Mining Airbnb Data in New York City

Justin Ho, *University of Nebraska-Lincoln*
Brian Nguyen, *University of Nebraska-Lincoln*
and Ben Buckwalter, *University of Nebraska-Lincoln*

1 Introduction

This is an exploration of a dataset of Airbnb listings in New York City from 2008 to 2021. This dataset is taken from InsideAirbnb, a website that is regularly updated by scraping data from available public Airbnb data. The dataset has a size of 36,923 instances that each contain many attributes but for the purposes of this paper, the ones that are important and used are:

1) Spatial Location
2) List of Amenities
3) Prices
4) Neighborhood
5) Room Type

Airbnb is a commonly used marketplace to find short-term lodging usually for vacation homestays and tourism activities. For the motivation, the question posed is, based on the data, what factors are common to appeal to customers? In an attempt to answer this question, the actions taken are to identify amenities and which other amenity or amenities are likely to be paired or included with those amenities. This paper will also try to cluster Airbnb data based on geospatial data and price and identify any correlations between the clusters, price, location, and other factors.

The objectives of this paper are:

1) **Data Processing**
   Processing the raw data into a usable state primarily for association analysis
2) **Association Analysis**
   Create associations between different amenities based on the cleaned data and find patterns that appear within the assocations.
3) **Clustering and Validity**
   Using DBSCAN, cluster the data based on geospatial data and price. Validate the clusters in visualization and metrics. Analyzing clusters and inferring patterns found between the clusters and other factors.

2 Data Preprocessing

2.1 Motivation

Like with most data sets, some amount of data preprocessing was necessary in order to convert the source Airbnb data into a usable format. In particular, the source format of the amenities for a given listing is not usable with the association analysis library we are using. Furthermore, as the amenity data is provided by the Airbnb hosts themselves, there are often different wording for similar types of amenities.

2.2 Methodology

We used a custom program written in the R programming language in order to do our data preprocessing. We chose this language as it was one we had some prior experience with and has convenient ways of expressing common data manipulations. The program can be broken down into three phases:

1) **Parsing**
   Read the data source file and retrieve the amenities and host identification lists.
2) **Simplification**
   Combine similar user provided amenities into more general wordings.
3) **Export**
   Write the processed data back to disk in a format that is usable by Weka and Python.

2.2.1 *Data Parsing*

For the association analysis objective, we are interested in generated rules based both amenities and host identification methods provided for a given Airbnb listing. Both of these attributes are provided in two separate columns as serialized JSON arrays. Tables 1 and 2 are excerpts demonstrating how this appears.

host_identification
[’email’, ’phone’, ’reviews’, ’kba’]
[’email’, ’phone’, ’facebook’, ’reviews’]

TABLE 1: Host identification examples

amenities
[”Wifi”, ”Kitchen”, ”Heating”, ”Air Conditioning”]
[”Wifi”, ”Heating”, ”Dryer”, ”Air Conditioning”]

TABLE 2: Amenities examples

After merging these two lists together, we transformed the data to be pairs of a listing_id and an amenity. This format allowed us to quickly perform modifications conditionally across all amenities.

2.2.2 Data Simplification

Now, that we had amenities in a usable format, we spent some time looking into what types of unique values were in the amenities. Counting the unique amenities, we found there to be a total of 568 different amenities. However, manually inspecting these attributes revealed that many of them were slight variations on the same amenity. One example of this was with WiFi. There were 11 different amenities corresponding to different internet speeds. To counteract this, we came up with some mappings of common substrings in the amenities and a corresponding general term to replace a matching amenity with. Table 3 provides the mappings we used in our cleaning.

Substrings	Generalized term
netflix,hbo,amazon	StreamingServices
roku,chromecast	DigitalMediaPlayer
sound,jbl,sonos	SoundSystem
shampoo,gel,conditioner,dial,aesop,bodysoap	ShowerAmenities
tv	tv
cable	cable
Dryer	Dryer
park	Parking
stove	stove
washer	washer
oven	oven
refrigerator	fridge
wifi	wifi
coffee	coffee
toiletries	toiletries
restaurant	restaurant
bar	bar
breakfast	breakfast

TABLE 3: Amentiy substring mappings

The terms with identical left and right wordings correspond to when we wanted to just remove any extra information such as brand information. After performing these string replacements on the amenities, we were left with 155 unique amenities.

2.2.3 Data Export

Finally, we needed to export the data into format that was usable for further analysis. For this, we recombined the pairs of listing_id and amenity back together into rows of listings with a binary encoding for the amenities. Each amenity had its own column and a 0 or a 1 was used to identify whether a listing had the given amenity. This format was then serialized to a CSV file, leaving us with a format that could be used for initial exploration in Weka.

Later, we also needed the data as a list of lists of amenities for use with a Python library. For this, we included a conversion from the binary encoded CSV at the beginning of our association analysis Python script.

2.3 Results

Overall, we were successful with our data preprocessing. The data was in a format usable with the Apriori algorithm library. Additionally, by merging over-specific amenities, we were able to reduce the total number of amenities from 568 to 155. Finally, we were able to come up 7 broad categories that amenities were part of:

1) Host Identification: 20
2) House Features: 50
3) Local Conveniences: 20
4) Kitchen: 25
5) Safety: 16
6) Technology: 15
7) Bedding: 9

3 Association Analysis on Amenities

3.1 Motivation

An important consideration for a prospective Airbnb client when browsing the listings is that the place they will be staying will provide the amenities they need for their trip. Furthermore, it can be important to clients that the host has provided some form of identification so the client can know if the host is legitimate. We are interested in the ways the amenities and host identification methods are part of larger patterns. In particular, we are interested in how certain groupings of amenities often imply other groupings. This information may be useful by hosts to provide more desirable listings.

