CITY OF BOSTON: TRANSIT AND PERFORMANCE - TEAM C DELIVERABLE 2

Analysis of MBTA Bus Performance and Demographics in Boston

Problem Statement

Public transportation plays a vital role in cities like Boston by providing mobility and access to jobs, services, healthcare, education, and amenities. However, historically not all communities have been served equally by public transit, often due to inequitable urban planning and infrastructure investments. This project aims to analyze the performance of MBTA bus routes across Boston in relation to the socioeconomic demographics of the communities they serve. The goal is to identify any disparities in service levels or quality and assess whether they may disproportionately impact marginalized groups. Specifically, this analysis will focus on examining MBTA bus operational data from January 2022, alongside Census demographic data aggregated at the neighborhood level in Boston. Key questions to be explored include: What are the end-to-end travel times for different bus routes across the city? Are there noticeable disparities in on-time performance and reliability between routes? What are the racial, economic, and age demographics of the communities primarily served by each bus route?

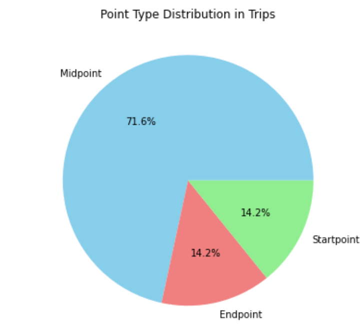
Public transit equity has broad implications for social justice, economic mobility, and environmental sustainability in diverse cities like Boston. This data-driven analysis aims to provide an objective, comprehensive understanding of where service gaps exist across bus routes, who they impact, and how they can be addressed. The goal is to promote discussion and progress towards an affordable, reliable transportation network that connects and serves all Boston communities.

Data Collection and Data Cleaning

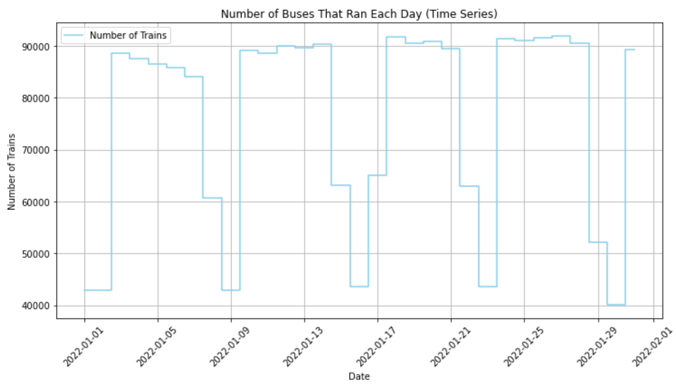
The first step in the data cleaning process was to join and filter the MBTA data to focus only on January 2022. This involved accessing the MBTA API to get real-time location, route, and schedule data for all bus trips during that month. The raw data was then filtered and aggregated to create a consolidated dataset for analysis. Next, key metrics like route travel times and on-time percentages were calculated. The scheduled and actual departure/arrival times in the MBTA data were used to compute total travel times for each route and trip. These travel times were averaged to get the mean time for each route. The scheduled versus actual times were also compared to calculate on-time performance.

When answering future questions, we will use spatial joins to associate the bus routes and stops with demographic data from the 2020 Census. We will first identify the census tracts that each stop is located in, and then we will link the tract-level population statistics to the stops and routes. This will allow us to append useful demographic data like income levels and race/ethnicity percentages to the MBTA data. Our MBTA dataset was checked for quality and missing values were handled prior to analysis.

