CS506 – DATA SCIENCE FUNDAMENTALS
NOVEMBER 20, 2023

DELIVERABLE 2 REPORT

CITY OF BOSTON TRANSIT PERFORMANCE - TEAM B

TEAM MEMBERS:
Vishvakishore Venkatesan, Suin Lee, Kevin Smith, Yufeng Song, Yu Han

PROBLEM STATEMENT:
Public transport in Boston, rooted in a rich history dating back to the 1600s, is currently managed by the MBTA whose network of ferries, trains, & buses serve over 1 million individuals daily and contribute an estimated economic value of 11.5 billion dollars annually to the greater Boston area. Recognizing its vital role in economic development, environmental sustainability, and equitable access, it is crucial to assess the performance of the MBTA services since they so directly impact the people of Boston.

The objective of this project is to conduct a data-driven analysis of the MBTA bus system's performance in the year 2022, aiming to uncover service quality trends in relation to geographical and demographic variance. Our data-driven insights will shed light on potential disparities among neighborhoods, providing a basis for informed decision-making. By examining the geographical distribution of bus performance, the analysis will offer valuable information to policymakers, city planners, and the MBTA itself, facilitating the identification of areas that may require targeted improvements.

OUR AVAILABLE DATA:
- **MBTA V3 API** (access to MBTA schedules, alerts, & real-time information)
- **MBTA Performance API** (wait times, station-to-station travel times, etc)
- **MBTA Historical Data Archive** (actual arrival & departure times from 2022)
- **2020 Boston Census** (demographics & geographical information)
- **MBTA_Systemwide_GTFS_Map** (containing extra information on disability access metics)

DATA COLLECTION & PREPROCESSING:
We began by collecting arrival & departure times over the entire year of 2022 from the MBTA Historical Data Archive as well as copies of the routes' schedules. We also pulled data on the stations in the network via calls to the V3 API - their locations, routes served, accessibility information, etc.

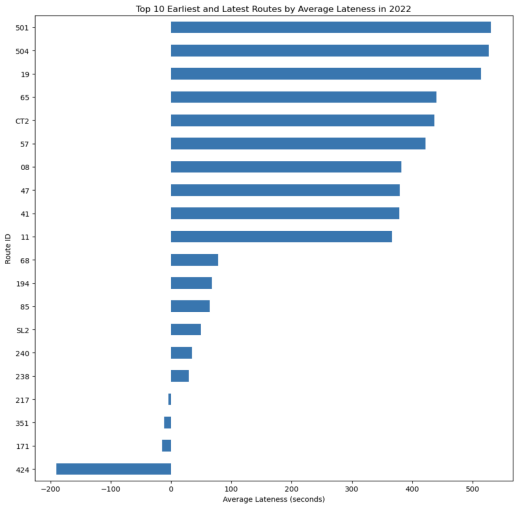
To clean the data ahead of analysis, we removed rows with mostly null values, then

filled remaining null values in numerical fields with route averages. Categorical or other text fields we left null and treated those nulls as unknowns in our calculations. This preprocessing step was essential not just in cleaning the data and making it easier to process, but it familiarized us with the data which we believe was essential to our understanding of the results of our analysis.

EXPLORATORY DATA ANALYSIS: Early Findings

Our team has successfully answered foundational questions by processing the "MBTA Bus Arrival Departure Times 2022" dataset from the MBTA official website and census data together with “Bus Stop” provided by MBTA. MBTA arrival and departure time offers detailed records of MBTA bus arrivals and departures up to the most recent month of 2022 and the census data provide more of the demographic information we need for the base questions.

"MBTA Bus Arrival Departure Times 2022" provides a meticulous account of the arrivals and departures of MBTA buses up to the latest complete month of 2022. Accompanying the dataset is a detailed data dictionary that elucidates the fields included, such as service date, route ID, direction ID, and various time points. The data spans several types, including date, time, string, and integer, offering a rich set of attributes for each trip. These attributes include both scheduled and actual departure times, as well as the headways — both scheduled and actual.



