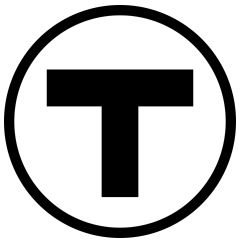
CS506 – DATA SCIENCE FUNDAMENTALS
DECEMBER 15, 2023

DELIVERABLE 5
FINAL REPORT



CITY OF BOSTON BUS TRANSIT PERFORMANCE
TEAM B

TEAM MEMBERS:

Vishvakishore Venkatesan ‘24, vishvav@bu.edu
Suin Lee ‘24, suinlee@bu.edu
Kevin Smith ‘24, smithkj@bu.edu
Yufeng Song ‘24, jyfsong@bu.edu
Yu Han ‘24, hanyurh@bu.edu

PROJECT GOAL & OVERVIEW:
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 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 opportunities for 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.

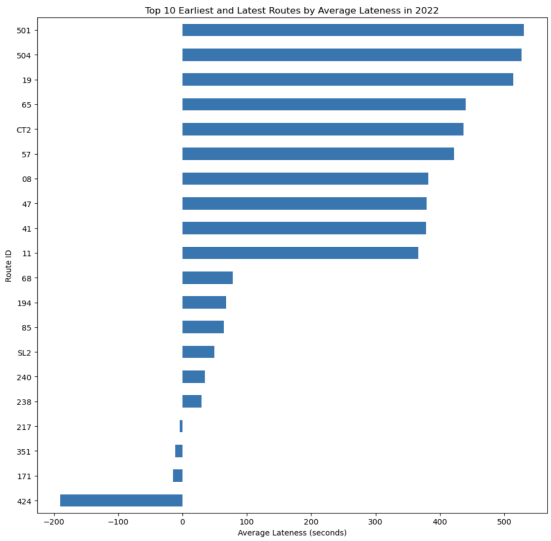


Figure 1: Top 10 Earliest and Latest Routes by Average Lateness in 2022

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

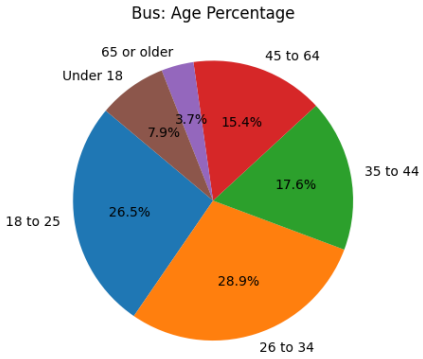


Figure 2: Age Percentage of Passengers

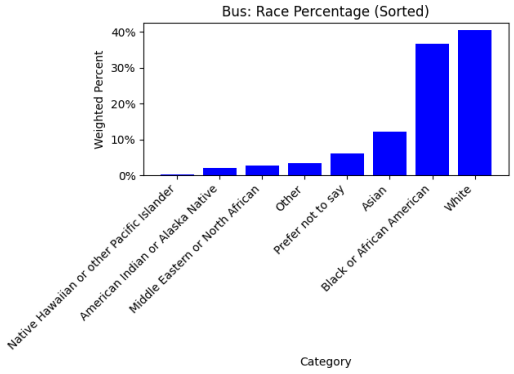


Figure 3: Race Percentage of Passengers

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, & INSIGHTS FOR KEY BASE QUESTIONS