Exploratory Data Analysis - MBTA Dataset

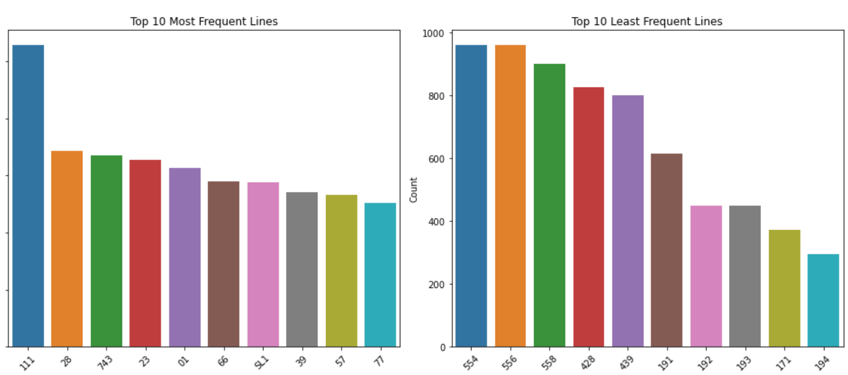


The graph on the left charts the proportions of measurements that are classified as either starting point, end points, or neither (midpoints). From this, we can see that over 28% of our collected data comes from either a starting or ending location. This possibly varies depending on the route (i.e., shorter routes may have less endpoints); however, it is possible that we have the same number of midpoints (5) for each route, with 1 start and 1 end point.



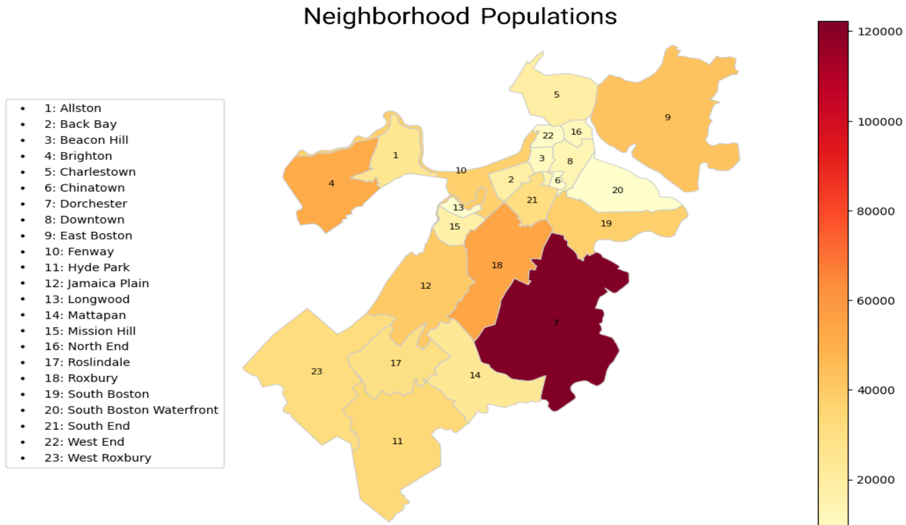
From the second graph given here, we can see the total number of bus routes that were run each day on the MBTA for the month of January. We can see a clear pattern where more bus routes (~90,000) are run during the week than on the weekends, and that Saturday (~60,000) has more bus routes than Sunday (~40,000).

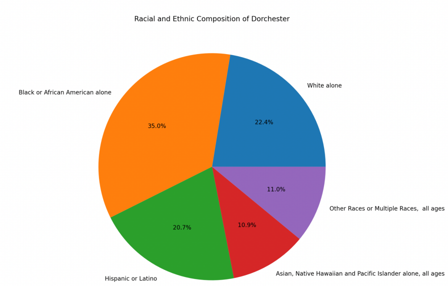
From the frequency plots given below, we can see the top 10 most frequent and least frequent lines by total count of routes run. We see that the "111" bus route is by far the most frequent, with a roughly 66% increase in total number of routes compared to the "28" route (the second-most frequent).. Additionally, we see that 192, 193, and 194 routes are listed as least frequent, only being run 350-400 times total.



Exploratory Data Analysis - Census Dataset

The plot given below is the population distribution in the 23 neighborhoods of the city of Boston. The darker portions represent a high density of population as compared to the lighter portions which represent lower population density. We can observe that the neighborhood of Dorchester has the highest population in the city of Boston.



