluster\_centers <- as.data.frame(kmeans\_model$centers)

rownames(cluster\_centers) <- paste0("Cluster\_", 1:nrow(cluster\_centers))

print(cluster\_centers)

```

```{r}

cluster\_summary <- station\_features %>%

group\_by(cluster) %>%

summarise(

avg\_total\_trips = mean(total\_trips),

avg\_peak\_trip\_proportion = mean(peak\_trip\_proportion),

avg\_member\_proportion = mean(member\_proportion),

avg\_trip\_duration = mean(avg\_trip\_duration),

avg\_total\_docks = mean(total\_docks),

pct\_close\_to\_mbta = mean(close\_to\_mbta) \* 100

)

print(cluster\_summary)

```

```{r}

# Melt the cluster centers for visualization

library(reshape2)

cluster\_centers <- as.data.frame(kmeans\_model$centers)

cluster\_centers$cluster <- rownames(cluster\_centers)

melted\_centers <- melt(cluster\_centers, id.vars = "cluster")

# Bar plot of cluster centers

ggplot(melted\_centers, aes(x = variable, y = value, fill = cluster)) +

geom\_bar(stat = "identity", position = "dodge") +

theme\_minimal() +

labs(

title = "Cluster Centers by Features",

x = "Feature",

y = "Scaled Value",

fill = "Cluster"

) +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

```

# LINEAR REGRESSION

```{r}

library(dplyr)

# Aggregate data by start station

station\_trip\_frequency <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_name, start\_station\_proximity) %>%

summarise(

total\_trips = n(),

avg\_hour = mean(hour, na.rm = TRUE),

avg\_weekday = mean(is\_weekday, na.rm = TRUE), # Proportion of weekday trips

member\_proportion = mean(member\_casual == "member", na.rm = TRUE), # Proportion of trips by members

electric\_proportion = mean(rideable\_type == "electric\_bike", na.rm = TRUE), # Proportion of electric bike trips

.groups = "drop"

)

# View the aggregated dataset

head(station\_trip\_frequency)

```

```{r}

# Aggregate data by hour and day of week

time\_trip\_frequency <- trip\_data\_with\_mbta %>%

group\_by(hour, day\_of\_week, is\_weekday) %>%

summarise(

total\_trips = n(),

member\_proportion = mean(member\_casual == "member", na.rm = TRUE),

electric\_proportion = mean(rideable\_type == "electric\_bike", na.rm = TRUE),

.groups = "drop"

)

# View the aggregated dataset

head(time\_trip\_frequency)

```

```{r}

library(ggplot2)

ggplot(time\_trip\_frequency, aes(x = hour, y = total\_trips, color = is\_weekday)) +

geom\_line(size = 1.2) +

theme\_minimal() +

labs(

title = "Hourly Trip Trends by Weekday vs. Weekend",

x = "Hour of Day",

y = "Total Trips",

color = "Weekday"

)

```

```{r}

ggplot(time\_trip\_frequency, aes(x = day\_of\_week, y = total\_trips, fill = is\_weekday)) +

geom\_bar(stat = "identity", position = "dodge") +

theme\_minimal() +

labs(

title = "Trip Frequency by Day of Week",

x = "Day of Week",

y = "Total Trips",

fill = "Weekday"

)

```

```{r}

# Ensure necessary libraries are loaded

library(dplyr)

# Convert proximity to MBTA, user type, and bike type to factors (if not already)

trip\_data\_with\_mbta <- trip\_data\_with\_mbta %>%

mutate(

start\_station\_proximity = as.factor(start\_station\_proximity), # Proximity to MBTA (e.g., Close or Far)

member\_casual = as.factor(member\_casual), # User type (e.g., member or casual)

rideable\_type = as.factor(rideable\_type) # Bike type (e.g., electric or manual)

)

# Fit the linear regression model

lm\_duration <- lm(trip\_duration ~ start\_station\_proximity + member\_casual + rideable\_type,

data = trip\_data\_with\_mbta)

# View the model summary

summary(lm\_duration)

```

# Near vs. Far Blue Bike Stops

```{r}

library(dplyr)

# Aggregate activity metrics by proximity

proximity\_analysis <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_proximity) %>%

summarise(

total\_trips = n(),

avg\_trip\_duration = mean(trip\_duration, na.rm = TRUE),

member\_proportion = mean(member\_casual == "member", na.rm = TRUE),

electric\_proportion = mean(rideable\_type == "electric\_bike", na.rm = TRUE),

peak\_hour\_proportion = mean(hour %in% c(7, 8, 9, 17, 18, 19), na.rm = TRUE) # Peak commuting hours

)

# View the summary

print(proximity\_analysis)

```

```{r}

ggplot(trip\_data\_with\_mbta, aes(x = start\_station\_proximity, fill = member\_casual)) +

geom\_bar(position = "fill") +

theme\_minimal() +

labs(

title = "User Type Distribution by Station Proximity",

x = "Station Proximity to MBTA Stops",

y = "Proportion",

fill = "User Type"

)

```

```{r}

# Normalized comparison - average activity

# Number of stations in each proximity category

n\_close <- 93 # Replace with the actual count of "Close" stations

n\_far <- 319 # Replace with the actual count of "Far" stations

# Normalize total trips by number of stations

proximity\_analysis <- proximity\_analysis %>%

mutate(

trips\_per\_station = case\_when(

start\_station\_proximity == "Close" ~ total\_trips / n\_close,

start\_station\_proximity == "Far" ~ total\_trips / n\_far

)

)

