# Airbnb Clustered Price Ranges

Daniel Forgosh, Brandon Ubiera, and Bemnut Nuru
Computer Science
Hood College
djf7@hood.edu, bu2@hood.edu, bn3@hood.edu

December 21, 2018

## Abstract

Clustering is a powerful method to data mine geographical data. It can group coordinate data points that consist of latitude and longitude. Aiming to use clustering as a basis, this paper data mines geographical data of all the Airbnbs in New York City to find the frequency distribution of the price within each cluster. The frequency distribution and clustering method used will find interesting data within the Airbnb data set. It is determined that K-means is to be used over other clustering strategies such as: density-based, hierarchy, and mobility. With K-means and choosing certain attributes within the data set, we were able to better understand the results by relating the data to geographical locations on the map of New York City.

## 1 Introduction

Airbnb is a online marketplace and hospitality service that allows its users to offer lodging in exchange for money. When someone offers an Airbnb listing, the individual must record numerous attributes of information for customers to see to help determine which Airbnb should be rented. Airbnb offers a data set that contains all the Airbnb listings in New York City. The data set contains 27392 Airbnb listings and 52 attributes of information for each Airbnb listing and is 1.74 GB of data. The full list of attributes can be found in the Appendix (section 10.1). Although having many attributes is helpful to provide more options when data mining large data, having high number of dimensions within data mining is bad because there will be many similarities found within the data. To solve this problem, only the latitude, longitude, and price attributes will be used. The reduced data set is 787 KB of data. The reduced data set is 1.739 GB smaller than the original data set. This will also help with processing time.

## 2 Problem Statement

While there are a large amount of attributes (Appendix 10.1), it was determined that three of the attributes to have a higher significance than the other attributes. Latitude, longitude, and prices will be incorporated into different clustering methods to provide visual feedback on Airbnb price ranges in NYC. These attributes are chosen because they will directly apply to clustering geographical data and finding density distribution of price. Latitude and longitude will be used to cluster the data. Price will be analyzed within each cluster to find the frequency distribution of price within each cluster. By clustering and analyzing these attributes, a solution can be found to help customers of Airbnbs find the most affordable Airbnb within different sections of New York. Although other attributes of the data set can provide addition information, they do not have a direct application to the goal of this project. Comparing price ranges within the clusters will provide a median, and the frequency distribution will provide the mode for prices in each cluster.

## 3 Objectives

By plotting the locations for all Airbnbs, it will be easier to visualize where large and complex clusters will appear on a map. Finding clusters of Airbnbs will be the main objective. In order to compare prices for specific areas of New York, the Airbnbs first need to be grouped by area. There are different approaches to preforming clustering on geographical data such as K-means clustering and density clustering.

After the clusters are made, the clusters will need to be plotted onto a map to help visualize each cluster. This will help visualize the location of each cluster instead of relying on viewing clusters by their geographical coordinates. The result of plotting clusters onto a map will be a map of New York City separated into sections using different colors.

Once the clusters are created, price ranges within each cluster will help determine competitive pricing for Airbnbs within specific sections of New York. After price ranges are found, the frequency distribution of price for each cluster can be found. Finding this information is important because the frequency distributions can show where the price values are most dense within the price range for each cluster.

# 4 Literature Review

## 4.1 Simplifying Clusters

Clustering geographical data is a way of simplifying many coordinate points that are in a similar space on a graph. Although clustering can be used to simplify data, clustering may still leave complex results. There are different methods to simplify clustering. One method to simplify clustering is to use heat maps. Heat maps can be used with clustering to visually show a spectrum of where there is heavy clustering through light clustering on a map; this is a visual way to simplify clustering [1]. Another way to simplify clustering is to separate complex clusters into smaller, simpler clusters. There is a method that takes complex clusters from many spatial objects, and it separates them into simple clusters from single spatial objects [2]. This method is used to simplify complex clusters into smaller, simpler clusters to make the clusters easier to understand. Both heat maps and separating complex clusters simplify clusters, but they do so in different approaches.

## 4.2 Simplifying Clusters Using Distance

Clusters can become large if not handled properly. It can be a difficult task to decide how many clusters to make and how to eliminate noise that can cause clusters to be skewed. The method that separates complex clusters to help with simplicity determines where to separate clusters based on the distance between coordinate data points [2]. This is similar to eliminating coordinate data points that are too far away to be included in a cluster to avoid noise. There is a method named Flow Hierarchical Density Based Clustering of Applications with Noise (flowHDBSCAN); this method successfully used distance to eliminate noise in clusters [3].