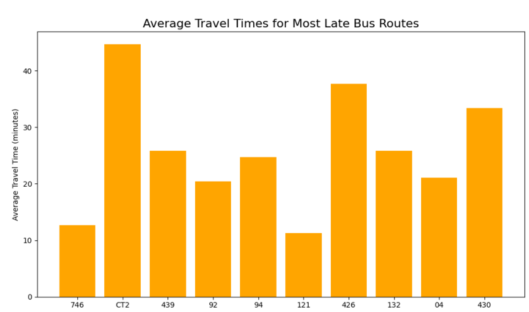


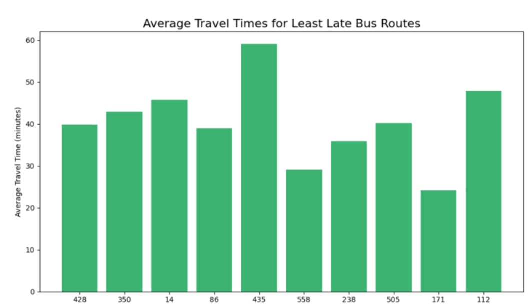
The pie chart here represents the ethnic distribution of the most populated neighborhood in Boston - Dorchester. We can infer from the pie chart that over 50% of the population is Black, White, or African American whereas the other 50% is composed of Hispanic, Latino, Asian or other races.

Visualizations and Key Insights for Addressing Base Questions

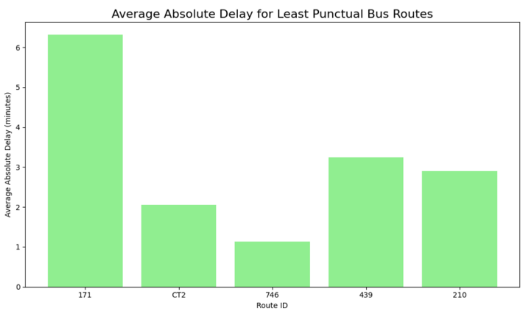
The end to end travel times and average travel times for the bus routes were calculated from the MBTA bus data. Two important factors for service disparities were defined - Lateness and Punctuality. Punctuality here is defined as the difference between actual departure time and scheduled departure time from start point. Lateness is measured as the difference between actual total travel time and scheduled total trip time.

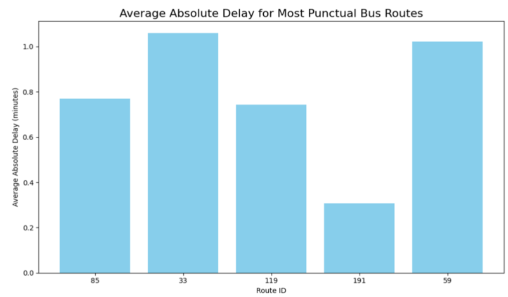
The average travel times for the least late and most late depicted in the bar plots given below. The average travel time for the most late bus routes ranges from 10 to 40 minutes whereas that of the least late bus routes ranges from around 25 to 60 minutes.



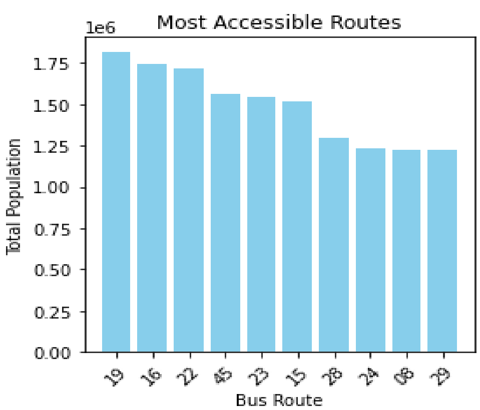


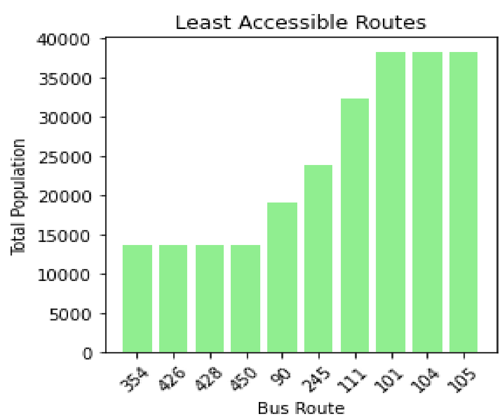
The service level disparities are visualized as the average absolute delay for the least punctual and most punctual bus routes in the below graphs. It can be inferred from the graphs that the average absolute delay for the least punctual bus routes is higher compared to the average absolute delay for the most punctual bus routes.