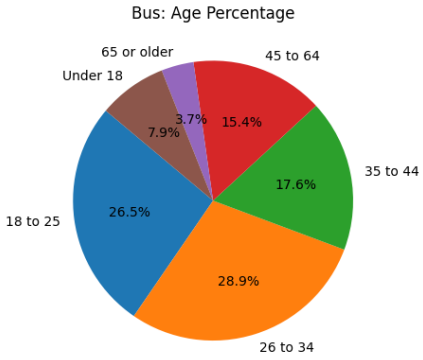
In our data processing, we focused on the following fields:

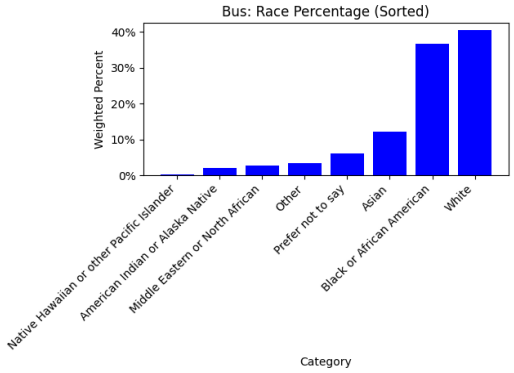
route_id: bus route
half_trip_id: a unique identifier for a specific bus trip, distinguishing between trips of the same route number that depart at different times
point_type: it denotes whether the bus is at the starting point, a midpoint, or the destination of its route
service_date: the date
actual: actual time
scheduled: scheduled time

Leveraging this dataset, we have unearthed critical insights into the complete travel times of different bus routes. We have pinpointed the top 10 most punctual and most delayed bus routes and have determined the average lateness percentage for each route.

Besides, to take a brief look at the demographic for the entire bus system, we went through a dataset called “MBTA 2022 System-Wide Passenger Survey”.

In this data processing, these fields were focused on:
service_mode: identify the MBTA service mode, and we got bus information from it.
measure_group: ethnicity, age.
weighted_percent: The proportion of riders

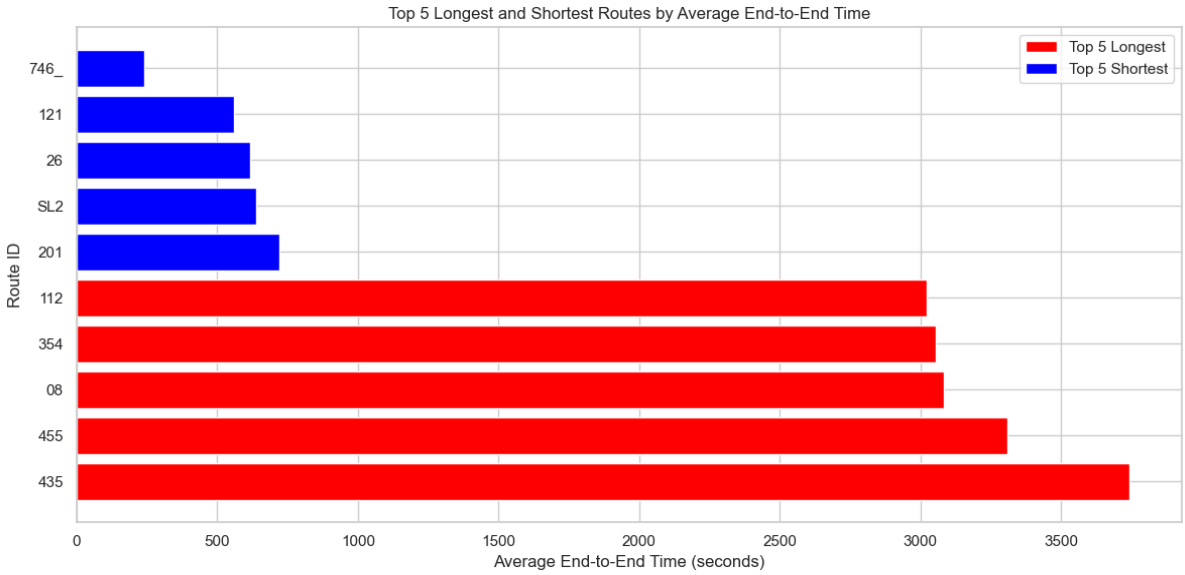




We plotted a pie chart for the distribution of age for the MBTA Bus system and a bar chart for the race distribution. From the pie chart, we see most of the bus riders are adults aged over 18, especially the group of people from 18 to 34 years old. From the bar chart, we can conclude that the majority of bus riders were white in 2022.
Overall, these findings led us to do more research and data processing on bus stop positions and bus routes' lateness and end-to-end time as well as how the distribution of groups with different ages and races is influenced by the bus routes.

Visualizations, methodology for finding underlying patterns, and insights for key base questions

1. What are the end-to-end travel times for different bus routes



Route ID 435

Route ID 746

From our data processing so far, we got some relevant information about the 2022 MBTA bus.