## 4.3 Use of Clustering Hierarchy

The hierarchical spatial flow clustering method is commonly used when determining clusters. The new flowHDBSCAN clustering method incorporates hierarchy clustering by converting density to hierarchy and then extract clusters through cluster stability [3]. Cluster stability finds only the furthest of outliers, which is important to find accurate clusters and determining which cluster a data point falls under. This differs from the Two-Steps Parameter Free Clustering Algorithm (TOSCA) clustering method which uses the hierarchy to extract (sub)clusters and reduce outliers even further [4]. This is because the TOSCA method is more focused on mobility and personal location detection by finding which real world map locations (stores, gyms, houses) correlated to the (sub)clusters. Both of these methods will be beneficial in clustering latitudes and longitudes into accurate clusters.

## 4.4 Center Based Outliers

Understanding people's behavior is a use of clustering and can help determine which locations are more prevalent. TOSCA is a user location detection clustering method that is used for GPS mobility data. Finding the center of the (sub)cluster will allow for an accurate exclusion of outliers, rather than comparing points that all together could be considered one cluster [4]. The use of heat maps in [1] will show relative ”hot spots” for the amount of data points in a certain location. TOSCA's (sub)group generation and being able to relate it to a real world location will provide meaning to the heat maps generated by geoTree [1, 4]. In order to find the exact center of each cluster to find the hot spots for heat maps the method K-means can be used; K-means is a clustering method. In this method, k number of centroids are created and are placed in the center of all k clusters created [6]. With these combined clustering algorithms, it is possible to use the Airbnbs data to find more prevalent Airbnbs. This has potential to relate to the price attribute as well because more used Airbnbs will likely be the best priced and should be found near the center of a cluster. If this relationship is not the case, it will provide that feedback as well, which would show that customers are choosing the most common Airbnbs rather than the best priced.

### 4.5 Density Based Clustering

Density based clustering of applications with noise (DBSCAN) is a technique that would be used to take two parameters with a minimum distance between each other .The two parameters would be: epsilon, which is the distance parameter that would be used to define radius to search near points and the minpoint, which is the least amount of point needed for the cluster. After the parameters are found then there would be a set which would have its clusters identified as a result of the nearest neighbor computations [5]. There would also be an algorithm called the flowHDBSCAN which would be used to combine the density based clustering and the hierarchical clustering. flowHDBSCAN is shown to be capable of inheriting density based methods strength to be able to extract the clusters and it is also capable of revealing how the data would flow based on the hierarchy, so that the viewers would be able to understand the relationship in the cluster. The flowHDBSCAN is shown to be built from two certain techniques: core distance, the distance of an object and the nearest neighbor and reachability distance, shows the clusters and noise of the dataset [3]. Unlike DBSCAN the flowHDBSCAN uses a tree like structure to create the hierarchial structure based on what was collected from the data. Unlike the DBSCAN which uses two parameters the flowHDBSCAN would happen to use one parameter, based on the minimum cluster size of MinFlows. FlowHDBSCAN can be shown to be an improvement towards DBSCAN because flowHDBSCAN is based on extended version of DBSCAN by converting it to a hierarchical cluster algorithm. The flowHDBSCAN is shown to work faster than the regular DBSCAN when performing the clustering of data.

### 4.6 Distance and Parameters

TOSCA is shown to be used to combine two clusters strategies and uses a certain tests for the parameters. The TOSCA technique would be used as an idea to detect the location of a set of users that won't result in losing the the cluster quality and any tuning phase for the parameters. A reason for the use of TOSCA is that DBSCAN is because the DBSCAN focuses on the the density around an individual point by considering the time to discover the individual. The DBSCAN shows that by this result it won't be to cluster data sets that have large amounts of density differences because combinations of the parameters called minpoints and epsilon would have a problem in being placed for all the clusters. In certain algorithms TOSCA would use a combination of two steps to solve the problem. The combination steps of TOSCA is: extracting clusters and to correspond the medoids by the means of center based methods and to cluster the medoids from the use of Single Linkage hierarchical algorithm. Unlike the DBSCAN, TOSCA gives a better detection of distance for the clusters that would be produced from the data. TOSCA also shows that it won't have the need for parameter tuning and an ad-hoc clustering for its performance unlike the other algorithms. Many people would prefer to use TOSCA in solving other types of clustering problems.

