Optimal Advertisement Placements in Boston

Kristel Tan, Nisa Gurung, Yao Zhang, Emily Hou

Boston University

CS591 L1 Data Mechanics, Fall 2016

I. Introduction

Companies spend millions of dollars on advertisements every year. Therefore, it is important to find optimal locations to place them for the most impact on consumers. Eliminating ineffective advertisements can minimize waste while increasing competition and boosting a city's economy. Accomplishing this not only helps businesses, but also reaps benefits for any given city. In this project, our main focus of research is looking into optimal zip codes in Boston. By looking into the following datasets, MBTA bus stops, MBTA T stops, Big Belly garbage locations, college campuses, and Hubway stations, as potential locations/landmarks for physical forms of advertisement around Boston, we created an optimization tool in the form of a web service to determine the best locations by zip code in Boston, adjusted to individual need.

II. Data

The following are the five publicly available data sets that we used to help determine the most effective advertisement placements in the city of Boston:

- [MBTA Bus] https://boston.opendatasoft.com/explore/dataset/mbta-bus-stops/
- [T Stops] http://erikdemaine.org/maps/mbta/mbta.yaml
- [Big Belly Locations] https://data.cityofboston.gov/City-Services/Big-Belly-Locations/42qi-w8d7
- [College/University Locations] https://boston.opendatasoft.com/explore/dataset/colleges-and-universities/
- [Hubway Locations] https://boston.opendatasoft.com/explore/dataset/hubway-stations-in-boston/

II.I. Algorithms

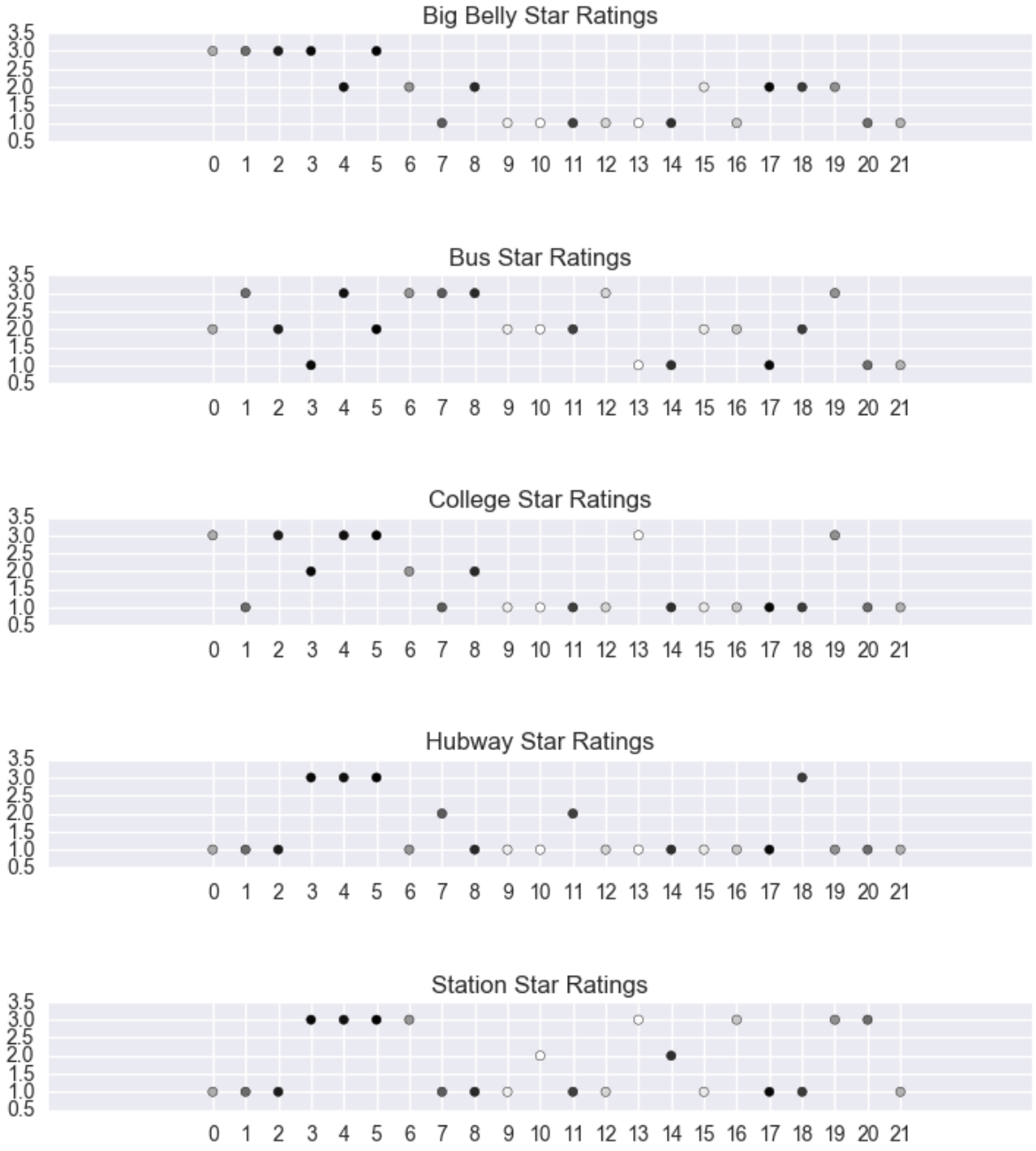
After retrieval the raw data from the five original datasets, we generated transformed datasets that intersected on zip codes so that each of the five landmarks had a count for each zip code.