# View normalized results

print(proximity\_analysis)

```

```{r}

library(dplyr)

# Aggregate trips by day of week and proximity

day\_of\_week\_analysis <- trip\_data\_with\_mbta %>%

group\_by(day\_of\_week, start\_station\_proximity) %>%

summarise(

total\_trips = n(), # Total trips for the group

num\_stations = ifelse(start\_station\_proximity == "Close", 93, 319), # Adjust for known station counts

trips\_per\_station = total\_trips / num\_stations, # Calculate trips per station

.groups = "drop"

)

library(ggplot2)

# Plot trips per station by day of the week and proximity

ggplot(day\_of\_week\_analysis, aes(x = day\_of\_week, y = trips\_per\_station, fill = start\_station\_proximity)) +

geom\_bar(stat = "identity", position = "dodge") +

theme\_minimal() +

labs(

title = "Daily Trips per Station by Proximity to MBTA Stops",

x = "Day of Week",

y = "Trips per Station",

fill = "Proximity"

)

```

# Casual Vs. Member Bike Usage

```{r}

library(dplyr)

# Aggregate metrics by user type

user\_type\_analysis <- trip\_data\_with\_mbta %>%

group\_by(member\_casual) %>%

summarise(

total\_trips = n(), # Total trips

avg\_trip\_duration = mean(trip\_duration, na.rm = TRUE), # Average trip duration

electric\_proportion = mean(rideable\_type == "electric\_bike", na.rm = TRUE), # Proportion of electric bike trips

peak\_hour\_proportion = mean(hour %in% c(7, 8, 9, 17, 18, 19), na.rm = TRUE), # Proportion of peak-hour trips

.groups = "drop"

)

# View the summary

print(user\_type\_analysis)

```

```{r}

library(dplyr)

# Aggregate total trips and metrics by user type

user\_type\_summary <- trip\_data\_with\_mbta %>%

group\_by(member\_casual) %>%

summarise(

total\_trips = n(), # Total number of trips

avg\_trip\_duration = mean(trip\_duration, na.rm = TRUE), # Average trip duration

electric\_proportion = mean(rideable\_type == "electric\_bike", na.rm = TRUE), # Proportion of electric bike trips

peak\_hour\_proportion = mean(hour %in% c(7, 8, 9, 17, 18, 19), na.rm = TRUE), # Proportion of peak-hour trips

.groups = "drop"

)

# View the results

print(user\_type\_summary)

```

# Aggregate Total Trips per station for GIS Analysis

```{r}

library(dplyr)

# Aggregate data by station name

aggregated\_station\_data <- trip\_data\_with\_mbta %>%

group\_by(start\_station\_name) %>%

summarise(

total\_trips = n(), # Count total trips for the station

avg\_lat = mean(start\_lat, na.rm = TRUE), # Calculate average latitude

avg\_lng = mean(start\_lng, na.rm = TRUE), # Calculate average longitude

.groups = "drop"

)

# View the aggregated dataset

head(aggregated\_station\_data)

# Write to a CSV file

write.csv(aggregated\_station\_data, "aggregated\_station\_data.csv", row.names = FALSE)

```

# Top MBTA Stops

```{r}

# Load necessary libraries

library(dplyr)

# Load the data

file\_path <- "aggregated\_station\_data\_w\_mbta.csv"

data <- read.csv(file\_path)

head(data)

# Summarize total trips by MBTA stop

top\_mbta\_stops <- data %>%

group\_by(STATION, LINE) %>%

summarise(total\_trips = sum(total\_trips, na.rm = TRUE)) %>%

arrange(desc(total\_trips)) %>%

head(10)

# View the top 10 MBTA stops with lines

print(top\_mbta\_stops)

```

```{r}

ggplot(top\_mbta\_stops, aes(x = reorder(STATION, total\_trips), y = total\_trips, fill = LINE)) +

geom\_bar(stat = "identity") +

coord\_flip() +

theme\_minimal() +

scale\_fill\_manual(

values = c(

"GREEN" = "green",

"RED" = "red"

)

) +

labs(

title = "Top 10 MBTA Stops by Blue Bike Trips",

x = "Total Trips",

y = "MBTA Stop",

fill = "MBTA Line"

)

```

```{r}

# Summarize total trips by MBTA stop

bottom\_mbta\_stops <- data %>%

group\_by(STATION, LINE) %>%

summarise(total\_trips = sum(total\_trips, na.rm = TRUE)) %>%

arrange(desc(total\_trips)) %>%

tail(10)

# View the top 10 MBTA stops with lines

print(bottom\_mbta\_stops)

```

1. D. Romm et al., “Differences in First-Mile and Last-Mile Behaviour in Candidate Multi-Modal Boston Bike-Share Micromobility Trips,” *Journal of Transport Geography*, Pergamon, June 7, 2022 [↑](#footnote-ref-1)
2. H. Tang and D. Zhou, “Understanding the Competition and Cooperation Between Dockless Bike-Sharing and Metro Systems in View of Mobility,” *Sustainability*, vol. 16, no. 13, 2024 [↑](#footnote-ref-2)
3. M. Lu et al., "Analysis of Spatial Interactions Among Shared E-Scooters, Shared Bikes, and Public Transit," *Journal of Intelligent Transportation Systems*, vol. 28, no. 4, 2023, pp. 587–603 [↑](#footnote-ref-3)
4. H. Tang and D. Zhou, “Understanding the Competition and Cooperation Between Dockless Bike-Sharing and Metro Systems in View of Mobility,” *Sustainability*, vol. 16, no. 13, 2024 [↑](#footnote-ref-4)