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

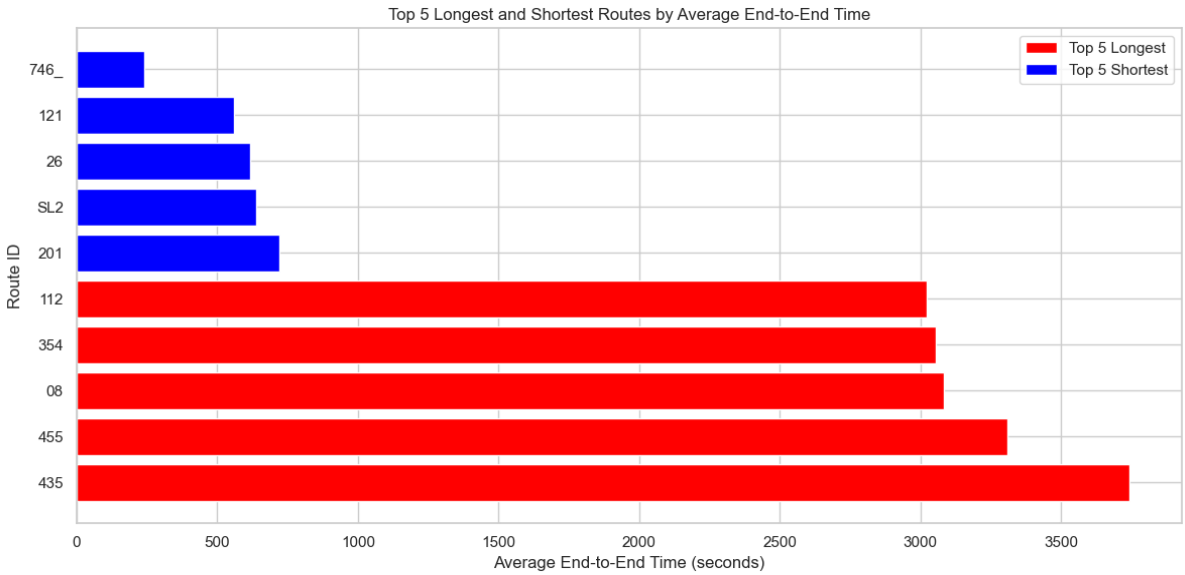


Figure 4: Top 5 Bus Routes with Longest or Shortest Average End-to-End Travel Time

Figure 5: Route 435 is the one with the longest end-to-end travel time

Figure 6: Route 476 is the one with the shortest end-to-end travel time

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?

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.

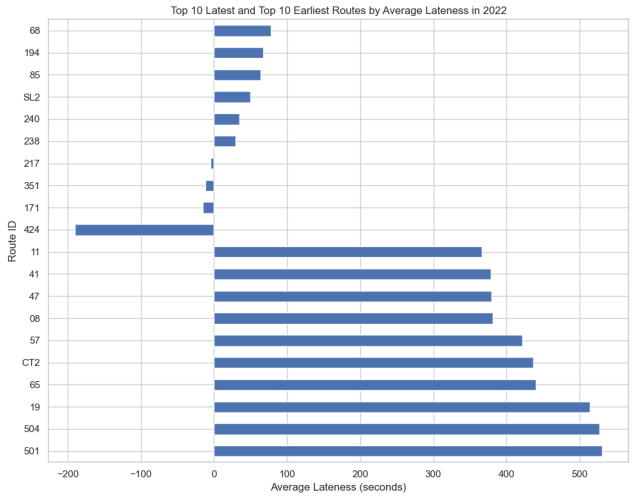


Figure 7: Top 10 Latest or Earliest Bus Routes by Average Lateness in 2022

Figure 8: Route CT2 as one of the latest bus route

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)?

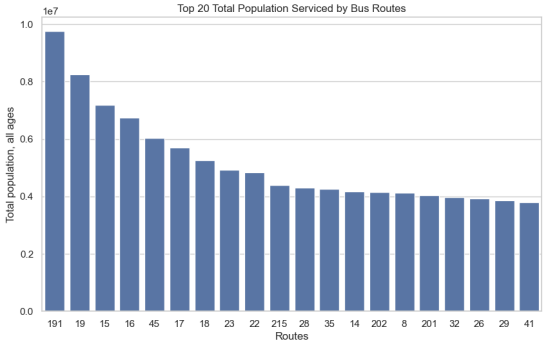


Figure 9: Top 20 Total Population Serviced by Boston Bus Routes

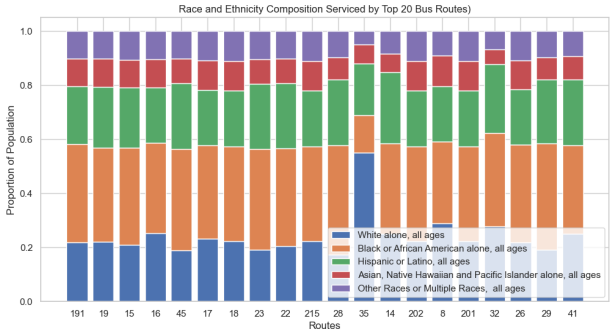


Figure 10: Race and Ethnicity Composition of Neighborhoods Serviced by Top 20 Bus Routes

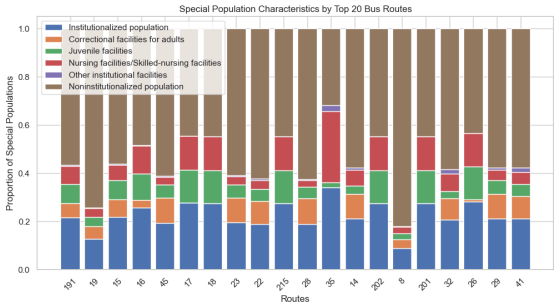


Figure 11: Special Population Characteristics of Neighborhoods Serviced Serviced by Top 20 Bus Routes

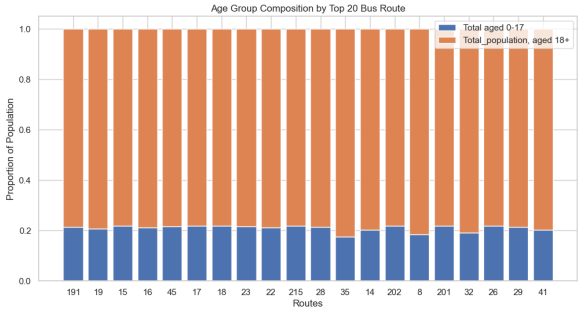


Figure 12: Age Group Composition of Neighborhoods Serviced by Top 20 Bus Routes

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?

