Optimal Restaurant Locater in Boston

Jesus Franco
Department of Computer Science
Boston University
jofranco@bu.edu

Lin Chen
Department of Computer Science
Boston University
lc546@bu.edu

1   Introduction

As one of the major cities in the United States, Boston has countless restaurants and different cuisines. While the quality of food is the first thing a restaurant owner wants to consider, the location of a restaurant also determines the chances a restaurant can thrive or not. Our initial goal of this project is to find the best possible locations in Boston for a restaurant. A good location needs to have high traffic to bring more business in, which can be measured by the accessibility of public transportation and the number of neighborhood clusters nearby. In the beginning, we implement non-trivial transformation like selection, intersection and aggregation to clean up the datasets that we found at MBTA, City of Boston, and the Hubway.

Later, we apply k-means to generate neighborhood clusters in Boston, and statistical techniques to prove there is no positive correlation coefficient relationship between the number of restaurants near Hubway stations, and the number of restaurants near the past criminal locations. It indicates that the customers don't care about if a place is safe or not when choosing dining option, or/ and high traffic doesn't lead to a higher crime rate. So always choose the higher traffic area in term of customers' safety, because people don't care when it comes to food. We also create an interactive heat map that when you click on a random point of the map, it will show the latitude and longitude with the criminal history of the area. The heater the area is, the higher density of crime in the area. Moreover, we have a map to show the clusters of the neighborhood and restaurants using k-means. By looking at the map, users can find the areas that have a big number of properties and people, while haven't been

occupied by a lot of restaurants yet.

2   Datasets and External Source

The six datasets we used are from three different sources, including MBTA, City of Boston, and the Hubway bike-sharing station.

They are:

- MBTA - Red line datasets
- City of Boston – Employee Earning Report (2016)
- City of Boston – Crime Incident Reports (July 2012 – August 2015)
- City of Boston – Active Food Establishment Licenses (2017)
- City of Boston – Property Assessment FY2016
- Hubway – The Hubway Station in Boston 2017

Most of the datasets contain accurate information of location, such as latitude and longitude, street name, and zipcode. All the datasets are from recent years and in JSON format. It shows the credibility and accuracy of the datasets. Many datasets have problems with unexpected white space between data, no value or wrong value in the blanks, and inconsistent format. Overall, the quality of datasets is decent.

3   Techniques and Analysis

In project 1, we choose three non-transformations to design the algorithms. First, we apply selection to filters out all the red line T stops that are not in Boston. Second, we apply intersection to find the restaurants that are on the same street where had crimes. Third, we apply aggregation to calculate the mean of the income of the population in a neighborhood based on zipcode.

In project 2, we figure out two non-trivial solutions for the problem, and implement

k-means and statistical analysis. First, we use the k-means to produce the neighborhood clusters in Boston based on their latitude and longitude, and we are able to show the clusters on the Boston map. Second, we calculate correlation coefficient relationship between the number of restaurants near Hubway stations and the number of restaurants near the past criminal locations.
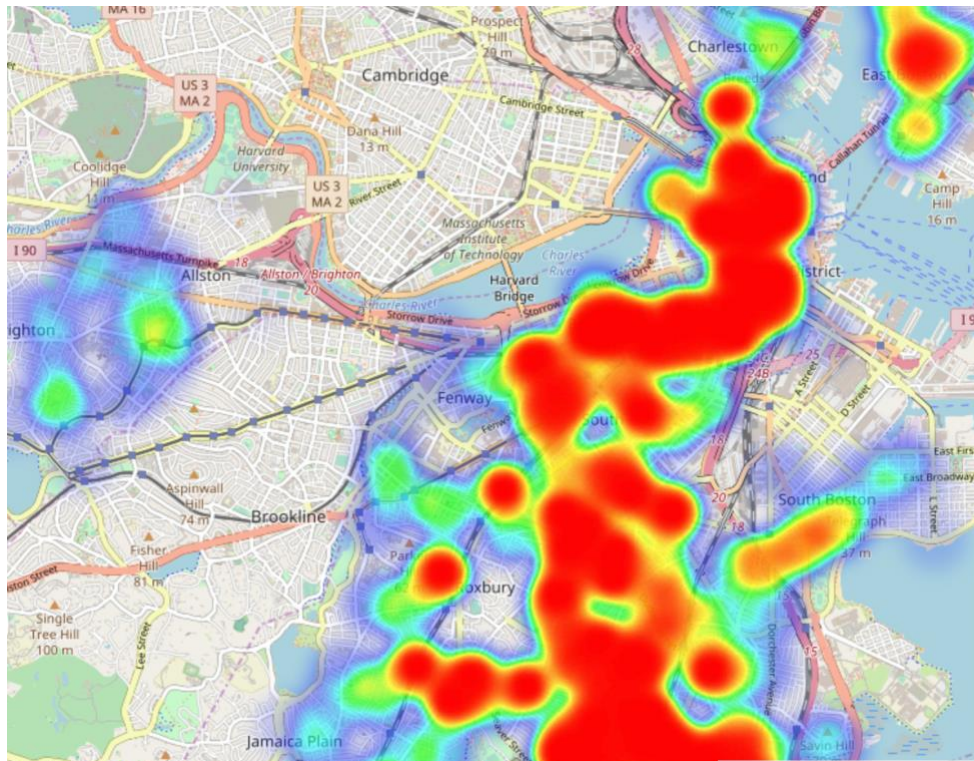

