Airbnb

Team 12B - Andrea Pan, Angie Pang, Carson Pimental, Jaya Sruthi Raj Perikala

Table of Contents

1. Data Description

1.1. Data Source

Airbnb is one of the largest accommodation platforms. People can book a place to stay directly from hosts around the world. On each listing page, people can get the house information, like host information, price, utilities, and review ratings etc. For the dataset, we downloaded New York City listings data from Inside Airbnb, which offers free downloadable datasets for each city, and no need for permission for academic purposes. The data on this website is updated monthly, so we collected data updated on October 1st.

1.2. Data Dictionary

The original dataset has 79 columns and 36111 rows. After selecting and converting columns, we have 53 columns (including clusters) for analysis.

Variable Name	Description	Variable Name	Description
id	Unique listing identifier.	estimated_revenue_l365d	Estimated revenue generated in the last 365 days.
host_id	Unique identifier for the host.	last_review	Date of the most recent review.
host_since	Date since the host first joined Airbnb.	review_scores_rating	Overall review rating score (0–100).
host_response_time	How quickly the host typically responds	review_scores_accuracy	Rating for listing accuracy.
host_response_rate	Percentage of inquiries the host responds to.	review_scores_cleanliness	Rating for cleanliness.
host_acceptance_rate	Percentage of booking requests the host accepts.	review_scores_checkin	Rating for smoothness of check-in.
host_is_superhost	Whether the host has Superhost status (T/F).	review_scores_communication	Rating for host communication quality.
host_listings_count	Number of listings shown on the host's profile.	review_scores_location	Rating for location.

host_total_listings_count	Total number of active and inactive listings host has on Airbnb.	review_scores_value	Rating for value for price.
host_has_profile_pic	Whether the host has a profile picture (T/F).	instant_bookable	Whether the listing can be booked instantly (T/F).
host_identity_verified	Whether host identity is verified (T/F).	calculated_host_listings_count	Count of host's active listings at that moment.
latitude	Latitude coordinate of the listing.	calculated_host_listings_count_entire_homes	Number of entire-home listings host owns.
longitude	Longitude coordinate of the listing.	calculated_host_listings_count_private_rooms	Number of private-room listings host owns.
accommodates	Maximum number of guests the listing can host.	calculated_host_listings_count_shared_rooms	Number of shared-room listings host owns.
bathrooms	Number of bathrooms.	reviews_per_month	Average number of reviews per month.
bedrooms	Number of bedrooms.	host_years_active	Number of years the host has been active (derived from host_since).
beds	Number of beds.	neighbourhood	Neighborhood name provided by the host.
amenities	List of provided amenities (JSON-like text).	neighbourhood_group_cleansed_Bronx	Dummy variable: 1 if listing is in Bronx, 0 otherwise.
price	Nightly price of the listing.	neighbourhood_group_cleansed_Brooklyn	Dummy variable: 1 if listing is in Brooklyn.
minimum_nights	Minimum nights required for booking.	neighbourhood_group_cleansed_Manhattan	Dummy variable: 1 if listing is in Manhattan.
maximum_nights	Maximum nights allowed for booking.	neighbourhood_group_cleansed_Queens	Dummy variable: 1 if listing is in Queens.

number_of_reviews	Total number of reviews the listing has received.	neighbourhood_group_cleansed_Staten Island	Dummy variable: 1 if listing is in Staten Island.
number_of_reviews_ltm	Number of reviews in the last 12 months.	room_type_Entire home/apt	Dummy variable for room type “Entire Home/Apt”.
number_of_reviews_l30d	Number of reviews in the last 30 days.	room_type_Hotel room	Dummy variable for room type “Hotel Room”.
number_of_reviews_ly	Number of reviews last year.	room_type_Private room	Dummy variable for “Private Room”.
estimated_occupancy_l365d	Estimated occupancy percentage over last 365 days.	room_type_Shared room	Dummy variable for “Shared Room”.
		review_cluster	Cluster label produced from clustering review score features.