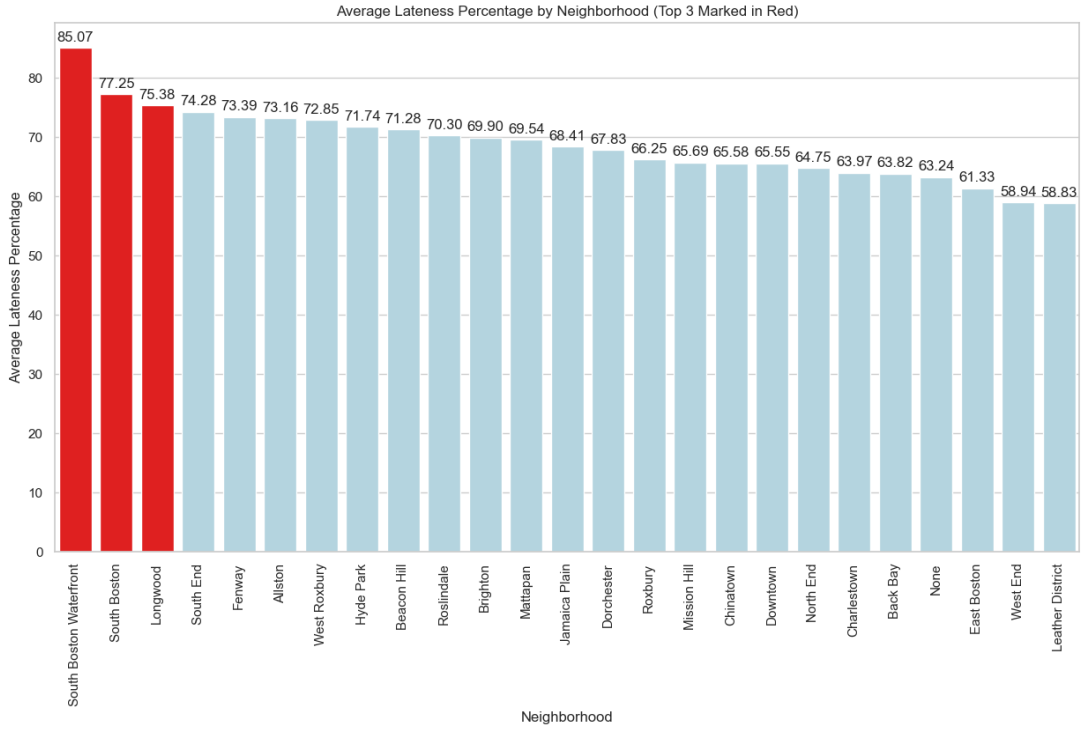


Figure 13: Average Lateness Percentage by Neighborhood

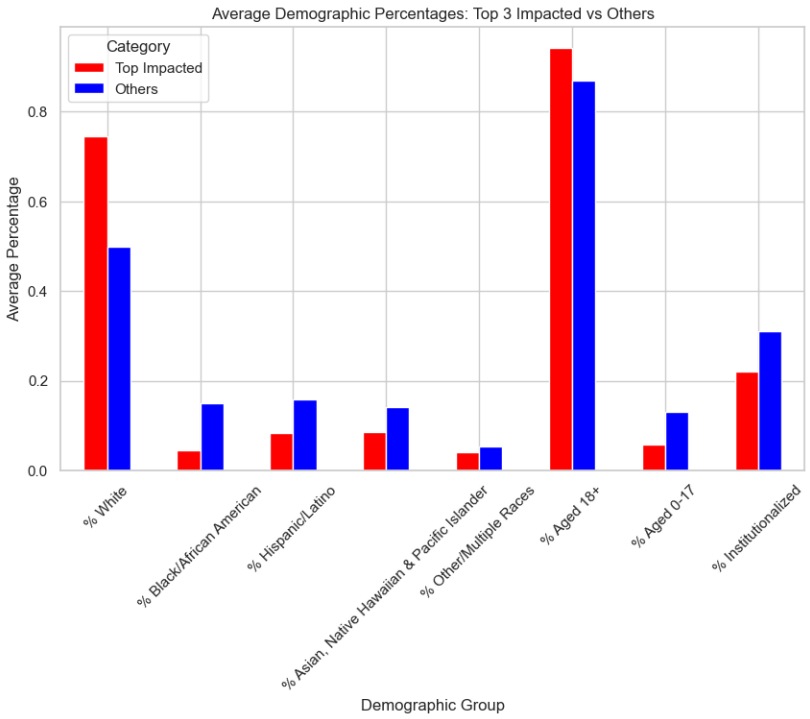


Figure 14: Average Demographic Percentages (Race & Age) by Neighborhood

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.