Figure 1: Crime heat map screen shot. The heater are there, the more crimes happened in the area. Detail information will be showed in the zone in version.
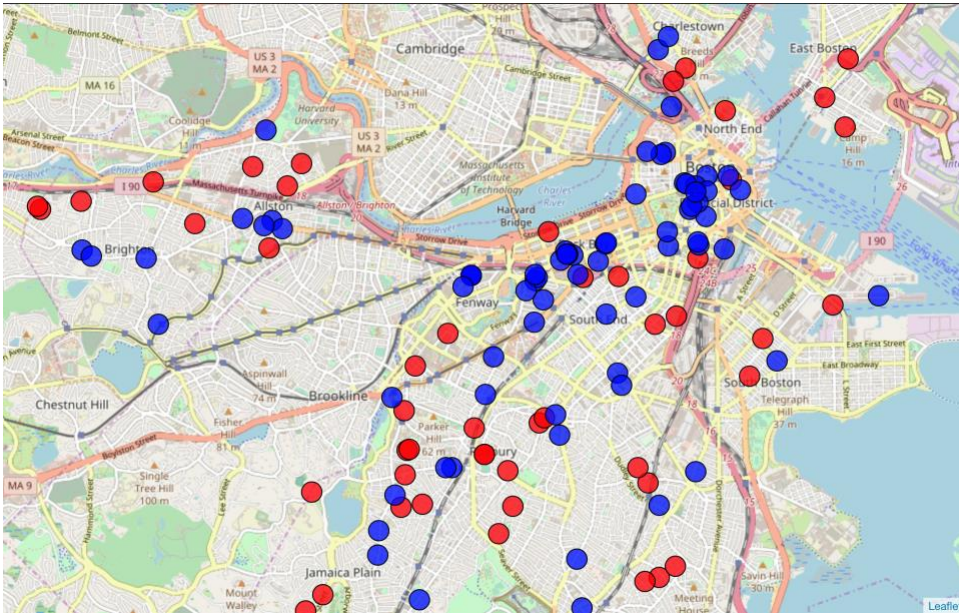
Figure 2: Restaurants and neighborhood map. The red points are the properties, and the blue points are the restaurants.

4   Summary

Data Mechanics is an effective approach for data mining. The problem we want to solve is to find the best location for restaurants in Boston. We consider various factors that would make a restaurant good, including local traffic, crime rate, and property. Using non-trivial transformations and statistical techniques to analyze the datasets in Python and MongoDB, we are able to find the locations with dense properties, and create a heat map to locate high crime areas. The results are presented in the data visualizations, which are the interactive maps on web services. By looking at the maps, users can find the great locations for their restaurants.

Overall, for the people who are interested in having a restaurant or bar in Boston, our data visualizations are great tools for them to choose the potential location for a restaurant.