SocioEconomic Effects on Airbnb Rates

by Jinghang Yang, Yuan Zhang, Xinyun Cao

Introduction

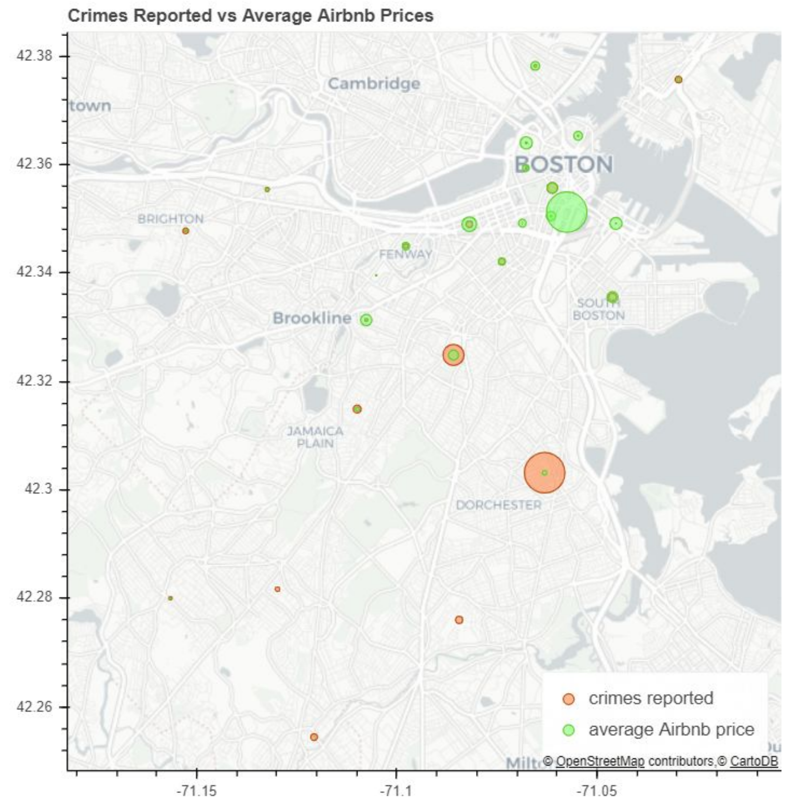
Although the city of Boston is filled with universities and young working professionals, the cost of living is known to be notoriously high. As students ourselves, we were curious about what factors contribute to this. To explore this topic, we gathered data sets pertaining to various socioeconomic qualities in Boston and compared them to the average Airbnb rates in every Boston neighborhood. Specifically, our data comprised of:

- Airbnb - listings data for Boston (last updated Feb. 2019)
- Boston.gov - Crime reports in Boston since 2015
- Analyze Boston - Boston community center locations (last updated May. 2018)
- Analyze Boston - Swimming pools managed by Boston Center for Youth and Families (last updated Jan. 2019)
- Analyze Boston - Boston Public Schools (last updated Sep. 2018)
- Analyze Boston - Boston Police Stations (last updated Dec. 2016)

Due to the varied contents of our datasets, we had to preprocess some of it before we could begin our analysis. Most of our data had longitude and latitude coordinates for each data point, but we were curious about the statistics pertaining to each neighborhood region. To reverse geocode the coordinates into a particular neighborhood, we used a shapefile containing the geographic boundaries of each of Boston's neighborhoods and geometrically compared it to each coordinate in the crimes reported data set. We also had to clean up the names of the neighborhoods so that the Airbnb data could be joined with crimes reported using the name as a key.

Method and Implementation

There were two main approaches we took to comparing our datasets. One, we directly grouped Airbnb entries by its neighborhood and compared those values to the number of crimes reported in each neighborhood. As stated above, we had to design our own way of geocoding the geographic coordinates to neighborhoods as there was no publicly available API for it. The revised “crimes reported” dataset took approximately an hour and a half to process due to its immense size. We also performed some data transformations including grouping the “Airbnb rates” data by neighborhood and calculating the average rate per neighborhood. To visualize these results, we generated the following Bokeh map:



Bokeh visualization of crimes reported vs average Airbnb rates.

The size of the circles correspond to the numerical values of each reported neighborhood. Each circle is drawn over the approximate center point of each neighborhood.

Another approach includes two clustering process. In order to cluster the data of all public places(schools, centers, center pools and police stations) according to the position of Airbnb houses, we used the k-means algorithm to cluster the Airbnb houses first. Based on the location, consists of latitude and longitude, we did the k-means clustering for all Airbnb data then we record the corresponding means of each cluster, as well as the average price of each cluster.

Now that we had all mean prices of every cluster for Airbnb houses, we used these clusters' locations to find the nearest public places and put the number of them into the result of that cluster. After this, we had a cluster of average Airbnb prices and total numbers of every public place. The following map is the Folium map we generated for Airbnb houses and public places clustering:

Folium visualization of Airbnb and public places clustering.

The marks of correspond to the centers of each cluster , by clicking on them Airbnb average price and numbers of all kinds of public places of that cluster will be shown.

Results

I. *Crimes vs Airbnb rates per neighborhood*

To our surprise, we found very little correlation between the number of crimes reported and the average rates per neighborhood when we calculate the Correlation Coefficient between average Airbnb prices and number of crimes reported. Despite some neighborhoods such as Dorchester having the expected result of high crime-low cost of living, the majority of neighborhoods did not produce such results.

	neighborhood	crime_rate	price	longitude	latitude
0	Allston	9569	132.296970	-71.132380	42.355413
1	Back Bay	14192	323.399602	-71.081876	42.349018
2	Bay Village	899	209.459459	-71.068641	42.349227
3	Beacon Hill	3132	190.785714	-71.067770	42.359357
4	Brighton	12394	112.487603	-71.152736	42.347797
5	Charlestown	6681	225.704403	-71.065396	42.378208
6	Chinatown	3387	235.566176	-71.061452	42.350523
7	Dorchester	85675	157.464158	-71.063050	42.303181
8	Downtown	22753	255.049569	-71.061154	42.355720
9	East Boston	14641	132.204620	-71.029625	42.375686
10	Fenway	9884	203.334842	-71.097719	42.344984
11	Hyde Park	14642	84.825397	-71.120675	42.254385
12	Jamaica Plain	17039	145.594747	-71.109880	42.314923
13	Leather District	503	751.714286	-71.057528	42.351256
14	Longwood	1580	105.666667	-71.105122	42.339593
15	Mattapan	15455	81.578947	-71.084424	42.276026
16	Mission Hill	6842	269.594470	-71.107621	42.331355
17	North End	3110	226.782946	-71.054702	42.365324
18	Roslindale	10105	84.371681	-71.129805	42.281681
19	Roxbury	44744	245.869186	-71.085826	42.324908

	Correlation Coefficient
Crimes reported	-0.16378839031967327

II. *Public places vs Airbnb rates per neighborhood*

What disappointed too is that after we calculated the correlation coefficient of all kinds of public places, there were not so much relation between price and public places, the two factors we consider are not really playing important roles.

	Correlation Coefficient
Number of centers	-0.15939130465865842
Number of swimming pools	0.019961379242068346
Number of public school	-0.33146333954281415
Number of police station	-0.122010212532738

Future Work

Unfortunately, we didn't prove that crimes and public places are playing important roles in determining the price of Airbnb rates. However, we will take other factors, such as transportation and the quality of Airbnb houses, into account so that we can see if they have significant impact over a person's decision to live/rent in a given neighborhood. We will keep using the techniques such as data transformation, statistic analysis and visualizations based on real world datasets so that we could get a reasonable conclusion.