Bluebikes: Who even uses them?

Gabrielle Dominique and Jose Velasquez

Table of contents

Introduction

“Bluebikes is public transportation by bike! With more than 5,300 bikes and 550 stations, it's a fast, fun, and affordable way to get around Metro Boston.”

We’ve all seen one, some even used one, but who is keeping this business afloat? Specifically, we aimed at analyze the relationship between Colleges, Universities, and Bluebikes.

Research Question

Is there a relationship among college campus characteristic and Blue Bike usage, if so, what characteristics drive Blue Bike usage?

We used a Multiple Linear Regression analysis to examine the relationship.

Data

Dataset	Details	Variables	Role in model
Blue Bike Trip Data	Around 4 million individual trips in the year 2024	Latitude, Longitude, Start Station ID, Data	We used this data to calculate the average monthly use (Y)
University Data	50 college campuses in the Greater-Boston Area, including Boston, Cambridge, Somerville, and Medford	Student Population, University Type, Public/Private, Latitude, and Longitude	We were able to calculate the distance from Start Stations to the nearest campus using euclidean distance. Moreover, we used college characteristics in our model.
MBTA T-Stops	Every train station in the Greater Boston Area	Station ID, Station name, Latitude and Longitude	We mapped these points to the nearest university campus using euclidean distance, and then created a binary variable.

Spatial Linking

Because universities generate, high student foot traffic, predictable commute patterns, strong demand for micro mobility, in order to study how universities influence Bluebike ridership, we need to know:

1. Which Bluebike station is closest to each university?

2. How close are university campuses to the nearest T stop?

Method

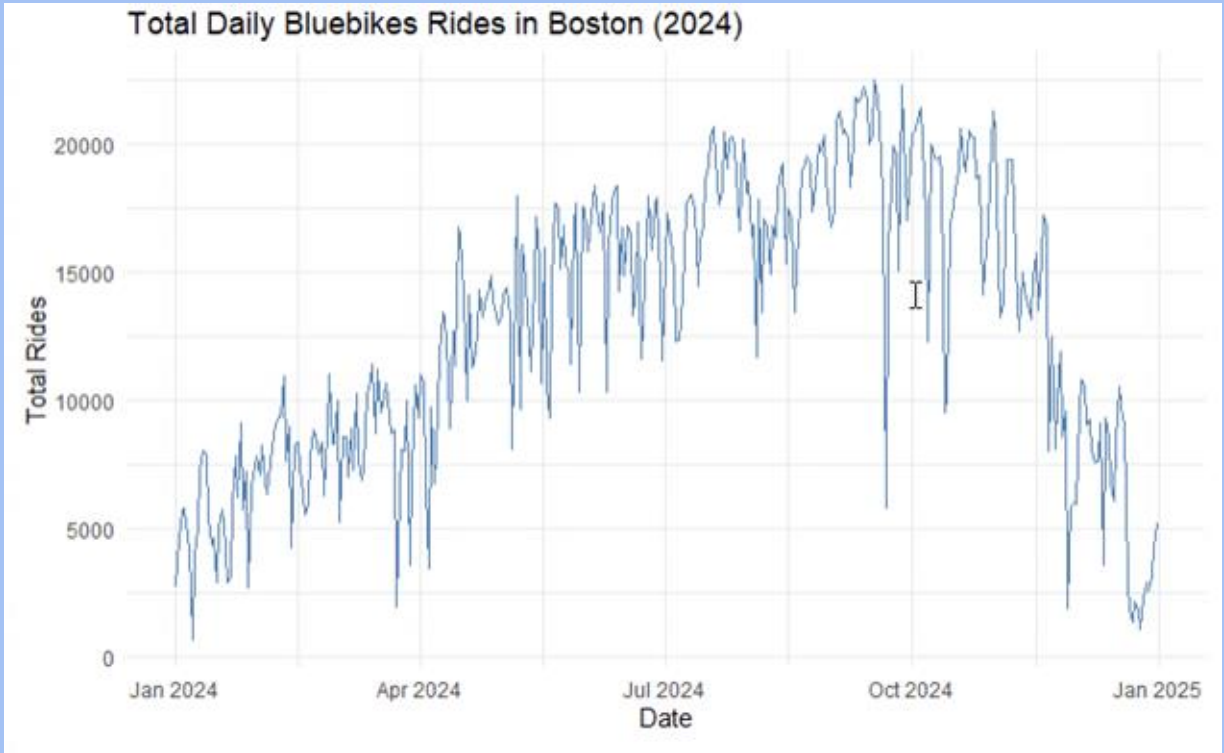
Converted universities and Bluebike stations into spatial objects (sf package)
Used st_nearest_feature() to find:

- the nearest Bluebike station to each university
- the nearest MBTA T stop to each university