## 5 Methodology

### 5.1 Clustering Parameters

It has been determined that clustering methods are more effective the less reliant they are on parameters. The clustering method flowHDBSCAN has the one parameter of "MinFlows" which is a determined minimum cluster size; this is used for the splitting of the hierarchy tree dendrogram. K-means uses one parameter as well; it sets the number of clusters that are desired. TOSCA is a parameter free clustering algorithm which allow it to automatically adapt and have both good accuracy and good efficiency. The less parameters, or exclusion of them, help to prevent the most common errors resulted by clustering. Since the max number of parameters for these methods is 1 parameter, they are each effective because they do not rely on many parameters. Due to this, the clustering method used must be determined based on other factors.

### 5.2 Clustering Approach

Forms of density clustering rely on having dense sections to find clusters. All of the data within the New York City Airbnb data set must be clustered to separate New York City into different sections regardless of density of coordinate data points. K-means can cluster all data into logical clusters based on the number of clusters specified which makes K-means a better option to cluster the Airbnb data. It would be used to cluster similar locations based on latitude and longitude.

### 5.3 Displaying Clustering

Using maps to help show the original Airbnb data and the clustered Airbnb data will be a great resource to help visualize the data. Since there are a vast number of data entries in the data set, having a better way to visualize the data would have a large, positive impact on the project. Large amounts of data can be difficult to understand, so having something easy to look at and understand is essential.

## 6 Discussion

The overall goal is to use K-Means to cluster the geographical data and can be plotted with heat maps; then use frequency distributions to provide accurate data on price ranges for each cluster. K-Means was determined to be the best for the data representation of latitude and longitude (Methodology 6.2). Heat maps will provide different colors for clusters, based off the price range of the cluster. Clustering commonly uses maps to provide visual feedback, which will be important in displaying which Airbnb locations are most frequently used and the price ranges of those locations. The feedback provided from this data analysis can be used for any region with data on the location of each Airbnb and the prices of those Airbnbs. In the future, more attributes (Appendix 10.1) can be implemented, such as: bedrooms, beds, and bathrooms to a comparison of the value of each Airbnb.

## 7 Pre-processing

This project only focuses on three attributes from the Airbnb data: latitude, longitude, and price. These three attributes are the only attributes that need to be pre-processed because the other attributes will not be used. Since latitude and longitude were already separated attributes and are a plain number, they did not need to be processed. The price attribute had a dollar sign ($) in front of the price value for all Airbnb entries. To make processing the price attribute easier, the dollar sign was removed using Microsoft Excel's ”replace all” feature.

## 8 Creating Clusters

The programming language R was used to separate the Airbnbs into clusters. The R code reads in all the Airbnb data as a dataframe; then it creates a new dataframe that only contains the latitude and longitude attributes from the Airbnb data; none of the other data is relevant when creating clusters of geographical data. Next, the clustering method K-means is performed on the new dataframe to create 30 clusters. The cluster number for each Airbnb is combined with all the original Airbnb data and is saved into a new CSV file.

## 9 Results

### 9.1 Cluster Locations

A centroid is the center location of a cluster. The following plot shows where the centroids are located for the 30 clusters created in a geospatial plane

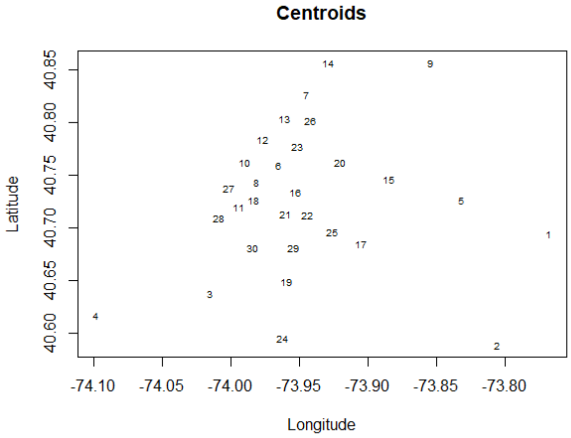

Figure 1: Plot of 30 centroids of Airbnbs in New York City

To add context to the centroid plot, we put a map of New York City behind the plot of centroids.

