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 Metro Boston expands access to sustainable transportation, Bluebikes—a public bike share program with over 3,000 bikes at 400+ stations across Boston, has become an integral part of the region's transit network [1]. **This capstone project explores demographic and geographic factors associated with the number of bike docks across municipalities in the Boston Area.**

Using a linear regression model, the study investigates the following questions:

(1) How are Bluebikes stations distributed across different municipalities, and what influences that distribution?

(2) Is there a relationship between the number of docks at a station and its location?

Data

Description

The main dataset, downloaded from the Bluebikes website, includes all Metro Boston Bluebike stations. For geographic and demographic factors, additional datasets were merged:

- **Colleges and university locations (**Mass.gov/MassGIS) [2]: to calculate distance to campuses.
- **MBTA station locations (**Mass.gov/MassGIS) [3]**:** to calculate distance to nearest MBTA station.
- **Municipality population (**MA Dept. of Revenue) [4]
- **Largest age group by municipality (**CensusReporter.org) [5]**:** manually compiled.

The final dataset supports analysis of how geographic and demographic features relate to station size.

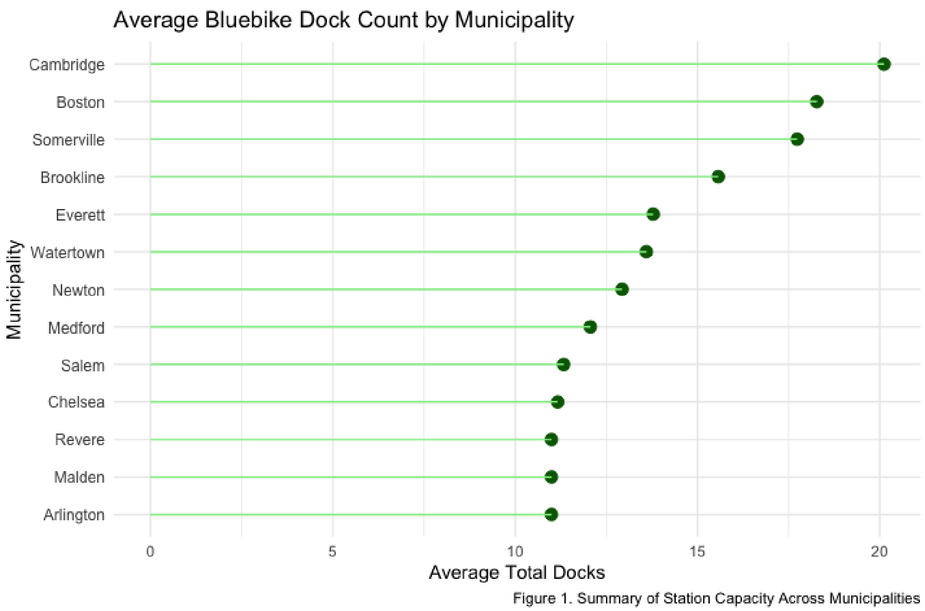

Figure 1. Summary of Station Capacity Across Municipalities

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

Cleaning

Renamed columns, removed extra headers, and converted variables (e.g., latitude, longitude, dock count) to numeric. “No ID pre-March 2023” entries were set to NA in station_ID. Distance to campuses and MBTA stations were calculated, and datasets merged by station or municipality.

Data Modeling: First-Order Model

A multiple linear regression model was fit using four predictors: 2023 population, largest age group, distances to the nearest college, and MBTA station. Stepwise selection via AIC refined the model. Due to mild non-constant variance, a Box-Cox transformation with $\lambda = -1$ was applied to the response. The final model ($R^2 = 0.385$, adj. $R^2 = 0.381$, F-test $p < 2.2 \times 10^{-16}$) showed no multicollinearity concerns. The best first-order model is:

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

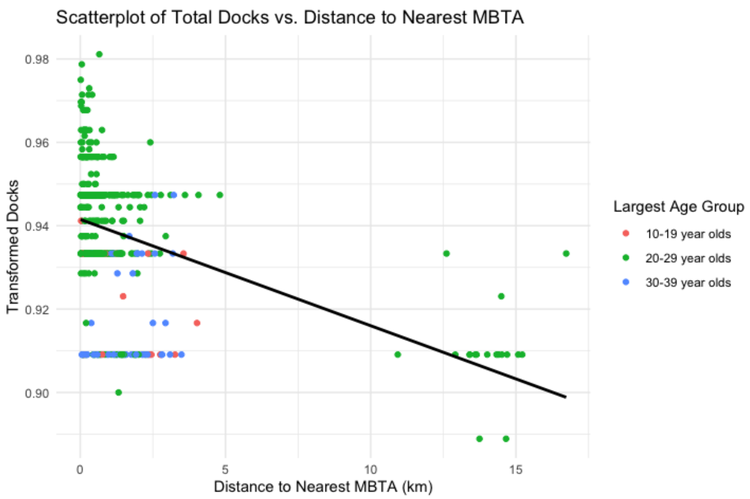

Figure 2. The scatterplot shows the distance to a MBTA station’s effect on docks.