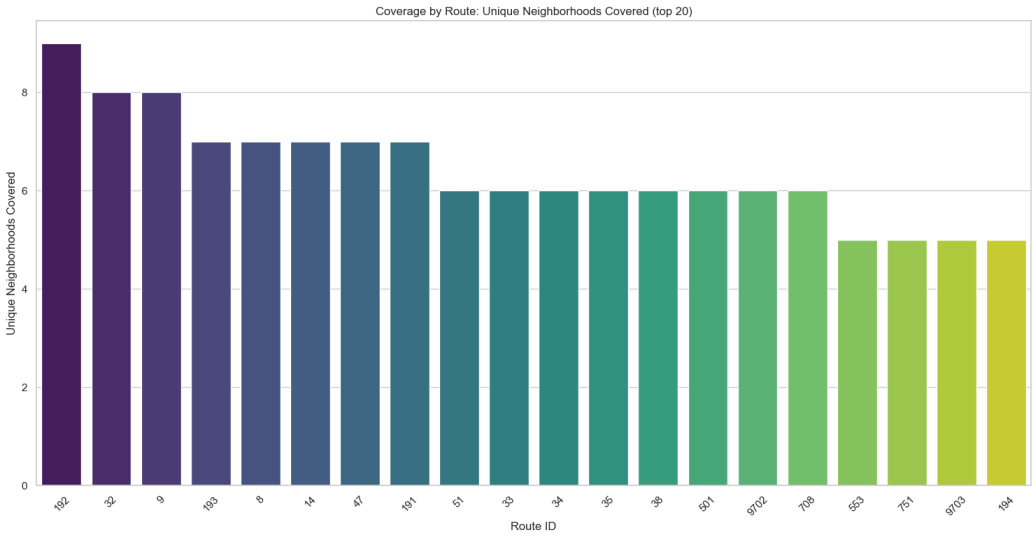


Figure 15: Unique Number of Neighborhoods Covered by Bus Routes that Service Top 20 Most Population

