Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Machine Learning-based Categorization of Airbnb Listings in NYC

Umar Farooque Syed Safdar Kadri

Peoplecert DevOps Institute Ambassador, United Kingdom

E-mail: farooqkadri@gmail.com

Abstract

This research focuses into creating a machine-learning-driven system to categorize Airbnb listings in New York City (NYC) based on neighborhood attributes and listing features. Utilizing data scraped from InsideAirbnb.com, including custom attributes such as median household income, craft beer and specialty coffee counts, and a connectivity score, KMeans clustering was applied to classify listings into four groups. These groups, named Normal People, The 2%, Central Action, and Hip Kids, offer insights into the city's diverse landscape of Airbnb offerings. The classification model's accuracy was validated using various semi-supervised learning techniques, resulting in 100% accuracy for some models. Dropping significant features like income in validation tests reduced accuracy to 66-78%, showing the importance of feature selection. The study demonstrates the potential of machine learning in enhancing Airbnb's understanding of customer preferences and refining inventory management.

Keywords: KMeans Clustering, Airbnb Listings, Geographic Analysis, Machine Learning Validation, Urban Data Analysis.

1. Introduction

Airbnb is an online marketplace that connects people seeking short-term accommodation with hosts who have available space in their homes or apartments. Since its founding in 2008, Airbnb has revolutionized the hospitality industry, enabling millions of residents from cities around the world to monetize their extra space by offering it to travelers. The platform boasts over 3 million listings in more than 65,000 cities globally, making it a

DOI: https://doi.org/10.36548/jscp.2024.3.006
Received: 13.08.2024, received in revised form: 03.09.2024, accepted: 21.09.2024, published: 30.09.2024

major player in the market for short-term lodging [1]. The concept of Airbnb is built around three main types of listings: shared spaces, private rooms, and entire homes or apartments.

In a shared space, guests often share common areas, such as the living room or kitchen, with the host and other guests. A common example would be renting a couch or a shared bedroom. Private rooms, on the other hand, offer guests a dedicated room with a door, typically a bedroom, while still sharing other common areas with the host. Finally, an entire home listing allows the guest to rent out the entire property, offering more privacy and independence [8].

Airbnb provides several features to enhance the guest experience. Hosts can list various attributes of their properties, such as photos, access to technology (e.g., Wi-Fi), parking availability, washer/dryer, kitchen appliances, and other amenities that might attract guests. Additionally, past guests are encouraged to leave reviews, offering valuable insights into the quality of the space and the overall experience. [16]. However, while the platform collects detailed information about the property itself, Airbnb does not directly provide ratings for the neighborhood or its surrounding amenities. For certain popular destinations, Airbnb has introduced neighborhood guides to offer general information about local attractions, transportation, and landmarks, but these guides are not consistently available across all locations. [15]

The motivation behind this research is to explore the feasibility of developing a machine-learning-driven classification system for Airbnb listings in New York City [6,5]. The goal is to create a rating system that would provide tourists with insights into the neighborhood amenities surrounding their rental. By analyzing neighborhood attributes in combination with specific listing features, this system could help guests make more informed decisions about their accommodation based on proximity to public transportation, restaurants, nightlife, and other key factors. [11][13].

Such a classification system could be valuable for both tourists and Airbnb. For tourists, a guidebook-style rating system based on both listing and neighborhood characteristics would offer an additional layer of transparency, helping them choose accommodations that align with their preferences and needs. For Airbnb, this system could provide a deeper understanding of customer behavior and preferences. By analyzing patterns in the data, Airbnb could better understand which types of listings are in the highest demand—whether there is a greater interest in budget accommodations, listings with superior public transit options, or those located near

specific types of amenities. This information could be used to enhance Airbnb's search algorithms and improve the overall user experience. [7].

In the course of the analysis, four distinct groups of listings were identified through clustering techniques. These groups were given descriptive names based on their unique characteristics. The first group, titled "Normal People", represents the largest category, characterized by more affordable listings with fewer amenities, catering primarily to budget-conscious travelers. The second group, "The 2%", consists of high-end, exclusive listings located in wealthier neighborhoods with premium prices and luxury amenities. The third group, "Central Action", includes listings situated in the heart of Manhattan, offering convenient access to major tourist attractions, popular coffee shops, and excellent public transportation options. Finally, the fourth group, "Hip Kids", reflects trendy, upcoming neighborhoods such as Williamsburg, where younger demographics flock to enjoy the vibrant nightlife, craft beer, and moderate prices. These groups not only reveal different accommodation options but also help to understand the social and economic landscape of Airbnb's offerings in New York City [14].