Here is the bar chart of the top 5 longest and shortest routes according to their average end-to-end time throughout the year 2022. From the chart, we can see that the shortest route id is 746 and the longest route id is 435.

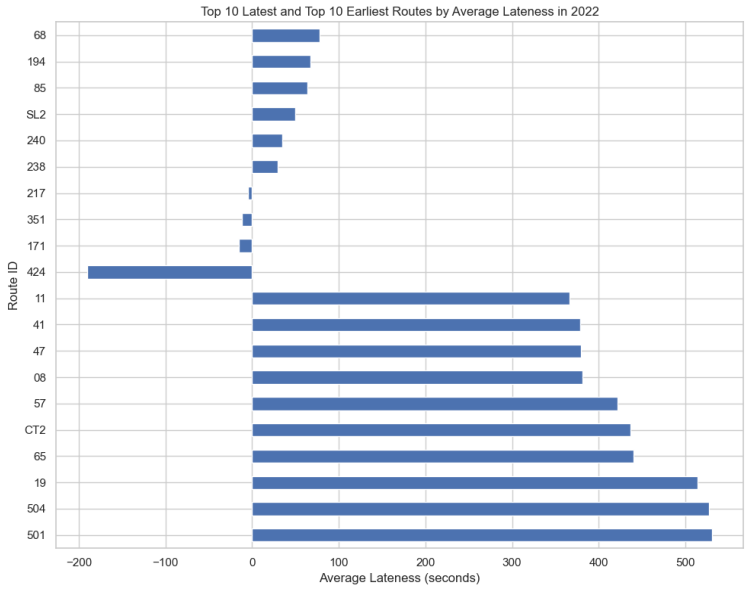
Based on the station locations for Route 435 combined with data provided by the Bus Stop dataset, it's evident that the route covers a significant distance. This accounts for its extended end-to-end travel time.

Conversely, Route 746 comprises just three stations, with minimal distance separating each. This proximity naturally results in a shorter end-to-end travel time.

2. Are there disparities in the service levels of different routes? (which lines are late more often than others)

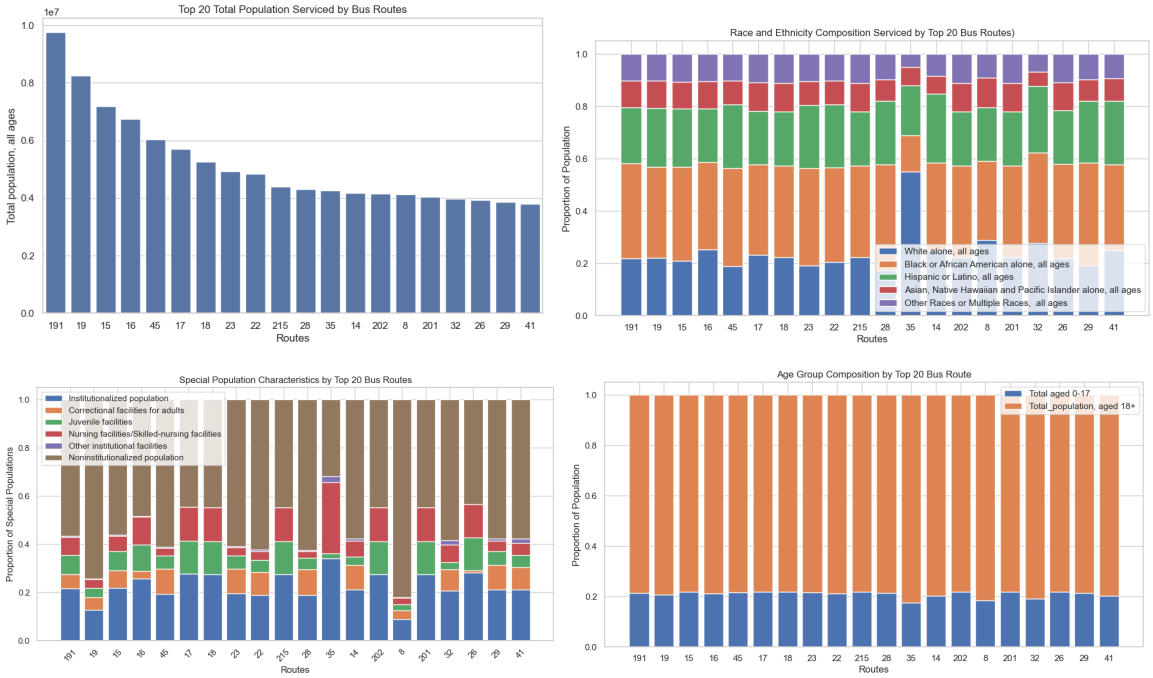
The data suggests that some routes such as 424 are consistently early, while 501 is consistently late for over 500 seconds. From the lateness percentage, we could see route 193 is late more often than others in 2022.