3.2 Choice of Algorithm

For our association analysis, we went with the Apriori algorithm. This algorithm is the most supported allowing us flexibility on choice of programming language. Using an anti-monotone property, the algorithm is able to efficiently calculate frequent item sets. This is essential for us given the large number of listings and attributes. Another anti-monotone property is used to efficiently calculate association rules from the frequent item sets. Overall, these properties were essential in quickly seeing the results of an adjustment to the configuration.

3.3 Parameter Determination

For the Apriori algorithm there are four main parameters to adjust.

1) Max item set length
2) Minimum Support
3) Minimum Confidence
4) Minimum Lift

3.3.1 Max item set length

Since some of the more interesting rules appear in larger item sets, we decided to leave the max item set length unlimited. Although, leaving this parameter unbounded significantly increases the computation time, we were able to find an efficient implementation where this was not an issue.

3.3.2 Minimum support

Minimum support is another very important parameter. Due to the variance of amenities provided by hosts, we decided to have a relatively low minimum support. This allowed us to find interesting rules even if they did not appear in the majority of listings. We settled on a minimum support of 0.35.

3.3.3 Minimum confidence

Minimum confidence is a useful parameter for controlling the threshold at which a rule is considered interesting. As this was not the primary parameter we were interested in examining rules by, we settled on a minimum confidence of 0.8.

3.3.4 Minimum lift

Minimum lift was the primary parameter we used for determining how good a rule is. After performing some iterations to find a good cutoff point that yielded a reasonable number of rules, we settled on a minimum lift of 2. This means that the RHS of a rule is twice as likely as would be expected given the LHS and suggests a strong correlation.

3.4 Analysis

There are two main aspects to consider when addressing the results from our association analysis. Primarily, it can be interesting to explore the total number of item sets and rules generated given our parameters. Additionally, it is important to explore the rules that were generated in order to try to understand why a rule exists.

3.4.1 Item set and rule totals

Table 4 shows the breakdown of item sets by length as well as the total number of rule and item sets generated. Overall, there were itemsets ranging from length 1 to length 11.

Category	Count
Total # of item sets	23055
Total # of rules	5972
Item sets of length 1	31
Item sets of length 2	291
Item sets of length 3	1277
Item sets of length 4	3284
Item sets of length 5	5399
Item sets of length 6	5914
Item sets of length 7	4281
Item sets of length 8	1981
Item sets of length 9	528
Item sets of length 10	66
Item sets of length 11	3

TABLE 4: Counts from Association Analysis

From this, we can make some observations. First of all, since every rule must be based on a frequent item set, all rules are made up from a pool of 31 different amenities despite the fact that 155 remained after data preprocessing. This suggests that most amenities are actually infrequent.

Another observation is the shape of the item set length frequencies. There seems to be a somewhat bell shape to the data. The growth at the beginning is due to the number of combinations that can be formed from the shorter item sets quickly increasing. Then, in the middle, the growth slows before heading back towards zero as the minimum support starts to play a larger role.

3.4.2 Rule exploration

For exploring the rules, we decided to sort the generated rules by lift. This allowed us to quickly find the most interesting rules. Table 5 shows a selection of the highest lift rules found in the data set.

Clearly, the relationship between multiple kitchen amenities is quite strong. Many of the highest lift rules are different combinations of various kitchen amenities. This is understandable as you would expect a location to have an oven if it has a stove. What is more interesting is seen in the second pair of rules from Table 5. Since these rules are essentially the previous two with an additional amenity tacked on, it suggests that this is probably just the result of phone having a very high support and as a result not having any real impact on the rule. The fact that support hardly changes between these two pairs of rules reflects this assumption.

Overall, most of the other rules follow this trend. There will be some grouping of logically related amenities such as heating and air conditioning, dryer and washer, or Dishes and Silverware and Fridge. Then, there will be a number of related rules derived from this initial rule that just include up to several higher support amenities.

3.5 Conclusion

In conclusion, we were able to successfully perform association analysis on the data set and generate rules between amenities. Additionally, these rules were usually sensible and reflect related amenities. Unfortunately, this information seems to have little value for hosts. Since the rules generated are common sense, hosts will likely already list related amenities together. While many of the rules appear novel at first glance, closer inspection shows that this is due to frequent amenities that are provided across nearly all listings.

4 Clustering of Airbnb

4.1 Motivation

As Airbnb has grown explosively over the years, we are interested in how this service is distributed across the city of New York. In particular, we are interested in how it is distributed spatially across the city but also the distribution of price with respect to the geography. This is because more densely populated boroughs such as Manhattan and Queens are likely to have higher housing costs in general. Clustering based on spatial data alone would not reveal extremely interesting results. However, the interaction between prices and geography, and any hidden patterns and interactions can be uncovered by clustering based on both of these features.

4.2 Choice of Algorithm

The choice of algorithm for clustering is `DBSCAN`, a density based clustering algorithm. This is because of two of the main advantages the algorithm has as compared to other clustering algorithms and they are:

1) **Arbitrary Shapes in Clusters**
Based on the distribution of Airbnb in New York City, it is unlikely that the clusters conform to specific shapes that is better clustered by algorithms

LHS	RHS	Supp	Conf	Lift
Kitchen, Oven	Fridge, Stove	0.352	0.932	2.433
Fridge, Stove	Kitchen, Oven	0.352	0.919	2.433
Kitchen, Oven, Phone	Fridge, Stove	0.351	0.932	2.433
Fridge, Stove	Kitchen, Oven, Phone	0.351	0.915	2.433

TABLE 5: High lift rules

such as k-means clustering. Therefore. we need an algorithm that can accommodate abritrarily-shaped clusters in our analysis [2].