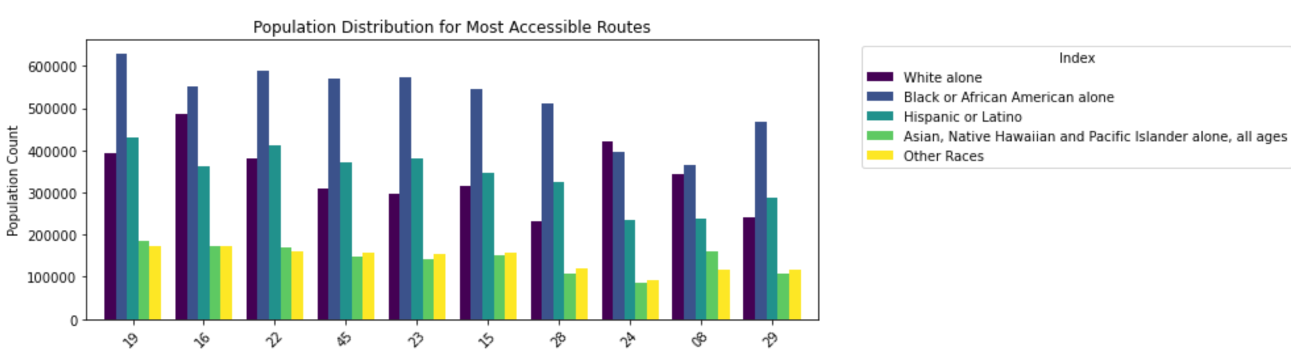


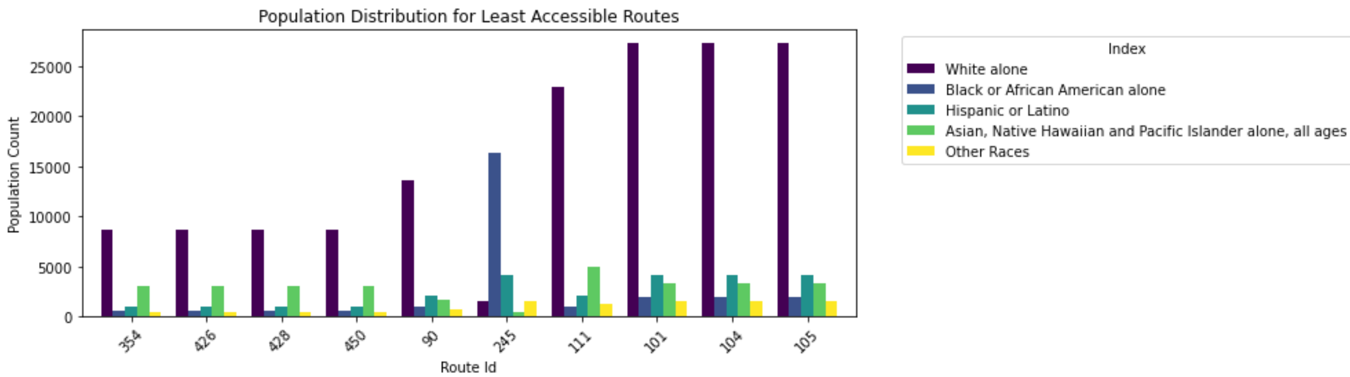


For further analysis, the MBTA data was correlated to the census data. The MBTA V3 API was used to get the latitude and longitude coordinates of the stops for all the routes. The stops were then mapped to the neighborhoods of the city of Boston using the City of Boston neighborhoods geojson file. Since the census data had information pertaining to the 23 neighborhoods of Boston city only while the MBTA data had route information for the entire state of Massachusetts, some stops belonging to other cities of Massachusetts could not be mapped to the neighborhoods. The graphs below depict total population and ethnic distribution of the most accessible and least accessible routes. Most accessible routes here are defined as the routes serving maximum number of people overall and least accessible routes mean the routes serving a small proportion of the population.