2. Data Processing

2.1. Data Cleaning

2.1.1. Drop columns and rows

Before diving into the data, we removed 36 columns that were not useful, which were unable to create insights into listings, including: 'listing_url', 'scrape_id', 'last_scraped', 'source', 'name', 'description', 'neighborhood_overview', 'picture_url','host_url', 'host_name', 'host_location', 'host_about', 'host_thumbnail_url', 'host_picture_url', 'host_neighbourhood', 'neighbourhood', 'minimum_minimum_nights', 'maximum_minimum_nights', 'minimum_maximum_nights', 'maximum_maximum_nights', 'minimum_nights_avg_ntm', 'maximum_nights_avg_ntm', 'calendar_updated', 'availability_30', 'availability_60', 'availability_90', 'availability_365', 'calendar_last_scraped', 'availability_eoy', 'first_review', 'license', 'has_availability', host_verifications', 'neighbourhood_cleansed', 'property_type', and 'bathrooms_text'.

Because our goal is to find the relation between estimated revenue and review scores and understand what drives higher review ratings, we deleted the listings with missing values in 'review_scores_rating' and 'estimated_revenue_l365d', as these are our primary analysis variables. For some other variables such as 'host_since', 'host_is_superhost', 'bathrooms', 'bedrooms', and 'beds', the missing values could not be reliably imputed using methods like mean, median, or mode. Since the proportion of missing data in these columns was small, we chose to exclude those records to maintain data quality.

2.1.2. Remove outliers

For the ‘estimated_revenue_l365d’ column, because the range of revenue was very wide, extreme values couldn’t reflect the overall real market. To address this, we applied the IQR method to remove outliers and kept only the rows within the 1.5 × IQR range.

2.2. Transformation Techniques

2.2.1. Data Type

We transformed several columns to make them more suitable for analysis. The ‘host_since’column was converted to ‘host_years_active’, by calculating how long each host has been on the platform, allowing us to measure host experience. The last_review column was simplified by extracting only the review year, since the full date was not needed. Finally, the price column was cleaned by removing the dollar sign so it could be treated as a numeric variable.

For the 'host_acceptance_rate' and 'host_response_rate', because the values include '%', we trimmed the percentage sign and divided it by 100 to convert the number to a scale of 0-1. We then filled the remaining missing values using the median for each column.

2.2.2. Dummy Variables

For the neighbourhood_group_cleansed column, which includes categories such as Manhattan, Bronx, Queens, Brooklyn, and Staten Island, and for the room_type column with values like Entire home/apt, Hotel room, Private room, and Shared room, we converted these categorical variables into dummy variables so they could be used in our analysis.

2.2.3. Binary

In 'instant_bookable', 'host_is_superhost','host_has_profile_pic', and 'host_identity_verified' columns, the values were stored as true/ false, yes/ no. These are converted into numeric values as 0/1 making them suitable for analysis.

2.2.4. Order

For the host_response_time column, which contains ordered categories such as "within an hour," "within a few hours," "within a day," and "a few days or more," we first filled the missing values using the mode. We then converted these categories into numerical values from 1 to 4 to reflect their order.

3. Analytical Methods

3.1. Software and Tools

In this Project, we used Python to do all the data preprocessing, model building, and visualization, because Python has lots of built-in libraries and functions for data analysis, and it can make a clear workflow for teammates to work on. We also used ChatGPT to help generate and debug parts of our analysis code. This support allowed us to work more efficiently while still ensuring that all analytical decisions, interpretations, and final outputs were reviewed and validated by our team.

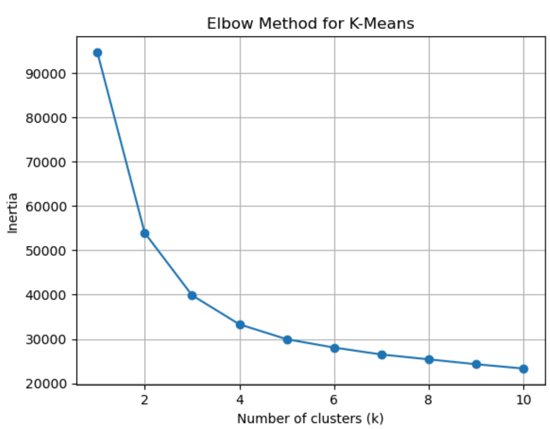
3.2. Techniques Employed and Model Building

