

Exploring House Renting Districts in Boston Based on multifarious features

Baoxiang YANG, Desheng Zhang
ybx@bu.edu tscheung@bu.edu

December 14, 2016

Abstract

Boston is famous for its high housing and renting price, especially in some particular areas like Back Bay. However, sometimes it is hard for us to determine where to rent a room since there are only trivial information available. Thus, there must be some hidden information beyond the price. In our project, we hypothesize that within a city, renting houses or apartments can be grouped into similar "clusters based on location and various features like price, crime rate, the number of nearby shopping markets and nearby restaurants. In order to classify great districts within Boston, which would be suitable for different kinds of people(some may focus more on having more restaurants and some may more like to have more cafe bars.), we use several different clustering methods, including K-means, Hierarchy Method and Gaussian Mixture Model. Furthermore, we use these models to display and label different clusters on a map and analyze the clustering result combine with some Heatmaps, which finally gives us a nice result of different areas for renting in Boston.

1 Introduction

Different districts in Boston vary a lot in the renting price and behind this, is lots of features which could combine together to determine it. Streets separated by just a few blocks may have a big difference in price and lifestyle. However, people might just have an implicitly impression on which area is expensive for renting and which is not. The reasons for this are coming from other people's experience living in that area. Thus, for a newcomer, it might be important for him or her to know the reasons behind the price not just by listening but knowing deeply about a particular district.

1.1 Motivation

There are lots of different people need to rent for a room in America, what they need also various a lot because each person would have its own most effective concern. For instance:

- 1) Boston is famous for its campus. Each season, there are plenty of students coming in, who need to solve the problem of renting a room in a quiet short time. Generally, for students, they might focus more on price and safety. Thus, knowing which area has a cheaper price and a lower crime rate would be very valuable for them.
- 2) For an employee or business people of a company, they might focus more on the surrounding environment, like how many restaurants and cafe shop nearby or is it convenient for shopping. Therefore, knowing which area has a better environment would be good for these people.
- 3) Actually for each individual, the demand is different. Even if two are both students, one of them might want to have more restaurants nearby, but the other would focus on a lower price. Thus, knowing deeply of a certain area would be far more efficient than just finding elements on an official map by themselves.
- 4) For Landlord of business or person who want to lease their house: they could see various of aspects of their houses by checking the area they belong to and therefore to make a better decision about the lease price based on that information.

1.2 Our Dataset

Our Dataset has 5203 renting houses data of the whole Boston area. Each of them is combined with 13 features includes room numbers coordinate, price and crime rate and neighborhood information. We got these data from the trulia and zillow websites, which are famous in the market of renting houses in the U.S.

2 Technique & Methods

2.1 Scraping Data

Since we cannot find a suitable dataset for our topic, we decided to scrape the data from our own. For the first, we tried to use the API provided by Zillow.com. However, we found that they need us to provide the address and zip code of a certain house or apartment. So we decided to get this information from Trulia.com for the first, then get more information through the Zillow API.

Finally, this part could be separated into three steps:

- 1) We implement BeautifulSoup library to scrape basic information includes address and zip code from Trulia.com
- 2) Then we collect some features data of housing rent through the Zillow API.
- 3) Finally, we use the Selenium package to scrape some other good features of neighbor-

hood through Trulia.com and combine these data together.

One of the challenge part here is using Selenium. Since this package actually just imitates the process of opening the browser and selecting the target information by clicking the page, there are some problems while using it. Exceptions may occur either when the network is not stable or the data might just disappear because that house has been rented just at the moment. To figure out this problem, we used the `wait_for` function and `try catch` method to detect exceptions. We set a delay time of 10 seconds and the Selenium would try again and again until this time-out.

2.2 Basic analysis on initial Data

At the first step, we use beautiful soup to scrape the basic house information, including address, zip code, price, number of bedrooms and bathrooms and the url of houses. In this part, sometimes we cannot exactly get number of bedrooms and bathrooms. In the second step, we try to get some more detailed information through the Zillow API. We get information about Zillow id, the accurate number of bedrooms and bathrooms, room type, longitude and latitude of houses. Lastly by selenium, scrape the crime information, school information, number of restaurants, cafes, groceries, nightlives and shops 1 miles near the houses.