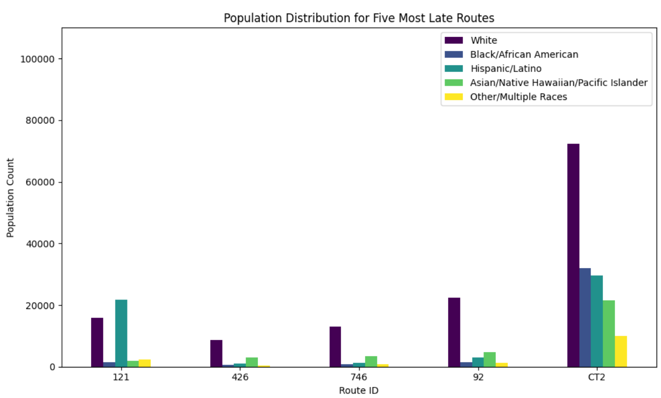


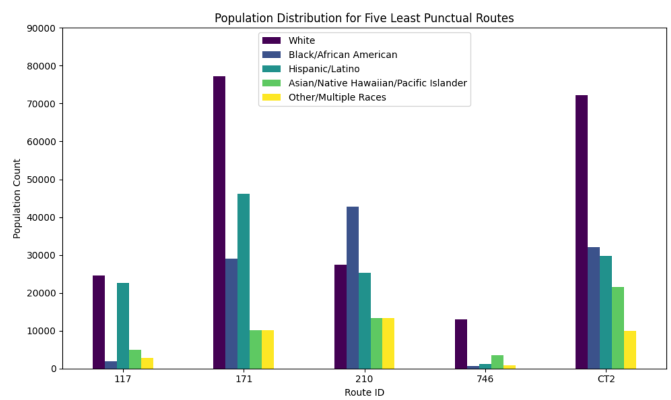




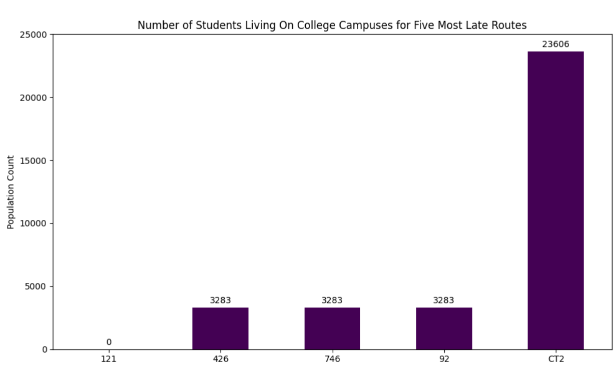


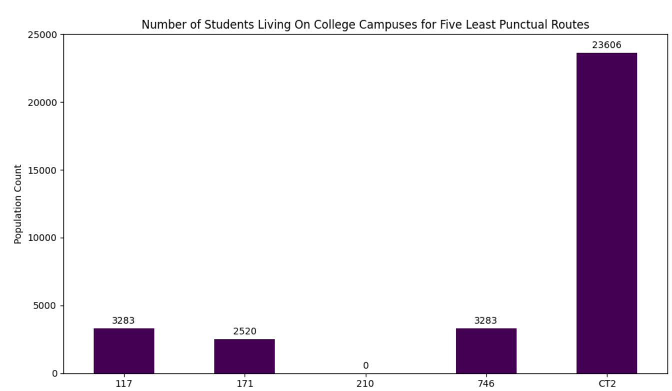
Further visualizations were done by analyzing the population sizes and characteristics of the communities affected by the most late and least punctual bus routes. The two grouped bar charts below depict the ethnic distribution of the population living in the areas served by the most late and least punctual bus routes.





The graphs below show the number of students who live on campus in areas served by the most late and least punctual bus routes. This correlation was done to analyze the extent of effect on the college students by the lateness and punctuality of the bus routes. The CT2 bus route is used by approximately 23,600 students, which is significantly higher than any other route in both the most late and least punctual categories. Routes 746, 92, and 117 each serve around 3,200 students and are present in both categories, suggesting high student usage regardless of the route's punctuality. Routes 121 and 210 have no students using them. The data highlights the importance of CT2 and could indicate a need to prioritize reliability improvements for routes with high ridership.