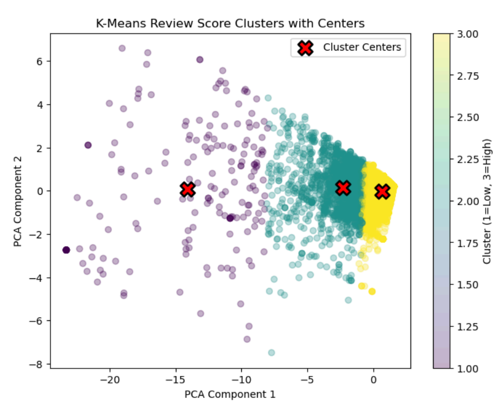
Overall, we imported "pandas" and "numpy" for data analysis; "StandardScaler from sklearn.preprocessing", "KMeans from sklearn.cluster", "PCA from sklearn.decomposition" for modeling; "seaborn" and "matplotlib.pyplot" for visualization. To conduct statistical tests, including ANOVA and regression analysis, we used the "scipy.stats" library and the "statsmodels" package. These tools allowed us to efficiently preprocess data, build clustering models, test statistical significance, and produce clear visual summaries of our findings.

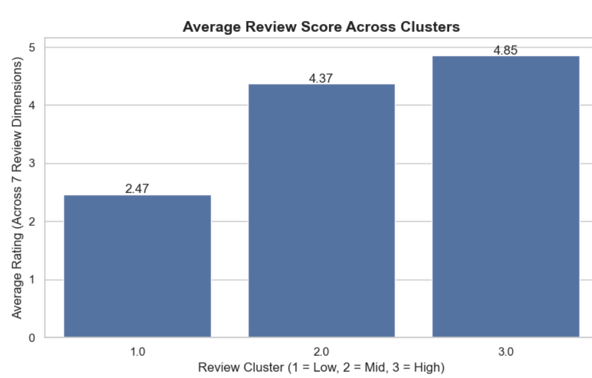
Because our goal is to increase Airbnb revenue, we assumed one way is to attract more people to book a place to stay. The biggest trigger that people want to make a reservation is reviews. Therefore, we want to understand the advantages of the high rating locations and the disadvantages of the low rating ones.

After we completed data preprocessing, we wanted to segment the listings to find the characteristics of each cluster for future targeting. Therefore, we chose all the review-related columns: 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness','review_scores_checkin','review_scores_communication', 'review_scores_location', and 'review_scores_value', for clustering.

We first standardized the review variables using StandardScaler so that all features contributed equally to the clustering algorithm. Since K-Means relies on distance, variables on larger scales would otherwise dominate the clustering process. After scaling, we applied the Elbow Method to determine the optimal number of clusters. The graph showed that k equals 3 provided a meaningful balance between simplicity and explanatory power. We also used the PCA method to do dimension reduction, so we could visualize the cluster distribution on a 2-D graph. Lastly, applying k = 3 to the k-means, we got 3 clear clusters. The average review scores were 2.47, 4.37, and 4.85 for Clusters 1, 2, and 3, respectively. Therefore, we named Cluster 1 as Low Rating, Cluster 2 as Mid Rating, and Cluster 3 as High Rating.





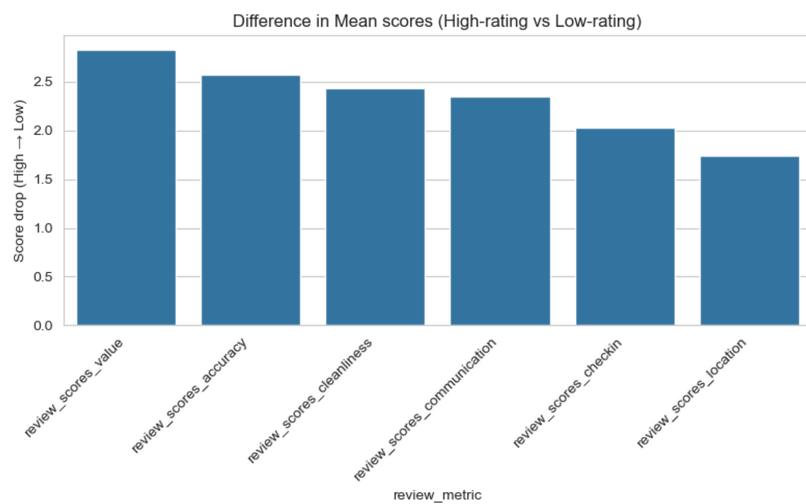


To further validate differences across clusters, we employed one-way ANOVA tests using scipy.stats.f_oneway to compare variables such as price, occupancy, and number of reviews. We also used Ordinary Least Squares (OLS) regression from the statsmodels package to test whether review clusters and host experience significantly predict listing performance. These modeling and statistical techniques together provided a robust framework for identifying the attributes that distinguish high-performing Airbnb listings from lower-performing ones.