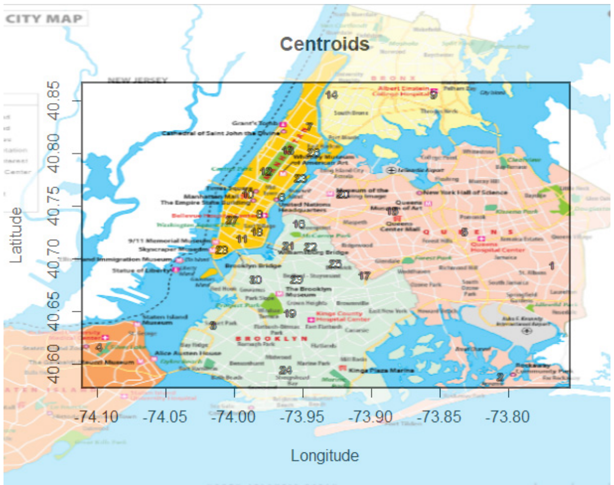

Figure 2: Map of 30 centroids of Airbnbs in New York City

## 9.2 Price frequency Distribution

The price frequency distribution results is an interesting graph that cannot be easily understood without taking a closer few at the larger prices in the data. With a reduced view, having a maximum of 200 frequency, the larger outliers can be noticed. After seeing the frequency of the higher prices a relationship is created with the individual cluster histograms and the histogram of all prices. The relationship is that if an individual cluster contained some of these larger outliers, or lack thereof, it would be interesting to take a look at why these clusters have such different price ranges while being close geographically.

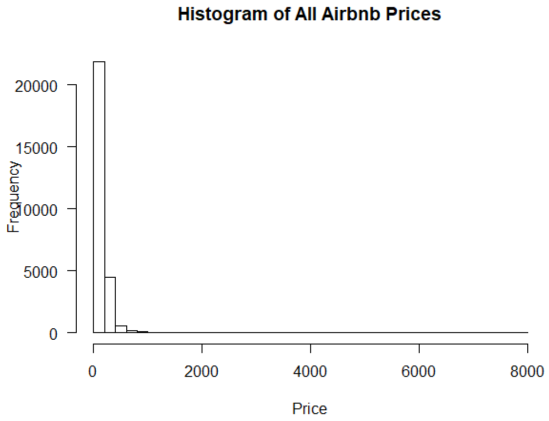

Figure 3: Original Price Frequency

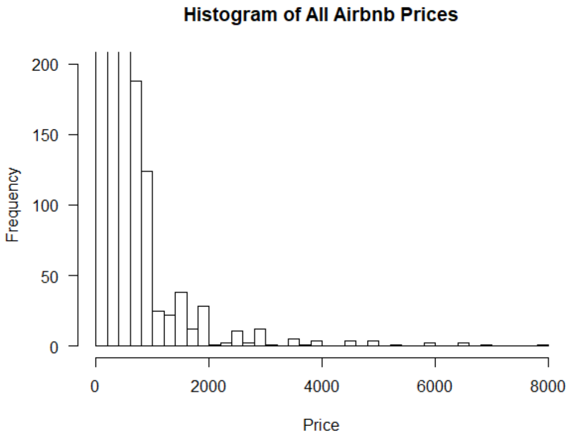

Figure 4: Reduced Frequency for Outliers

## 9.3 Frequency Distribution of Cluster Size

Each cluster is given individuality by having their own centroid. The decided centriod number of 30 corresponds to a cluster so that they can be looked at in their specific areas (Figure 5). This is the large benefit in finding interesting data using frequency distributions and K-Means because its visual representation shows the prevalence of Airbnbs in areas of New York. Cluster 2 has the minimum frequency of 52 Airbnbs and cluster 29 has a maximum frequency of 2147 Airbnbs. When looked at the mapped centroids, the range between clusters is understandable because cluster 29 is created in a central area of Brooklyn, near parks and museums, whereas cluster 2 is on an island part of Queens (Figure 2).