route_id	Average Lateness Percentage
193	88.888889
CT2	87.512371
11	85.070365
24	84.741800
21	84.418016
65	84.052068
52	82.754777
08	82.654756
74	82.425094
50	82.115385



As the lateness image of the 2022 bus shown above, we've selected the first ten that are the longest late and the first ten that are the longest early. CT2 has a high probability of being late. Perhaps the probability of traffic congestion has increased because the routes of the bus pass through areas where there is a greater concentration of residents according to the bus stop map shown above.

3. What are the population sizes and characteristics of the communities serviced by different bus routes (e.g. race, ethnicity, age, people with disabilities/ vulnerabilities)?



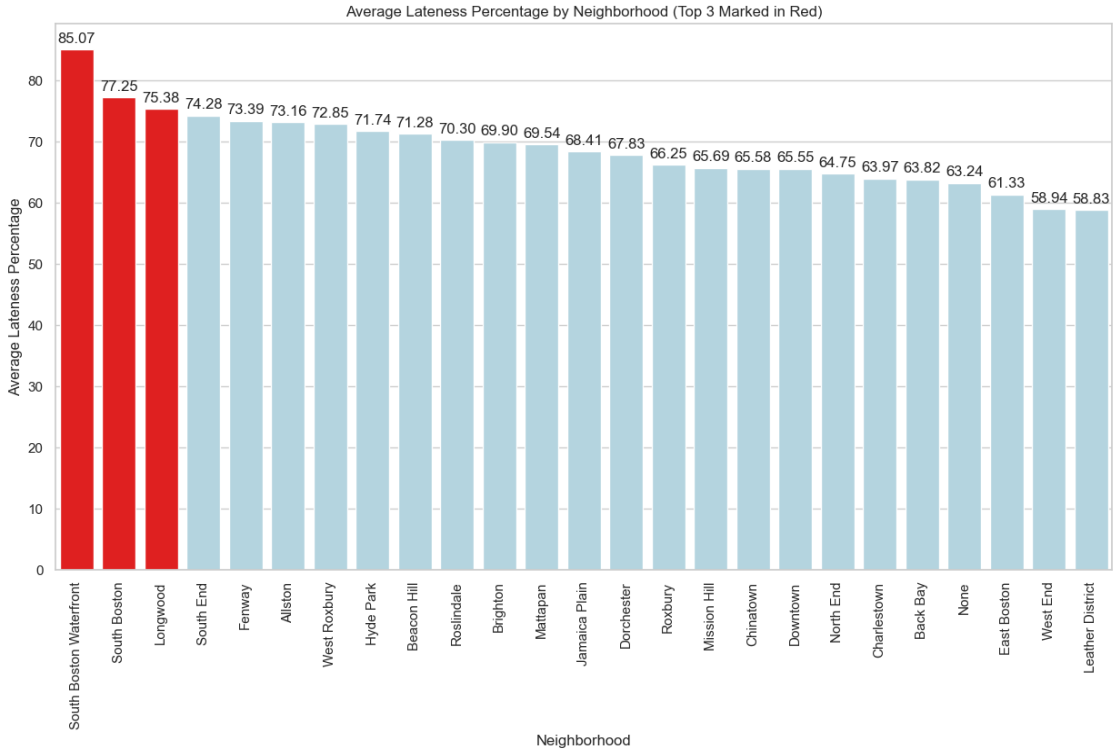
I. *Population Size and Age Characteristics*: Analysis of the top 20 bus routes, based on the highest serviced population, reveals a consistent age group distribution. On average, the ratio of adults (aged 18 and above) to young individuals (aged below 18) across these routes is approximately 8:2. This indicates a predominant adult population being serviced by these key routes.