2) **Outliers**
Outliers are expected from this dataset. This includes geographic outliers, which can arise from Airbnb listings that are extremely far from other listings and also price outliers, which are underpriced or overpriced listings relative to their neighbours. Outliers can also exist as a result of both features. DBSCAN removes outliers and hence is suitable to our analysis [2].

4.3 Choice of Features

The choice of features are chosen based on subjective interest in the dataset. These are *spatial* and *price* since these are important aspects for any Airbnb as a host or a customer. The other reason, and arguably the more important reason, is the interpretability of the clusters. This is because we are interested in comparing these cluster with respect to ground truth data to see if it reveals interesting patterns to be understood further by future research.

4.4 Parameter Determination

DBSCAN requires two main parameters, and they are ε, which is the radius with respect to other points and $minPts$, which is the minimum points required to be considered as a dense region. In this analysis, $minPts$ is fixed at $minPts = 30$ whereas we vary the value of ε. This can be interpreted as the minimum cluster that is allowed is 30 Airbnb listings and the determination of cluster is determined by the two features we are interested in. This is ensure that the combination of parameters are linear at $O(n)$.

DBSCAN is iteratively ran with varying parameter ε. This is then checked against the ground truth spatial and price distribution to validate the clusters in context with the problem at hand. At the same time, while it is important for us to capture "micro-patterns" within the data, the number of clusters also informs us on the choice of parameter in order to be able to interpret the results found.

4.4.1 Range of ε

A pertinent issue as we iteratively apply DBSCAN on the data is that the data is extremely sensitive to the value of ε. This is because a small variation would result to a huge change in the clusters.

The range of ε that is used is as follow

$$\varepsilon \in \{1.0, 1.0 + 1 \times 10^{-7}, 1.0 + 2 \times 10^{-7}, \dots, 1.0 + 1 \times 10^{-6}\}$$

This result to more variation in the clusters as other range choices of ε resulted a domination of a single cluster in the distribution of listings, which is problematic for the analysis and usefulness of the cluster that we obtained. This resulted to the following plot on the value of ε with increments of 1×10^{-7} and the number of clusters formed as a result of DBSCAN.

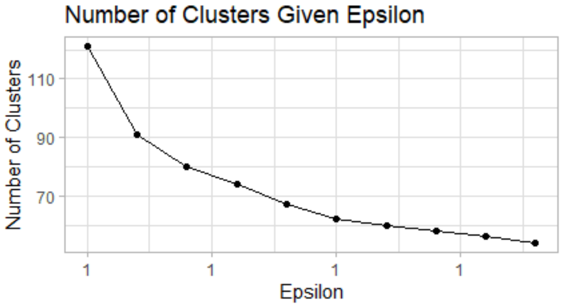

Fig. 1: Number of clusters with increments of 1×10^{-7}

4.5 Validation of Clusters

Cluster validation is an important next step for our analysis to ensure that the clusters obtained from our algorithm reflects the ground truth spatial and price distribution. It is also important for us to combine both these features and obtain clusters that reflect both features. Hence, two systematic methods are developed in the validation of clusters and they are as follows

1) **Visual Validation**
Visual validation is used to ensure that the price and spatial distribution of our Airbnb listings. This is done by plotting GIS heat maps and relating ground truth distributions of said maps.

2) **Metric Validation**
After obtaining reasonable visual validations of our clusters, using metrics that measures goodness of fit for clusters is our next step in the process. This is done in a way to ensure that the goodness of fit of both features are taken into account.

4.6 Visualizing the Data

Given that we have a total of $36,923$ instances in our dataset, it becomes abundantly clear to use that plotting points on the map of New York City proves to be extremely unwieldy and impossible to perform any reasonable inferences on them. Hence, a few rules are established in order to make the analysis possible.

4.6.1 Dominant Cluster in Neighbourhood

Based on neighbourhood level GIS data obtained from Open NYC, we are able to assign Airbnb listings to specific neighbourhoods based on the latitude and longitude provided by the dataset. This allows us to group these listings into the neighbourhoods to produce relevant heat maps that shows the spatial distribution of Airbnb Listings in New York City [1].

From the map above, we can see that most of the Airbnb listings are primarily concentrated in Manhattan, and the border between Queens and Brooklyn. From there, we want to ensure our clusters reflect such spatial distribution as well [1].

Hence, the ”Dominant Cluster in Neighbourhood” is proposed as a means to visualize the clusters.

The Dominant Cluster in Neighbourhood is defined as the cluster c_i is the dominant cluster of the neighbourhood (as determined by the GIS polygon NOT the neighbourhood of DBSCAN) such that $|c_i| = max(|c|)$, $\forall c \in C_N$, whereby C_N is the set of all clusters in a given neighbourhood.

Simply put, this refers to the largest cluster in a given neighbourhood. For example, if the $C_{\textit{Upper West Side}} = \{c_1, c_2, c_2\}$, and $|c_1| = 100$, $|c_2| = 200$, $|c_3| = 300$, then the **dominant cluster** of N = *Upper West Side* is c_3. This produced the following visuals as seen below.

