Bluebike Summer Usage Pattern: Determinants of a Station's Popularity

Ashley Kang Wellesley College Class of 2025 Data Science Major Capstone

Introduction

Background:

Bluebike is a public transportation system centered in the Metro Boston area that offers users over 4,500 bikes from more than 480 stations, ranging from Arlington to Watertown.

Research Question:

What station characteristics and demographic factors contribute to the popularity of Bluebike stations in Massachusetts during the summer months?

Data Cleaning

Sources:

- Bluebike Station Data
  - Indiv. station information: total docks, municipality, seasonal status, location (lat. / long.)
- Bluebike 2020-2024 trip history for the summer months (June through Sept.)
  - Indiv. trip information: total duration, user's membership status, start/end station
- Higher Educational Institution Data
  - US institution information: name, location (lat. / long.), state abbreviations
- MBTA Train Station Data
  - Individual station information: name, location (lat. / long.)
- Massachusetts Weather Data
  - avg. minimum temperature, avg. maximum temperature, avg. precipitation
  - Used separate datasets (above/below the Charles) due to limited datasets
- New variables:
  - Trip Duration (Numeric)
  - Trip Displacement (Numeric)
  - Population (Numeric)
  - n_institution_1km (Numeric): institutions within 1km of the station
  - n_mbta_1km (Numeric): MBTA stations within 1km of the station

Data Manipulation:

- Dimensions: 8800 observations x (27 predictors + 1 response)
- Aggregated monthly
- Removed rows with missing and non-imputable station information
- Tested Full, AIC, and BIC first-order models for model selection

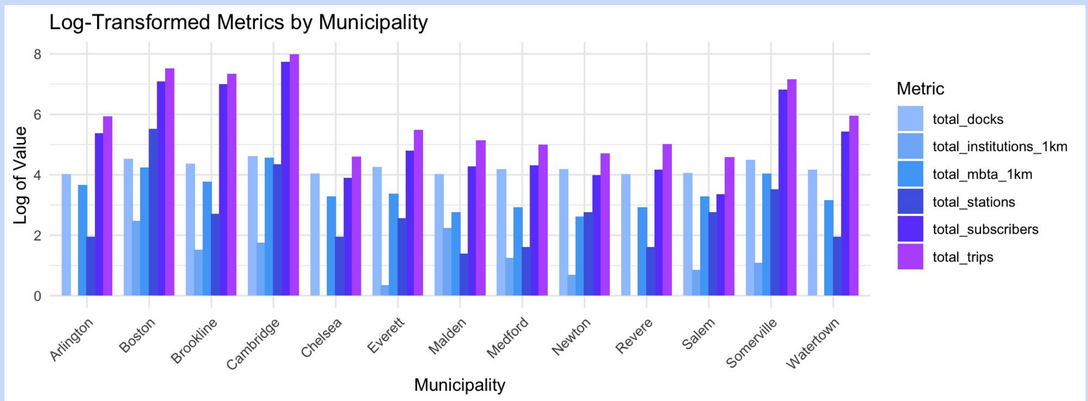

Figure 1: *Barplot by municipality showing the distribution of five key factors logged along with total trips as the response variable. We can see that most of the factors are positively correlated with the response variable total_trips.*

Methodology

STEP 1: Check multicollinearity: remove the 'population' variable due to the VIF score being 53

STEP 2: Identify the best first order model by comparing multiple metrics, including the 10-fold CV

	Full Linear Model	AIC	BIC	BIC outliers Removed
RMSE	18.798	18.829	18.831	**18.721**
R squared	0.965	0.965	0.964	**0.965**
MAE	11.539	11.544	11.515	**11.476**
p	26	23	21	**21**

Table 1: *Table for comparing the four models. We can see that the all three metrics don't vary much by model, but the BIC model with the significant outliers removed tends to be slightly better.*

STEP 3: Check the assumptions for the best model

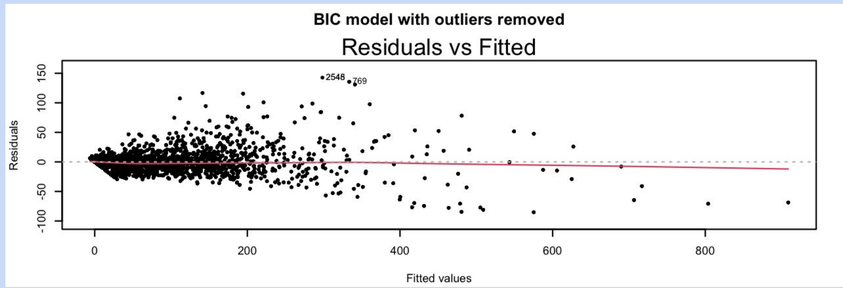

Figure 2: *The larger cluster observed in the plot corresponds to a higher concentration of observations in Boston, Cambridge, and Somerville. This grouping is expected, given the greater number of data points from these municipalities. Despite this, the residuals appear to be evenly distributed around the zero line across the range of fitted values, suggesting that the model satisfies the constant variance assumption.*

Results and Model

Based on the observations, the linear model appears to be the best model for predicting the popularity of a station, which I have defined as the total number of trips that started and ended at that station.

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
Intercept	**-926.9**	Municipality - Revere	**-14.25**	Municipality - Watertown	**7.274**
Seasonal status: winter storage	**-2.011**	Municipality - Salem	**-41.79**	Total Docks	**0.1746**
Seasonal status: year round currently stored	**-19.32**	Municipality - Somerville	**-2.642**	Institutions within 1km of the Station	**2.181**
Municipality - Cambridge	**-7.383**	MBTA Stations within 1km of the Station	**-0.1379**	Latitude	**161.7**
Municipality - Chelsea	**-11.05**	Seasonal status: year round	**0.4774**	Longitude	**83.31**
Municipality - Everett	**-9.042**	Municipality - Boston	**11**	Number of Trips made by Subscribers	**1.296**
Municipality - Malden	**-14.79**	Municipality - Brookline	**8.073**		
Municipality - Medford	**-8.434**	Municipality - Newton	**10.74**		

Table 2: *Table of my final model's predictors and coefficients. Positive coefficients are highlighted in blue, and negative coefficients are highlighted.*

Discussion and Conclusion

- **Discussion:** The weather related variables do not have an impact on the model. This could be due to the consistent summer temperatures over the years, as the temperature during the summer does not vary significantly over years.
- **Limitation & Future Work:** The relationship between the predictors and the outcome may be non-linear, which standard linear models may not fully capture. I will propose the Random Forests method to better capture complex, non-linear patterns in my predictors. Also for a better prediction, I can add tourist attractions as a factor since people tend to visit Boston to tour the area.