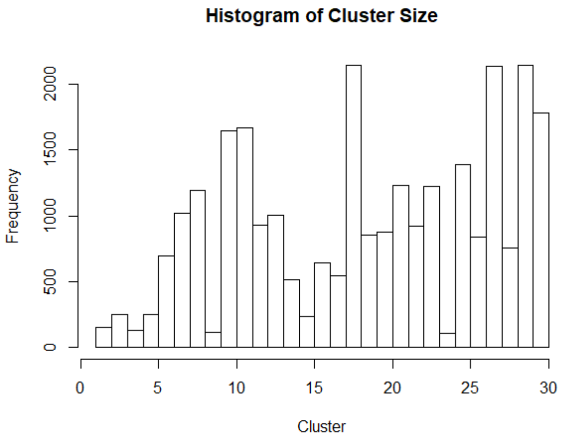

Figure 5: Airbnb Frequency per Cluster

## 9.4 Cluster 10

Cluster 10 is located in Manhattan. The following histogram shows the price range of the Airbnb listings in cluster 10. This cluster had the largest price range because it includes an outlier of $8000.

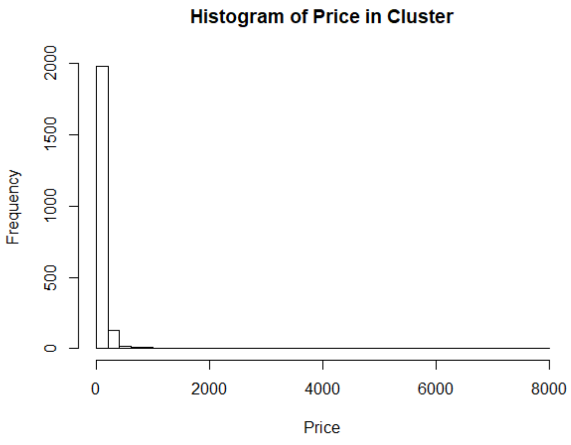

Figure 6: Cluster 10 Price Range

## 9.5 Cluster 26

Cluster 10 is located in Manhattan. The following histogram shows the price range of Airbnb listings in cluster 26. The price range is significantly less than cluster 10, with the range between maxes being over $7500.

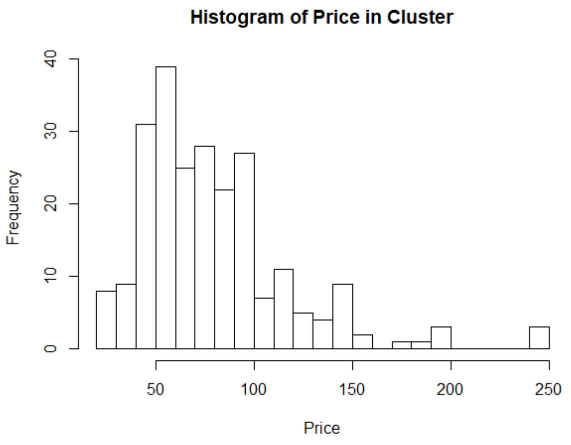

Figure 7: Cluster 26 Price Range

## 9.6 Reduced Data

All Airbnbs that have a nightly price over $100 were removed from the data set to produce further results and analysis. This is done to analyze the dense pricing sections of each cluster.

### 9.6.1 Reduced Cluster Size

The following histogram shows the size of all 30 clusters from the Airbnb data after all Airbnbs that have a nightly price over $100 were removed from the data set.

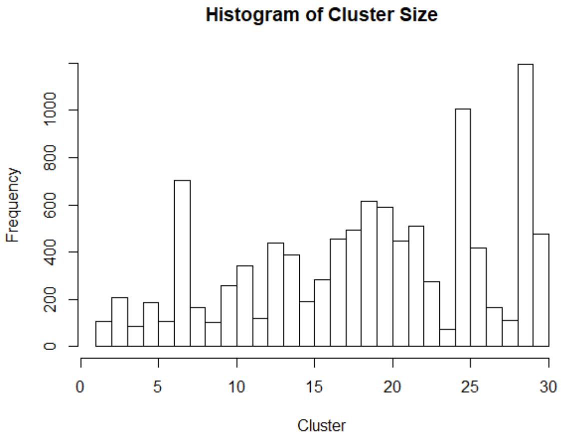

Figure 8: Reduced Cluster Sizes

### 9.6.2 Cluster 10

The following histogram shows the individual dollar price ranges within cluster 10 after all Airbnbs that have a nightly price over $100 were removed from the data set.