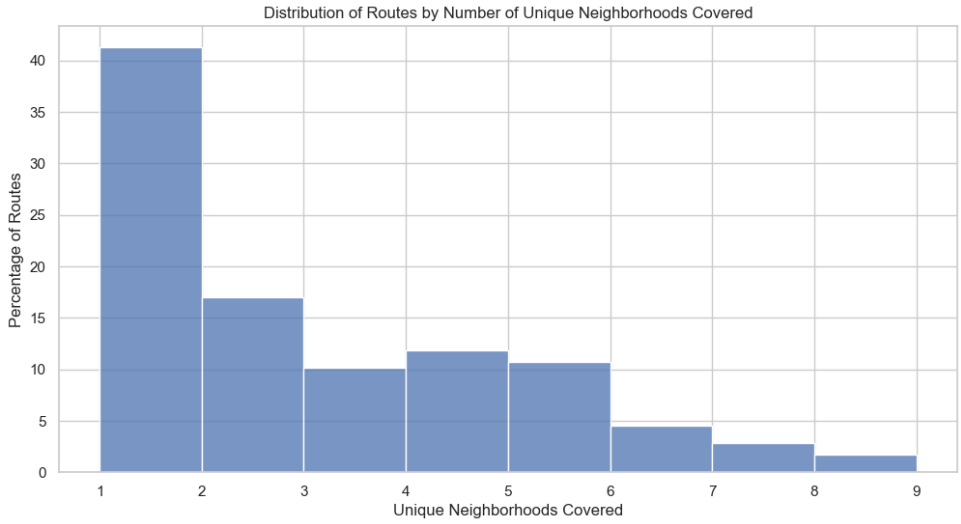


Figure 16: Distribution of Routes by Number of Unique Neighborhoods Covered

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

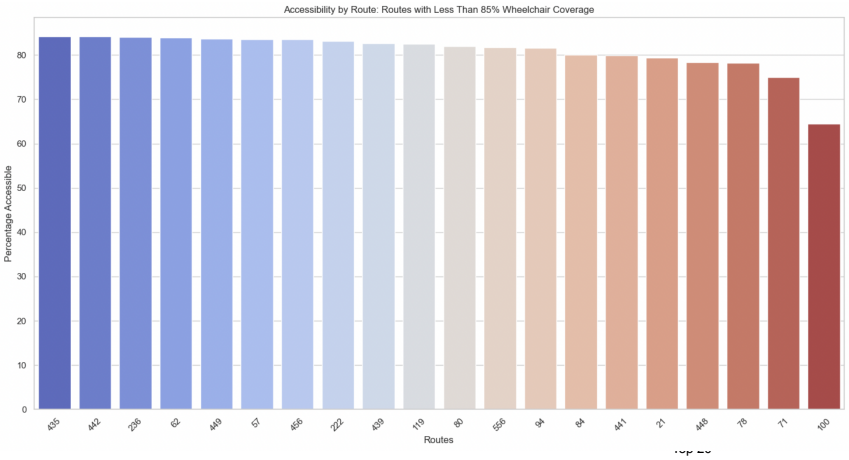
The City of Boston stands committed to inclusivity and equal access to public services, and with this extension proposal, we aim to enhance the city's bus transportation system to better serve individuals with disabilities. Recognizing the critical role that public transit plays in community life, this 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.



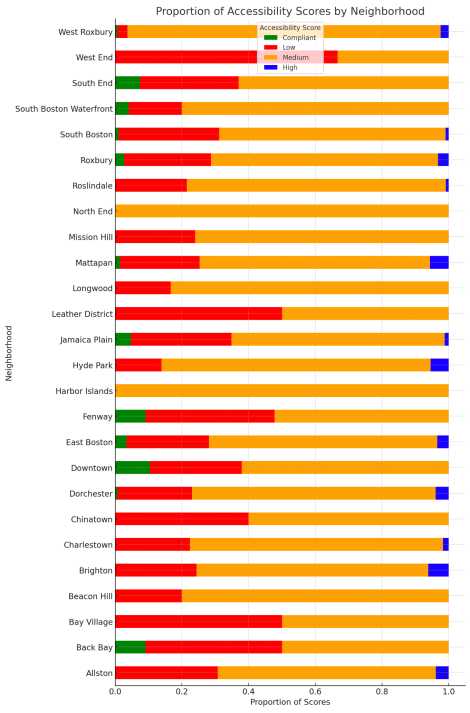


Figure 17 (Up): Accessibility by Route: Routes with Less Than 85% Wheelchair Coverage
Figure 18 (Right): Proportion of Accessibility Scores by Neighborhood

EXTENSION PROJECT KEY QUESTIONS:

1. **How many and which bus stops are equipped for wheelchair access?**

From the map we generated, it is clear to see that there are 6030 stops equipped for wheelchair access.

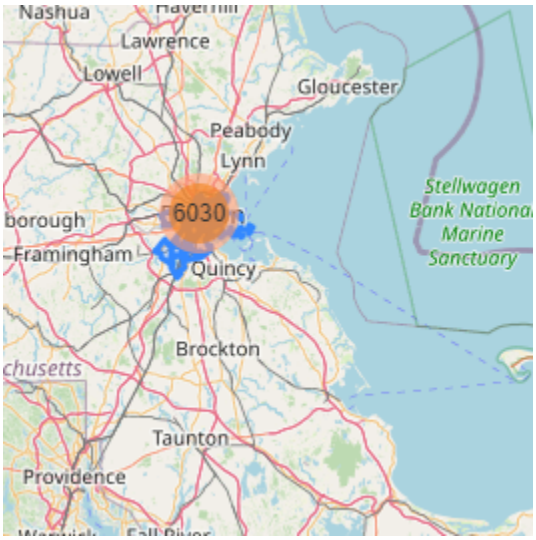


Figure 19: Visualization of Total Number of Bus Stops with Wheelchair Access

2. **Are there neighborhoods or areas that are currently underserved by accessible services?**

Accessible services are currently underserved on Harbor Islands.