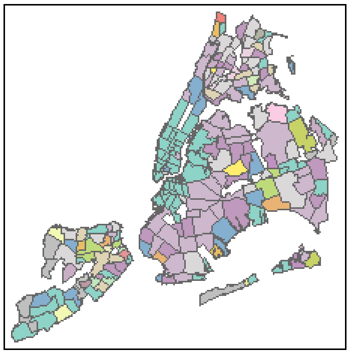

Fig. 2: Dominant Cluster Map of $\epsilon = 1 + 4 \times 10^{-7}$

In general, we want the dominant clusters reflect to the ground truth spatial distribution of Airbnb listings. Using $\varepsilon = 1.2$, we can see that this is not true.

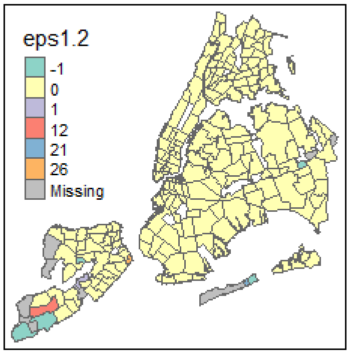

Fig. 3: Dominant Cluster Map of $\epsilon = 1.2$

This is largely due to the fact one cluster, which is cluster 0 has nearly 80% of the Airbnb listings in our dataset. Therefore, refining our parameter ε, we are partially informed by the viable range of ε by examining the heat map of the Airbnb listings and the clusters formed.

For instance, for $\varepsilon = 1 + 4.0 \times 10^{-7}$, we can see that the distribution closely match to the heat map of the number of Airbnb in Figure 4. Other ε in the proposed range in the previous section exhibit similar clustering patters.

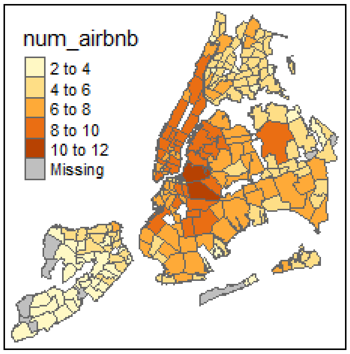

Fig. 4: Heat Map of Number of Airbnb (log transformed)

4.6.2 Price as a Dimension

The other feature that we are concerned is price, which is a critical dimension that has to be taken into account in our clusters. The heat map in Figure 5 shows the price distribution with respect to location.

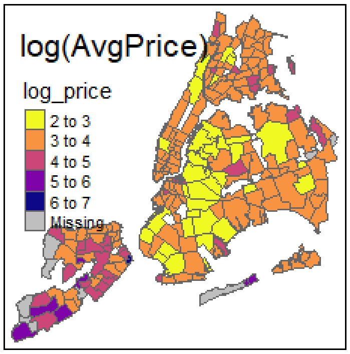

Fig. 5: Heat Map of Price of Airbnb (log transformed)

While it is difficult to visually inspect the goodness of fit of price to the extent of the analysis thus far, we would be able to better visually validate the goodness of fit with respect to price and spatial distribution with the selected cluster in a later section.

4.7 Metric Validation of Clusters

The main measure that we are interested in in this analysis is the intracluster differences, which is better if minimized. This is calculated by the definition of a centroid, which is calculated based on the mean of the features. The intracluster difference is then the average difference between each listing and the centroid of the cluster the listing belongs to.

However, a naive approach would not suffice in our analysis. This is because we observe a different goodness of fit with respect to price and spatial distribution of Airbnb listings. This is potentially a huge problem if we take the naive approach which might reflect one or the other feature in our clusters.

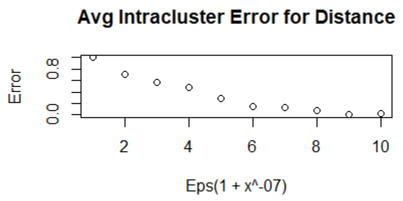

Fig. 6: Average Intracluster Distance Error

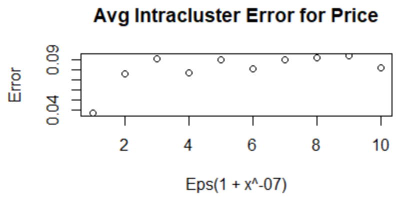

Fig. 7: Average Intracluster Price Error

4.7.1 Arithmetic Mean

Our first approach was to combine both of the features into a single metric using the arithmetic mean. Due to scale differences, we normalized the values.

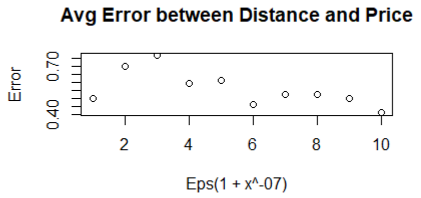

Fig. 8: Average Error with Arithmetic Mean

From the graph in Figure 8, it looks like $\varepsilon = 1 + 6 \times 10^{-7}$ is the ”elbow” of the curve, whereas $\varepsilon = 1 + 1 \times 10^{-6}$ is the minimum point to be selected as the cluster parameter.

However, the arithmetic mean as a metric would not take into account the balance of both of the features of concern. For this example, since there is more variation in distance, the arithmetic mean would favour smaller distance errors over price errors.

4.7.2 Harmonic Mean

Due to the limitations of the arithmetic mean. We propose to use the arithmetic mean as a different measure in order to account for the trade-off of both features.

Applying the harmonic mean on both features, we derive the following intra-cluster error curve in Figure 9

This shows that $\varepsilon = 1$ is the best fit. However, due to the number of clusters when $\varepsilon = 1$, which is approximately 125 clusters, it is unviable for us to choose this parameter. $\varepsilon = 1 + 9 \times 10^{-7}$ turned out to be best fit since there are only 53 clusters with this parameter and it is close to the errors derived from $\varepsilon = 1$.