4. Results

4.1. Insight 1: Differences in Review Scores

In order to best understand what differentiates low-performing hosts from high-performing hosts, we first calculated the mean review scores for each cluster. After identifying the highest-rated and lowest-rated clusters, we compared their customer experience metrics and visualized the score differences in descending order. The largest score drop occurred in value, indicating that listing quality and perceived worth are the strongest key factor in highlighting the differences between high-rated hosts compared to lowly-rated hosts. This was followed by differences in accuracy, cleanliness, communication, check-in, and location, suggesting that logistical factors and alignment with expectations drives host performance.



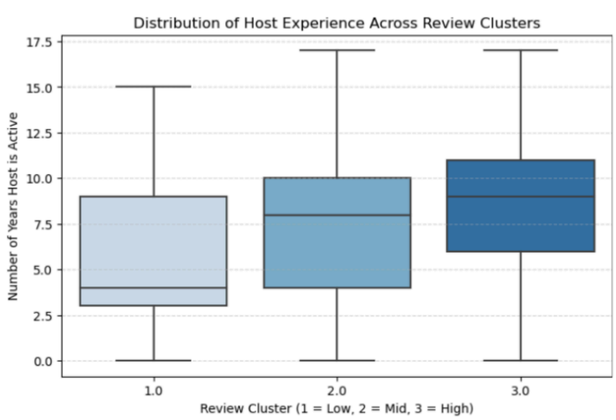
4.2. Insight 2: Leverage Hosting Experience