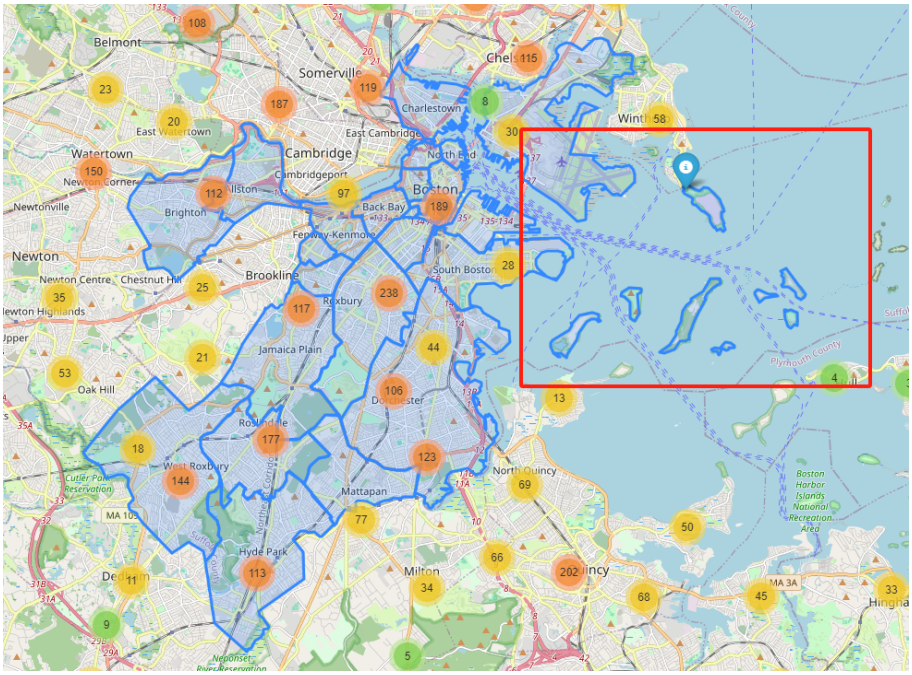


Figure 20: Visualization of Harbor Islands as One of the Most Underserved Region with the Least Wheelchair Accessibility

3. **What are the correlations between areas with high concentrations of disabled individuals and the availability of accessible buses?**

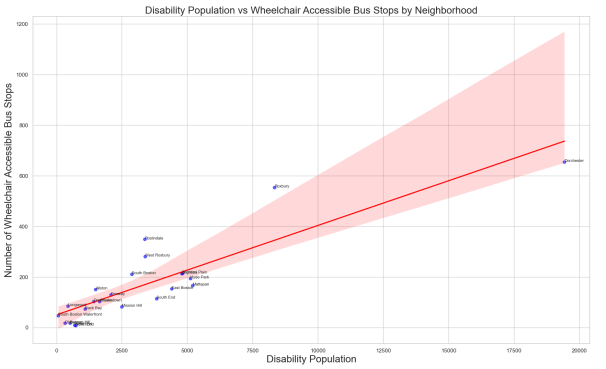


Figure 21: Disability Population vs Wheelchair Accessible Bus Stops by Neighborhood

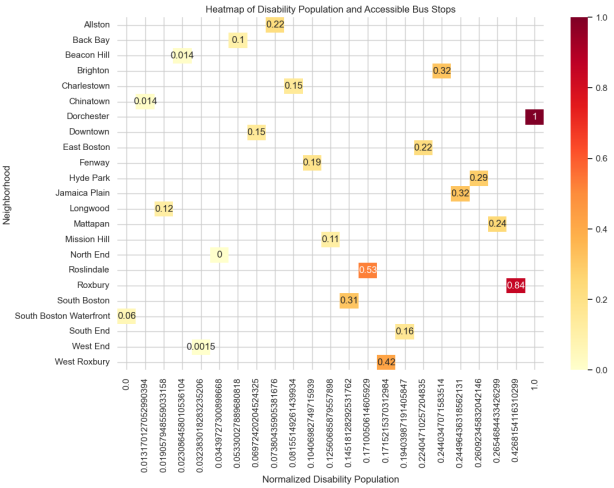


Figure 22: Heatmap of Disability Population and Accessible Bus Stops

Disability population in Boston neighborhoods is positively correlated with the number of wheelchair-accessible bus stops, which means as the disabled population increases, so does the number of bus stops with accessibility features. The heatmap provides us with a closer look at this correlation. In neighborhoods with fewer disabled residents, the connection appears to be less strong. However, in areas like Roxbury and Dorchester, where there's a significant disabled population, we can observe a notably strong positive correlation. This suggests that Boston's bus system is performing well in prioritizing accessibility in neighborhoods where it's most needed for disabled individuals.

More on Accessibility Score That We Use to Build the Interactive Map

```python
# Define a scoring system for sidewalk condition
condition_scores = {
    'Excellent': 5,
    'Good': 4,
    'Medium': 3,
    'Fair': 2,
    'Poor': 1,
    'Cracked': 1,
    'N/A': 0  # assuming N/A is the worst case or unknown
}

# Define a scoring system for sidewalk material
material_scores = {
    'Concrete': 3,
    'Asphalt': 2,
    'Brick': 1,
    'Other': 0
}

# Define a scoring system for shelter
shelter_scores = {
    'JCD': 2,
    'MBTA': 1,
    '0': 0  # Assuming '0' means no shelter
}

# Define the scoring function
def calculate_accessibility_score(row):
    width_score = float(row['Sidewalk_Width__ft_']) / 10  # Assuming 10 ft is the max desirable width
    condition_score = condition_scores.get(row['Sidewalk_Condition'], 0)
    material_score = material_scores.get(row['Sidewalk_Material'], 0)
    shelter_score = shelter_scores.get(row['Current_Shelter'], 0)

    # Combine these scores
    total_score = width_score + condition_score + material_score + shelter_score
    return total_score
```

Figure 23: Computation of Accessibility Score

We used a new dataset from the MBTA, featuring bus route IDs and an existing 'Accessibility score' column, was utilized. The scores, ranging from Compliant, Low, Medium, to High, were further analyzed. Additional data from columns like Sidewalk Width, Condition, Material, and Current Shelter presence was incorporated. A Python script assigned point values to each metric, quantifying the Accessibility score more precisely. This enhanced scoring system allowed for an updated Folium map, which now clusters bus routes based on their accessibility scores. On this map, darker red areas indicate clusters with higher accessibility scores, while green spots signify areas with lower scores as seen below.