Data Modeling: Interaction Model

To explore how demographics and geography relate to bike dock counts, I fit a linear model including an interaction between largest age group and distance to the nearest MBTA station. This model builds on the first-order model, which included these same two predictors, allowing for testing whether adding interaction terms improves model performance.

A partial F-test comparing the interaction model to the simpler first-order model found a statistically significant improvement in fit ($p = 0.004$), supporting the inclusion of interaction terms. The final model is:

$$\widehat{Total\_Docks}^{-1} = 0.92 + 0.02 * Largest\ Age\ Group(20-29) - 0.02 * Largest\ Age\ Group\ (30-39) - 0.003 * Distance\_to\_Nearest\_MBTA\_km + 0.0008 * Largest\ Age\ Group(20-29) * Distance\_to\_Nearest\_MBTA\_km + 0.006 * Largest\ Age\ Group(30-39) * Distance\_to\_Nearest\_MBTA\_km + error$$

This model explained ~39.74% of the variance in transformed dock counts (Adj $R^2 = 0.3917$).

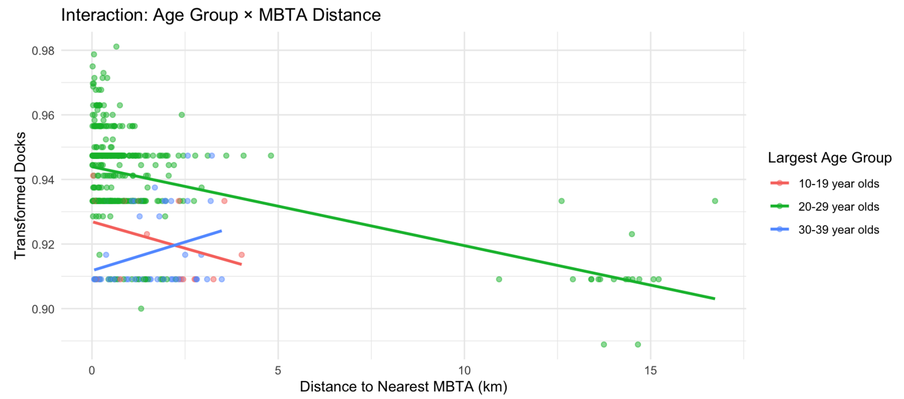

Figure 3. These visualizations highlight the ”significant” interaction effects in the final model.

Results

- The **first-order model** showed a positive association between the 20–29 age group and transformed dock count and a slight negative association with MBTA distance.
- The **interaction model** reveals that in areas where 30-39 year-olds are the largest age group, more distance from an MBTA station is linked to more docks.
- These results highlight the importance of age demographics and proximity to infrastructure in dock placement.

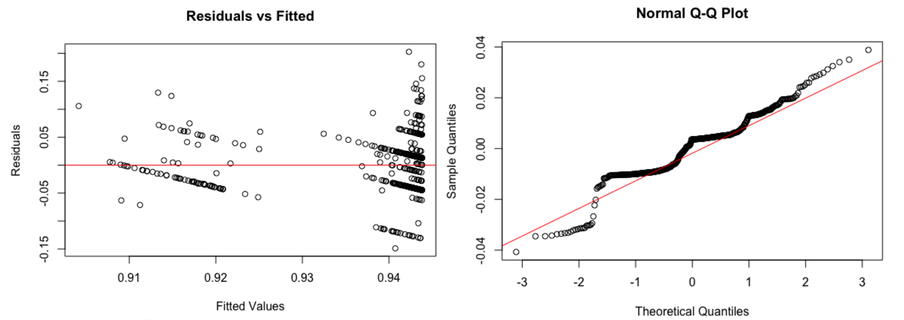

Figure 4 & 5. Diagnostic plots for the first-order model show no strong violations of regression assumptions. Residuals appear roughly normally distributed.

Data Ethics & Limitations

This project explores factors linked to Bluebike station placement and usage, but it’s important to consider ethical concerns. Relying on existing infrastructure and data may overlook underserved areas, reinforcing access gaps. Prioritizing proximity to transit may miss communities most in need of improved service.

Limitations include:

•Spatial data may violate independence assumptions.

•Manually compiled age group data may contain errors.

•Some datasets may be outdated or not reflect recent 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/