The address and zip code are basically used for Zillow API. We try to classify the houses or departments by their home type. However, the description and types of home type are complicated. If we use them, we may discard a large part of data, thus we decide not to use it. The coordinates can be used for clustering and plotting. Price, number of bedrooms and bathrooms can be used to determine the average price per room of a house. Price of a house can be a range, we usually get the median of the price as the renting price of this house or apartment, then we decide the average price per room by this price and numbers of bedrooms. The rest of the data are the features of houses, we will filter some useless features and use the rest of them to cluster and find implicit boundaries.

2.3 Cleaning the Data

As mentioned above, the number of bedrooms and bathrooms scraped from Trulia.com may not be accurate, thus we discard them get use the information from Zillow API. Actually, address scraped at first step can be useless, some of the addresses are pretty rough (eg. It may just an area like the South End, Fenway or Kenmore). Thus we get nothing from Zillow API by these addresses. We discard this part of houses. Also, if the return number of bedrooms and bathrooms are none from Zillow API, we discard them. At the last step, we lose some data because the houses had been rented when we tried to use selenium to scrape them. Finally, I discard the data with incorrect price (Some houses have no price,

and just label ?Contact for price?). We scrape about 7500 houses using beautiful soup. Then by using Zillow API, we have about 5700 left. Finally, after using selenium and cleaning unreasonably data, we get information of 5203 renting houses in Boston.

2.4 Pre-processing the Data

After getting the data, we could find that some data belongs to different dimensions and some are in different. Therefore, we need to do some preprocessing work on the data to be prepared before fit it to our model.

1) Default value calculation:

Some particular points do not have the value on a particular dimension, which shows "NAN". We used the Imputer class in sklearn.preprocessing library to deal with this problem.

2) Dealing with none-value data:

For some features, there are only, descriptions like "Below Average", "Average" and "Above Average". To deal with these features, we simply go back to the Trulia website and see their evaluation of what makes an "Average" words and then translate that word into the mean of the range that it stands for.

3) Scaling:

One of the most important parts of preparation is to get a nice scale. Thus, for this part, I used three ways to test which would be good for the input data and they are StandardScaler, MinMaxScaler and Normalizer. Finally, I used the StandardScaler class because it used the method of z-score and then make the features followed as a standard normal distribution with a Mean of 0 and STD in 1:

$$x' = \frac{x - \bar{x}}{S}$$

On the other hand, the Normalizer translated all the features into "unit vectors" and make it into the same dimensions. This is good for the feature selection, but not for an input data to fit the model:

$$x' = \frac{x}{\sqrt{\sum(x[j]^2)}}$$

2.5 Determine the number of clusters

Before doing k-means, we need to determine how many clusters we are going to use for the clustering. Thus, firstly we simply implement with the error function.

