Demographic and Geographic Factors Associated with Number of Docks at a Blue Bike Station

Cianna Salvatora ‘25, Data Science Major Capstone

Background and Research Questions

As the Metro Boston area continues to expand access to sustainable transportation, Bluebikes—a public bike share program with more than 3,000 bikes at over 400 stations across Boston, Brookline, Cambridge, Somerville, and Everett—has become an integral part of the region's transportation network [1]. **This capstone project seeks to identify demographic and geographic factors that are associated with the number of bike docks across municipalities in the Boston Area.**

This study combines demographic, location, and transit-related statistics for various municipalities in the Greater Boston area to fit a linear regression model, with particular interest in the following questions:

(1) How are Bluebikes stations distributed across different municipalities, and what factors might influence where stations are placed?

(2) Is there a relationship between the number of docks at a station and its location, such as being near public transportation or educational campuses?

Data

Description

The primary dataset for this project was downloaded from the Bluebikes website and includes data on every Bluebike station across the Metro Boston area.

To enhance the analysis and include potential demographic and geographic predictors, several additional datasets were merged with the Bluebike station data:

- **Colleges and universities** from Mass.gov/MassGIS [2]: used to calculate the distance between each Bluebike station and the nearest college or university .
- **MBTA stations** from Mass.gov/MassGIS [3]**:** used to determine the distance from each Bluebike station to the nearest MBTA station.
- **Population by municipality** from MA Department of Revenue [4]: provided total population information for each municipality.
- **Age group data** from CensusReporter.org [5]**:** manually compiled to determine the dominant age group in each municipality.

The final merged dataset allows for exploration of how geographic and demographic features relate to station size and municipality.

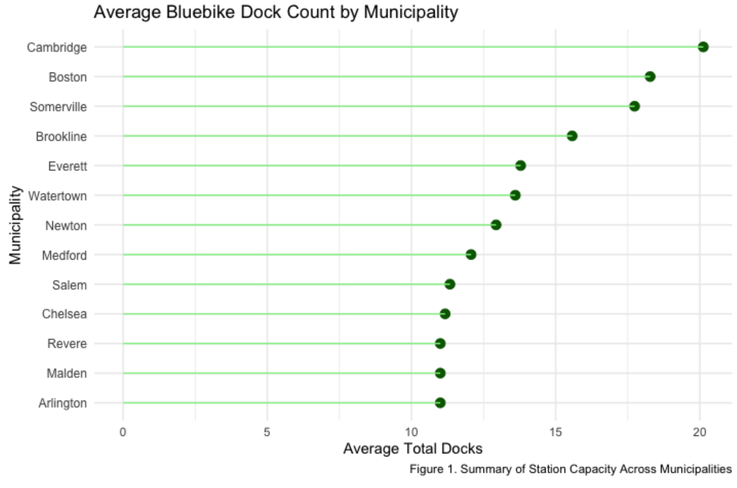

Figure 1. overview of how dock availability differs by area.

Cleaning

Initial steps included renaming columns, removing a misread header row, and converting relevant variables (e.g., latitude, longitude, dock counts) to numeric. For the station_ID variable, “No ID pre-March 2023” entries were set to NA to preserve station records. Distance calculations were then added using longitude/latitude coordinates, and all datasets were merged by station or municipality.

Data Modeling: First-Order Model

A multiple linear regression model was initially fit using four predictors: 2023 population, largest age group, distance to nearest college, and distance to nearest MBTA station. Stepwise model selection was then performed using AIC.

Initial residual plots showed mild non-constant variance, so a Box-Cox transformation was applied. As a result, a Box-Cox transformation with λ = -1 was applied to the response variable. The final model had an R-squared of 0.385 and an adjusted R-squared of 0.381 (F-test $p < 2.2 \times 10^{-16}$), with no multicollinearity concerns. The best first-order model is:

$$\overline{Total\_Docks}^{-1} = 0.93 + 0.019 * Largest\ Age\ Group(20-29) - 0.004 * Largest\ Age\ Group\ (30-39) - 0.0024 * Distance\_to\_Nearest\_MBTA\_km + error$$

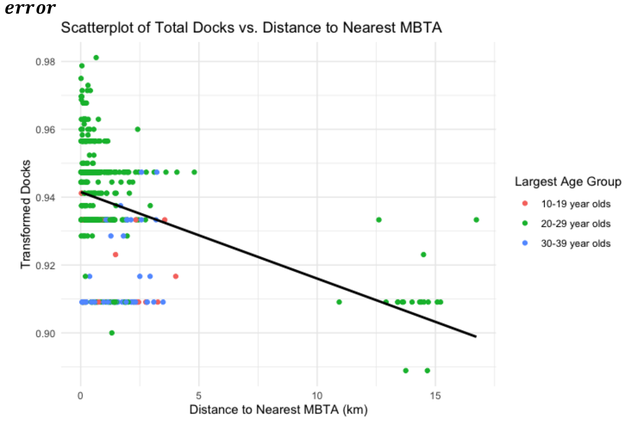

Figure 2.

Data Modeling: First-Order Model

To explore how demographic and geographic features relate to the number of bike docks. I first fit a linear regression model including an interaction between the largest age group in each municipality and distance to the nearest MBTA station. This revealed that the relationship between MBTA distance and bike docks differs by age group: specifically in municipalities where 30-39 year olds are the largest age group, greater distance from MBTA is associated with a larger increase in transformed dock counts compared to other groups.

