Bluebikes: recommendation of locations for new bike stations

1. Introduction

a) Problem statement

Bluebikes is a continuously growing bike share system serving Boston and the four nearby cities, Somerville, Cambridge, Brookline, and Everett. Our goal is to recommend sensible locations for new Bluebikes stations for their future expansion.

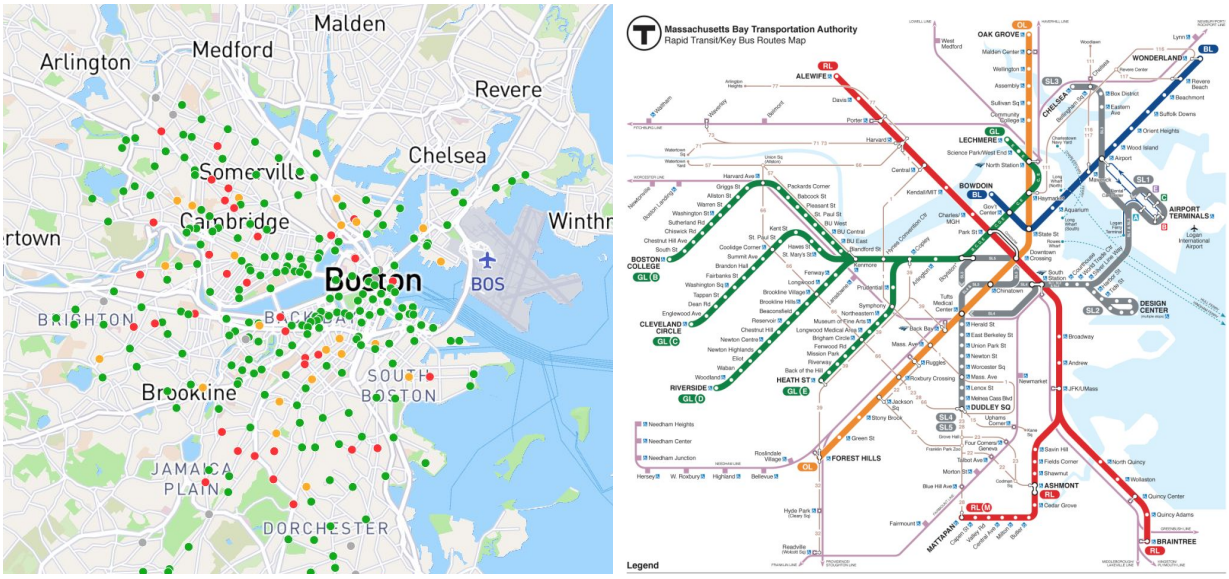


Figure 1:
(Left) Bluebikes stations with colors indicating the availability of bikes (image source: Bluebikes webpage).
(Right) Public metro 'MBTA' lines and stations (image source: MBTA webpage).

b) Background

Bluebikes (formerly Hubway) launched in Boston city proper in 2011 with 60 stations and 600 bikes. Since then, they have grown incrementally over 8 years. The company partnered with Cambridge, Somerville, and Brookline in 2012, had a major expansion to residential areas of Boston in 2015-2017, and is starting their service in Everett this summer of 2019. The system now deploys over 260+ stations with more than 2500+ bikes in operation.

Through their website, Bluebikes takes suggestions for new bike station locations. Anyone can drop a pin on the map and add a short reasoning why one would like a station there. People can also upvote the existing pins that other people have created.

c) Goal

This project aims to provide a more thorough proposal for locations for new bike stations backed up by data-driven analysis. We combine bike trip history data, the MBTA station information, and some other intrinsic features of the Boston metropolitan area to understand Bluebikes usage patterns and estimate user volumes. Also taken into account are the MBTA Green line extension project[1)] spanning Cambridge, Somerville, and Medford scheduled for completion in 2021 and the new casino Encore[2)] that will open in summer 2019 in Everett. Based on our prediction on user volumes, we recommend new bike station locations that we believe will be both profitable for the company and beneficial to customers.