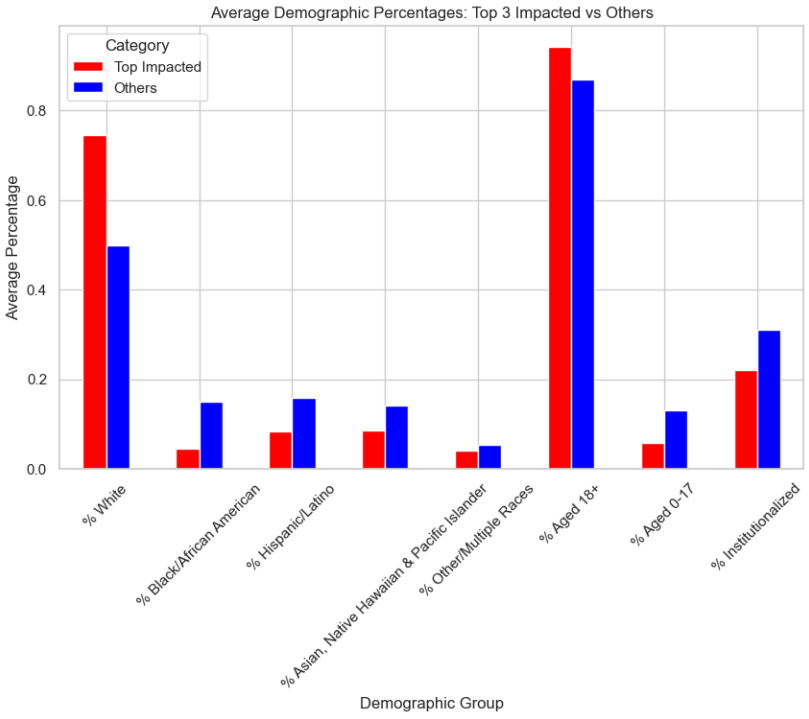
II. *Race and Ethnicity Composition*: When examining the racial and ethnic composition serviced by these top 20 bus routes, a remarkably stable ratio emerges across the majority of the routes. With age held constant, the proportion

of African American, White, Hispanic/Latino, Asian, Native Hawaiian, and Pacific Islander, and Other/Multiple races typically adheres to a ratio of 4:2:2:1:1. However, an exception is observed in Route 35, which services a predominantly White population, with significantly fewer African Americans. The representation of Asian, Native Hawaiian and Pacific Islander, and Other/Multiple races also decreases on this route, though not as sharply as the decrease in African American population.

III. People with Disabilities/Vulnerabilities: The distribution of special populations, such as individuals with disabilities or vulnerabilities, among the top 20 routes does not display a clear correlation with the total number of people serviced. This suggests a diverse and non-uniform distribution of special populations across these bus routes.

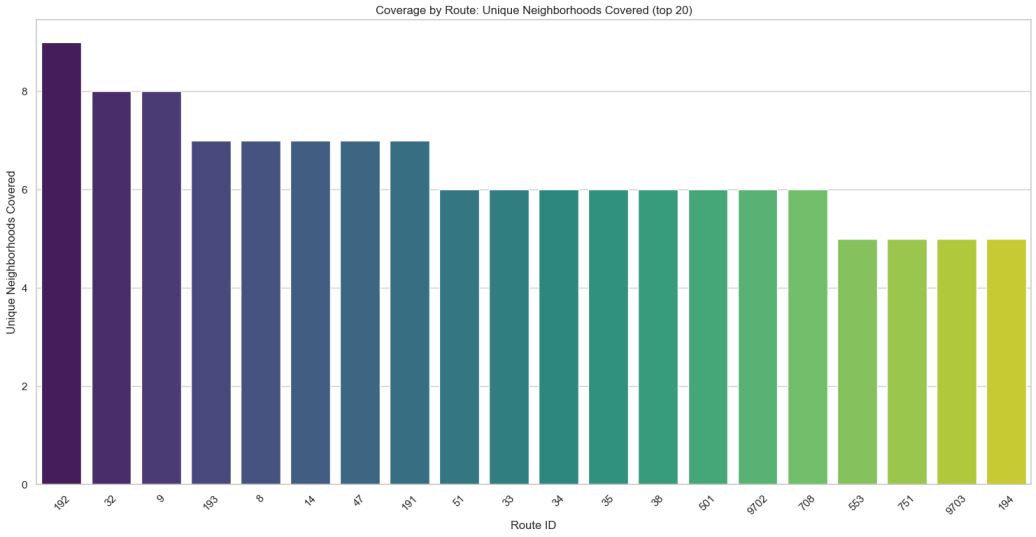
4. If there are service level disparities, are there differences in the characteristics of the people most impacted? Which neighborhoods are served better/worse by the MBTA bus system? Which routes are better/worse? differences in quality of service by class/race, contributing variables, ect.

