Yilin Feng

Xi Chen

CS591

May 2nd 2017

Boston Property Evaluation

We aimed to determine the factors that affect property value in Boston. We chose dining and access to schools, and analyzed how they can have an impact on the value of real estate properties. Using data provided by City of Boston, we determined the correlation between property tax (a measure of value) and distance to the nearest restaurants and schools is 0.145.

**Data**

We used 3 datasets including restaurant establishments, properties, and schools. Restaurant dataset contains information on the location and type of establishments. Property dataset includes location and taxes paid by the owner/s. School dataset contains information on the location and level of education the school offers. These datasets are from the City of Boston. We assembled a result dataset which contains the centroids of the all the properties, restaurants and school from K-means algorithm. It also contains the correlation coefficient between property tax and distance to nearest restaurants and school and P-value.

Initially we decided to use another national wide dataset on tax returns from the IRS. We wanted to use tax return is an indicator of income which could be correlated with property value. The dataset only contains a few records from cities within the greater Boston area, so only a few properties could be paired up with the income tax records. We decided not to use it as it does not provide sufficient data.

**Methods**

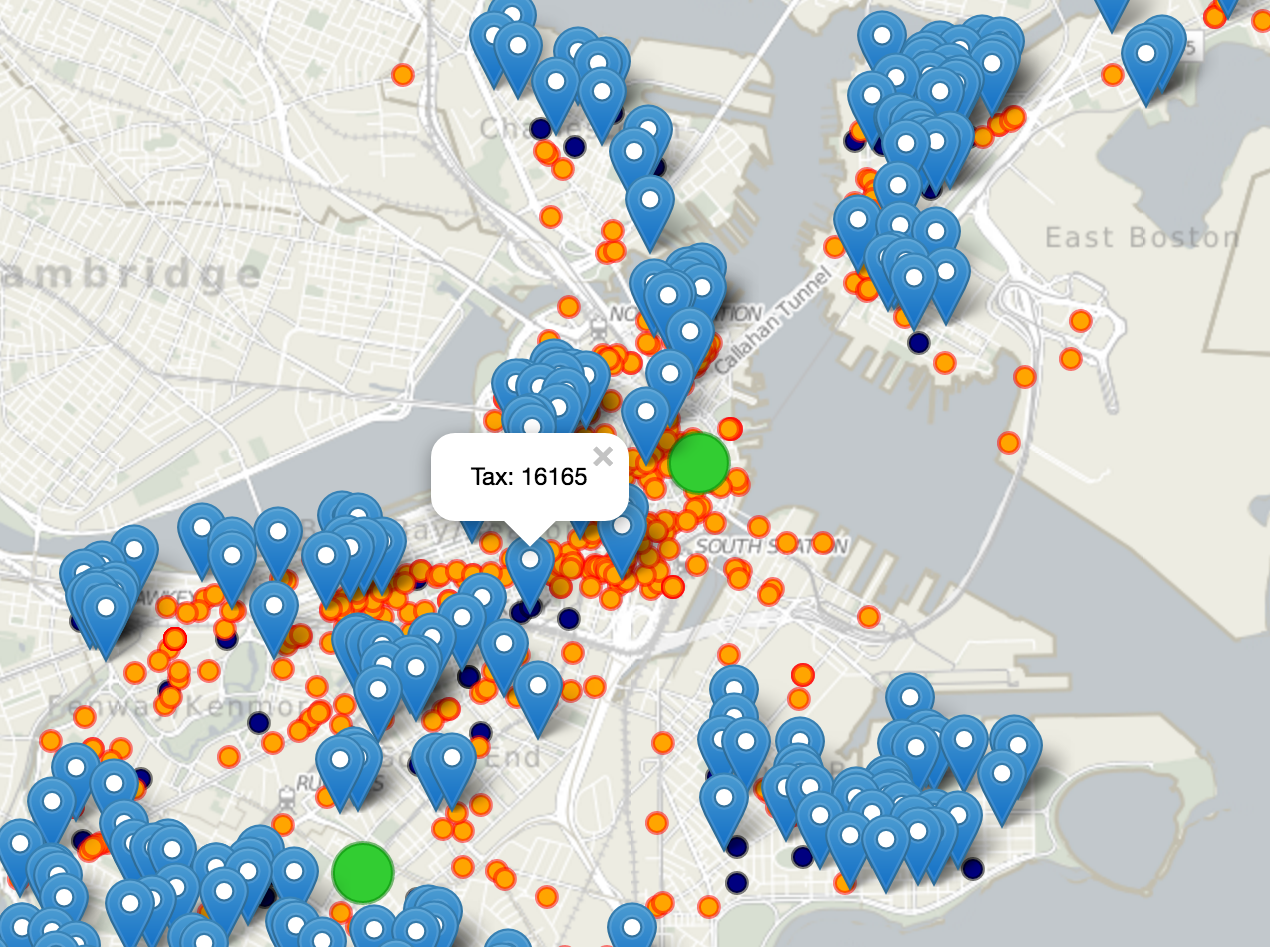
The two main methods we used are K-means clustering algorithm and statistical analysis. We first filtered out data points with (0,0) as coordinates. Then we combined the coordinates of restaurants and schools. We want to measure the distance between the centroids of these restaurants and schools to properties. The property dataset’s coordinates were misplaced, so we switched them in our “stringprocess” function. The coordinates were strings we also changed them to floats. We also made changes to the K-means algorithm. Coordinates are floats so when the algorithm calculates the coordinates of centroids, they change by a very unnoticeable difference every iteration. Then we added a loop to check if the difference between previous centroids and centroids from the current iteration is less than 0.001 then stop the algorithm. We then calculated the distance from each property to the nearest centroid from the K-means algorithm.