In our optimization algorithm, we wanted to address the ultimate goal of finding the optimal zip code(s) for physical advertisement based on a rating system. For this system, we decided to implement a 3-star scale. If a zip code receives a 1-star rating, it means that the area with that zip code has a low outreach in terms of viewing population. If a zip code receives a 3-star rating, it means that the area with that zip code has a high outreach. To determine the rating for each zip code, we calculated the standard deviations for each of the landmark counts' data, stored all in their own separate dictionaries. We took into account the square miles of each zip code area as well as the population density to create a fair evaluation for each zip code rating score. The standard deviations set the range for each numeric rating per landmark. The same process was repeated and applied to generate an overall rating for each zip code, taking into account for the weighted rating scores by landmark. A new dataset that contains a finalized list of all of the overall ratings for each zip code is then stored in the MongoDB database, to be used in our optimization tool to recommend optimal locations for advertisement placement based on user query.

In our optimization model, the backbone of our web service, potential users are able to query desired ratings for each landmark and specify the number of optimal zip code(s) that they would like returned from their query. The state space is $2^{(\#\ of\ zip\ codes)} = \{0, 1\}^{(\#\ of\ zip\ codes)}$, where 0 represents the choice to not advertise in a certain area and 1 represents the choice to advertise. Depending on the users' query, the objective function, which is the sum of all permutations of the subset from the original state space that correlates most with the user's query multiplied by population density, will be maximized, thereby returning the zip code(s) with the highest population density. In other words, our algorithm will generate the optimal result given the user's constraints by focusing on maximizing the viewing population (highest population density).

II.II. Analysis

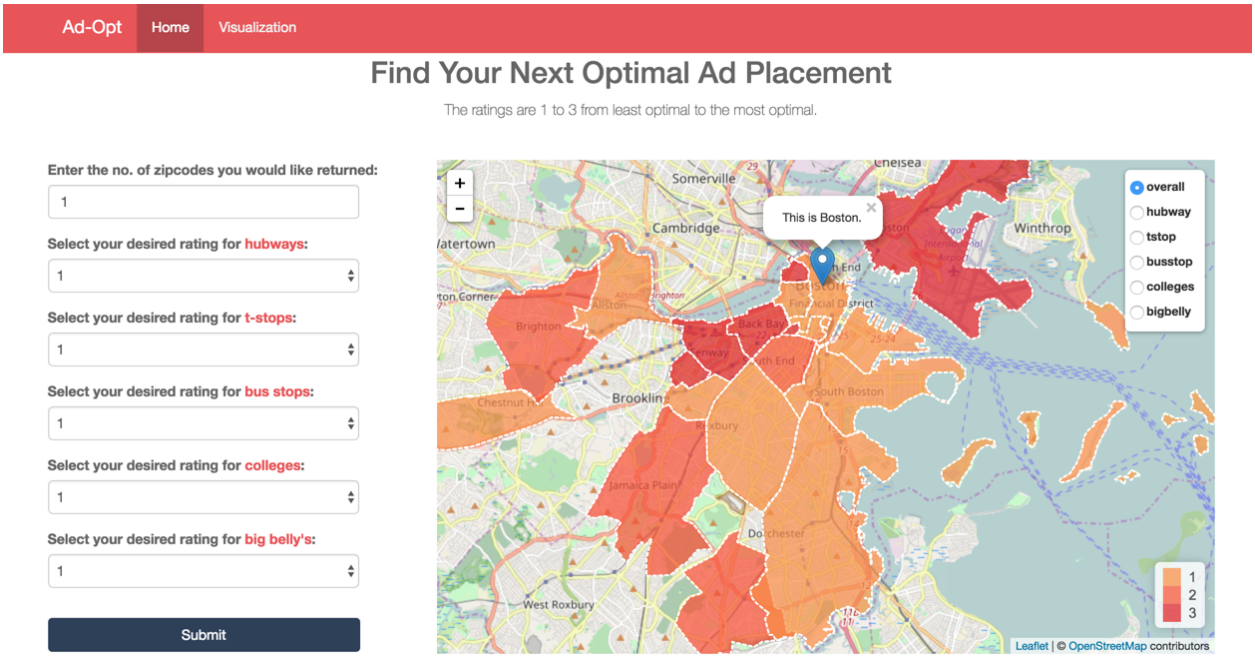
We were interested in potential trends that may exist in the correlation between any two landmark ratings. By interpreting the similarities, if a majority of the comparisons yield a positive strong correlation, similar datasets could be combined/eliminated or one could potentially be chosen over another. We wanted to know: would this be beneficial or detrimental to our optimization tool?