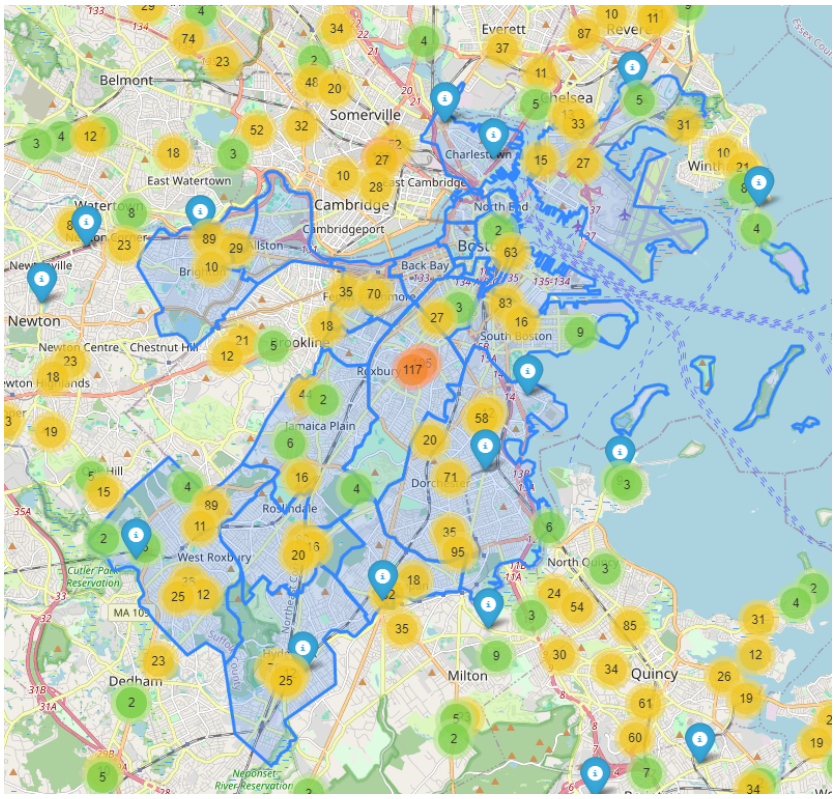


Figure 24: Bus Stops Clustered by Accessibility Score, with Redder Color Means Well-served Region

LIMITATIONS AND CHALLENGES

We faced a plethora of challenges and limitations that are essential to consider for a comprehensive understanding of the project's findings and implications. One of the most significant hurdles was data limitations, particularly in the MBTA's APIs and Historical Data Archive, where gaps in data, especially in real-time tracking and historical performance metrics, could have led to incomplete or skewed insights into the system's overall performance.

Additionally, the complexity involved in preprocessing and analyzing large datasets

presented a considerable challenge. This task, vital for ensuring data validity, required extensive time and technical skills, potentially limiting the depth of analysis in certain areas. The team also navigated the complexities of accurately representing diverse populations using demographic data from sources like the 2020 Boston Census. This raised concerns about potential oversimplification or misrepresentation in the demographic analysis. Gathering all this data was also a challenge, as much of this data was sparse and incorrectly labeled across the MBTA's data platforms. This showed that the MBTA needs to improve their data collection and data governance practices.

Methodological constraints were another aspect that the team had to contend with. Relying on specific tools and algorithms, such as Python for data analysis and clustering algorithms for the extension project, might have influenced the outcomes and insights. These choices, while robust, may not have captured all variables impacting the bus system's performance and usage.

The extension project aimed at enhancing bus accessibility in Boston brought its unique set of challenges. Accurately mapping and analyzing the distribution of accessible bus stops and services, especially in underserved areas, required a nuanced understanding of urban planning and disability access. Aligning the project's goals with ADA guidelines and best practices demanded a detailed and comprehensive approach.

Moreover, the team dynamics, involving collaboration and coordination among members, posed its challenges. Ensuring consistent communication, integrating individual contributions into a cohesive report, and balancing the workload effectively were critical for the project's success.

These challenges show the complexity of public transit analysis and the need for ongoing refinement in methodologies, data collection, and collaborative strategies. They also highlight the importance of considering a broad range of factors, from technical data limitations to the complexities of representing diverse urban populations, in urban planning, data governance, and public policy analysis.

