Boston Airbnb:
Clustering Analysis

Clustering
Dr. Matt Wheeler

Orange Team 2
Qing Feng, Andrew Moolenaar, Carlos Chávez, Julie Huang, Bill Jenista

Executive Summary

Tasked with finding locations in Boston, Massachusetts to rent on Airbnb for top dollar, our team has identified six distinct segments of listings based on the renters' sentiment towards the property from reviews, the average cost to rent the property each night per bed, and the average distance the property is to eight local attractions. We believe it would be best to invest in properties that come from or are similar to the sixth segmentation, as these 34 listings have above average reviews, a low price per bed, and a closer average distance to eight popular tourist attractions compared to the average listing. As can be seen in Figure 1, a large portion of these properties (red circles) are located near the heart of the city where many tourist attractions are located (large black circles). Furthermore, we can find the average number of nights the listings were rented out between September 6, 2016, and September 5, 2017, to see which listings are most frequently rented within each cluster. By delving into each cluster, we can show the listings that contain all of the following: best reviews, the lowest price per bed, closest distance to tourist attractions, and the highest frequency of rents in the year.

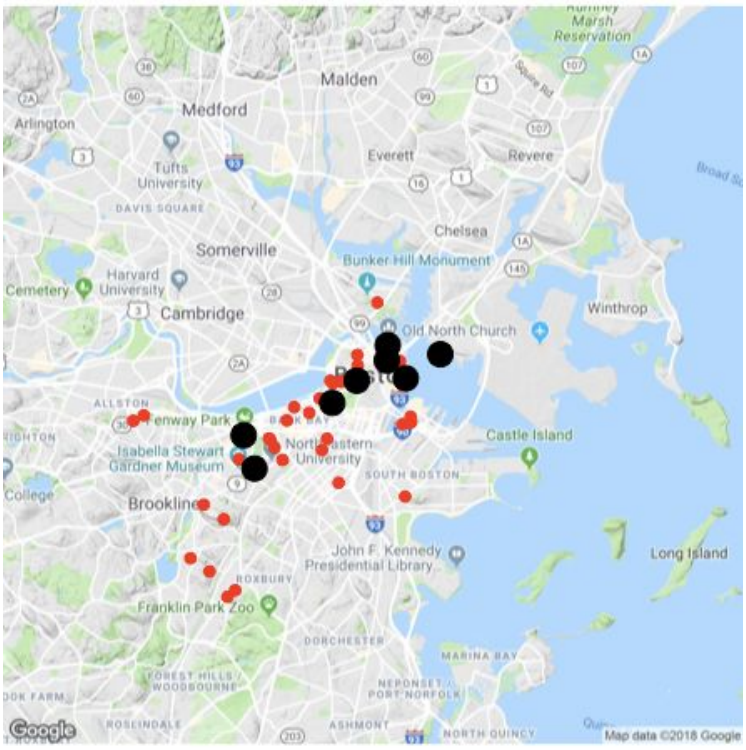


Figure 1: Segment 6 (most desired) with attractions

Background

Our team was tasked with finding segments of Airbnb listings in Boston, Massachusetts that would be considered worthy of investment. To do this, we used public Airbnb data sources that included listing information, property reviews, and the renting calendar of the properties between September 6, 2016, and September 5, 2017. Listings that had less than 4 reviews were removed, as we did not feel we could get an accurate measure of the reviewers' overall sentiment with so little reviews. In addition to these three datasets, we identified the top eight tourist attractions in Boston and included their latitude and longitude to find their distance to each of the listings. These attractions were the Museum of Fine Arts, Fenway Park, Boston Harbor, Old North Church, New England Aquarium, Massachusetts State House, Freedom Trail, and Boston Common and Public Garden. We decided on these attractions because not only did we find them to be popular according to our research of the Boston area, but also because they are all located within the busy parts of the city while being somewhat spread out from each other (Figure 2).