2. Data and Methods

2.1 Airbnb Listings

The Airbnb listings for New York City were collected from InsideAirbnb.com,[2] a website run by a New York City-based housing activist named Murray Cox. The website scrapes Airbnb's website for many cities worldwide, creating snapshots of all listings on the site in a town on a given day. The data used in this analysis was from the scrape of New York City on March 2nd, 2017.

The scraped data includes a lot of relevant information that was used in the analysis including the price of the listing, how many reviews have been posted on it, its approximate location (Airbnb does not publish the exact location of listings for security reasons), the minimum number of nights per booking, the room type (shared, private, entire), and many more

2.2 Outliers

To make sure the analysis was performed on Airbnb listings that are being rented, certain outliers were removed. This left us with the listings with the below attributes:

- Minimum of 7 or less nights per booking
- Listing price of $500 or less
- At least 1 review

The listing requiring a booking of greater than 7 nights begins to be considered a sublet rather than short-term housing similar to a hotel room. Some listings had absurdly high listing prices. $500 a night is the equivalent of a high-end hotel. Finally, by requiring at least one review, The listings that had likely never been booked was removed. After eliminating all outliers, the listings endeavors up with 28,970 listings.

2.3 Custom Attributes

In addition to the attributes collected by inside Airbnb, four custom attributes were added to each listing. This was developed using publicly available data. They are:

- Median Household Income
- Craft Beer Count
- Specialty Coffee Count
- Connectivity Score

2.4 Median Income

The median income of an area is a good indicator of its safety, apart from being a class marker [12]. Higher-income neighborhoods don't necessarily translate into better access to amenities like cafes, restaurants, and public transit (for instance, Upper East Side), which might be critical for a tourist visiting the city for a short duration.

Median income information was sourced from the American Community Survey of 2015. Information was collected at the Census block group level, and each Airbnb listing was assigned the median income score of the census block group in which it was located.

2.5 Craft Beer and Specialty Coffee Counts

One way to distinguish a neighborhood is to identify the kinds of businesses it can support. Certain kinds of companies cater to specific tastes. Two kinds of establishments that have become popular in what could be called the taste making young professionals' class are specialty coffee (also referred to as third-wave coffee) and craft beer. The density of these kinds

of establishments in a neighborhood should indicate the clientele a specific neighborhood attracts.

The location of specialty coffee shops and breweries were collected using Yelp's API.[4] All coffee shops in New York City returned from the API when searching for the string 'third wave coffee' were collected. Similarly, all breweries returned from the API when searching for 'brewery' were also collected. This resulted in lists of 375 coffee shops and 67 breweries.

These lists were then run through Mapzen's Isochrone API.[3] This endpoint takes a point and returns a polygon representing all the areas one can travel for a specific time and using a particular mode. Every coffee shop and brewery was then merged on a polygon that represented the area one could access in 10 minutes while walking (also known as a walkshed).

Finally, using these walksheds, the total number of specialty coffee shops and breweries within a 10-minute walk was calculated for every Airbnb listing. These totals gave the 'beer count' and 'cofee count.' The distributions of these attributes can be seen in Figure 1 and Figure 2 respectively.

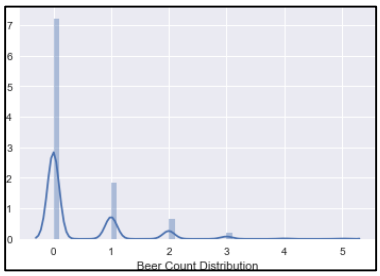

Figure 1. Distribution of Beer Count

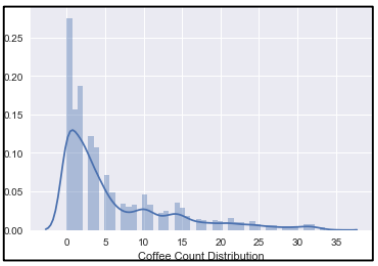

Figure 2. Distribution of Coffee Count

2.6 Connectivity Score

a transportation connectivity score was created based on the listing's access to the subway system. The first task was the creation of a network representation of the NYC subway system by using the Metropolitan Transportation Authority (MTA) [9] data that describes the service in each line. Links between the stations were created following the MTA's rules for running the trains during weekday rush hour. As shown in Figure 3 after the network of the subway system was complete, the average shortest path length was calculated for every station. In other words, from a single origin, the shortest path length to all the other possible stations

was calculated and then averaged and assigned to that station. The smaller the average shortest path length, the closest the station was to all the other stations in the system. Then, the value was reinterpreted to create a score from 0 to 29, with 29 being the closest station in the system.