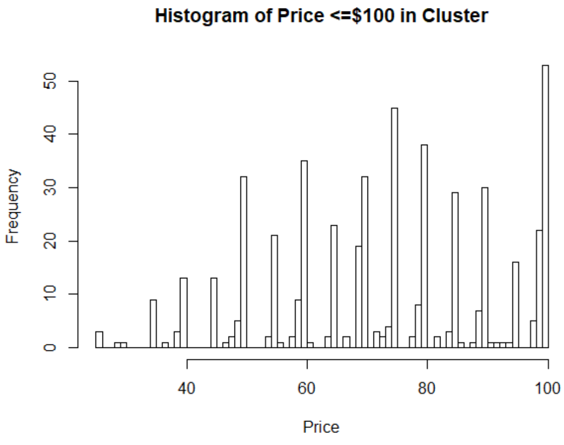

Figure 9: Reduced Price Range for Cluster 10

### 9.6.3 Cluster 26

The following histogram shows the individual dollar price ranges within cluster 26 after all Airbnbs that have a nightly price over $100 were removed from the data set.

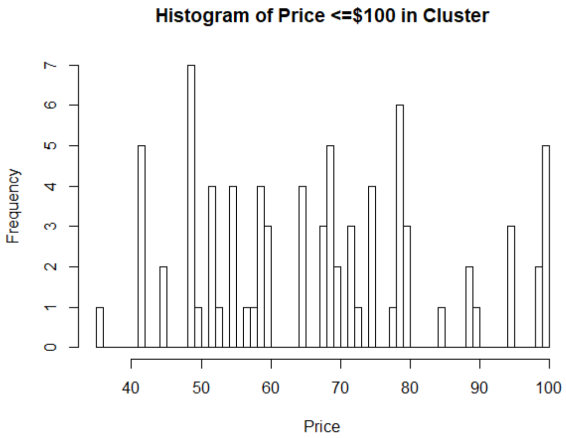

Figure 10: Reduced Price Range for Cluster 26

# 10 Analysis

When using K-Means to create clusters from the Airbnb data, 10 clusters were originally created. The clusters were large and did not retain one specific area; for example, one cluster would consist on the island of Queens and the lower portion of Brooklyn. The number of clusters was increased to 20; however, there were still problems with cluster size being too large and clusters not staying in one area. The number of clusters created was increased again to 30 clusters. Using this number of clusters made no cluster too large, and each cluster retained one specific area. Using 30 clusters was determined to be optimal for clustering New York City geographical data.

## 10.1 Cluster Sizes

The largest cluster is cluster 29 which has 2147 Airbnbs located in it, and the smallest cluster is cluster 2 which has 52 Airbnbs located in it. This is a major difference in cluster size; however, it is easily understandable. The clustering method K-Means will create cluster where it can find clusters; it is not worried about trying to make clusters equal size.

## 10.2 Cluster 10

Our goal is to find interesting data, and the individual clusters provided that. Cluster 10 is the one cluster that contained the large $8000 dollar Airbnb. This cluster is created in Manhattan where prices were no were near as high as this particular clusters price range (Figure 6). To better understand why there is a larger price range, a Google Map street view of the area is used. The area consisted of homes, open space, nicer apartments and an overall nicer area. The Airbnb listings likely has more beds, bathrooms, and bedrooms to offer which would impact the larger price range of this cluster

## 10.3 Cluster 26

Cluster 26 is chosen because geographically it is near cluster 10, but the max price Airbnb listing is $250 (Figure 7). The same approach of looking at a Google street view of the area shows that the area is not family friendly. The area consisted of only apartments which meant less to offer for the Airbnb listing, which would impact the low prices. It is likely these Airbnbs could only offer a bed or possibly one bedroom.

## 10.4 Reduced Data

In the frequency distribution of all the prices from the Airbnbs, the most frequent category is from $90-100. Since the outliers in pricing are high dollar amounts, it was decided to focus on all Airbnbs that have a price less than $100 since the nightly cost of $100 is the most frequent category. All data for Airbnbs that have a nightly cost that is above $100 was removed from the data set to focus on the distribution of the lower prices.

### 10.4.1 Reduced Cluster Size

When all Airbnbs with a nightly price over $100, the number of Airbnbs in each cluster drastically drops. The largest cluster is cluster 29; it contains 1197 Airbnbs. Cluster 29 is the same cluster that had the largest number of Airbnbs in the complete data set before any Airbnbs were removed. Since cluster 29 had 2147 Airbnbs in it before any Airbnbs were removed, 55.8% of the Airbnbs in this cluster have a nightly price that is less than $100. The smallest cluster is cluster 2; it contains 27 Airbnbs. Cluster 2 is the same cluster that had the smallest number of Airbnbs in the complete data set before any Airbnbs were removed. Since cluster 2 had 52 Airbnbs in it before any Airbnbs were removed, 51.9% of the Airbnbs in this cluster have a nightly price that is less than $100.