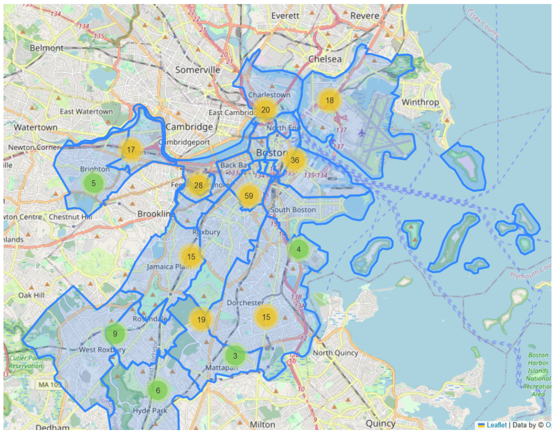


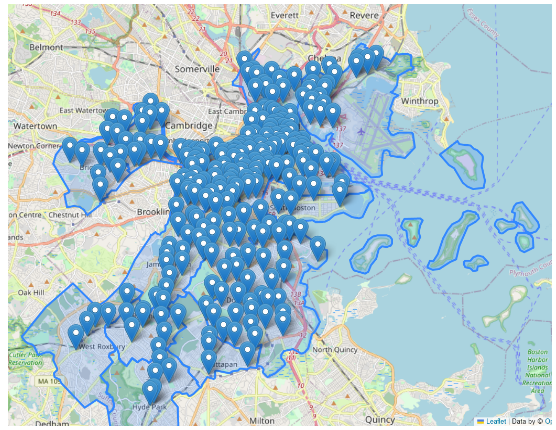


Extension Proposal

Extension Proposal	
Extension Pitch	*With our original topic, we analyzed MBTA trip data and Boston census data to draw conclusions about service disparities in different neighborhoods. For our extension, we will analyze potential disparities in Bluebike availability by neighborhood, and determine if there are enough stations to meet potential demand from different communities.*
Rationale	*After analyzing availability of MBTA bus data, it is important to consider alternative transportation options. Bluebikes are a public option that people can rent at an affordable price.*
Questions for Analysis	*We are curious about the number and locations of different stations, whether they are dispersed evenly throughout the city and/or concentrated in areas of high-population density, or if there are some areas with a lack of stations. H*
Datasets & Sources	*Bluebikes Stations* - *https://s3.amazonaws.com/hubway-data/current_bluebikes_stations.csv* *Blue Bikes Comprehensive Trip Histories* - *https://s3.amazonaws.com/hubway-data/202310-bluebikes-tripdata.zip*
Data Visualizations	*Proposed graphs include:* ● *Map of locations of each station within Boston area* ● *Bar graphs of number of stations per neighborhood* ● *Clustering of stations and/or trip data*
Additional Information	*Information on Bluebikes data - https://bluebikes.com/system-data* *Example references:* ● *https://cambridge-intelligence.com/geospatial-data-visualization-regraph-redwoodjs/* ● *https://www.chaossearch.io/blog/blue-bikes-data-dive-part-1*