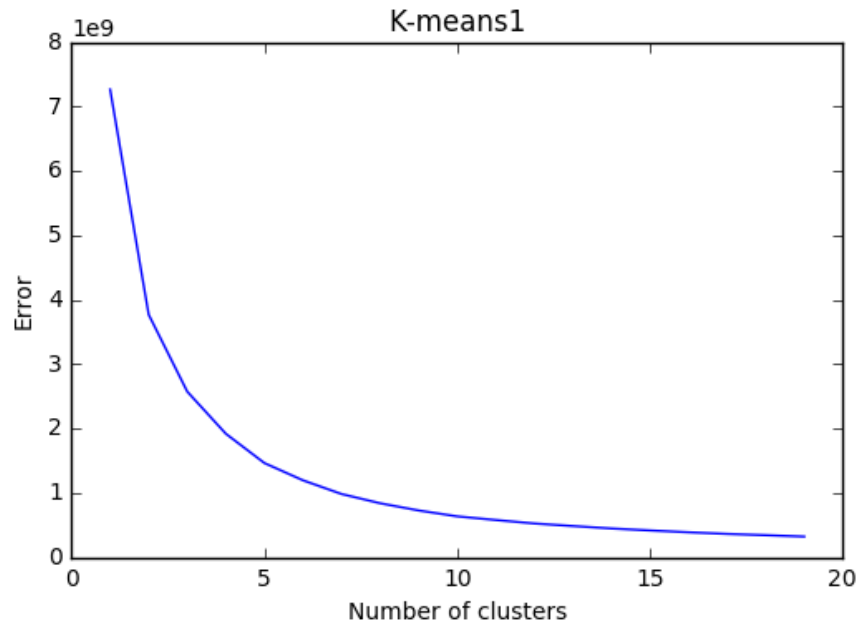


Figure 1: Elbow Graph

Then we implement the Silhouette Coefficient Score:

$$s = \frac{b - a}{\max(a, b)}$$

Let a be the mean distance between a data point and all other points in the same cluster. Let b be the mean distance between a data point and all other points in the next nearest cluster.

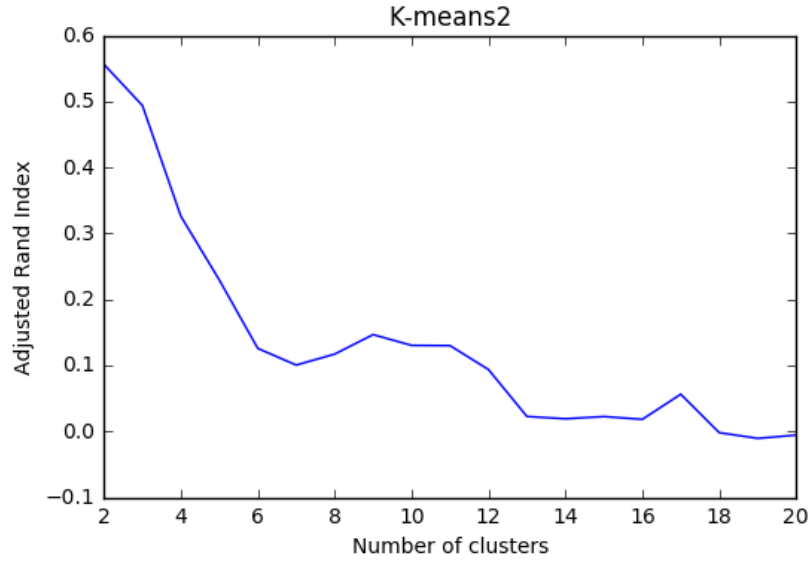


Figure 2: Silhouette Coefficient

By looking at the result, we decided $C=12$ as our K-means clusters' number.

2.6 Clustering

1) K-means: K-means clustering is a clustering algorithm that will find k centroids with distributions that would minimize the sum of squares of distance. By using the K-means class by Scikit-learn, we performed K-means clustering on all our renting data with random K++ means initialization and 300 iterations. The clustering is 12 by the Silhouette score and we used a Euclidean distance metric for our clustering, which finally get the result like:

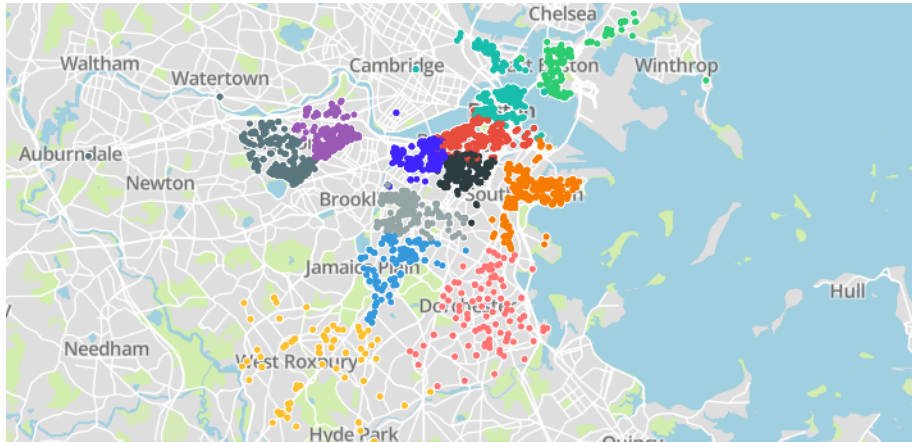


Figure 3: K-means Cluster

2) Hierarchy Method: Next, we used Hierarchy Method to do the test clustering work. The result seems to be similar to the k-means.