As seen above, the ratings plotted for all the landmarks to check for correlations, suggesting that no strong correlations exist among any of two landmark ratings. Therefore, we decided to keep all of the datasets in our final analysis for predicting optimal zip code(s).

II.III. Web Service & Visualization

To better visualize, we created a web service that allows users to test out our optimization algorithm:

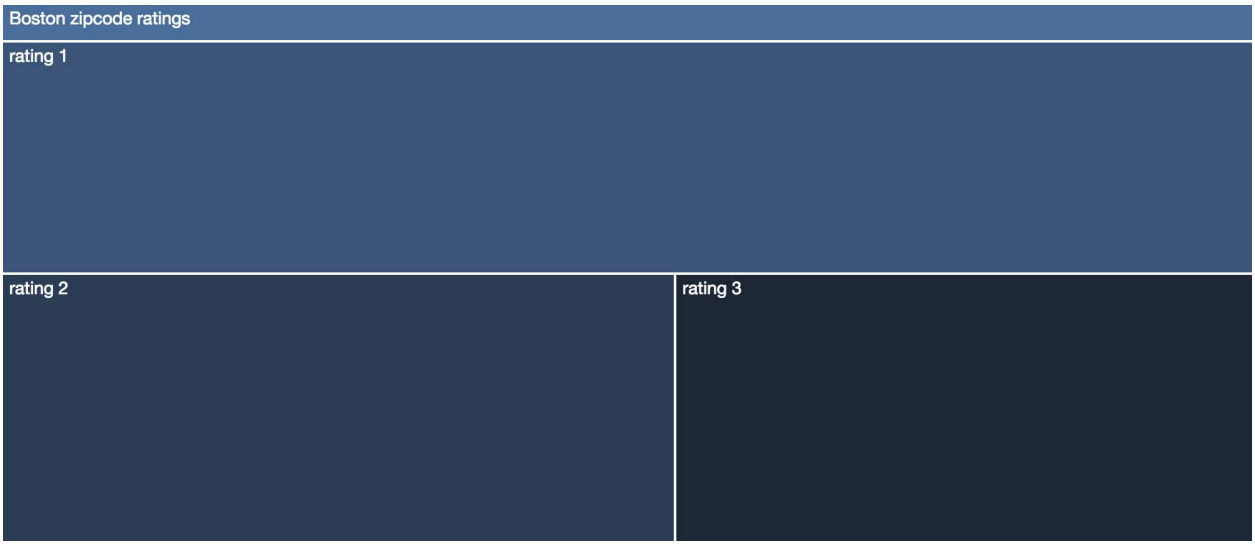


The web service incorporates Leaflet, an open-source JavaScript library that displays an interactive map. Our map illustrates the zip codes by area, colored by gradients. The gradients are defined by the calculated ratings for each zip code (darker color represents a higher rating and lighter color represents a lower rating). The map is initialized to show one pin in the heart of Boston and shading by overall zip code rating. After users input their queries for the optimal zip code(s), the map will zoom in/out to frame new pin(s) that are dropped on the optimal zip code location area(s).