### 10.4.2 Cluster 10

Cluster 10 is the cluster that originally had major outliers within the data; the largest outlier in this cluster was an Airbnb that has a nightly price of $8000. With all Airbnbs with a nightly price over $100 removed, the lower price range can be easily analyzed. There is a pattern of the most frequent price options to be an increment of $10. The second most frequent price option is $1 less than an increment of $10. This is most likely done to make those Airbnbs less expensive. The most frequent price in this cluster is $100. This shows this cluster has relatively higher value Airbnbs.

### 10.4.3 Cluster 26

Cluster 26 is the cluster that originally had the least amount of outliers within the data; the largest outlier in this cluster was an Airbnb that has a nightly price of $250. With all Airbnbs with a nightly price over $100 removed, the lower price range can be easily analyzed. Unlike cluster 10, there is not a pattern of the most frequent price options to be an increment of $10. There does not appear to be a pattern in how Airbnbs are priced in this cluster. The most frequent price in this cluster is $49. This shows this cluster has relatively lower value Airbnbs.

## 11 Conclusion

The New York Airbnb data set provided interesting data when price ranges were associated with geographical locations. The data mining performed also shows that using frequency distributions and K-Means can be preformed to any Airbnb data set. The results also show that New York has a wide variety of price ranges, with most of the Airbnb listing being reasonably low priced. Some parts of New York City, such as nicer areas, show patterns in pricing whereas in other parts of New York City, such as run-down parts of the city, there do not seem to be patterns in pricing.

## 12 Future Work

While this project was able to provide interesting data, for the price frequency distributions to make more sense it required a street view of the area. To automate this and improve the understanding for the results, additional attributes can be included to look at the value rather than the price of the Airbnb listing. Potential attributes that could be added to provided a value are: beds, bedrooms, and bathrooms. Of the attribute list (Appendix 13.1) these appear to be the most reliable attributes that would accurately show a Airbnb listings value.

# 13 Appendix

## 13.1 Data Set Attributes

There are 52 attributes in the data set for all the Airbnb listings in New York City.

id
scrape_id
last_scraped
name
picture_url
host_id
host_name
host_since
host_picture_url
street
neighbourhood
neighbourhood_cleansed
city
state
zipcode
market
country
latitude
longitude
is_location_exact
property_type
room_type
accommodates
bathrooms
bedrooms
beds
bed_type
square_feet
price
weekly_price
monthly_price
guests_included
extra_people
minimum_nights
maximum_nights
calendar_updated
availability_30
availability_60
availability_90
availability_365
calendar_last_scraped
number_of_reviews
first_review
last_review
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
host_listing_count

# References

[1] Che-An Lu, Chin-Hui Chen, and Pu-Jen Cheng. 2011. *Clustering and Visualizing Geographic Data Using Geo-tree.* In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT '11). ACM, New York, NY, USA, 479-482.

[2] Eun-Jeong Son, In-Soo Kang, Tae-Wan Kim, and Ki-Joune Li. 1998. *A spatial data mining method by clustering analysis.* In Proceedings of 6th ACM international symposium on Advances in geographic information systems (GIS '98). ACM, New York, NY, USA, 157-158.

[3] Ran Tao, Jean-Claude Thill, Craig Depken, II, Mona Kashiha. 2017 *flowHDBSCAN: A Hierarchical and Density-Based Spatial Flow Clustering Method* Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics (UrbanGIS'17). ACM, New York, NY, Article No. 11.

[4] Riccardo Guidotti, Roberto Trasarti, Mirco Nanni. 2015. *TOSCA: two-steps clustering algorithm for personal locations detection* Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15). ACM, New York, NY, Article No. 38.

[5] Joachim Gudmundsson, Andreas Thom, and Jan Vahrenhold. 2012. *Of motifs and goals: mining trajectory data* Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12). ACM, New York, NY, 129-138.

[6] Viet-Hoang Le and Sung-Ryul Kim. 2015. *K-strings algorithm, a new approach based on Kmeans* Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS). ACM, New York, NY, 15-20.