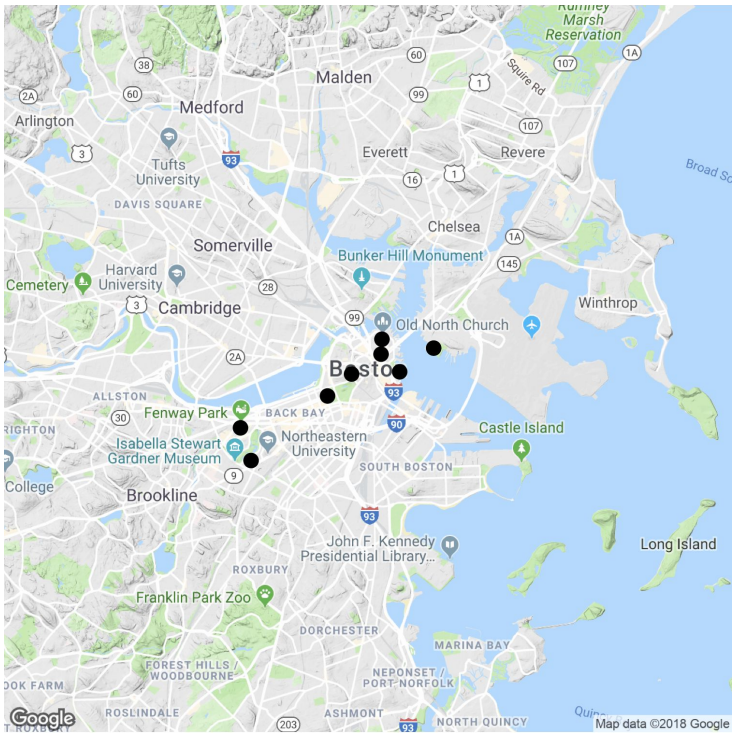


Figure 2: Top 8 Attractions

Using these datasets, we sought out to answer three main questions: Which listings and locations would we recommend investing in based on our segmentation, which listings seem to have a lot of rentals based on a close location to tourist attractions regardless of the quality of its reviews, and what are the qualities of a "good" Airbnb review?

Analysis

In order to build our segments, we incorporated the reviewers' sentiments towards a property, the price to rent per bed, and the average distance the property was to all eight attractions chosen. The sentiment of the reviews was categorized as either positive, negative, or neutral and then given a score as to how positive or negative a review was. Distance from the property to all of the attractions was calculated by taking the average of the eight separate distances. The price per bed was calculated by dividing the rental price per night by the number of beds.

Our algorithm produced six segments that allowed us to interpret the distinct characteristics of the listings. The fifth segment contains only four listing which we believe to be a very niche group. A detailed chart of each of the six segments and their interpretations can be seen in Table 1.

	Average Distance to Attractions	Average Sentiment	Price per Bed	# of Listings	Popularity (Average # of Nights Rented)
Segment 1 (Sets the bar)	Neutral	Neutral	Neutral	1388	173
Segment 2 (I'll take Uber)	Far	Neutral	Cheaper	150	133
Segment 3 (I hate Uber, I'll pay to live close)	Close	Neutral	Expensive	379	157
Segment 4 (You don't want to be like this)	Somewhat Far	Negative	Cheaper	61	157
Segment 5 (I'm rich but not satisfied!)	Close	Negative	Very Expensive	4	38
Segment 6 (Most desired by review)	Close	Positive	Cheaper	34	222

Table 1. Six Segments and Interpretations

After analyzing these segments, we determined that Cluster 6 is ideal. This cluster is composed of listings that have good reviews, are close to tourist attractions, and have a low average price per bed. This segment also has the highest popularity in terms of the average number of nights rented for a year. In our opinion, these listings give the customers the best "bang for their buck." We recommend looking deeper at the 12 listings in Table 2, as they were rented every day between September 6, 2016, and September 5, 2017, and seem to be the most popular listings within Segment 6. Any property invested in should share similar characteristics to these listings.