Another method we used was correlation analysis. We used the distance from a property to a closest mean of restaurants and schools and the property tax paid by the owner/s. We also checked the P-value with different number of permutations. We started at 5000 to 10000 permutations.

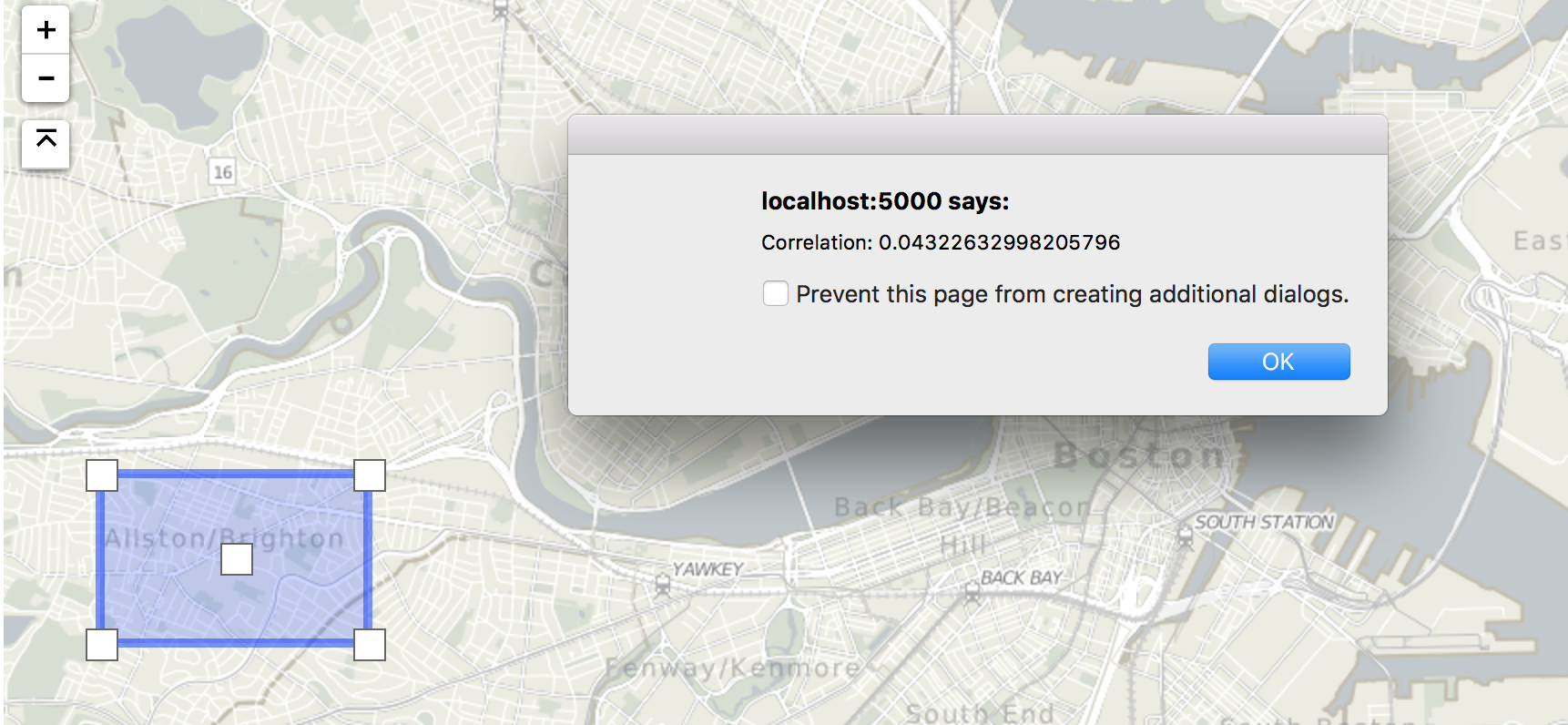
**Results**

Our K-means algorithm produces the following centroids of restaurants and schools are: [ 42.282258809566706, -71.1182603497138 ], [ 42.331928721684456, -71.08253018639779 ], [ 42.35047284414537, -71.1397570779497 ], [ 42.35791784088177, -71.05376542116292 ]. The centroids are located in in Back Bay, South Boston, Allston and Weld Hills.

The correlation between property value and distance to the nearest centroids is 0.145. We also calculated the P-value for this correlation coefficient. From 5000 to 20000 permutations, the P-value went from 0.024 to 0.019. Below is a screen shot of an area in Boston.



These amenities are not a good indicator of property value in Boston. We also provide a way for users to choose an area of properties. The rectangle is drag-able and resizable. When the user selects an area, and click on the correlation button, there will be a popup window showing the correlation coefficient.

****

**Future work**

We should try to find more factors that affect real estate value in Boston. Amenities are not a good determinant of value, so perhaps neighborhood, quality of the property, the year it was built, lot size or other factors are correlated with property value. We could also try to use a subset of the restaurant or schools if we can separate them into groups depending on ranking or scores. For example, property value could be correlated with school ranking or the quality of the nearby restaurants.