Bluebike Subscriber 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 and Methods

Sources

- Bluebike Trips for the summer months from 2020 to 2024
  - Attributes: start/end stations (station id, trip duration, trip displacement)
- Bluebike Station Data
  - Attributes: station information (total docks, municipality, seasonal status)
- Institution Location Data
  - Attribution: name, location (Lat, Long, State Abbreviations)
- MBTA Train Station Data
  - Attribution: name, location (Lat, Long)
- Massachusetts Weather Data
  - Attributions: TMAX, TMIN, PRCP

Data Manipulation

- Merged all the user datasets and added/calculated necessary columns (e.g. year, month, trip duration, trip displacement)
- Using the station data, for each station, calculated the number of trips that started/ended at that station, number of trips that were made by subscribers, number of institutions within the 1km boundary, and number of MBTA stations within the 1km boundary
- The data is aggregated monthly.
- Tested models using a full model and compared them using AIC and BIC for model selection.

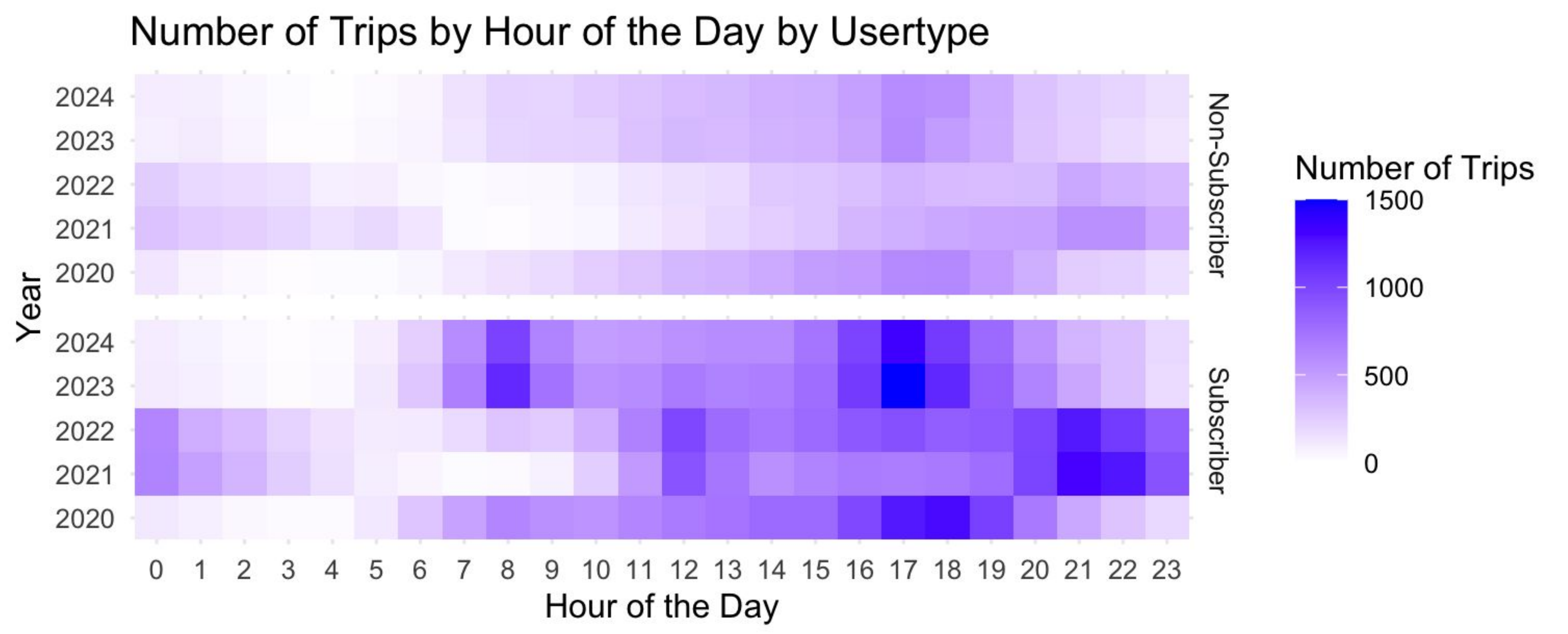

Figure 1: Heatmap of the number of trips by hour of the day by user type (subscriber vs. non-subscriber)

