Study on Optimal Location and Amount for Placement of New Bike Rental Stations in 5 Major US Cities

Introduction

Our website is designed to analyze current data on bike sharing programs throughout the country to figure out the efficacy of the system and ways to improve. We analyzed the efficiency of each program by city by analyzing trip data for each of 5 cities (Boston, MA; Washington, DC; New York City, NY; Chicago, IL; and San Francisco, CA) to determine a ranking of how utilized each system is based on population. Each city studied had its own data portal to source this information from (see **Data Sources**). We then offered the functionality in our web application to specify a budget and desired size of additional stations in order to determine the optimal placement location for new stations, utilizing current station data for each city.

Method

We planned to study both the optimal location and amount for addition of new bike rental stations to maximize profits for rental bike companies in different cities. For each city, we obtained trip data for an arbitrary recent month (09-2018) to analyze the usage per city. This data varied in its schema city by city, so we utilized different functions to parse the data into MongoDB. We also obtained data on the existing station placement for each city, which was provided in General Bikeshare Feed Specification (GBFS), making it relatively consistent across each city's data portal.

We then used a summation of the time spent using the bike sharing program across each city as relating to the population to determine a ratio of time spent using the program versus the population for each city. This information gave us an idea of which city would be best to add new bike stations to.

To determine the amount of stations to add, we first utilized the budget data provided by the user, as well as the cost for the type of station they chose to build, to determine the maximum amount of stations the city could build (see **References**). We then implemented a linear regression using this max constraint in addition to the data relating to the user's specified city; including station amount, total bike time, and population. Viewing the data in scatter plot form, it was clear that a linear fit was the best option, as opposed to quadratic, cubic, etc. From this regression, it was clear that adding additional stations could only increase profits.

Knowing k=max_constraint, we inputted it into a K-means algorithm as k, with the points being represented by the latitude/longitude pairs for each station in the city. This gave us k cluster centers, which we claim are the optimal locations to add k new stations for the user's city.

This method is not perfect, as depending on the city's geographic features, cluster centers could land anywhere from the optimal location on a city block, to in the middle of a body of water, to overlapping with a pre-existing station. Still, it should give a decent estimate of high traffic locations in a city where an additional station could increase profits.

Note: as each city's data ranged in the length of time it covered, we picked an arbitrary recent month to collect data for (09-2018)

Data Sets Utilized

- Boston Bike/Station Data (https://www.bluebikes.com/system-data)
- New York City Bike/Station Data (https://www.citibikenyc.com/system-data)
- Washington/Station Bike Data (https://www.capitalbikeshare.com/system-data)
- Chicago Bike/Station Data (https://www.divvybikes.com/system-data)
- San Francisco Bike/Station Data (https://www.fordgobike.com/system-data)
- Census Data (https://data.cdc.gov/api/views/dxpw-cm5u/rows.csv?accessType=DOWNLOAD)

Note: While the station and bike data are similar to that of each city, they had differences in syntax and download format.

Data Sets Constructed

- AggBikeData (Sum Aggregate of Bike Data per city)
    - Data table consisting of one entry per city, including city name and total time the population utilized the system
    - Generated from Bike data sets for each city
- UnionPopBike (Union of Population data and Bike data)
    - Data table consisting of one entry per city, including city name, total time the population utilized the system, population, and amount of bike stations
    - Generated from AggBikeData, Station data sets for each city, and census data
- OptStationNum (Optimal Station Number)
    - Data table consisting of one entry per city, including the optimal number of stations to add, as well as data specific to the regression
    - Generated from UnionPopBike
- KMeans
    - Data table consisting of one entry per city, including city name and optimal locations

Algorithms

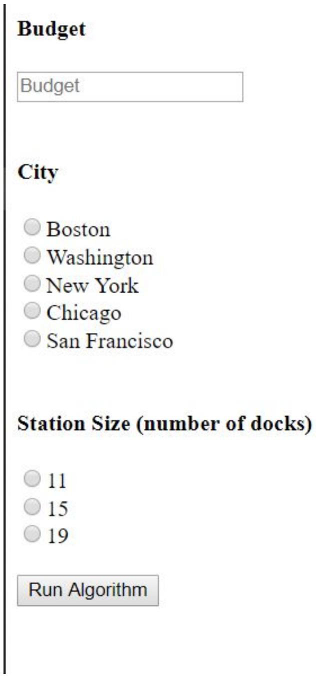
- Statistical Analysis: Linear Regression
    - Used linear regression on a ratio of time the population spent on bikes to the population
    - Utilized to determine if there is an upper limit on adding stations where additional stations would not increase efficacy
    - Outputted optimal number of stations to add in range of 0<x<max (see **Summary** for max calculation)
- Optimization: K-means
    - Utilized K-means clustering algorithm on existing station locations to determine optimal placement of new stations (k obtained from linear regression)

Tools

- Basemap - used to map the lon/lat pairs used to represent stations

Visualizations

Input:



Budget

Budget

City

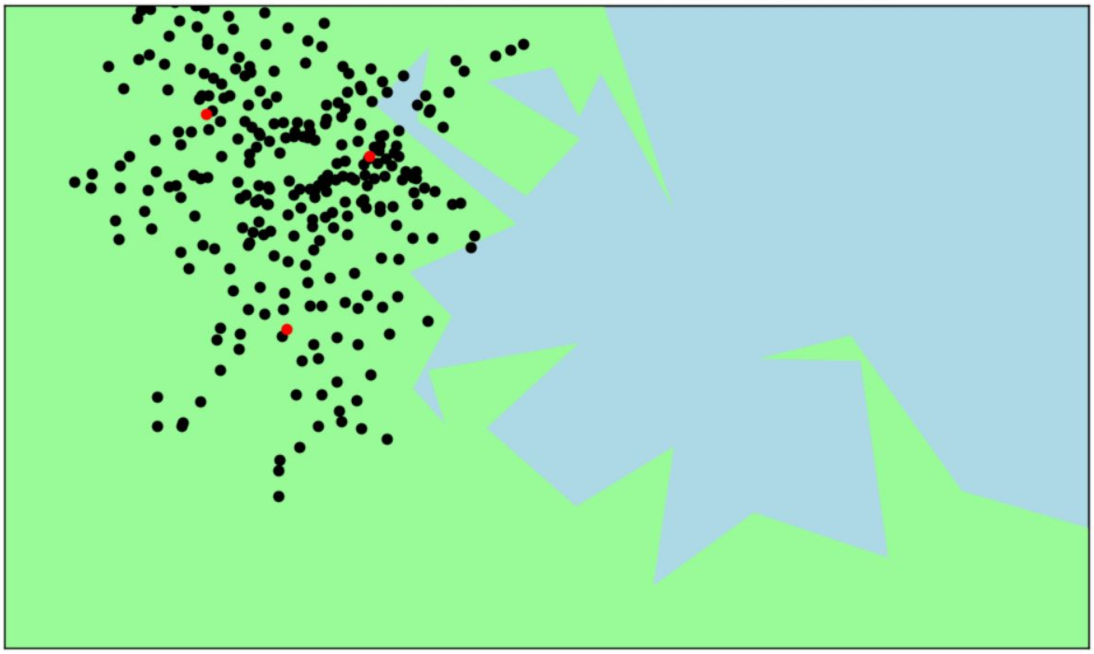
○ Boston
○ Washington
○ New York
○ Chicago
○ San Francisco

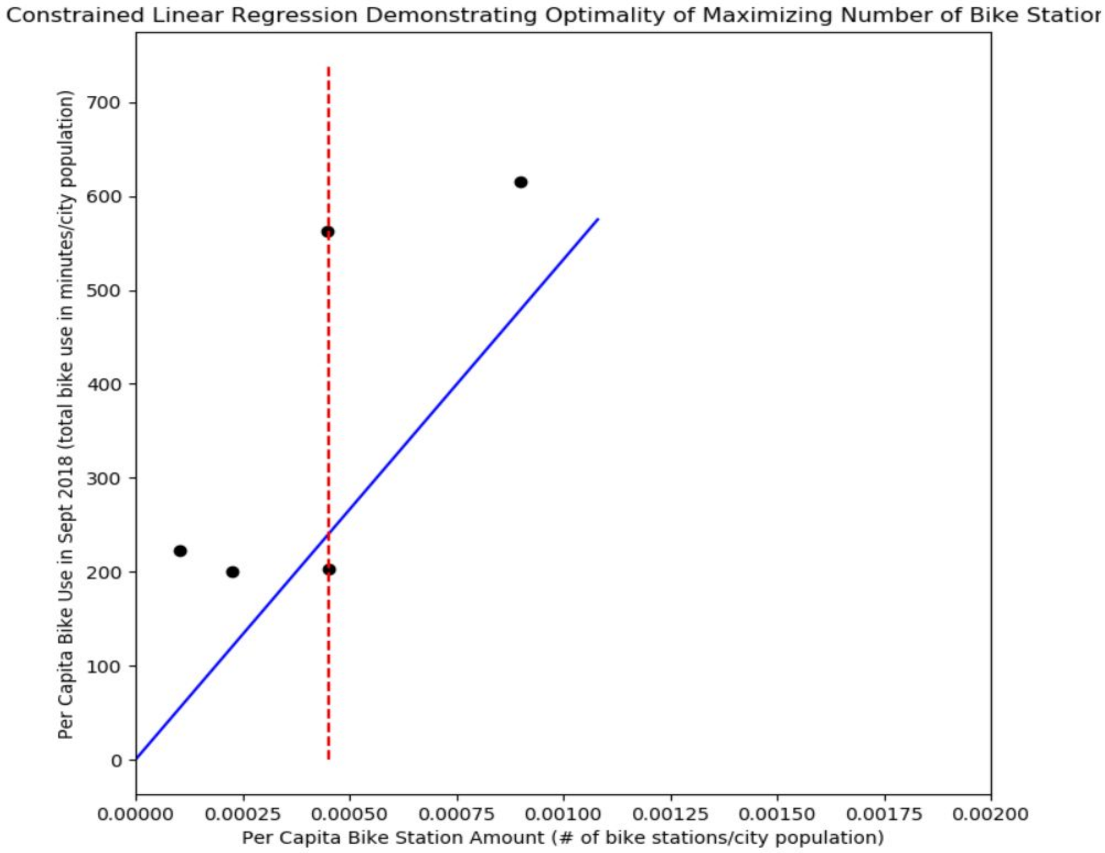
Station Size (number of docks)

○ 11
○ 15
○ 19

Graphs:

Existing Station Locations (black) and Suggested Station Locations (red)





Future Work

This project could logically be extended to include any city in the world with this sort of bike sharing program. More data could allow for a more accurate linear regression formula to determine an optimal station number and to determine if there is a max limit.

References

Cost to build varying sizes of stations:

https://engineering.wustl.edu/current-students/student-services/ecc/documents/heda.pdf