Visualizations and Insights of Extension Proposal



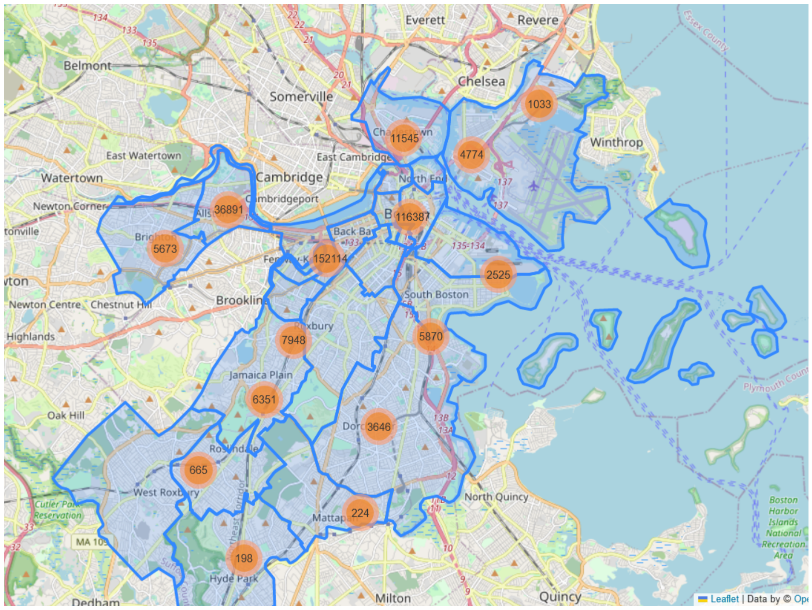


The first image depicts a map of Boston with markers representing the locations of Bluebikes stations throughout the city. At initial glance, it is apparent that the distribution of stations is not even across all neighborhoods. There is a high concentration of Bluebikes stations clustered in the downtown and central neighborhoods of Boston. Specifically, the area encompassing major attractions like the Boston Common, the waterfront, Fenway Park, and various universities has a very high density of stations. Within this central zone, the stations appear evenly dispersed along major roads and paths, providing convenient access for residents, workers, and tourists in these parts of the city.

However, the more residential outskirts and neighborhoods farther from downtown Boston have significantly fewer Bluebikes stations. Areas like East Boston, Roxbury, Mattapan, and Hyde Park have large swaths that lack bike share access entirely. Huge sections of these neighborhoods do not have a single station located nearby, which could severely limit transportation options for residents.

This imbalance in the location of Bluebikes stations across Boston neighborhoods can help us visualize and analyze potential disparities in availability and access in different communities. The map visualization makes it abundantly clear that large coverage gaps exist between downtown Boston and the outskirts. The distribution raises concerns about equity, as residents in certain areas may be excluded from utilizing the Bluebikes service simply due to lack of access. By displaying the precise geographic distribution, it will allow for clear identification of communities that are underserved by existing stations. This visualization and analysis sets the foundation for investigating why such access gaps exist, and how to best expand coverage to provide more equitable transportation options.

It appears that this map paints a picture of a bike share system highly concentrated in central commercial districts, with surrounding residential communities unable to enjoy the same benefits. This exemplifies why analyzing Bluebikes availability by neighborhood is an important undertaking, to ensure equal access across all of Boston.

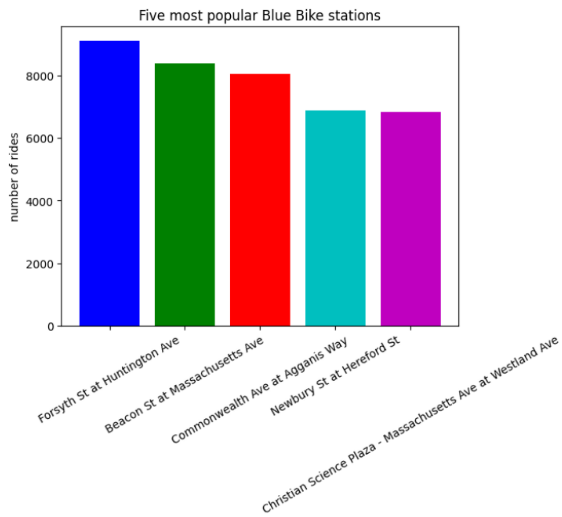


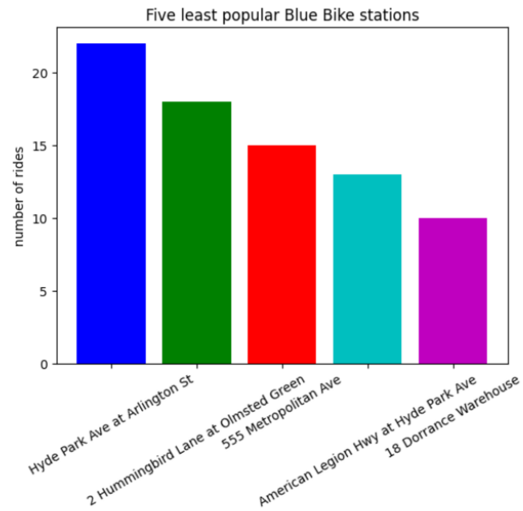
The map above provides a summary of Bluebikes ridership data for the month of September 2022. It shows that over 200,000 rides were taken on Bluebikes during that period, indicating that it is a widely-used and popular transportation option in Boston. With such high demand, it is important to analyze exactly where these rides are occurring.