Listing ID	Average Sentiment Score	Average Distance to Attractions (km)	Price per Bed	Popularity (Total # of Nights Rented)
2016500	2.93	2.65	$150.00	365
6355733	2.70	2.87	$80.00	365
6908672	2.67	1.75	$99.50	365
6927063	2.56	2.67	$69.00	365
7698631	2.52	2.21	$100.00	365
8481291	2.71	5.75	$69.00	365
9840919	2.48	2.36	$87.50	365
11624428	2.48	1.75	$75.00	365
12825412	2.54	2.60	$66.67	365
13073078	2.57	2.74	$60.00	365
13677640	2.80	5.97	$53.00	365
14572515	2.75	5.18	$50.00	365

Table 2: Top 12 Listings from Segment 6

Additionally, we would also recommend looking more into the listings from Cluster 3. These 379 listings, seen in Figure 3, are frequently rented out due to their close location to tourist attractions despite the fact they are only neutrally reviewed. If solely interested in investing in properties that are close to the heart of the city, regardless of the reviews or the price, then we recommend listings from this cluster.

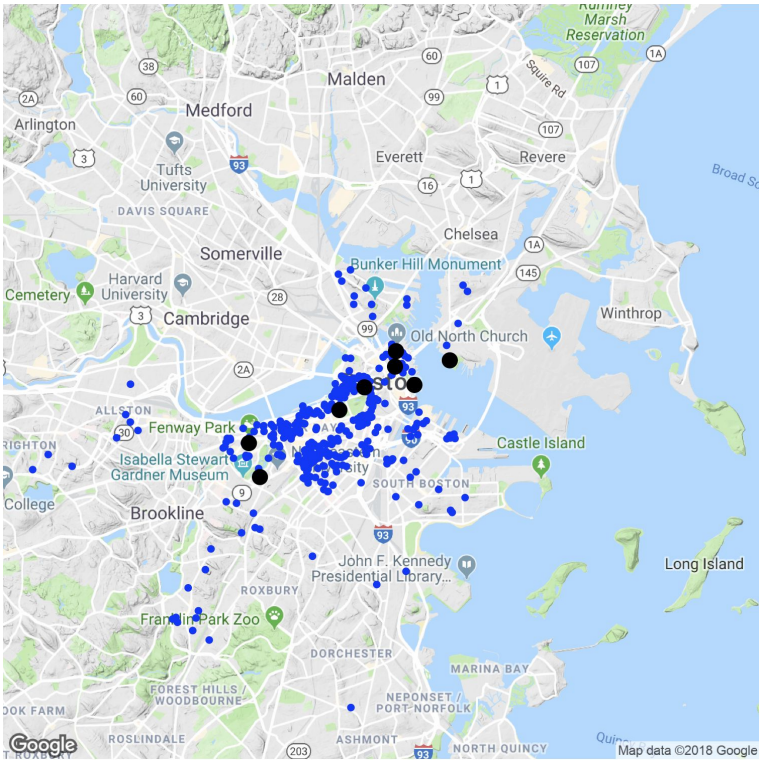
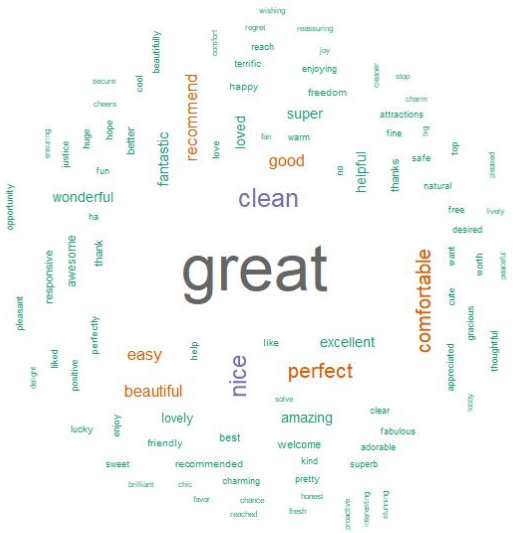


Figure 3: Segment 3 (pay to live close) with attractions

Finally, in order to find the qualities of a "good" Airbnb review, we looked at a word cloud of reviews from Cluster 6 and Cluster 4. Cluster 6 had the most positive of reviews, while Cluster 4 had the most negative of reviews. Using these word clouds, we established three topics that distinguished a positive review from a negative review. First, location does seem to affect the renter's experience. Not only does it appear that they like listings close by "attractions", but also areas that are "safe" and "secure". Second, an "easy" and "honest" booking process is preferred. Users want to rent from people who are "responsive" and can make the booking process as seamless as possible, while not lying about their property. Lastly, cost does matter. Renters want to feel they got their money's "worth" out of the listing, instead of feeling like they "paid" for something they didn't get. We recommend focusing on these areas when renting out new properties, as well as focusing on other word topics within these good and bad clusters.