We also tried $F_{\beta=2}$ with the emphasis placed on distance, which is calculated by

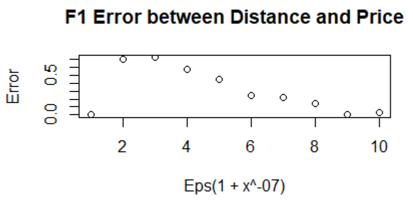

Fig. 9: Average Error with Harmonic Mean

$$F_{\beta=2} = \frac{1 + \beta^2 \times price \times distance}{(\beta^2 \times price) + distance}$$

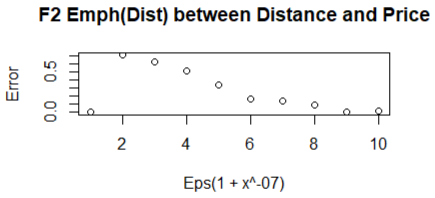

Fig. 10: Average Error with Harmonic Mean with Distance Twice as Important

The results in Figure 10 shows that $\varepsilon = 1 + 9 \times 10^{-7}$ is the best choice. This can be interpreted as the errors from distance are twice as important in the error measurement relative to price. Using $\beta = 3$, similar results are obtained.

Placing the emphasis on price, we calculated the F-score with the following formula

$$F_{\beta=2} = \frac{1 + \beta^2 \times price \times distance}{(\beta^2 \times distance) + price}$$

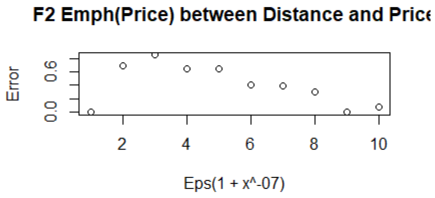

Fig. 11: Average Error with Harmonic Mean with Price Twice as Important

The results in Figure 11 shows that $\varepsilon = 1 + 9 \times 10^{-7}$ is the best choice. This can be interpreted as the errors of price are twice as important in the error measurement as compared to the errors in distnace. Using $\beta = 3$, similar results are obtained.

The results was relatively surprising given that the emphasis on either price nor distance retained the cluster results for $\varepsilon = 1 + 9 \times 10^{-7}$, which means that this parameter likely reflects both spatial and price distribution of Airbnb listings relatively well.

4.8 Choice of Parameter ε

Based on the analysis from the previous section, the choice for $\varepsilon = 1 + 9 \times 10^{-7}$.

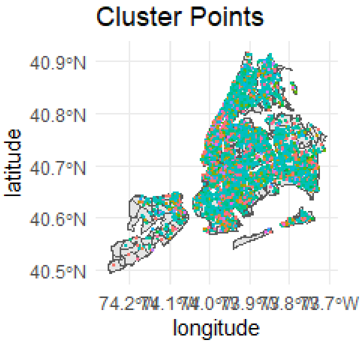

Fig. 12: Map of Airbnb Listings by Clusters of Choice

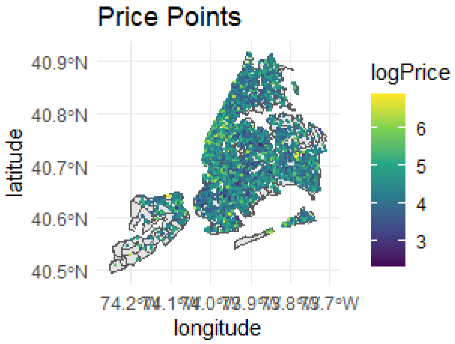

Fig. 13: Map of Ground Truth Price and Spatial Distribution of Airbnb Listings

Visually, when the Airbnb listings are plotted spatially and a price gradient is applied, it closely resembles to the clusters that is formed with our selected parameter. For instance, higher price (reflected on the lighter yellow dots in Figure 13) corresponds with locations of pink points scattered across the map in Figure 12. This provides some reinforcement of the goodness of fit for the parameter selected.

For the next section, all clusters used for the analysis and interpretation are based on the parameters $\varepsilon = 1 + 9 \times 10^{-7}$ and $minPts = 30$.

4.9 Cluster Analysis

Any data point with a price listed at $0 or less or above $1000 are excluded from the initial formation of clusters and are considered outliers. 53 clusters are formed based on the attributes price and location and meet the requirement of $minPts = 30$. All points that are not included in the 53 clusters were part of clusters that did not meet the $minPts = 30$ requirement and therefore will be considered noise.

4.9.1 Point Distribution in the Dataset

There are a total of 36,923 points in the dataset. As seen in Figure 14, the majority of prices fall between the range of $50 and $200. The average price is $145.92. The standard deviation is calculated to be 268.3845645 with a bucket size of $50 and max price at $1000.

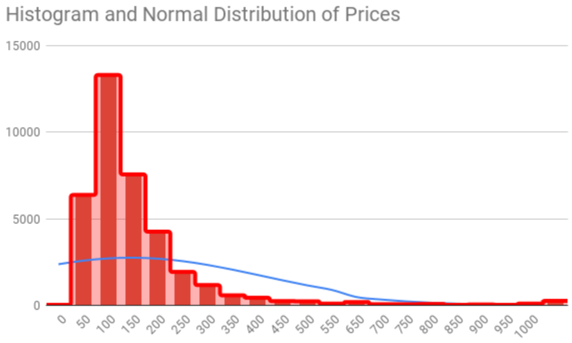

Fig. 14: Histogram and Normal Distribution of Prices

4.9.2 Point Distribution in Clusters

Out of 36,923 total points, 32,152 points are clustered, 318 points are considered outliers, and 4,453 points are not clustered and considered noise. As seen in Figure 15, the majority of clusters, 47 out of 53, contain less than 1,214 points. The average amount of points per cluster is 606.6415094 points.