References:

1) https://www.mass.gov/info-details/about-the-green-line-extension-project
2) https://en.wikipedia.org/wiki/Encore_Boston_Harbor

2. Datasets

a) Bluebikes trip history from Bluebikes

The trip history dataset was retrieved from Bluebikes website.Trip history is organized by month and comes in a csv format. We read in data from July 2017 to May 2019 as a DataFrame called tripdf.

The 15 attributes of this dataset tripdf can be categorized in the following way.

- Trip-time information: trip duration (in seconds), starttime, stoptime
- Start station information: station id, station name, latitude, longitude
- End station information: station id, station name, latitude, longitude
- User information: bike id, user type(annual or monthly subscriber or not), birth year, gender

b) Bluebikes station data from Bluebikes

Bluebikes publishes a list of bike station in a csv file format. Their latest release, however, is from July 2017 and is not up to date. Alternatively, Bluebikes releases real-time data of bike stations in a json file format. We use this json file and store in a DataFrame station.

The columns in station are labeled 'b ba bk bl bx d da dx id la lc lo lu m n s st su', which are not reader friendly. Fortunately, the real-time data of bike stations is also available in an xml file format. The xml feed has

interpretable tag names-for example, <nbEmptyDocks> instead of 'da' and <latestUpdateTime> instead of 'lu'. From there, we deduce what each column in station represents. We update station to only contain the following six attributes:

- station id
- station name
- latitude, longitude
- district (city name)
- the number of bikes available
- the number of empty dock available.

c) Zip codes of the bike stations via Google Map API

(This part of analysis was completed before Blueblikes started their service in Everett.)

Originally, there are only four districts in station: Boston, Brookline, Cambridge, and Somerville. Although these are correct and official city labels, district divisions can be improved based on the characteristics of neighborhoods of Boston.

As a first step, we make use of Google Map API to get the zip codes of each bike station in station. The zip codes are used to split Boston into eastern Boston, southern residential Boston, and downtown Boston. West side of Boston is grouped with Brookline as that part of the community shares more similarities with Brookline (see Figure 3 on Page 6).

d) Weather data

We obtain Boston daily weather summaries from National Centers for Environmental Information and parse it as a DataFrame daily. The attributes of the data includes

- weather station name (136 distinct observatories in the vicinity of Boston, MA)
- observation date
- geographic data of the station
- temperature
- precipitation, etc.

The full documentation for this dataset can be found under Documentation and Samples on this page.

e) Other shapefiles for visualization

We plot Bluebike stations along with MBTA lines and MBTA station location. We also identify rivers, channels, bays and other water body features. The relevant shapefiles are obtained from the following:

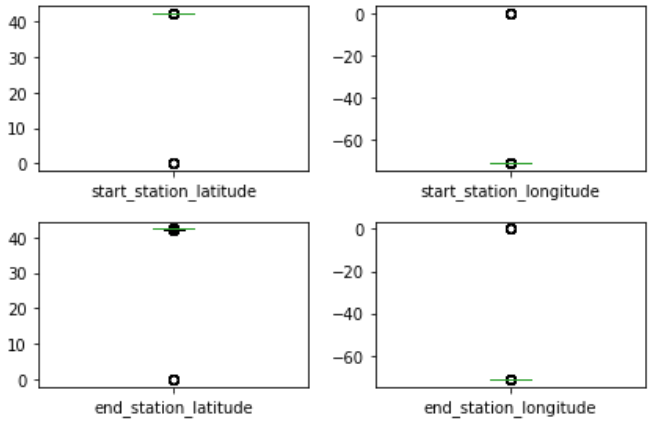
- MBTA line arcs and MBTA station nodes from MassGIS (Bureau of Geographic Information)
- Water body features from Boston Open Data

3. Data Cleaning and Data Wrangling

a) Matching the stations in trip history data and the stations in station.

In tripdf, the number of unique features of stations ranges from 321 to 366.

- start station names: 347
- start station id's: 321
- latitude, longitude of start stations: (365, 366)
- end station names: 348
- end station id's: 321
- latitude, longitude of end stations: (365, 366)



Box plots of latitudes and longitudes of the stations show that there are outliers, the ones taking values of zeroes. We conclude that the station id's are the most reliable entries to identify stations. We drop station names and geolocations and take only the useful portion of tripdf and store it as trip_simp.