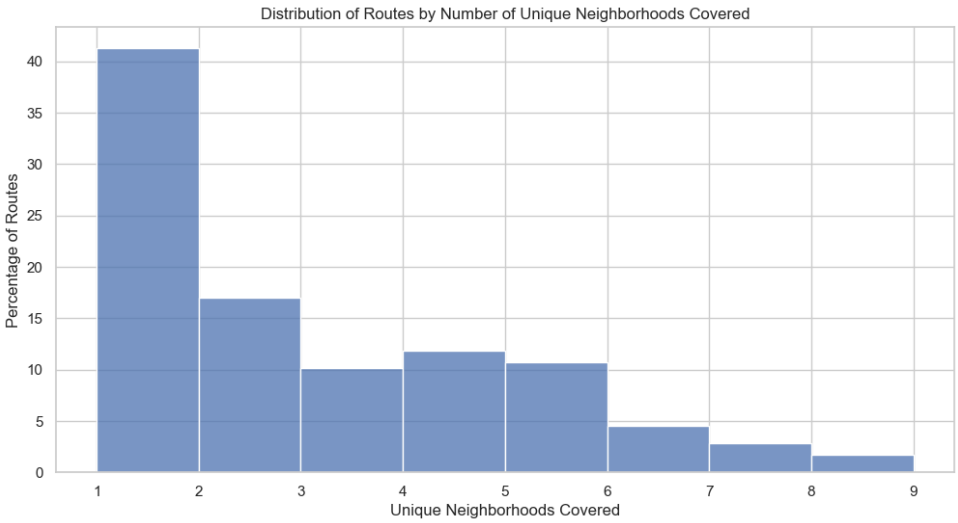




The investigation into service disparities reveals significant differences in service levels experienced by various demographic groups, particularly in terms of race, ethnicity, and individuals with special conditions. Focusing on the neighborhoods experiencing the most pronounced average lateness in bus services, namely South Boston Waterfront, South Boston, and Longwood, a pattern emerges.

The demographic analysis of these neighborhoods, as presented in the "Average Demographic Percentages: Top 3 Impacted Neighborhoods vs Other Neighborhoods" plot, highlights that the most affected group has a featured identity combination of white adults aged 18 and over. This finding suggests a specific demographic skew in those facing the greatest delays in bus service punctuality, indicating an area of concern for transit authorities in terms of service equality and effectiveness.





We conducted a comparative analysis of MBTA bus routes in Boston and evaluated the quality of service using three key metrics: punctuality (average lateness), population size serviced, and the number of unique neighborhoods covered.

I. *Punctuality* (Average Lateness): The top-performing routes in terms of punctuality, exhibiting the least average lateness, are Routes 424, 171, 351, 217, and 238. These routes demonstrate higher reliability in adhering to their scheduled times.

II. *Population Size Serviced*: When considering the total number of people serviced by each route, the most impactful routes are 191, 19, 15, 16, and 45. These routes cater to a larger segment of Boston's population, indicating their vital role in the city's public transportation network.

III. *Unique Neighborhoods Covered*: Routes 192, 32, 9, 193, and 8 cover the highest number of unique neighborhoods. This metric highlights the routes that offer broader geographical accessibility across Boston.

It is noteworthy that there is minimal overlap in the top routes when evaluated across these different metrics. Additionally, due to data limitations, other potentially effective metrics such as reliability and ridership could not be analyzed. This lack of overlapping suggests a diverse range of strengths across different routes, underscoring the complexity in defining a 'best' route without considering the specific needs and priorities of different user groups.

EXTENSION PROPOSAL: Enhancing Bus Accessibility in Boston

The City of Boston stands committed to inclusivity and equal access to public services, and with this extension proposal, we aim to enhance the city's bus transportation system to better serve individuals with disabilities. Recognizing the critical role that public transit plays in community life, this project seeks to assess and improve the accessibility of wheelchair-accessible buses and stops across the city, especially in

underrepresented neighborhoods. Our objective is to map the distribution of these resources, identify areas that are underserved, and recommend improvements to ensure citywide accessibility.

