College Use of Bluebikes in Boston

Kaitlyn O’Donnell

Northeastern University, Boston, MA, USA

# Abstract

In writing up your report, follow this template exactly. Your report will be compiled into the database design class proceedings. Future students will learn from your work. In the abstract, provide a summary overview of your project, its goals and accomplishments.

# Introduction

Summarize your project goals. Explain the motivation and significance of your project. Outline the main features or use cases.

The goal of this project was to analyze possible relationships between Bluebike station trips and proximity to various colleges in the city of Boston. It is unclear from just trip data if Bluebike trips numbers are influenced by nearby colleges. Depending on the results, Bluebike might want to install more bike locations close to colleges or advertise more to students. Because Bluebike is owned by the municipalities in the Boston area, this demographic analysis can help inform city policy as well.

The main feature of this database is the ability to retrieve all stations within a distance of meters to parcels owned by colleges in the city of Boston. This can be used by selecting a college, or by selecting a specific parcel.

# Database Design

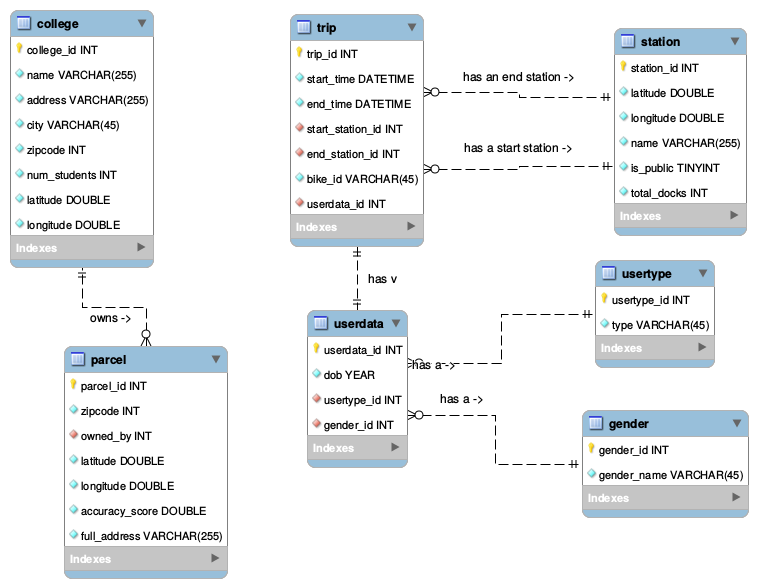


Figure 1: EER Diagram of Database

The key entities in the database are College, Parcel, Trip, and Station. A College can own zero or more Parcels. A Parcel must be owned by one College. Because the goal of this database is to analyze the relationship between Colleges and Bluebike trips, having Parcels not owned by a College would bloat the database unnecessarily.

The remaining key entities are Trip and Station. A Trip has exactly two stations associated with it. These are required because the original trip data did not have any entries with missing or extra stations. This is likely because a “trip” is characterized by these start and end locations. In contrast to the College-Parcel relationship, a Station can exist without any trips associated. This was mainly due to Bluebikes allowing access to their list of Stations. A secondary reason was that, logically, a Station can exist within our scope without an associated Trip. If for some reason, a Station had no Trips for a month, that would be worthy of note.

The non-key entities include the dictionary tables Gender and Usertype, and the table Userdata. Gender and Usertype are a product of normalization. It is interesting to note that the original values for “gender” from the trip data were one of the numbers 0, 1, and 2. There were no other indicators about what these might correspond to. I assigned them values based partially on their counts and the United States being a traditionally patriarchal society. Because 0 had the least values associated with it, it was assumed that meant undefined/not specified. This could also mean “other” to be inclusive of non-binary people, but it is rare that a company does this. Then it was assumed that 1 was “male” and 2 was “female”. This is usually how people who grow up in a patriarchal society order gender. Because of all these assumptions about what the gender values meant, this report does not try to draw any conclusions based off of gender.

Usertype is a straightforward table with the values “Subscriber” and “Customer”. Subscriber probably means that the user has an account with Bluebikes rather than being an occasional user.

Userdata is the collection of fields about the user of a Trip. This is a separate entity in the ER Diagram, but is not a separate table in the actual database. This is due to the one-to-one relationship between Userdata and Trip. Because the Userdata does not contain unique data for each user (birth year, usertype, and gender is not enough for uniqueness), this database cannot specify different users and associate them with multiple Trips.

# Data Sources and Methods

The Bluebike trip data for July 2019, September 2019, and stations was pulled from the Bluebikes system data page on their website[1]. July and September of 2019 were used because colleges generally have more students taking classes in September. August was not used because some colleges start their fall semesters in August, but others start in early September. This database does not contain data for other months or years because of storage and time constraints. A possible expansion of this project would be analyzing if the trends persist over different years.

The data on colleges and parcel ownership in Boston was gathered from the Analyze Boston website [2, 3].

The database uses the haversine formula on latitude and longitude values to determine distance between stations and parcels.

Explain step-by-step how you acquired your data. Document all data sources. Use numbered citations like this [1] or like this [2, 3] and list any references at the end of your report following a consistent style. Describe any work you did to modify or clean the data prior to being loaded into the database. Provide sufficient detail so that the reader could, in theory, fully reproduce your results.

# User Cases

List specific non-trivial questions that users could ask of your database. Summarize the question in English, provide a corresponding SQL query, and display your tabular output. Where appropriate, you should supplement your tabular output with charts, graphs, or other types of visualizations in order to better convey key insights.

# Conclusions

Summarize the results of your project. Be concrete about your accomplishments as well as the limitations of your work.

# References

1. Bluebikes. 2020. *Bluebikes System Data*. [online] Available at: <https://www.bluebikes.com/system-data> [Accessed 18 June 2020].

2. Analyze Boston. 2020. *Parcels 2017 Data Lite*. [online] Available at: <https://data.boston.gov/dataset/parcels-2017-data-lite> [Accessed 18 June 2020].

3. Analyze Boston. 2020. *Colleges And Universities*. [online] Available at: <https://data.boston.gov/dataset/colleges-and-universities> [Accessed 18 June 2020].

https://www.movable-type.co.uk/scripts/latlong.html

I just want to note my process:

- the csv version of the parcel data does not include lat and long data.

- so I downloaded the geoJson version and used ogr2ogr to convert the geoJson to a csv **with** the geometric data so I could properly calculate nearby stations

- this did not work because the geojson data has polygonal property areas and I really didn’t want to figure out this file format

- turns out geocoding hundreds of thousands of files costs money

- ended up filtering the files down to “like concat(“%”, colleges.name, “%”)” to get less than 500 entries and I put that into geocodio to get lat and long data