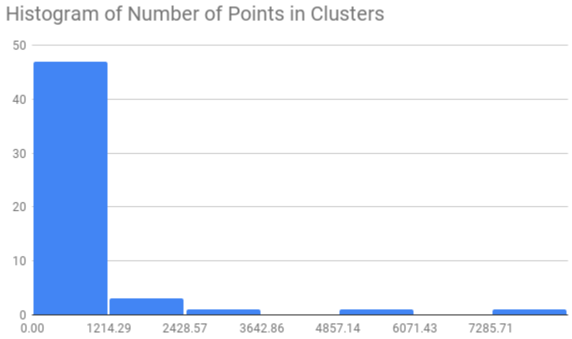

Fig. 15: Histogram of Number of Points in All Clusters

Figure 16 shows the number of clusters with less than 1,214 points. 28 out of 47 graphed clusters contain less than 166 points. The average amount of points per cluster is 238.4468085 points.

Figure 17 shows the number of clusters with less than 166 points. 13 out of 28 graphed clusters contain between 53 and 76 points. Other ranges contain relatively the same amount of clusters. The average amount of points per cluster is 79.33 points.

4.9.3 Distribution of Clusters

Although two attributes were used to make clusters, location and price, price seems to be the dominant attribute. Location seem to just cover the entire area of New York City. Clusters are formed based on a range of prices, usually in $5 increments. Figure 12 shows all points in each cluster. As seen, it is really hard to interpret if all the points in each clusters are plotted. In the next few sections, Cluster 31 will

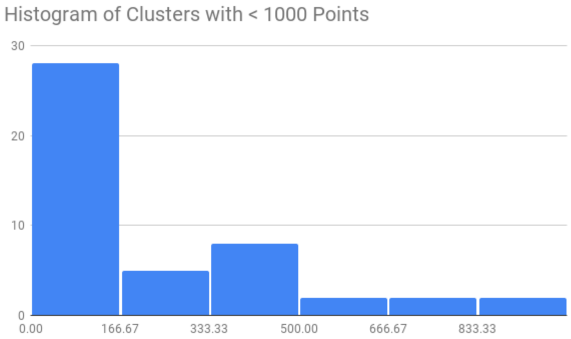

Fig. 16: Histogram of Number of Points in Clusters with ¡ 1,214 Points

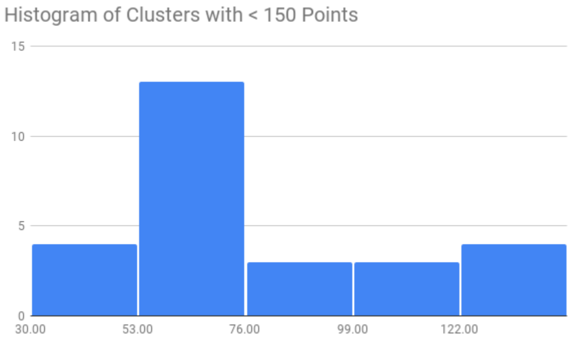

Fig. 17: Histogram of Number of Points in Clusters with ¡ 166 Points

be used as the primary cluster in which patterns are inferred for each category. Cluster 31 is used because it is the cluster with the largest amount of points. If another cluster is used, it will be stated.

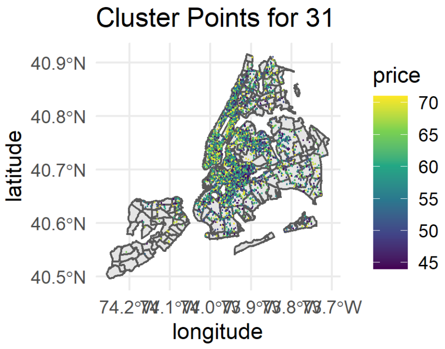

Fig. 18: Map of Cluster 31 Based on Price

4.9.4 Room Type

Room type vary by price. Prices can be split into ranges based on what room types are available at each price.

Room with low prices are defined as rooms types that cost $100 or less. Cluster 31 is an example of rooms in this price range, with the range of $45 to $70. Figure 19 shows the room types of Cluster 31. The overwhelming majority of available rooms in this range are private rooms.

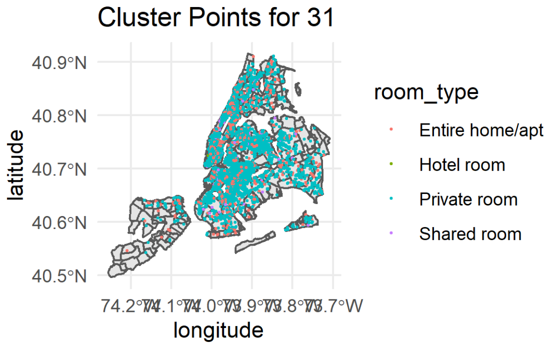

Fig. 19: Map of Cluster 31 Based on Room Type

Room with medium prices are defined as rooms that cost $100 to $175. Cluster 7 is an example of rooms in this prices range, with the range of $149 to $151. Figure 21 shows the room types of Cluster 7. It shows that most of the available rooms are entire homes or apartments with some private rooms mixed in.