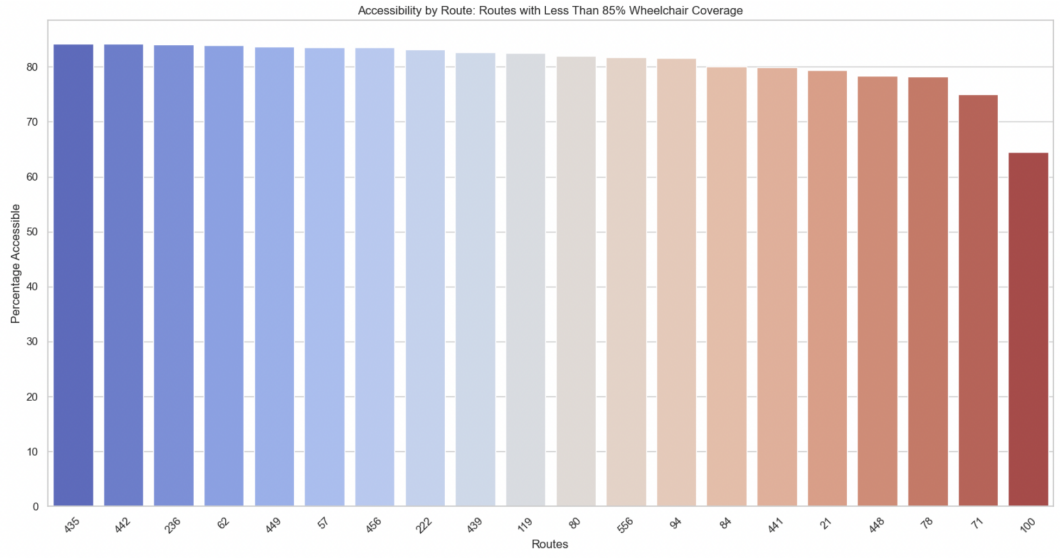
We begin with a rationale rooted in our commitment to fostering a more inclusive urban environment where mobility challenges do not hinder full participation in civic life. To this end, we will analyze key questions concerning the availability of equipped bus stops, the frequency of wheelchair-accessible buses on various routes, the accessibility services in different neighborhoods, and the correlation between disabled population concentrations and bus accessibility. Our data will be meticulously sourced from Boston's MBTA service data, demographic information from the Census, city planning data, and the MBTA's V3 API, which provides information on facilities with elevators and escalators. We will also use the MBTA_Systemwide_GTFS_Map dataset which contains information such as an accessibility score, wheelchair_board metrics, and sidewalk distance and condition to assess accessibility at specific route ID’s.

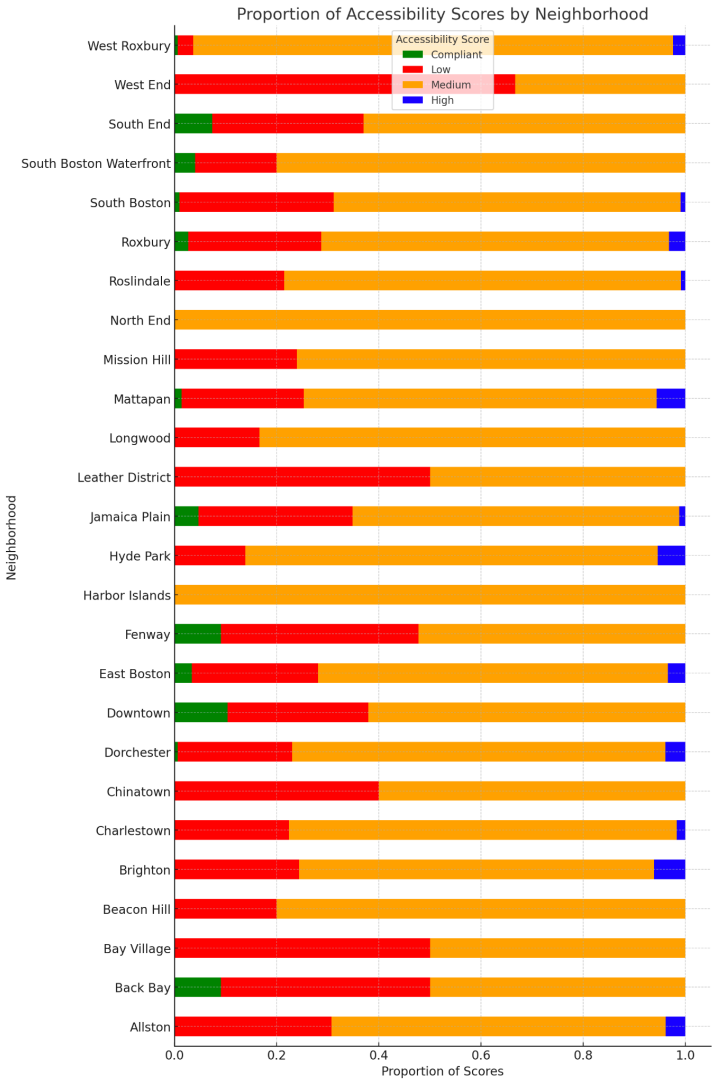
The data visualization component of our analysis will include heat maps to display the concentration of accessible bus stops, bar graphs to compare the number of accessible buses across different routes, and scatter plots to explore the relationship between the disabled population density and accessibility scores. Through a mixed-methods approach that blends quantitative data analysis with qualitative insights, we will not only map out current accessibility but also pinpoint priority areas for improvement and offer evidence-based recommendations. We also plan to use some data science model to look at Segmentation of Stops or Routes using clustering algorithms like K-means or DBSCAN to segment bus stops or routes based on ridership patterns, demographics, or geographic location. We are also considering performing an Accessibility Gap Analysis where we identify areas with gaps in accessibility for disabled passengers and propose infrastructure improvements.