To assign the score to the listings, Mapzen's Isochrone API.[3] was used again to create a 10-minute walking distance area from each station. The listings that fall inside the station's 10-minute walking distance generated area were assigned the station's connectivity score, if the listings fall within two or more stations ranges the score was added upas shown in Figure 4

Figure 3. Subway Network

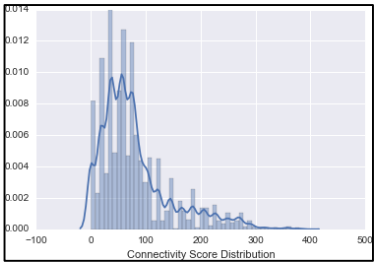

Figure 4. Connectivity Score Distribution

3. Clustering

KMeans clustering was used for four groups on the outlier reduced listings dataset to build an Airbnb listing classification model [10]. The clustering was performed on five attributes: Median Household Income, Craft Beer Count, Specialty Coffee Count, Connectivity Score, and Listing Daily Price. The size, in percentage of the total listings, of the clusters broke down as follows:

3.1 Listing Profiles

Analyzing the four groupings given by the unsupervised clustering technique, The listing profiles were created each as an example of how Airbnb might group and average the clustered attributes. Table 1 shows the percent breakdown of listings per KMeans classification.

Table 1. Percent Breakdown of Listings Per K-Means Classification

Group	Percent of Listings
0	55

1	2
2	12
3	31

similar technique to have a better idea of the kinds of listings on their site and to provide customers with an automated way to understand the neighborhood in which a listing is located. The four groupings were given unique names that captured their characteristics in a fun and memorable way. The four groups are listed below, along with the color of their mapping in Table 2.Figure 5 shows the map of Airbnb clusters.

Table 2. Groupings and Average Values of Clustered Attributes

Group Name	Color	Price	Income	Coffee	Beer	Conn.
Normal People	blue	$99	$25556	2.95	0.23	55.58
The 2%	green	$203	$183092	10.80	0.29	133.48
Central Action	red	$184	$111362	11.74	0.38	134.60
Hip Kids	purple	$153	$61687	9.86	0.73	98.10

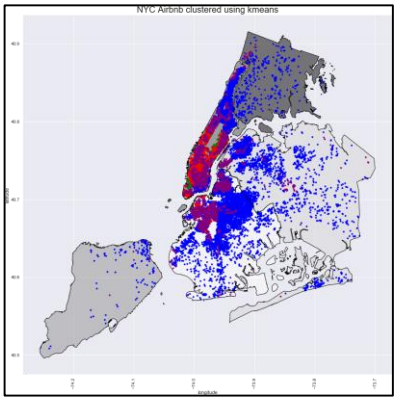

Figure 5. Map of Airbnb Clusters

Normal People: The Normal People cluster represents 55% of listings and has the lowest of all indicators, with a median income of $25,556 and an average price of $99 per night. It also has the weakest of the three custom indicators: coffee count, beer count, and connectivity score. As is evident on the map (Figure 7), the Normal People are clustered outside of the more popular centers of Manhattan and Brooklyn (though Chinatown and the Lower East Side are

still present). This is even more striking when looking at the density map of the Normal People (Figure 6). As the map reveals the strongest cluster of Normal People are in Brooklyn in the adjoining neighborhoods to the hippest parts of the Borough.