On the other hand, the number of unique stations in station is 277. This should be the correct number as this data reflects real-time station information. We suspect that trip_simp contains some duplicate stations or stations that are no longer in operation.

We choose to outer merge two datasets trip_simp and station. Since we do not know that the unique id's in one dataset is a proper subset of that in another, outer merge is a suitable choice to find any mismatches. The merge result in total 50 unmatched station id's in trip_simp. Further investigation confirms that

- Some stations were relocated, but very little. For example, stations were moved diagonally across the same intersection. The original station location was listed under one id, and the new station location was listed under the id. In this case, we will consider two stations the same and will map the old station id's to new station id's.
- Some stations are test stations.We remove any trip observations that start or end at these test stations.

We construct a dictionary {old id's : new id's} based on the findings, and apply the dictionary to clean up trip_simp accordingly.

We also find that there are five stations in station that never show up in the trip history dataset trip_simp. We verify on Bluebikes site that four out of five such stations are not yet in operation. The fifth station is running, but does not show up in trip history. It may be that this station started operating very recently in May 2019. When this initial analysis is performed, our trip history data spanned up to April 2019.

b) Creating zones

We use Reverse Geocoding function of Google Map API to obtain zip codes of stations in station. Upon feeding the latitudes and longitudes of stations, it returns the full address, among with many other entities, of the location. We extract zip codes from the full address and add it in a new column "zip" in station.

Most of the stations had their full addresses formatted

Street name, City, MA 00000, USA.

We execute a code that extracts the second to last word in the full address. However, there are a few incidents where Google API returned the full addresses

Street name, City, MA 00000, United States.

We manually handle these cases.

Based on zip column of stations, we split the Bluebikes service area into 6 zones (see Figure 3 on Page 6).

- Downtown Boston
- Residential south Boston
- East (and northeast) Boston
- West Boston (including Brookline)
- Cambridge
- Somerville

4. Exploratory Data Analysis and Initial Findings

a) Popular origins and destinations

We plot the locations of all Bluebikes stations on top of layers of MBTA lines and MBTA stations. We provide two figures--one to illustrate the popularity of start stations and another for the end stations. The marker size denotes the popularity of each station. The figures show that the popularity goes down as we move outward from the center.

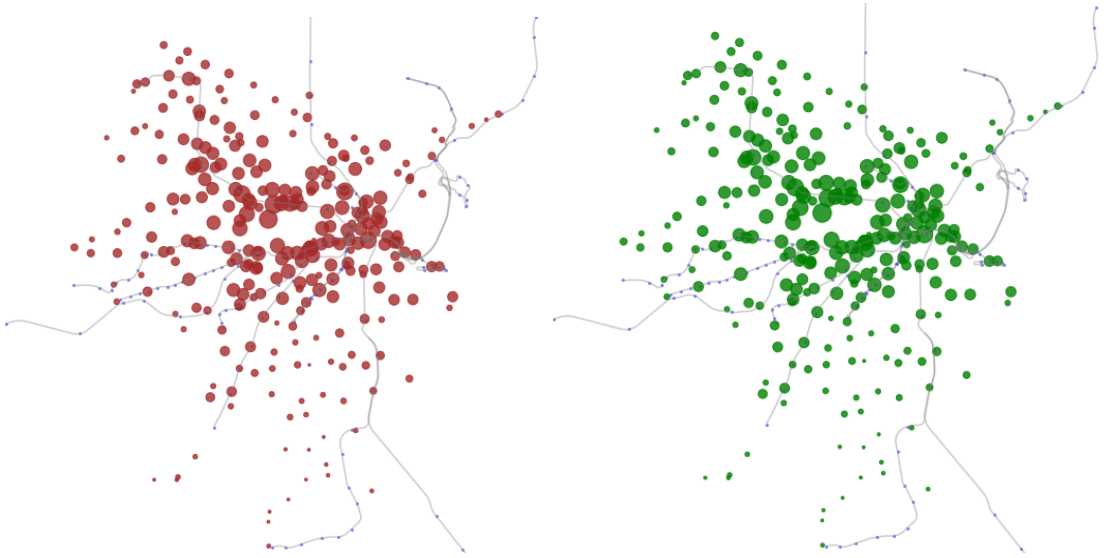


Figure 2: Popular start stations (brown) and end stations (green).