CONCLUSION

Our project on the City of Boston Bus Transit Performance has provided a comprehensive analysis of the MBTA bus system, offering significant insights into service quality, demographic impacts, and geographical disparities in the year 2022. Through our collaborative effort, we have successfully combined and processed diverse datasets, including MBTA performance data, census demographics, and bus stop information. Our findings reveal critical patterns in end-to-end travel times, service level

disparities, and population characteristics served by different bus routes.

Our analysis shows that while some bus routes excel in punctuality, others face significant delays, highlighting the need for targeted improvements in certain areas. The demographic analysis of bus routes reveals that while the distribution of age groups and racial demographics is generally consistent across most routes, certain routes like 35 show notable deviations. Particularly, the impact of service level disparities is most pronounced in neighborhoods like South Boston Waterfront, South Boston, and Longwood, predominantly affecting white adults aged 18 and above.

The extension project focusing on enhancing bus accessibility in Boston has shed light on the correlation between areas with high concentrations of disabled individuals and the availability of accessible buses. Our advanced mapping and clustering techniques have identified neighborhoods that are currently underserved by accessible services, such as Harbor Islands. Moreover, the project has underscored the importance of inclusive urban planning, ensuring that mobility challenges do not impede access to public services.

Moving forward, there are several directions for further research and development. An immediate focus could be on addressing the disparities in service levels, particularly in the most impacted neighborhoods. Exploring the feasibility of route optimizations and increased frequency of services in these areas could be beneficial. Additionally, expanding our accessibility analysis to include more comprehensive metrics and broader geographical areas would enhance our understanding of the overall accessibility landscape in Boston's public transportation network.

Moreover, integrating real-time data analysis could provide more dynamic insights into service quality and accessibility. This could involve developing predictive models for bus punctuality and identifying potential areas of service breakdown. Such models could help in proactive service management, ensuring more reliable and efficient transportation for all residents of Boston.

In addition to the comprehensive insights and recommendations highlighted in our project on the City of Boston Bus Transit Performance, we are proud to share a significant milestone that underscores the impact and relevance of our work. Our team was selected as the representative for the project titled "City of Boston: Transit and Performance" and had the opportunity to showcase our findings and analyses at the Spark Demo Day. This event was not only a platform for us to present our rigorous research and innovative solutions but also a testament to the practical implications and potential of our project in addressing real-world challenges in urban transit systems.

The success of our presentation at Spark Demo Day was further amplified when our work caught the attention of a journalist from WCVB, a prominent local news outlet. This interest from a reputable media source reflects the broader societal and community relevance of our project. The journalist expressed a desire to reach out to us for an in-depth discussion on our findings, with the potential of featuring our work in a news report. This opportunity to share our insights with a wider audience is not just an honor for our team but also a powerful reminder of the impact academic research can have beyond the confines of academia.

Our engagement with WCVB highlights the potential for our findings to inform public discourse, contribute to policy debates, and perhaps even influence decision-making processes within local government and public transportation authorities. It is a compelling example of how data-driven research can transcend academic boundaries and become a catalyst for real-world change.

This recognition and interest from external parties validate our efforts and the significance of our project. It encourages us to continue our pursuit of excellence in research and reinforces our commitment to contributing meaningful solutions to urban transit challenges. As we move forward, we are inspired to delve deeper into our research, explore new avenues for analysis, and strive to make an even more significant impact on public transportation systems and urban planning. Our experience at Spark Demo Day and the subsequent interest from WCVB are pivotal moments for our team, marking our project as not only academically successful but also socially and practically relevant.

Our project has laid a solid foundation for further research and development in public transportation analysis. The insights gained not only highlight the current state of Boston's bus system but also pave the way for future enhancements, ensuring that the MBTA remains a cornerstone of Boston's commitment to sustainable, equitable, and efficient urban transportation.

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, Yufeng 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 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.

Delving more into the extension project, through meticulous exploratory data analysis (EDA), Yufeng concentrated on exploring the relationship between neighborhood accessibility and the disabled population and successfully highlighted key insights into how the distribution of accessible public transportation aligns with the needs of disabled residents in various neighborhoods. Vishvakishore and Yu focused on enhancing our understanding of individual bus stops' accessibility. They developed a comprehensive scoring system for each bus stop, integrating various factors that contribute to overall accessibility. This scoring system became the basis for clustering bus stops, enabling us to categorize them effectively based on their accessibility levels. Moreover, they leveraged this data to create interactive maps. These maps are not only visually striking but also serve as an invaluable tool for visually representing our findings, making the data more accessible and understandable for all stakeholders. Their work significantly contributes to our project by offering a detailed, user-friendly overview of accessibility across the bus network.

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.