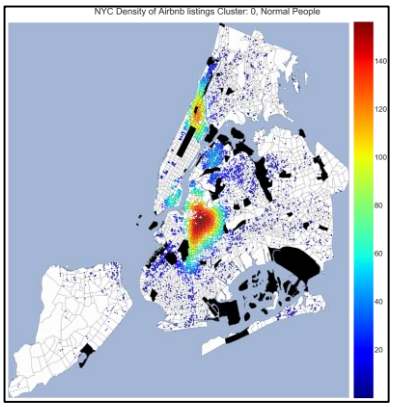

Figure 6. Density Graph of Normal People Listings

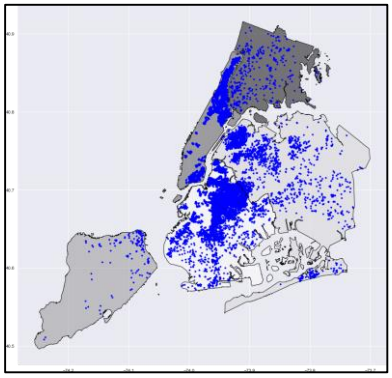

Figure 7. Map of Normal People Listings

The 2% The 2%, as the name implies, represents only 2% of the listings. These rarefied locations are also the most expensive, with an average price of $203 a night. They are also clustered in the wealthiest neighborhoods with an average median income of around $183,000. Interestingly, this high cost does not result in the highest scores in the custom indicators, with the 2% having the second highest coffee (Figure 9) and connectivity scores and second-to-last beer scores. The density map (Figure 8) of listings shows the highest concentration of these listings in Midtown East (curiously, not far from Trump Tower).

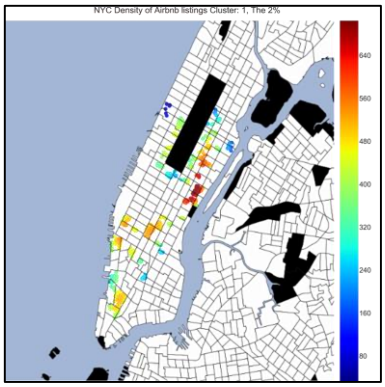

Figure 8. Cluster 1: The 2%

Figure 9. Map of the 2% Listings

Central Action: The Central Action grouping, as the name implies, is very centrally located at the core of Manhattan and is one of the city's known tourist attractions. This location no doubt contributes to the groupings of high price and general wealth, with the second highest

average nightly price and median income at $183 and about $111,000, respectively. This central clustering also puts these listings very close to coffee and connectivity, representing the highest coffee score and connectivity scores of the model. The density map in Figure 10 reveals a robust clustering in Soho and Chelsea. Figure. 11depicts the map of central action listings.

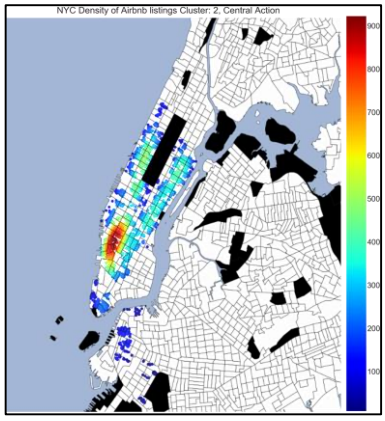

Figure 10. Cluster 2: Central Action

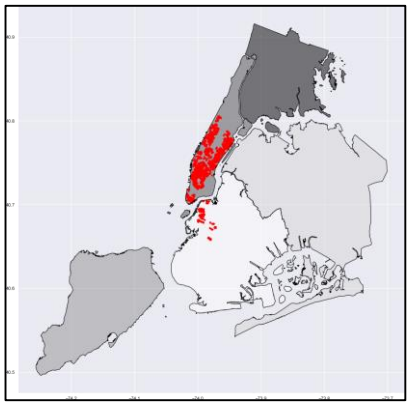

Figure 11. Map Of Central Action Listings

Hip Kids: The Hip Kids are looking for the cheapest tradeoff between accessibility, price, and amenities. Though this grouping has only the third highest price and median income (at $152 a night and about $62,000 a year, respectively), it has the highest beer score (by far) and a relatively high connectivity score and coffee score, although it ranks third in both. As the density map reveals (Figure 12), these listings are concentrated in the known hip centers of the Lower East Side and Williamsburg across the East River. As such, there is likely a correlation between the lower density of areas like Williamsburg and the significantly higher beer score, as breweries require expensive space to come to buy in Manhattan.