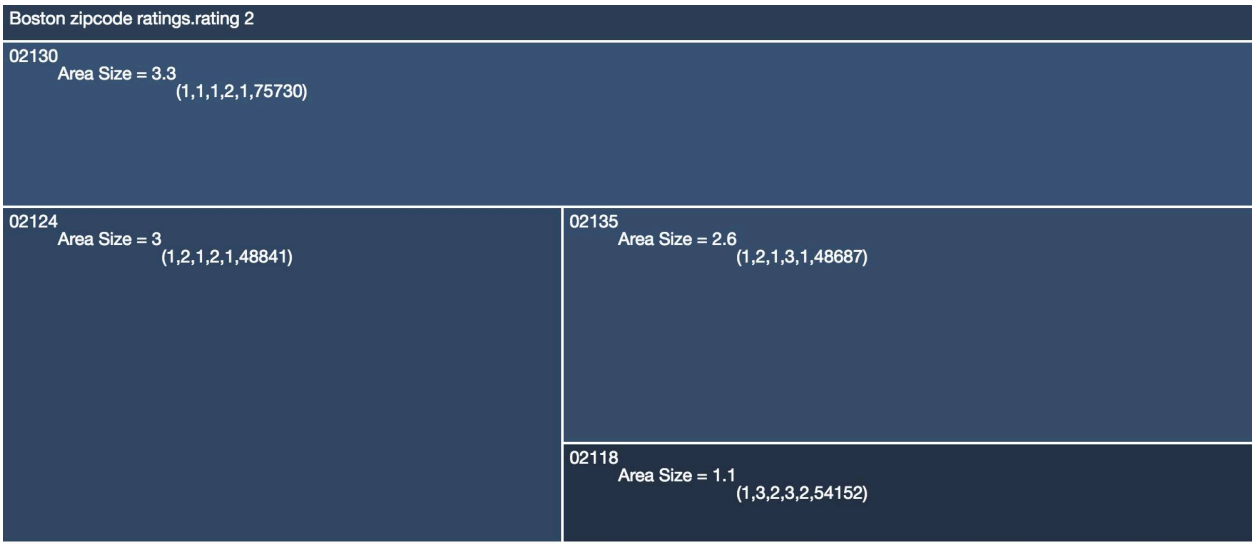
On the top right corner of our map, there is a toggle that allows users to select different "views". This is to give users flexibility and different perspectives in analyzing their target audience.

- **"overall"** represents overall zip code ratings
- **"hubway"** represents zip code ratings for hubway stations
- **"tstop**" represents zip code ratings for T-stops
- **"busstop"** represents zip code ratings for bus stops
- **"colleges"** represents zip code ratings for college campuses
- **"bigbelly"** represents zip code ratings for Big Belly locations

For our interactive visualization, we chose to implement a zoomable tree map. Tree maps are used to illustrate hierarchical data, which we represent with our rating system. It is implemented in D3.js, allowing for users to view the full dataset of overall rating scores in a compact way (by zooming in and out).



The sizes of these rating sections are determined by number of zip codes that possess that numeric rating. Clicking on a rating section will result in a zoom that shows the next level in the hierarchy.



The sizes of each zip code section are determined by zip code area size and the gradient color of each section is determined by population density. This means that the darker a section is, the more densely populated that zip code is. In this view, we display the median household income, which was data that we scraped, for each zip code in the users' results to provide a better idea for their target audience.

III. Conclusion

From the web service, we were able to verify logical reasoning just by playing with the toggle options on the map. Our algorithm was able to illustrate the relationship between accessibility and zip code ratings. For example, zip codes with poor ratings for T-stops had higher ratings for bus stops. This meant that bus stops were compensating for the lack of accessibility via the MBTA T-stops. During the transformation of data, however, we encountered an issue with overlapping zip codes. We noted that some zip codes were "embedded" in the same areas as other zip codes. Since the smaller zip code areas were a part of a more densely populated area, we decided to eliminate those zip codes with smaller areas so their ratings wouldn't skew the data. These smaller sectored-off areas were so small and insignificant that we decided to assume that it would encompass the same rating as its surrounding zip code area.

When transforming the raw data, there were some limitations due to the Google Maps API quota and geocoder query limit, but the necessary responses from the API were stored into a variable (hard-coded dictionary) to help create the collection.

Overall, the web service serves as a solid foundation for predicting optimal zip code(s) based on the five landmarks. The accuracy of our algorithm was a bit limited by the Google Maps API standard quota (2,500 requests per day). Currently, the sizes of each zip code area are not standardized – some zip codes cover more square miles, assigning a broad rating for the entire area, which may not always be accurate. Therefore, this project idea has potential to grow after proper refinement.

IV. Future Work

For future endeavors of this project, changing the definition of "optimal" could take advantage of the opportunity to determine whether the chosen landmarks were good determinants of optimal ad placements or which landmarks work better than others. For example, maximizing profit, amount of attention (views), viewers' interest in advertisement subject, etc. Additionally, the project could explore utilizing more landmarks to improve the optimized results for a given query, refining/running the algorithm on larger dimensional datasets (which in turn could allow further statistical analysis to beneficially impact any algorithmic changes), and incorporating revenue measurement of advertisements placed from recommendation of our optimization tool by user query. Revenue could open new doors to altering our rating system; it could integrate monetization and/or use real-time data to verify how accurate the results of our optimization tool are. Another interesting perspective could be to escape from the realm of physical advertisement and incorporate data from social media platforms.