Segment 6: Positive Reviews



Segment 4: Negative Reviews

Conclusion

Based on our analysis, we recommend looking into investing in Airbnb properties that fall into Segment 6 or share the same characteristics as these listings. These properties have positive reviews, a close distance to tourist attractions, and a low average price per bed. Specifically, we recommend looking into the 12 properties from Segment 6 specified in Table 2, as they are most frequently booked. Additionally, listings from Segment 3 can be of explored if only interested in the properties that are closest to attractions, without worrying about the sentiment of the reviews or the price to rent. When ready to invest in properties and post the listings, it is important that the property is in a good location, has a user-friendly booking and check-in processes, and is priced appropriately. Using the segments we've identified, along with the keywords from the

good reviews, analytically driven decisions on which properties and locations to invest in can be made.

Appendix

Section 1: Analysis Plan

Our client ("she") wants to invest in properties in the greater Boston area, and our goal is to identify locations that can be rented on Airbnb for top dollar. More specifically, she expresses her interest in two types of properties: 1) areas that are close to local attractions and have good reviews; 2) areas that are close to local attractions and have poor reviews. To explore these questions, we also analyzed the attributes of good Airbnb reviews. Understanding that she is interested in more than one type of property and each type is composed of multiple attributes, we decided to use a hierarchical clustering algorithm with three variables.

The datasets we used were: renters' reviews on each property (reviews.csv), property information (listings.csv), properties' daily renting status (calendar.csv), and major attractions in Boston (attractions.csv). The primary variables/features we considered for clustering were: sentiment score on each property (score), the average distance to each attraction (average distance), and price per bed (price/bed). After throwing the three variables together into the clustering function and determining the number of clusters to keep, we listed the top 12 properties for the top cluster of interest. All the analyses were completed using R studio.

Section 2: Rational

While deciding between hierarchical and k-means clustering, we chose the hierarchical clustering because we were not sure about the number of clusters we needed. With the hierarchical clustering, we can visually determine the number of clusters to keep. Instead of running the cluster function on each of the variables separately, we found that performing a multivariate clustering generated segments that are more interpretable and informative. For example, multivariate clustering can explicitly tell us that "cluster 6 has above average reviews, below average to neutral price, and below average to neutral distance"; however, three separate univariate clusterings can each create 6 clusters, but the clusters for one variable do not relate in any way to the clusters for the other variables.

In preparation for the clustering analysis, we calculated three variables from the above mentioned four datasets. Firstly, sentiment score for each property listed on reviews.csv. To ensure the validity of reviews, we only focused on the properties that have at least four reviews. By tokenizing each review and then merging with the afinn library, we were able to get a sentiment score for each word. We then added up all individual scores by listing id and took the average, resulting in the average sentiment per property. This generated an output of 2023 properties with corresponding average sentiment scores.

Second, price per bed from listings.csv. This was simply determined by dividing the daily price by the number of beds. However, due to the fact that a few properties claimed to have zero beds or NA beds, these properties were excluded from the analysis. Reasons for such exclusion include 1) the resulting infinity or NaN value for price per bed will fail the clustering computation; 2) these properties weren't contributing to clustering thus were useless. A merge with the sentiment score output leads into a matrix with 2016 properties with corresponding average sentiment scores and price per bed.

Last but not least, geographic distance to local attractions. Through online research, we identified eight major tourist attractions in the Boston area, mainly located in the middle of the city or along the harbor. A 2016-by-8 matrix was computed with properties' listing id as the rows, and the eight attractions as columns. Each data point within the matrix measures the geographic distance between the property and attraction. By taking the average of each row (property), we obtained the average distance to the major attractions. While we admit that the average distance is not always the best measurement, in this case, since most of the attractions are located close together but still spread out enough to not create bias, using the median wouldn't create that much of a difference. A final merge with the above 2 variables produced a final matrix of 2016 properties along with the average sentiment score, price per bed, and distance to local attractions. Once these values were computed, we standardized each of them to equally weight the variables.