I then tested whether distance to the nearest college interacted similarly with age group, but found no meaningful interaction effects or improvement in fit. To determine the most informative model, I used a stepwise AIC selection process that considered all two-way interactions between ‘Largest Age Group’, Distance to Nearest MBTA (km), and Distance to Nearest College (km). The final model is:

$$\widehat{Total\_Docks}^{-1} = 0.92 + 0.02 * Largest\ Age\ Group(20-29) - 0.002 * Largest\ Age\ Group\ (30-39) - 0.008 * Distance\_to\_Nearest\_MBTA\_km + 0.009 * Distance\_to\_Nearest\_College\_km + 0.004 * Largest\ Age\ Group(20-29) * Distance\_to\_Nearest\_MBTA\_km + 0.01 * Largest\ Age\ Group(30-39) * Distance\_to\_Nearest\_MBTA\_km - 0.009 * Largest\ Age\ Group(20-29) * Distance\_to\_Nearest\_College\_km - 0.012 * Largest\ Age\ Group(30-39) * Distance\_to\_Nearest\_College\_km + 0.0005 * Distance\_to\_Nearest\_MBTA\_km * Distance\_to\_Nearest\_College\_km + error$$

This model explained approximately 40.8% of the variability in the transformed dock counts (Adj $R^2 = 0.398$), and showed that in addition to age-modified MBTA effects, there may also be compounding effects between MBTA and college proximity.

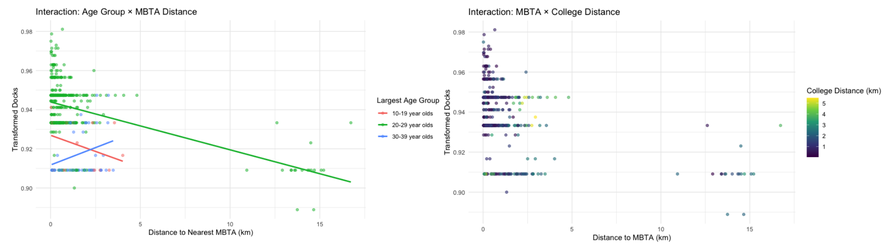

Figure 3&4.

Results

- The **first-order model** included main effects for the largest age group and distance to the nearest MBTA stop. It showed a positive association between the 20–29 year old group and transformed dock count, and a slight negative association with MBTA distance. No interaction terms were included in this model.
- A **stepwise selection process** identified a final model that added distance to the nearest college and several two-way interactions. Notably, the effect of MBTA distance was stronger in areas where the largest age group was 30–39. There was also a significant interaction between MBTA and college distance, suggesting their combined influence on dock placement.
- These results suggest that both age demographics and proximity to infrastructure play important roles in where docks are placed.

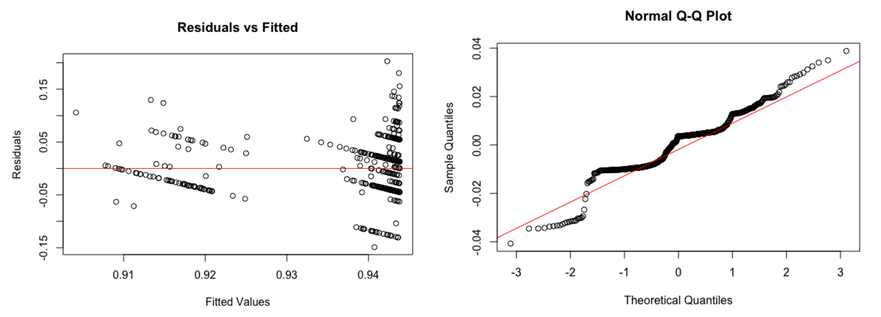

Figure 5. Diagnostic plots for the first-order model show no strong violations of regression assumptions. Residuals appear roughly normally distributed.

Data Ethics & Limitations

While this project aims to identify characteristics associated with Bluebike station placement and usage, it is important to acknowledge the ethical implications of relying on existing infrastructure and datasets. Areas currently underserved may be underrepresented in the data, potentially reinforcing existing access gaps. Prioritizing proximity to colleges or MBTA stations may also overlook communities that could benefit most from improved access.

Limitations include:

•The use of spatial data may violate independence assumptions.

•Age group data was compiled manually, introducing potential for error.

•Some datasets may not reflect recent population or infrastructure changes.

References

[1] Bluebikes. (n.d.). *About Bluebikes*. Retrieved March 2025, from https://bluebikes.com/about

[2] Massachusetts Bureau of Geographic Information (MassGIS). (n.d.). *MassGIS Data: Colleges and Universities*. Retrieved March 2025, from https://www.mass.gov/info-details/massgis-data-colleges-and-universities

[3] Massachusetts Bureau of Geographic Information (MassGIS). (n.d.). *MassGIS Data: MBTA Rapid Transit*. Retrieved March 2025, from https://www.mass.gov/info-details/massgis-data-mbta-rapid-transit

[4] Massachusetts Department of Revenue, Division of Local Services. (n.d.). *Population by Municipality*. Retrieved March 2025, from https://dlsgateway.dor.state.ma.us/reports/rdPage.aspx?rdReport=Socioeconomic.Population.population_main

[5] Census Reporter. (n.d.). *Arlington, MA Profile*. Retrieved March 2025, from https://censusreporter.org/profiles/16000US2501640-arlington-ma/