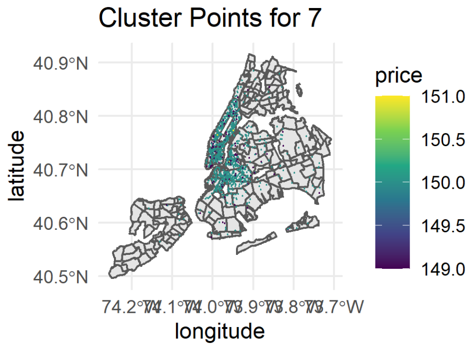

Fig. 20: Map of Cluster 7 Based on Price

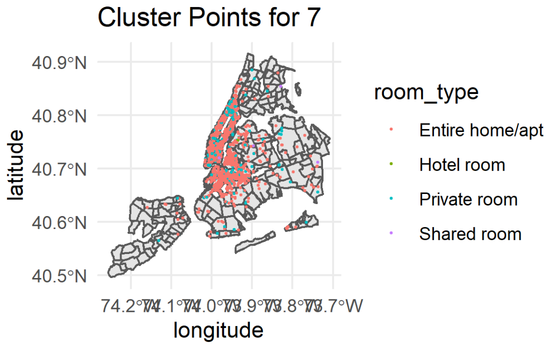

Fig. 21: Map of Cluster 7 Based on Room Type

Room with high prices are defined as rooms types that cost $175 or more. Cluster 52 is an example of rooms in this price range, with the range of $500. Figure 23 shows the room types of Cluster 52. The overwhelming majority of available rooms in this range are entire houses or apartments.

We can see a correlation between prices and room types with the higher the price, it is assumed that the room would be of higher quality, space, and/or given amenities. It can

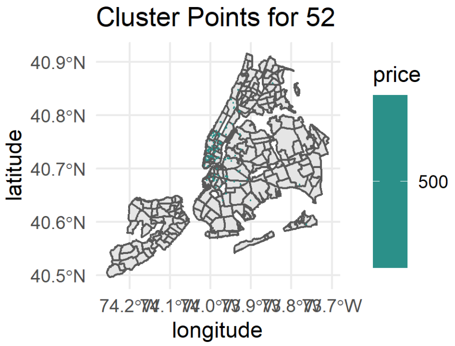

Fig. 22: Map of Cluster 52 Based on Price

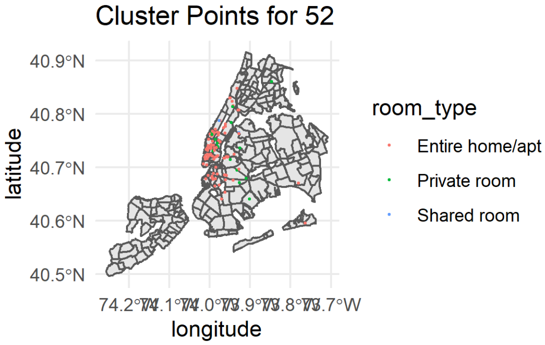

Fig. 23: Map of Cluster 52 Based on Room Type

also be inferred that the higher the price, the more limited the location. Lower prices tend to be more widespread and accessible.

4.9.5 Boroughs

Location within the boroughs and prices are correlated. Overall, the majority of locations are located in Manhattan, Northwestern Brooklyn, and Western Queens. Lower prices tend to be more widespread and across more boroughs, covering the Bronx, Manhattan, Brooklyn, and Queens as seen in Figure 18. Medium prices cover a more exclusive area, covering Manhattan, Northwestern Brooklyn, and Western Queens as seen in Figure 20. High prices are the most exclusive, mostly being in the Lower Manhattan area as seen in Figure 22.

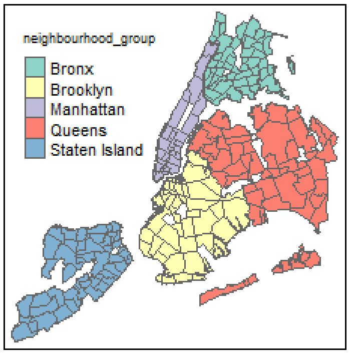

Fig. 24: Map of Boroughs of New York City

4.9.6 Subway Entrances

Subway entrances, as seen in Figure 25, are mainly concentrated in Manhattan and Brooklyn but do spread out across all boroughs. Comparing Cluster 31 to Figure 25, there seems to be a correlation between Airbnb location and the location of subway entrances. This makes sense as public transit in New York is the most common way to travel and since the majority of Airbnb users are assumed to be tourists, they need close access to some kind of public transportation to travel.

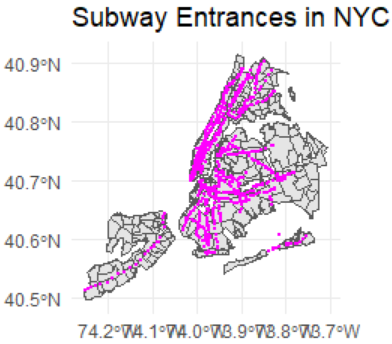

Fig. 25: Map of Subway Entrances in New York City

4.9.7 Commercial Zones

Commercial Zones are defined as shopping centers, restaurants, and tourist attractions. Essentially, any place that contains an exchange of goods. Comparing Cluster 31 to Figure 26, there seems to be a correlation between Airbnb location and commercial zones. This is intuitive, as previous mentioned, most Airbnb users are tourists and the most probable actions taken are to exchange money for some goods, whether they are souvenirs, food, or other items.