After performing multivariate hierarchical clustering algorithm, we found that keeping 6 clusters was most reasonable. Using the Dendrogram (Figure 4), splitting the tree at 6 resulted in clusters that were all interpretable. Among the clusters, the 3rd and 6th cluster appeared to meet the needs of our client. Recognizing that she doesn't have an analytics background, we decided to visualize the clusters of interest. Through the map, she can easily see the physical distribution of the relevant properties; with the word cloud, she can quickly capture the keywords renters have used in their reviews; from the list of top 12 properties, calculated by the number of days rented, she can get more detailed characteristics of the properties of interest.

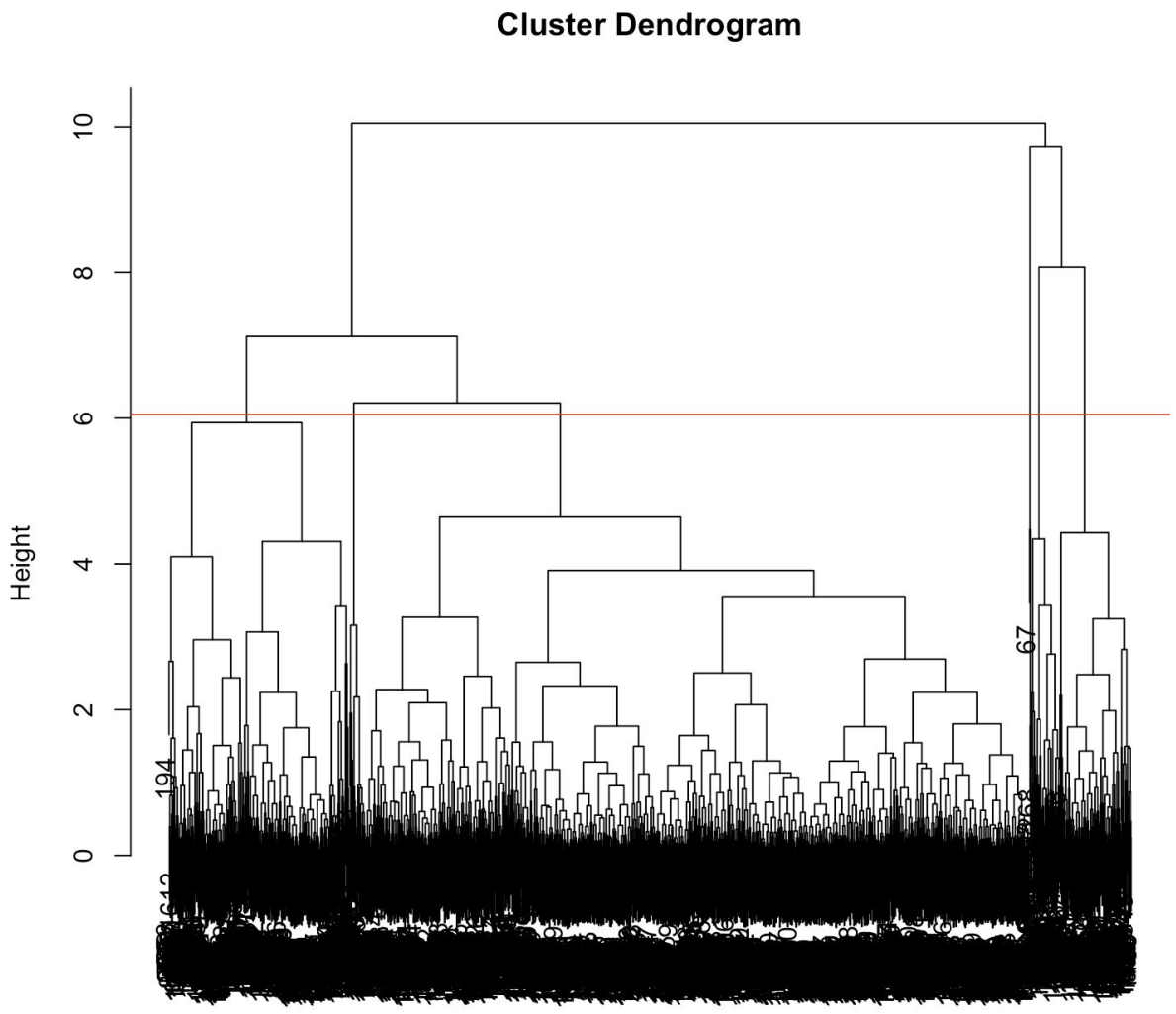


Figure 4. Dendrogram with A Cutoff Line

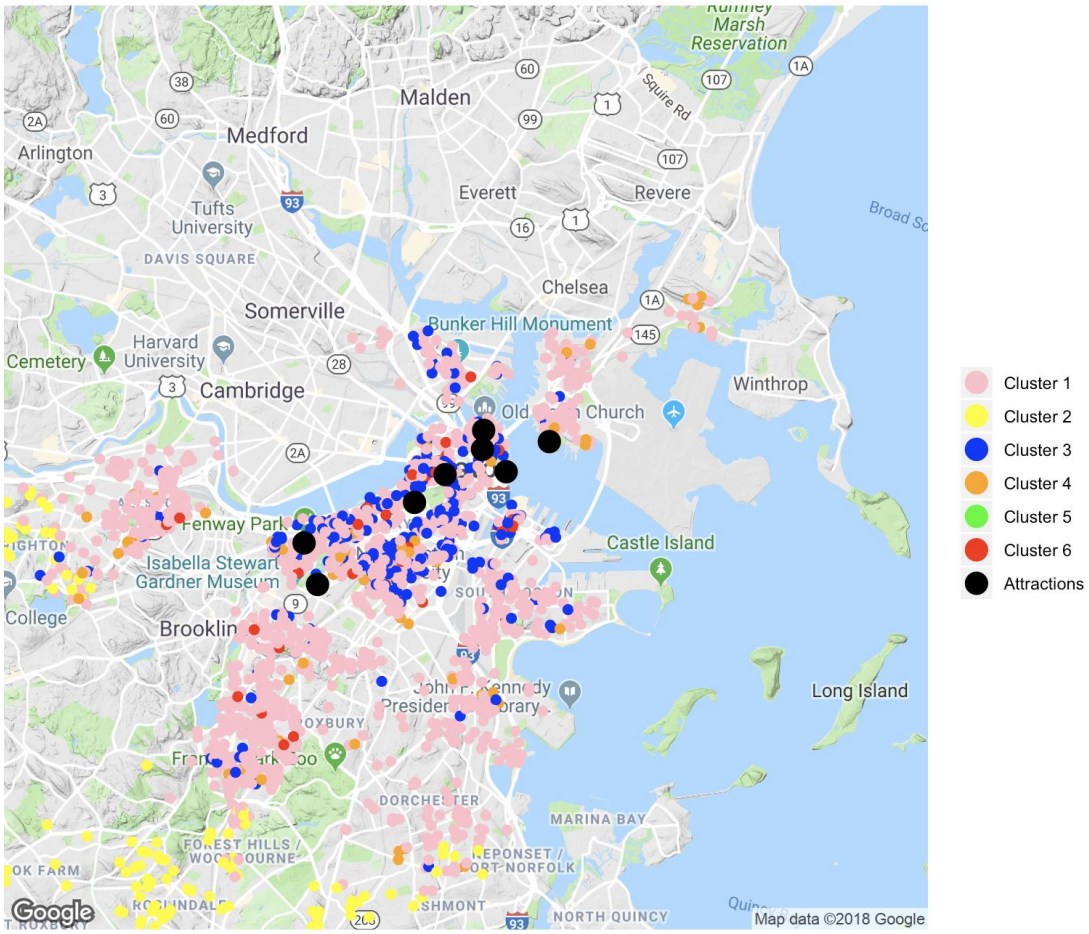


Figure 5: 6 Segments with attractions