We will also plan to review ADA guidelines, historical trends in service expansion, and best practices from other cities renowned for their high accessibility standards. Our ultimate goal is a comprehensive and actionable plan that will lead to a significant enhancement of Boston's bus services for those with disabilities, ensuring that all residents have access to reliable, dignified, and inclusive transportation. Even though BU should make many changes to be more ADA compliant, there seems to be a large issue with disability access throughout Boston that needs to be addressed.

Extension Proposal Template	
Extension Pitch	*The proposal seeks to enhance the accessibility of Boston's bus services for individuals with disabilities by assessing the current state of wheelchair-accessible buses and stops. We aim to map the distribution of these resources and identify areas that are underserved, creating a foundation for recommendations to improve citywide accessibility.*
Rationale	*This project underscores the commitment to inclusivity and equal access for all individuals, especially those with mobility challenges. By identifying gaps in service, we can advocate for targeted improvements, fostering a more inclusive urban environment that enables all citizens to participate fully in community life.*
Questions for Analysis	*How many and which bus stops are equipped for wheelchair access?* *What is the frequency of wheelchair-accessible buses on various routes?* *Are there neighborhoods or areas that are currently underserved by accessible services?* *What are the correlations between areas with high concentrations of disabled individuals and the availability of accessible buses?*
Data Sets & Sources	*Boston's MBTA bus service data on accessible vehicles and stops.* *Census data for demographic information on individuals with disabilities.* *City planning data to correlate infrastructure with accessibility.* *API calls from V3 API by MBTA to find if facilities have elevators, escalators, etc.*
Data Visualizations	*A heat map showing the concentration of wheelchair-accessible bus stops.* *Bar graphs comparing the number of accessible buses across different routes (possibly comparing to the subway system?).* *Scatter plots correlating areas with higher populations of disabled individuals against accessibility scores.*
Additional Information	*ADA requirements and guidelines for public transportation.* *Historical data on the expansion of accessible services in Boston.* *Best practices from other cities known for high accessibility standards.*

VISUALIZATIONS & INSIGHTS FOR EXTENSION PROJECT:





INDIVIDUAL CONTRIBUTION:

In this project, the collaborative efforts of the team members contributed to our success. Each individual's contributions played a significant role in driving the project forward.

Yu and Vishvakishore laid the foundation for the project with their meticulous work in preliminary data cleaning and analysis. Their efforts included sifting through the data, standardizing formats, and addressing any missing or outlier values. This initial phase set the tone for the project, ensuring that the data was reliable and well-structured for subsequent analysis.

Yu, Vishvakishore, and Kevin delved into the more complex aspects of computing and

coding. Their efforts culminated in the successful visualization of the end-to-end travel time and lateness of bus routes, a task they accomplished using a Python notebook. This work was not only technical in nature but also required a deep understanding of the subject matter to ensure that the visualizations were accurate and insightful.

Parallel to these efforts, Suin and Yufeng focused on the task of analyzing and writing up the deliverables. Their work involved a detailed examination of the methods used, the results obtained, and the initial conclusions that could be drawn from the data. Their analytical skills and attention to detail ensured that the project's findings were well-documented and clearly presented.

Vishvakishore took on the responsibility of working on the initial extension proposal. This task required not only a deep understanding of the project's core objectives but also the ability to envision how these could be expanded and enhanced.

Further demonstrating the collaborative spirit of the team, Vishavakishore, Suin, and Kevin collaborated to work on the computing, coding, and visualization aspects of the extension project. Their combined expertise ensured that this additional phase of the project was executed with the same level of precision and attention to detail as the main project.

Despite each member having specific tasks and responsibilities, the team consistently came together in meetings to collaborate and assist each other. This collective approach was not just about sharing the workload, but it was also about leveraging each other's strengths and learning from one another.