b) Station by zones & the most centered station

We also plot the stations by 6 zones that we have created

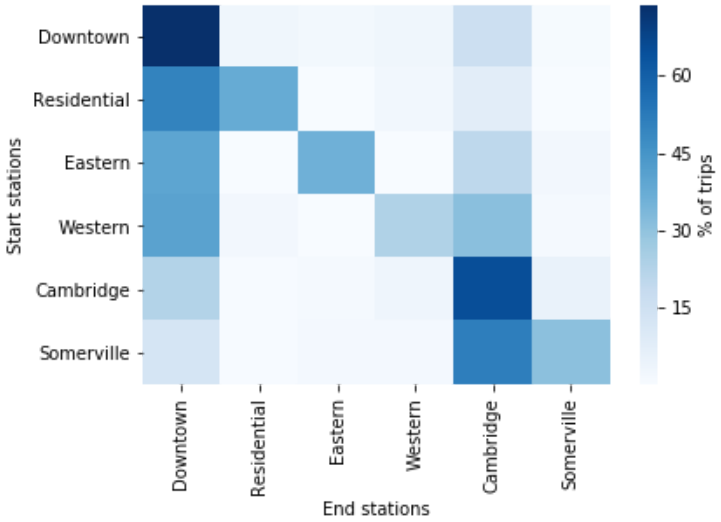
- Black: Downtown Boston
- Green: Residential south Boston
- Blue: East (and northeast) Boston
- Red: West Boston (including Brookline)
- Yellow: Cambridge
- Brown: Somerville

Figure 3: Bluebikes stations split into 6 zones.

We identify the most center station, the station whose average distance to all other stations is the smallest. The center station is marked with a teal colored star.

c) Net flow between zones

We see that diagonal entries, and the first and fifth columns have darker colors. This implies that the majority of bike trips are staying within same zone or converging to Downtown or Cambridge. Specifically, 74% of bike trips starting from Downtown ends in Downtown and 65% of trips starting from Cambridge ends in Cambridge.



d) Usage pattern for weekdays and non-weekdays (non-working days)

More trips are made during the weekdays.

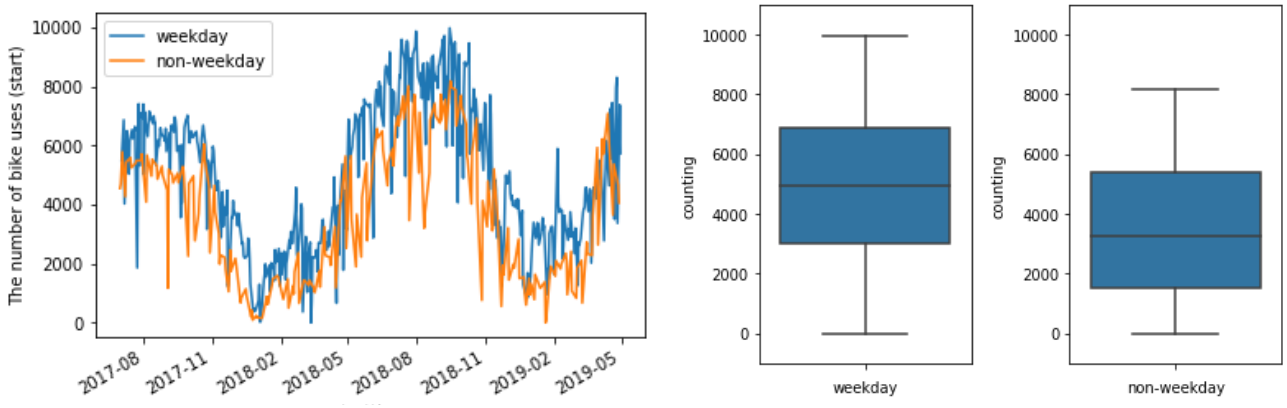


Figure 4: (Left) The number of bike uses for weekday and non-weekday.
(Right) Average number of bike uses for weekday and non-weekday.