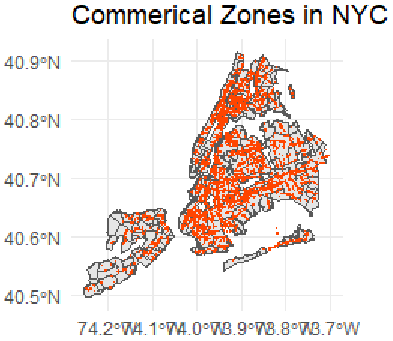

Fig. 26: Map of Commercial Zones in New York City

4.9.8 Special Zones

Special Zones are defined as areas set aside for special purposes and include schools, hospitals, government buildings, parks, and protected areas. Comparing Cluster 31 to Figure 27, there seems to be no correlation between Airbnb location and special zones. There are Airbnb locations near special zones but there is a large special zone in Staten Island that has virtually zero Airbnb locations near it. Most special zones where there are many Airbnb locations can be assumed to be residential or commercial zone adjacent areas like schools, hospitals, and parks. The large zone on Staten Island with no Airbnb locations is a protected nature reserve and so no residential area is allowed there.

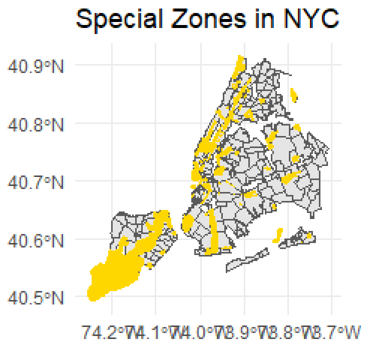

Fig. 27: Map of Special Zones in New York City

4.9.9 Eviction Rates

New York City has an Airbnb law that prohibits short term renting of apartments or renting of apartments for less than 30 days. Although this is a law, it is rarely enforced and so is commonly broken. However, when it is enforced, may affect eviction rates. Eviction rates seem to correlate to different price ranges similar to room type. Figure 29 shows the eviction rate by neighborhood.

Cluster 32 (Figure 28 is used instead of Cluster 31 to represent low price ranges as Cluster 32 contains a lower price range. Lower prices are correlated to areas with higher evictions rates, with the majority of locations in the upper Manhattan and Central Brooklyn areas.

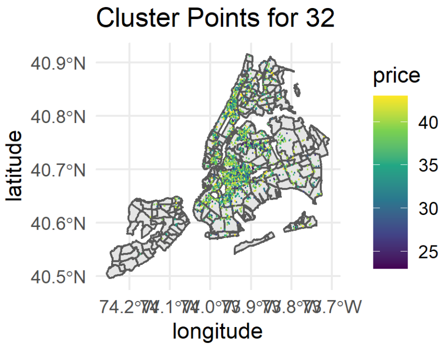

Fig. 28: Map of Cluster 32 Based on Price

Medium price ranges, represented by Cluster 7 in Figure 20, cover a smaller area than locations with low price ranges and are correlated to mid-level eviction rates in Upper Manhattan and Central Brooklyn.

High price ranges, represented by Cluster 52 in Figure 22, cover the smallest areas and are correlated to very low eviction rates in Lower Manhattan.

4.9.10 Noise

Noise points are defined as all non-outlier points that do not belong to any cluster with the $minPts = 30$ requirement. In general, these points follow the same patterns as the other clusters. They are generally found in the same boroughs, seem to correspond to subway entrances, special zones, and eviction rates and not correspond to special zones. It seems the only differentiating factor between noise points and clustered points is the range at which the price of each points fall in. Figure 30 shows all the noise points based on price.

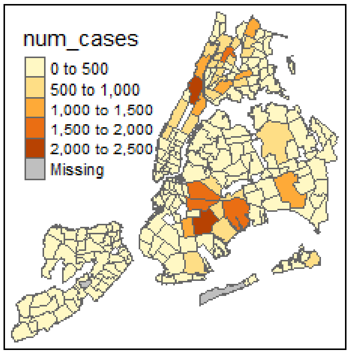

Fig. 29: Map of Eviction Rates in New York City

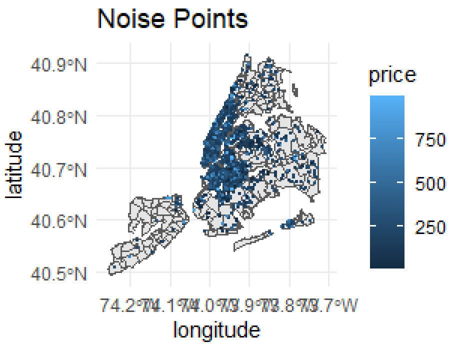

Fig. 30: Map of Noise Points Based on Price

Overall analysis seems to confirm the original hypothesis on what factors correlate to which other factors. The assumed correlations, based on economics and social science, match the actual analysis and inference of the correlations (or non correlations) between the clusters and different factors.

5 Further Work

Although correlations between different clusters and factors were founds, correlation does not that it is the reason or cause for this relationship. Research into if the causes are the result the correlation or not and why it this the case.

Association were found but no applications were applied to them. Future work can work to find and predict different type of users that are likely per association or per amenity.

6 Conclusion

In conclusion, all three objectives were successfully accomplished but further work and research must be done in order to answer to primary question of what factors are common to appeal to customers, especially the reasoning for such factors. Each objective may be successfully completed but some problems arose in the process.

Data Processing. Data Processing was successfully done but failed to reverse geocode the data due to time constraints and problems with the Google API.

Association Analysis. Association Analysis found patterns within the amenities but due to time constraints, more complex rules could not be processed and so the data was not very interesting.

Clustering and Validity. Clustering and Validity based on geospatial data and price was successful but data was more noisy than expected, making it harder to analyze and infer patterns when compared with other factors.

References

[1] Robin Lovelace, James Cheshire, and Rachel Oldroyd. *Introduction to visualising spatial data in R*. Mar. 2017. URL: https://cran.r-project.org/doc/contrib/intro-spatial-rl.pdf.

[2] Michael Steinbach, Pang-Ning Tan, and Vipin Kumar. *Data Mining*. Addison-Wesley, 2005.