3) Gaussian Mixture Model: Finally, we used GMM to do the test. This is a probabilistic generative model that assumes all the data points are generated from a Gaussian distribution. This method performs some kind of overlapping of the cluster. This might be because that Gaussian distributions define its clusters are shaped by variations in each dimension.

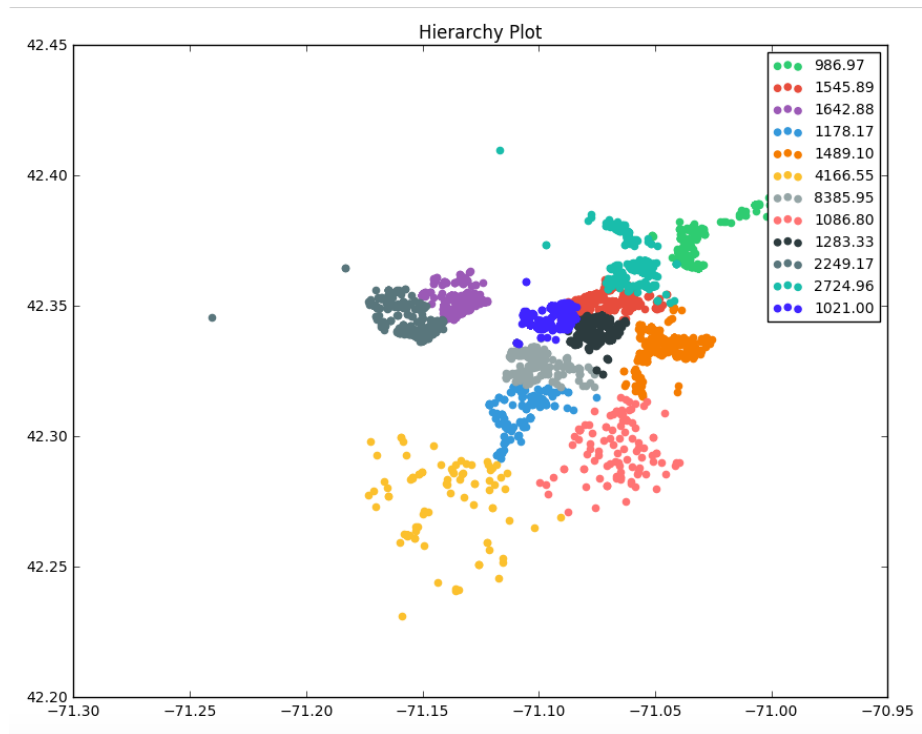
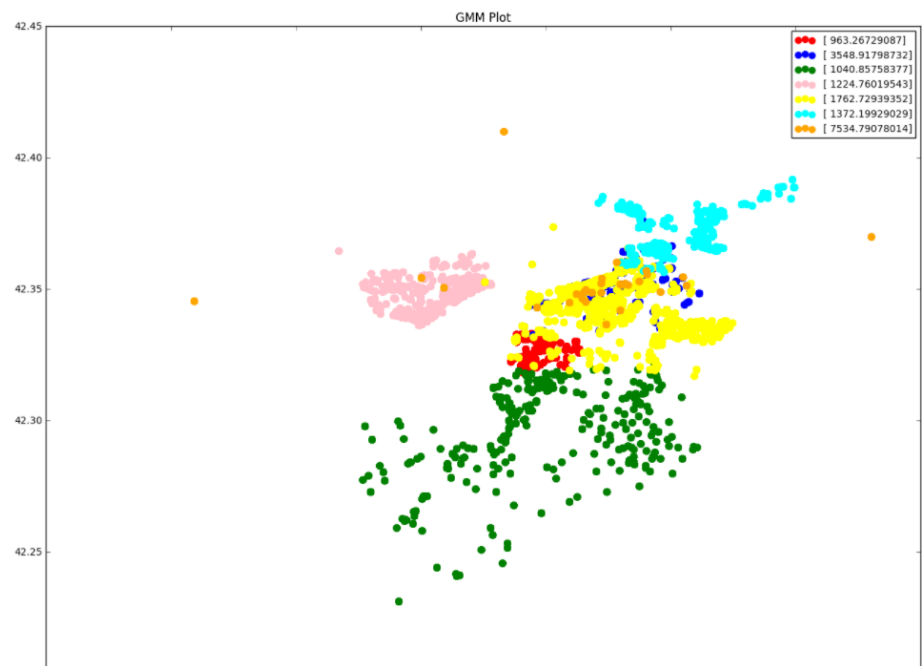