e) Usage pattern by day of week and by week of year

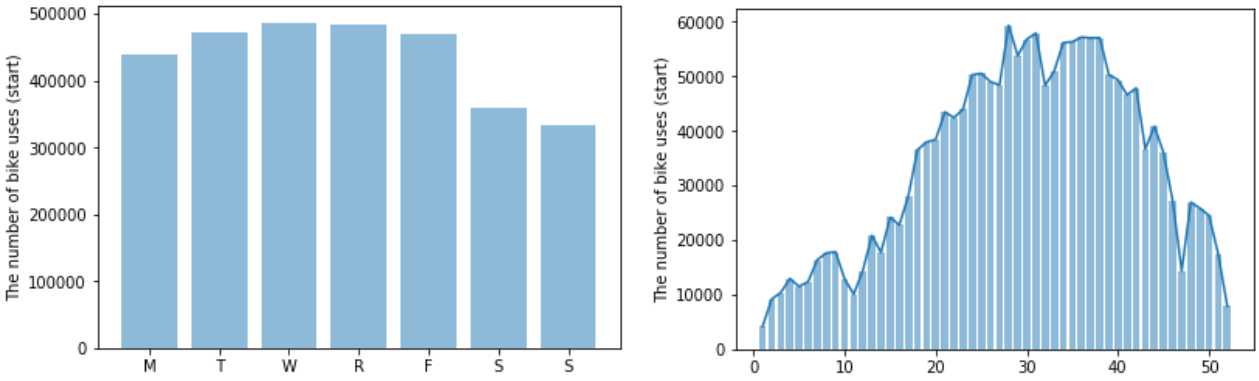
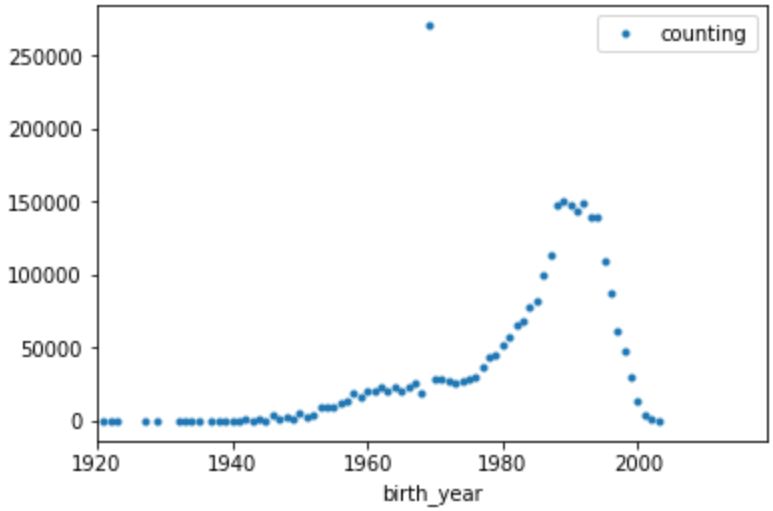


Figure 5: (Left) The average number of bike uses by day of the week.
(Right) The average number of bike uses by week of the year.

g) Usage pattern by birth year

Most of the users are born between 1985-1995.



5. Future Work

- We will explore the relationships between the exiting bike docking stations and the metro stations. Is the bike user volume high near metro stations?
- For each bike station, where is the nearest bike station? Investigate minimum, maximum, and mean distance. Can we write the mean as a function of the distance from the city center?
- Determine the portions of bike trips that are made for leisure and for commuting.

- Inter-metroline connections are possible only in downtown. For example, there is no short metro transfer if one wishes to travel from the end of Orange line to the end of Red line. We would like to see if Bluebikes can fill such gaps.
- The new Green extension will go through Cambridge and Somerville. How can we update stations in these regions?
- We will investigate which stations are worth opening during the winter. Currently, only Cambridge and Downtown Boston locations stay open through the winter. To seek answers, we will look into other columns (precipitation, snow) of weather dataset daily. We can also think about importing an hourly weather dataset.