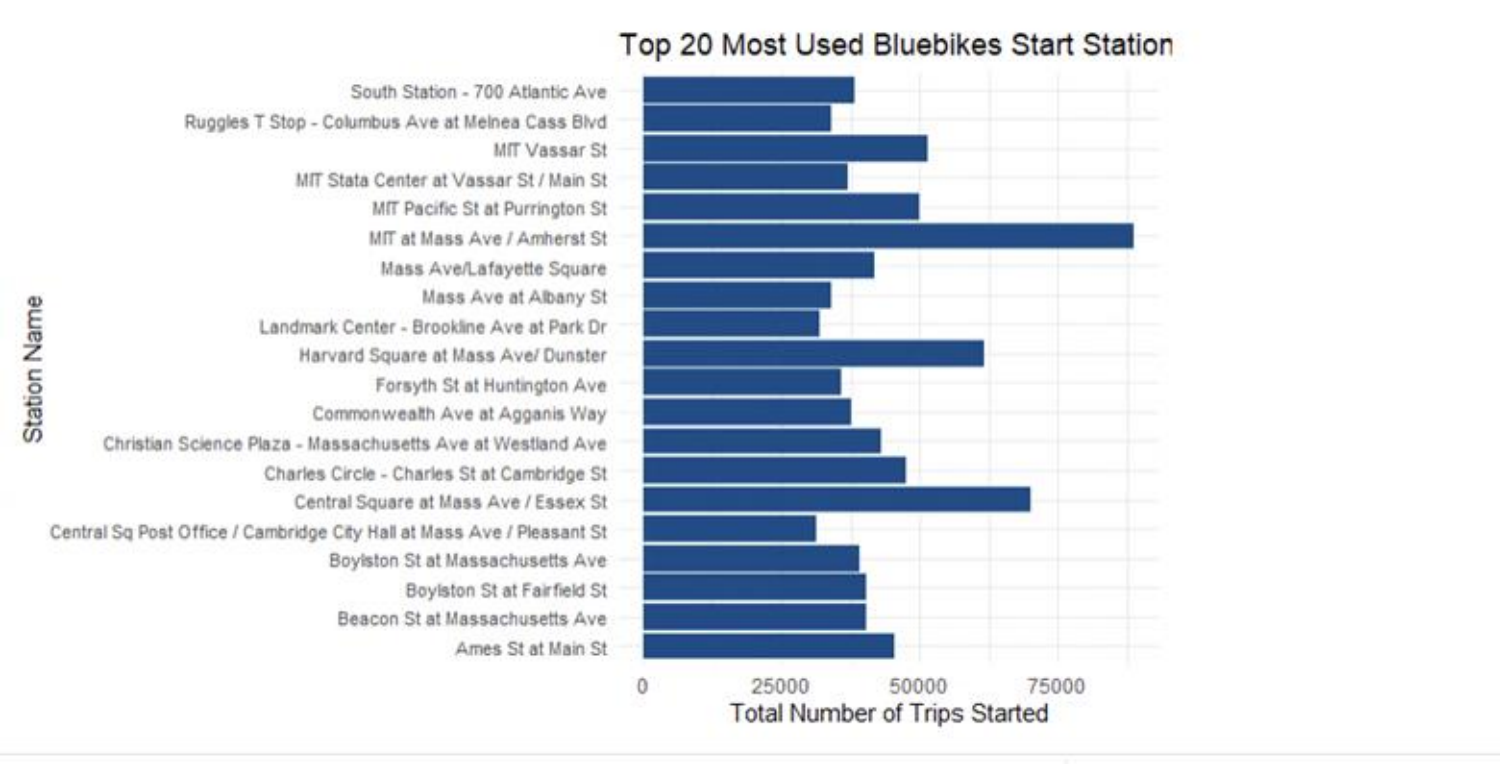
Calculated:

- Distance_to_BlueBikeStation_miles
- Distance_to_TStop_miles

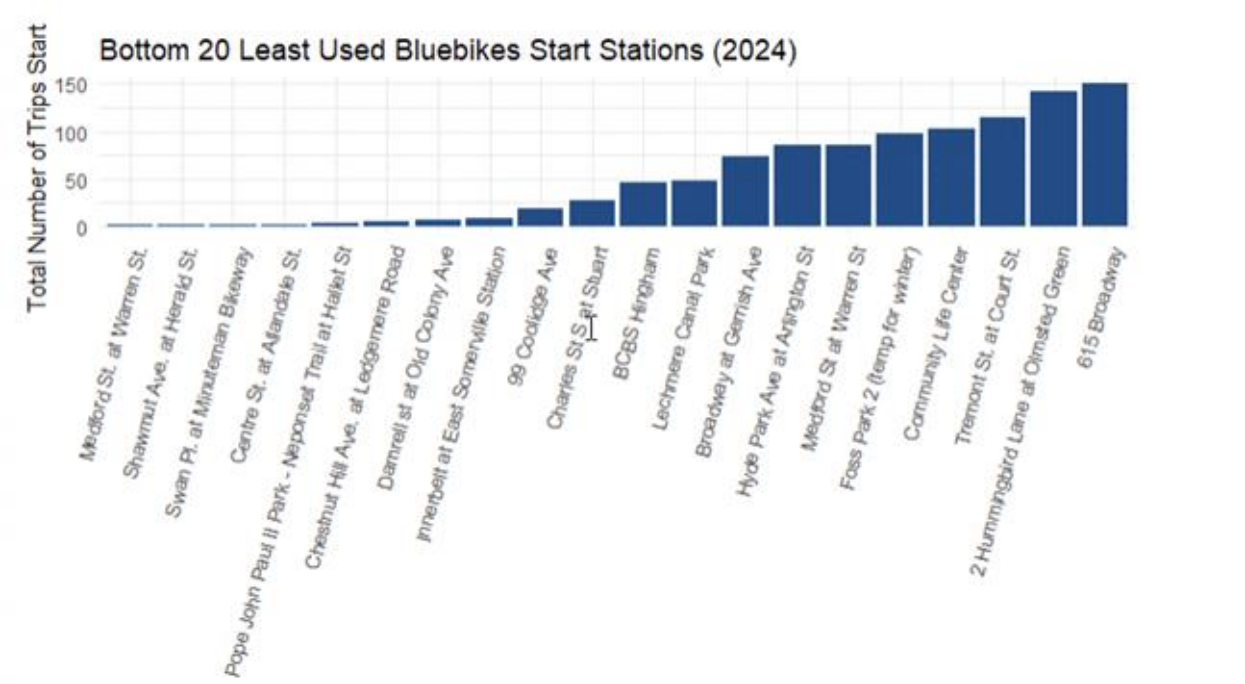
Bluebike Usage 2024



Most used stations 2024



Least used stations 2024



Model 1: Academic and Non Academic Months

Regression:

$$\text{Avg Monthly Rides}_i = \beta_0+\beta_1(\text{Student Population}_i)+\beta_2(\text{Distance to Bluebike Station}_i)+\beta_3(\text{Distance to T Stop}_i)+\beta_4(\text{City Dummies}_i)+\beta_5(\text{School Type Dummies}_i)+\beta_6(\text{Category Dummies}_i)+\epsilon_i$$

Where:

- i indexes each university
- Dummy variables represent CITY, TYPE, and CATEGORY(research, law school, liberal arts, etc)

Model 1 Results

Residual Standard Error ≈ 13,080
→ The model's predictions are off by about **13k rides** on average.

Multiple R^2 = 0.3067 (~31%)
→ The model explains only ~31% of the variation in average monthly rides.

Adjusted R^2 = –0.6177
→ The added predictors do not actually improve prediction. Useless predictors were strongly penalized.

F-statistic p-value = 0.9965
→ The model is NOT statistically significant overall, the predictors do not meaningfully explain average monthly ridership.

Conclusion
→ The full model with many categorical variables (CITY, TYPE, CATEGORY) is overfitting
→ There is no statistical evidence that universities closer to Bluebike stations or T stops have higher Bluebike usage

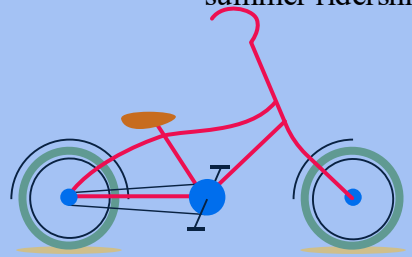
Model 2: Solely Academic Months

Academic Months Regression Model

- Uses the same variables as our full model

  (Student Population, Distance to Bluebike Station, Distance to T-Stop, CITY, TYPE, CATEGORY)
- But only includes months where universities are in session:

  Jan, Feb, Mar, Apr, Sept, Oct, Nov, Dec
- Purpose: To see whether university characteristics explain Bluebikes usage during the school year, excluding summer ridership.



Model 2 Results

Residual Standard Error ≈ 11,070
→ The model's predictions are off by about **11k rides** on average.

Multiple R^2 = 0.3014 (~31%)
→ The model explains only ~31% of the variation in average monthly rides.

Adjusted R^2 ≈ –0.63
→ The added predictors do not actually improve prediction. Useless predictors were strongly penalized.

F-statistic p-value = 0.9971
→ The model is NOT statistically significant overall, the predictors do not meaningfully explain average monthly ridership.

Conclusion
→ The full model with many categorical variables (CITY, TYPE, CATEGORY) is overfitting
→ There is no statistical evidence that universities closer to Bluebike stations or T stops have higher Bluebike usage

Interpretation

Our analysis found that campus characteristics alone are insufficient predictors ($p > 0.05$). This argues that Bluebikes demand is not driven primarily by our predictions, but likely by dynamic factors like daily weather and time-series trends, which requires the more advanced models.

Thanks/ Questions/ Feedback

Data

Data Sources

- **Boston Colleges & Universities Dataset**
  - University name, address, city, ZIP
  - Latitude & longitude
  - Enrollment size
  - Campus housing
  - School type (public/private)
  - Campus setting (city/suburb, small/midsize/large)
- **Bluebikes 2024 Ridership**
  - ~4 million trips
  - Station coordinates (latitude/longitude)
- **Boston T Stops**
  - Latitude & longitude of subway stations