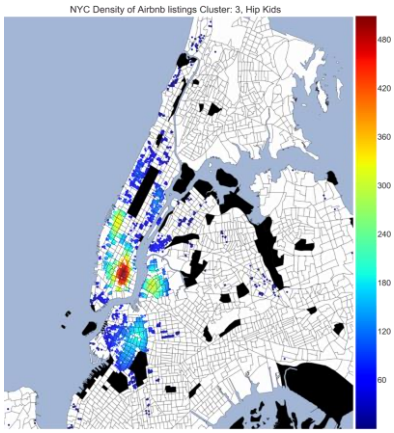

Figure 12. Cluster 3: Hip Kids

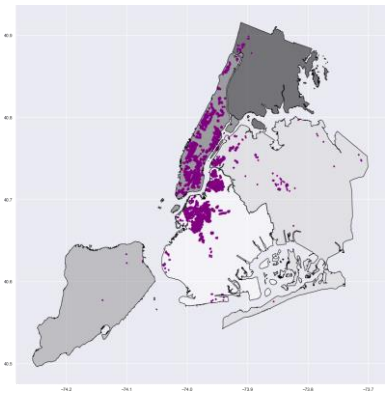

Figure 13. Map of Hip Kids Listings

3. 2 Validation

The clustering results were cross-validated using multiple semi-supervised classification techniques, namely, Naive Bayes, Decision Trees, Bagging, Adaboosting, Gradient Boosting, and Random Forest. The training data was defined so that these models could be used to predict the classes of new listings given the nature of the neighborhood they are being offered in.

Assessing the accuracy of each of the semi-supervised learning models to predict the classes defined by K-means and Hierarchical clustering techniques helped understand the quality of the clusters/classes produced by the two methods (See Table 3). Each model was run 100 epochs, and the results reported in the Tables 3 and 4 are the mean of all the 100 tests.

Table 3. Validation Results: Percent Misclassified with all the Features

Percent misclassified		
	Kmeans	Agglomerative
Decision Tree	0.00%	0.00%
Bagging	0.00%	0.00%
Adaboost	0.00%	0.00%
Gradient Boosting	0.00%	0.00%
Random Forest	0.0017%	0.0009%
Naïve Bayes	6.8755%	7.4523%

Another level of testing was carried out by dropping the most significant feature determining the initial clustering while carrying out cross-validation to check how the performance of the selected semi-supervised validation techniques gets affected (See Table 4). The degree of success in this test was expected to provide robustness of the whole methodology as in real life, all the four features used in the model might not always be available. Therefore allowing the user to make a reasonably sound prediction by feeding only 75% of the features to the model.

Results: Five out of six validation models yielded 100% accuracy in predicting the new listing class based on both K-Means and Hierarchical clustering. Naive Bayes model misclassified 6.8755% of test data for K-Means and 7.4523% of Hierarchical clustering. In the second round of validation, median income was dropped as one of the input features, and the accuracy of cross-validation models was found to be 66-78% for K-Means and 62-74% for Hierarchical clustering. The naive Bayes model was the worst performer for both clustering techniques, while Random Forest with Gradient Boosting was the best but computationally the most taxing.

Table 4. Validation Results: Percent Misclassified Without Median Income as a Feature

Percent misclassified		
	Kmeans	Agglomerative
Decision Tree	25.38%	28.43%
Bagging	23.76%	27.10%
Adaboost	33.87%	36.27%
Gradient Boosting	22.35%	25.27%
Random Forest	24.45%	27.85%
Naïve Bayes	33.27%	36.35%

3.3 K-means Clustering Methodology

K-means clustering was employed to categorize Airbnb listings into distinct groups based on neighborhood and listing attributes. The K-means algorithm partitions the dataset by minimizing the variance within each cluster. It operates as follows:

1. Initialization: Four initial centroids are randomly selected to represent the initial clusters.

2. Assignment: Each listing is assigned to the nearest centroid based on the Euclidean distance between the listing's attributes and the centroid's coordinates.

3. Update: After all listings are assigned, the centroids are recalculated as the mean of all listings within each cluster.

4. Iteration: Steps 2 and 3 are repeated iteratively until the centroids stabilize and there are no further changes in cluster assignments.

To determine the optimal number of clusters, the elbow method was used. This involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the point where WCSS starts to level off. Based on this method, four clusters were chosen as the optimal number, providing distinct and meaningful groupings for the dataset.

3.4 Validation of Clustering Results

To validate the results of the K-means clustering, several machine learning methods were employed to assess the robustness and accuracy of the clusters. The following tools were used:

1. **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, used to assess the predictive accuracy of the clusters.

2. **Decision Trees:** A model that uses a tree-like graph of decisions and their possible consequences, including chance event outcomes. It helps in understanding how features contribute to cluster classification.

3. **Bagging (Bootstrap Aggregating):** A technique that improves the stability and accuracy of machine learning algorithms by combining the results of multiple models.

4. **AdaBoosting (Adaptive Boosting):** An ensemble method that combines the results of multiple weak classifiers to improve classification accuracy.

5. **Gradient Boosting:** An ensemble technique that builds models sequentially, with each model correcting the errors of the previous one.

6. **Random Forest:** An ensemble method that combines multiple decision trees to improve classification performance and reduce overfitting.