Results and Model

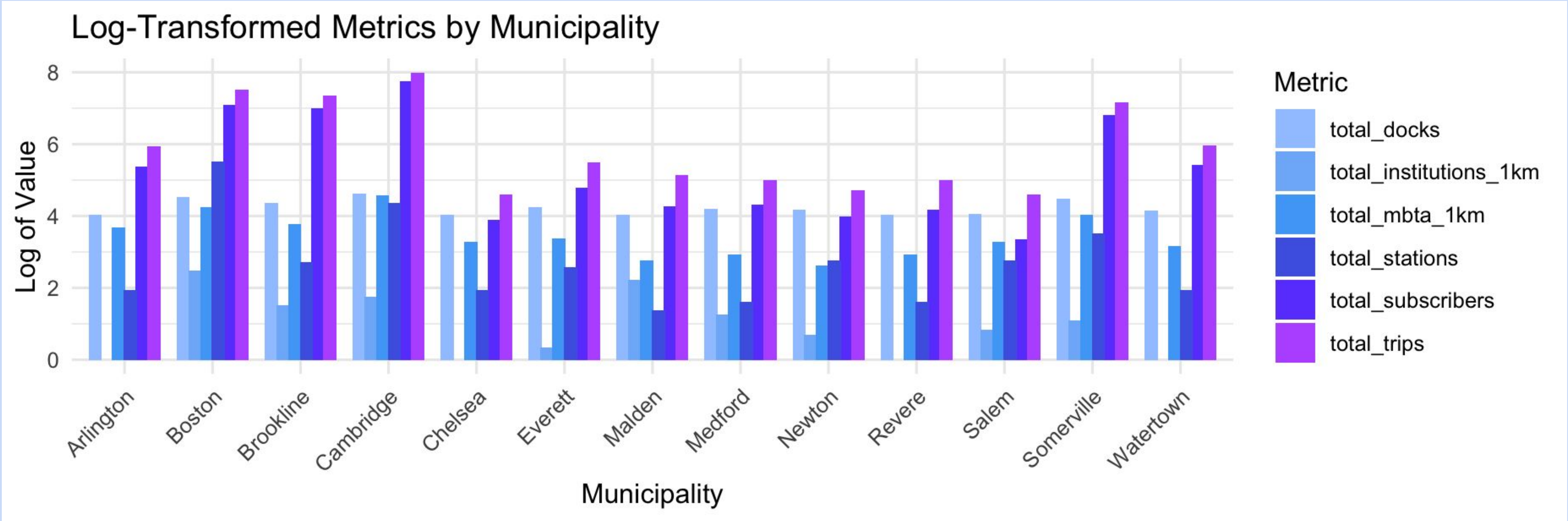

Figure 2: Barplot by municipality showing the distribution of five key factors along with total trips as the response variable. We can see that most of the factors are positively correlated with the response variable total_trips.

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.

$$\begin{aligned}\widehat{\text{total\_docks}} = &-926.9 - 2.011 * \text{seasonal.status.Winter\_Storage} + 0.4774 * \text{seasonal.status.Year\_Round}\\ &- 19.32 * \text{seasonal.status.Year\_Round\_Currently\_Stored} + 11.00 * \text{Municipality.Boston}\\ &+ 8.073 * \text{Municipality.Brookline} - 7.383 * \text{Municipality.Cambridge} - 11.05 * \text{Municipality.Chelsea}\\ &- 9.042 * \text{Municipality.Everett} - 14.79 * \text{Municipality.Malden}\\ &- 8.434 * \text{Municipality.Medford} + 10.74 * \text{Municipality.Newton}\\ &- 14.25 * \text{Municipality.Revere} - 41.79 * \text{Municipality.Salem} - 2.642 * \text{Municipality.Somerville}\\ &+ 7.274 * \text{Municipality.Watertown} + 0.1746 * \text{total\_docks} + 2.181 * \text{n\_institutions\_1km}\\ &- 0.1379 * \text{n\_mbta\_1km} + 161.7 * \text{Lat} + 83.31 * \text{Long} + 1.296 * \text{membership\_total}\end{aligned}$$

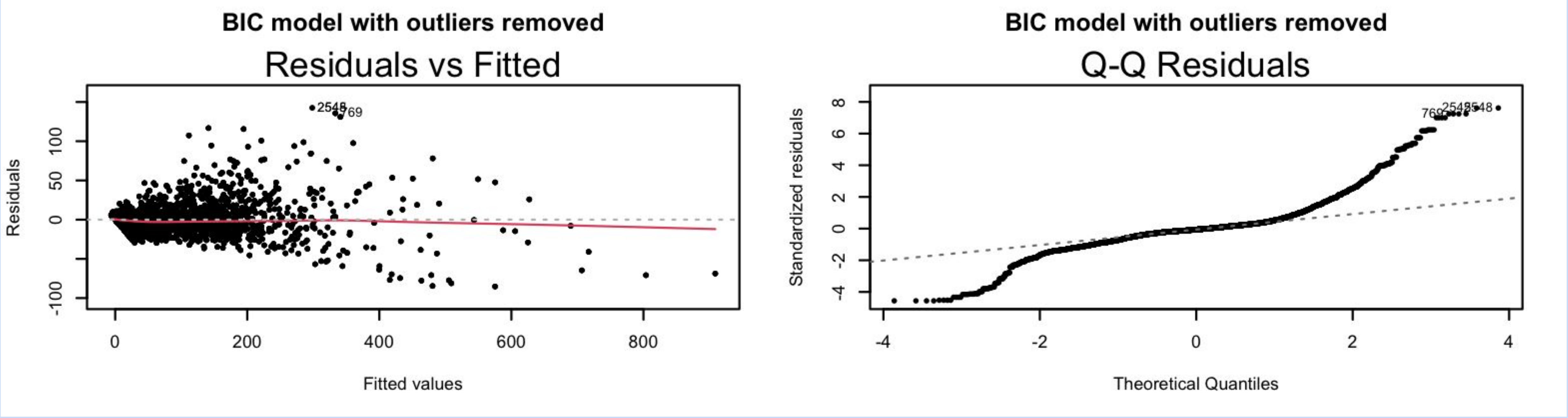

Figure 3: Residual v. Fitted plot for the final model. This plot shows that the linearity assumption meets.

Discussion and Conclusion

- The average minimum and maximum temperatures, as well as precipitation 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: The relationship between the predictors and the outcome may be non-linear, which standard linear models may not fully capture.
- Future work: I can add tourist attractions as a factor since people tend to visit Boston to tour the area.