Figure 4: Hierarchy Method















8

Figure 5: GMM Method

2.7 Labeling

After finishing the clusters, we need to do the labeling work with these clusters. For this purpose, we simply calculate the average value of each feature of the source data, and by using the method of MinMaxScaler, we get each score (from 2.5 to 5) for each point to describe features. Furthermore, we could do works like human-labels, which could give us a better conclusion of a particular area. But for directly seeing what is in a district, we just choose to show the original output here.

Scores Labeling

	Average Price	Café	Restaurant	Shop	Crime
• 	\$2921.48	5.00	5.00	5.00	0.29
• 	\$2018.92	4.25	4.27	4.60	0.28
• 	\$1932.64	4.50	4.37	4.60	0.28
• 	\$1751.90	4.75	4.80	4.80	0.28
• 	\$1694.05	3.00	3.03	2.93	0.28
• 	\$1369.08	2.75	2.77	2.70	0.26
• 	\$1212.04	2.50	2.50	2.50	0.24
• 	\$1157.96	3.25	3.30	2.79	0.30
• 	\$1139.38	2.75	2.77	2.93	0.24
• 	\$1118.01	3.62	3.40	3.19	0.27
• 	\$1084.29	2.88	2.77	2.88	0.27
• 	\$894.77	3.50	3.63	3.86	0.28

* The average price means the average price per room for all the house in the cluster

* For café, restaurant and shop are scaled to a score between 2 to 5

* For crime, we should the original score because it is too close. However, it is still and important data, so we show it on this table.

Figure 6: labeling: Score Values

2.8 Analysis

For the result, when we looked at the areas of the top 3, we could find they are districts near Back Bay, Fenway and Kenmore. We are quiet expected to see that these areas actu-

ally have the highest price among Boston but at the same time, they also got a high score on Cafe, Restaurant and Shopping.

Another example on the other hand would be the two green clusters, which located in Allston and Brighton. We are also expected to see that these areas enjoy a much lower price but at the same get a relatively lower score on other features.

2.9 HeatMap

During the clustering, we found that the price and other features have a nice distribution based on location, but it seems like it does not influence by a crime rate on a map. For checking this, we used the heatmap to show the result.