Criteria for Selecting Machine Learning Methods: The selection of these machine learning methods for validating the clustering results was based on the following criteria:

1. Model Diversity: A mix of probabilistic, tree-based, and ensemble methods ensures a comprehensive evaluation of clustering performance from different perspectives.

2. Accuracy: These methods are known for their strong performance in classification tasks and provide reliable metrics to evaluate the accuracy of clustering.

3. Robustness: Techniques like Bagging and AdaBoosting are robust to variations in data and help assess the stability of the clustering results.

4. Computational Efficiency: The chosen methods balance accuracy and computational efficiency, allowing for practical validation without excessive resource usage.

4. Discussion

This study demonstrates how K-means clustering effectively segments Airbnb listings into meaningful groups based on attributes like price, amenities, and location. For example, the ”Normal People” group comprises lower-priced listings in less central neighborhoods, while ”The 2%” group includes high-end listings with premium amenities in affluent areas. Geographic location plays a key role, with groups like ”Central Action” found in tourist hotspots and ”Hip Kids” in trendy neighborhoods. Future enhancements could incorporate user reviews, seasonal trends, or alternative clustering methods to refine the analysis and improve Airbnb’s strategic planning.

5. Conclusion

This study successfully applied K-means clustering to categorize Airbnb listings in New York City based on a combination of neighborhood attributes and listing features. By analyzing attributes such as median household income, craft beer and specialty coffee counts, connectivity score, and listing price, the K-means clustering method effectively identified four distinct groups of listings. These groups, named Normal People, The 2%, Central Action, and Hip Kids, provide valuable insights into the diversity of Airbnb offerings across the city.

The validation of clustering results using various machine learning techniques, including Naive Bayes, Decision Trees, Bagging, AdaBoosting, Gradient Boosting, and Random Forest, demonstrated high accuracy and robustness. The models achieved up to 100% accuracy in predicting the cluster classes, confirming the reliability of the clustering process.

Future work could explore additional attributes to enhance the classification, such as amenities suited for families, noise levels, or space size. Additionally, incorporating temporal factors, like seasonal variations in listing popularity, could further refine the classification model. The approach presented in this study highlights the potential for machine learning to enrich Airbnb’s inventory management and improve customer experience by providing a deeper understanding of listing attributes and neighborhood dynamics.

References

[1] About Us - Airbnb, https://www.airbnb.com/about/about-us

[2] Inside airbnb. adding data to the debate., http://insideairbnb.com/about.html

[3] Mapzen: Isochrone api, https://mapzen.com/documentation/mobility/ isochrone/api-reference/

[4] Yelp fusion api, https://www.yelp.com/developers/documentation/v3

[5] Alharbi, Zahyah H. "A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis." Sustainability 15, no. 17 (2023): 13159.

[6] Choudhary, P., Jain, A., Baijal, R.: Unravelling airbnb predicting price for new listing. arXiv preprint arXiv:1805.12101 (2018)

[7] Ghosh, I., Jana, R.K., Abedin, M.Z.: An ensemble machine learning framework for airbnb rental price modeling without using amenity-driven features. International Journal of Contemporary Hospitality Management 35(10), 3592–3611 (2023)

[8] Guttentag, D.: Progress on airbnb: a literature review. Journal of Hospitality and Tourism Technology 10(4), 814–844 (2019)

[9] Justice, J.B., Miller, G.J.: Accountability and debt management: The case of new york's metropolitan transportation authority. The American Review of Public Ad- ministration 41(3), 313–328 (2011)

[10] Sinaga, K.P., Yang, M.S.: Unsupervised k-means clustering algorithm. IEEE access 8, 80716–80727 (2020)

[11] Wang, D., Nicolau, J.L.: Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management 62, 120–131 (2017)

[12] Wilkinson, R.G., Pickett, K.E.: Income inequality and population health: a review and explanation of the evidence. Social science & medicine 62(7), 1768–1784 (2006)

[13] Wyatt, K.: Airbnb Valuation: A Machine Learning Approach. Master's thesis, University of Arkansas (2023)

[14] Yang, Y.: Predicting us airbnb listing prices by machine learning models. Highlights in Business, Economics and Management 24, 1408–1417 (2024)

[15] Zervas, G., Proserpio, D., Byers, J.W.: The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. Journal of marketing research 54(5), 687–705 (2017)

[16] Zervas, G., Proserpio, D., Byers, J.W.: A first look at online reputation on airbnb, where every stay is above average. Marketing Letters 32, 1–16 (2021)