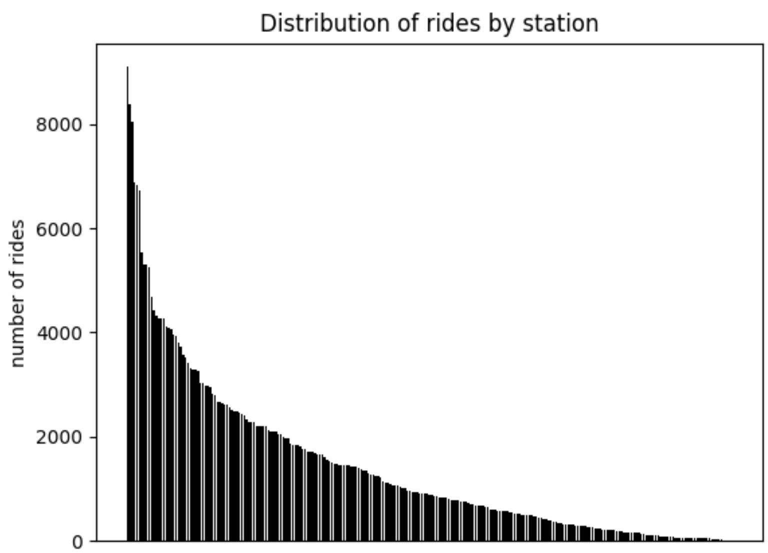
Mapping the origin and destination coordinates of each ride will illuminate which Boston neighborhoods generate the highest Bluebikes trip demand. Areas with a large number of rides originating could signify that residents rely heavily on bike share for transportation needs. Alternatively, neighborhoods with many ride endings may correlate to attractions or employment hubs. Comparing ridership volume to station locations will reveal if current infrastructure meets demand across all parts of the city.

Clustering rides by origin location would also help demonstrate if disparities exist between neighborhoods. Groupings of starting points may emerge in areas surrounding established stations. Meanwhile, communities with insufficient bike share access could generate fewer originating rides. This visualization can make inequality in ridership clear.

The comprehensive trip data enables multiple analyses to uncover potential inequities. Mapping precise ride locations, categorizing by rider status, and clustering by origins can quantify disparities in Bluebikes usage across Boston's diverse neighborhoods. This will pinpoint areas of unmet demand to guide future expansion and increased bike share accessibility.







The images and proposal context make it evident that while Bluebikes are widely used in Boston, their accessibility and ridership likely varies across different city neighborhoods. The planned analyses aim to quantify these potential disparities through several key steps.

First, mapping the locations of all Bluebikes stations will highlight any spatial inconsistencies in access between downtown, central, and outer residential areas. Clustering ridership data will then reveal which communities generate high demand versus low adoption. Joining these visuals with granular census demographics can uncover correlations between station availability, ride volume, and neighborhood factors like income levels.

Importantly, the proposal outlines filtering ride data by subscriber status and time period. This will provide valuable nuance, as neighborhoods with more members able to afford subscriptions may have higher average incomes. Analyzing different seasons could also reveal trends, like tourism boosting summer ridership in downtown.

Effective interactive data visualizations will be pivotal for conveying these complex disparities to audiences in a digestible way. Animated heat maps of ridership over time, charts comparing station counts, and clustered Point density maps can spotlight inequalities. The visual medium is key for impactful storytelling with data.

If clear disparities emerge from the analyses, the city should consider expanding Bluebikes to provide more equitable transportation access. But additional qualitative research through surveys or interviews may first be needed to understand why certain communities use or cannot access bike sharing. This can inform how best to increase adoption in underserved neighborhoods through changes to station locations, fee structures, or marketing outreach.

Overall, the thoughtful methodology proposed lays the groundwork for a comprehensive analysis of Bluebikes equity across all of Boston's diverse communities. Combining quantitative data, visualizations, and qualitative insights can enable both pinpointing and explaining any disparities that exist.

Team Member Contributions

Chandrahas and Manushi

- Data cleaning and preprocessing
- EDA of the MBTA dataset
- Correlating bus data and census data
- Addressing base questions 3,4,5

Munir and Eason

- EDA of census data
- Correlating bus data and census data
- Addressing base questions 1,2,3

Patrick

- Addressing base question 1
- Extension proposal
- Visualizations and EDA of extension project
- Preparation of results (report and presentation)