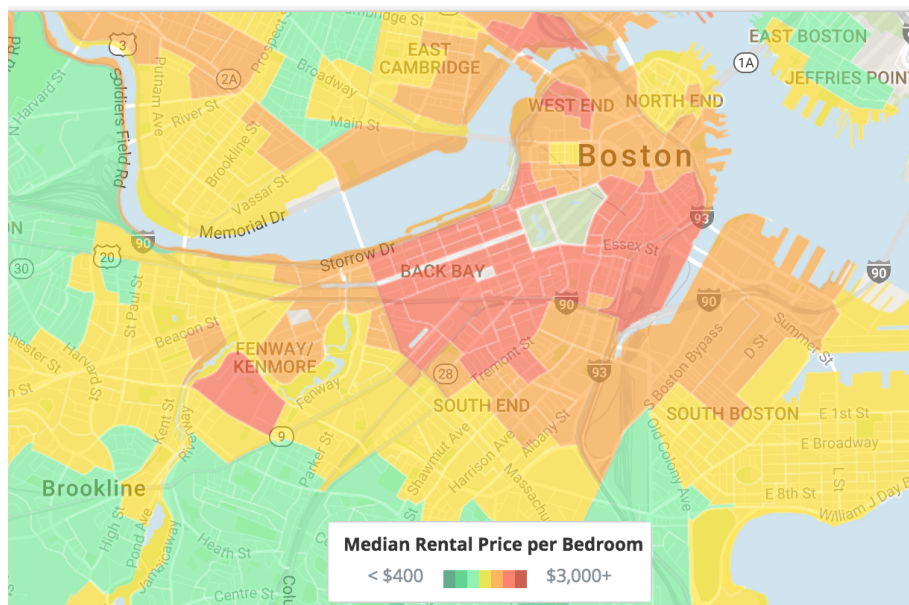


Figure 7: labeling: Price HeatMap

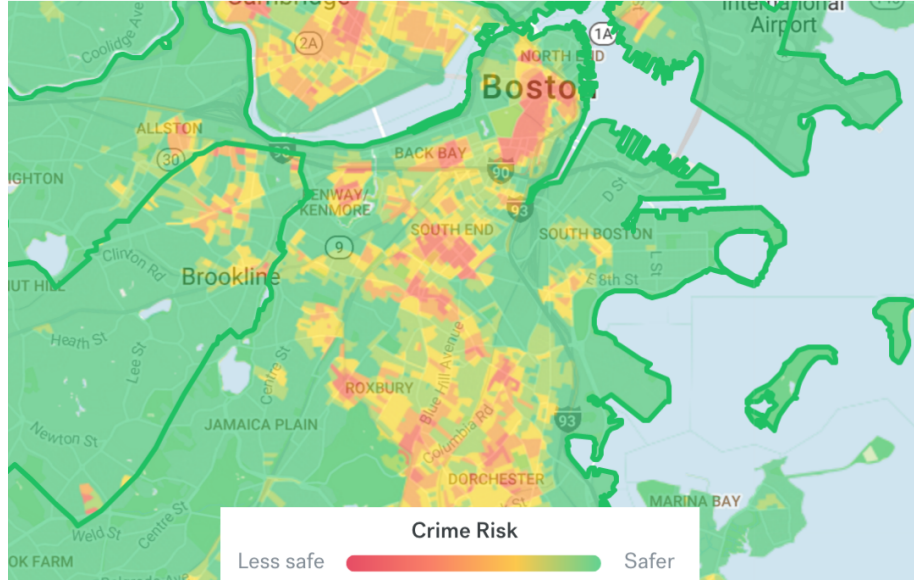


Figure 8: labeling: Crime Heatmap

By these heat maps, we can see that the price varies based on spatial location, but the crime data seems to distribute all around the Boston area, which is that nearly all housing areas are enjoying a similar crime rate on average. So the reason for this could be some future works which should not be the main purpose of this project. But it is easy to assume that nearly every housing area would have some non-significant events such as theft problem, which also counts for the crime rate and this would lead to a result that we cannot access to the big event like "rob" and "homicide", which would actually impress people a lot of on "what defines an unsafe area".

3 Discussion

Looking at the final results, we could actually see what behind each cluster based on location. For a higher price, you could get a better environment (a high score on cafe restaurant, etc) and with a lower price, the scores are relatively lower. But you may also find that the purple cluster, which located up to the China Town, gets the second highest on lots of the features with a slightly lower price.

The reason for this may be various, but we could easily think about another feature might make this happened, which is whether it nears a university. Since there are tons of students need to rent for a house every year, but not for local residents. Thus, the area with higher scores on restaurants and cafe might still have a lower price because it just a little far from

the university and a group of students would give it a lower price. And we also could see some overlap on the map since this stands for a real world map.

4 Future works

Furthermore, this multifarious features could be implemented by other features only if there are good datasets. But we should notice that clustering method does not perform well on high sparse high dimensional space, which means we might cannot get a good cluster on all the features combined together. Thus, developing a system that could let each person to choose its own concern parts would be better for each one may get a different cluster based on what it looks for and by this means, easier to find the most suitable house or apartment for it.