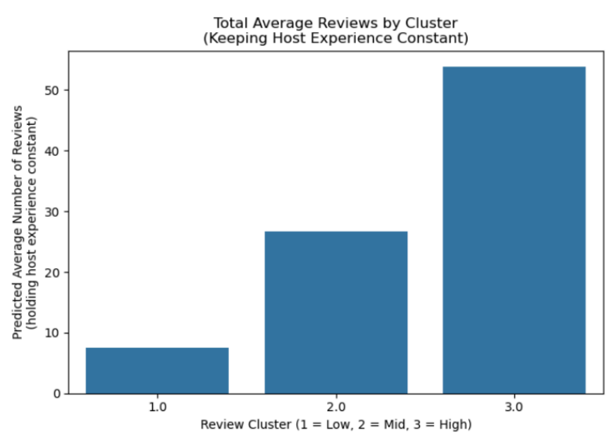
Through ANOVA test for number_of_reviews, we found that the F-statistic = 126.110, p-value = 5.441e-55, demonstrates statistically significant differences across clusters. The boxplot serves to visualize the distribution of active hosts and the total average experience within each cluster. Higher-reviewed listings tend to both accept more booking requests and achieve higher occupancy, indicating that highly-reviewed listings convert demand more effectively and are booked for more nights. However, this relationship is correlational rather than causal; the metrics move together, but one does not necessarily cause the other. Higher-rated listings also attract significantly higher review totals which helps to build long-term trust, increase visibility, and boost overall performance. More review also functions as a way to decrease the impact of negative outliers reviews

```
ANOVA for number_of_reviews:
  F-statistic = 126.110, p-value = 5.441e-55
  → Statistically significant differences across clusters.

=== OLS: number_of_reviews ~ Cluster + Experience ===
                            OLS Regression Results
==============================================================================
Dep. Variable:      number_of_reviews   R-squared:                       0.029
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     132.8
Date:                Mon, 24 Nov 2025   Prob (F-statistic):           8.06e-85
Time:                        16:09:49   Log-Likelihood:                -79886.
No. Observations:               13528   AIC:                         1.598e+05
Df Residuals:                   13524   BIC:                         1.598e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                -12.7485      6.551     -1.946      0.052     -25.589       0.092
C(review_cluster)[T.2.0]  19.1769      6.725      2.852      0.004       5.996      32.358
C(review_cluster)[T.3.0]  46.2896      6.520      7.100      0.000      33.510      59.069
host_years_active          2.4882      0.208     11.986      0.000       2.081       2.895
==============================================================================
Omnibus:                    23199.833   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         70153176.773
Skew:                          11.423   Prob(JB):                         0.00
Kurtosis:                     355.046   Cond. No.                         133.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

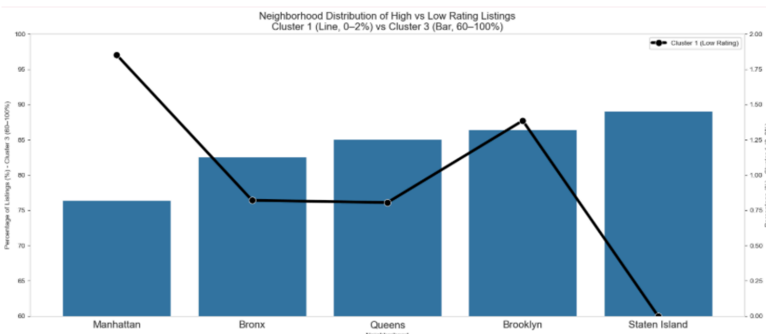


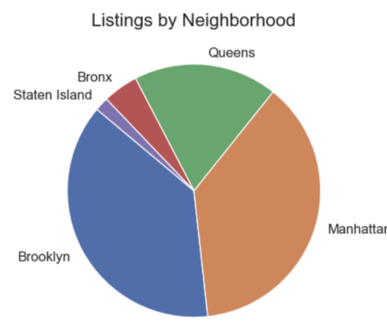


4.3. Insight 3: Target Neighborhoods

We computed percentages for Cluster 1 & Cluster 3 and then plotted them in the same bar & line chart with dual axis. This figure serves to highlight how the highest and lowest performing hosts are distributed across neighborhoods. We found that of all the neighborhoods, Manhattan has the lowest % of listings in Cluster 3 (the highly reviewed listings), but the largest % of its listings in Cluster 1 (poorly reviewed listings). Alternatively, Staten Island has the largest % of highly reviewed listings across all the regions, and the smallest % of poor reviews.

We also calculated the number of listings for each neighborhood and presented the results in a pie chart to illustrate the concentration of listings within each borough.





4.4. Insight 4: Demand Follows Quality, not Price

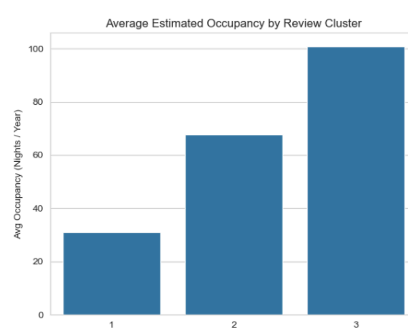
After conducting the ANOVA test, we found that "Price" does not meaningfully differ across review clusters ($p = 0.68$), but "Occupancy" does ($p < 10^{-63}$). This means higher-performing listings aren't charging more than lower-performing listings, but are being booked more, hence generating more revenue overall. This underscores that improving listing quality would be significantly more impactful for improving review scores than adjusting nightly prices.

```
Price - Cluster sizes: 190 2200 11138

ANOVA results - Price:
F-statistic: 0.3856, p-value: 6.8002e-01
=> Fail to reject H0: Mean prices are not significantly different across cluster
s.

Occupancy - Cluster sizes: 190 2200 11138

ANOVA results - Occupancy:
F-statistic: 146.6640, p-value: 9.6721e-64
=> Reject H0: Mean occupancy differs significantly across clusters.
```



```
Average Price by Cluster:
 Cluster  price
     1.0 140.88
     2.0 209.62
     3.0 213.62

ANOVA Test:
F-statistic = 0.39
P-value = 0.68

Prices are NOT statistically different across clusters (fail to reject H0).
```

5. References

Airbnb. (2024, November 11). *New report finds NYC's short-term rental law takes toll on outer boroughs*. Airbnb Newsroom. https://news.airbnb.com/new-report-finds-nycs-short-term-rental-law-takes-toll-on-outer-boroughs/

Inside Airbnb. (2024): Adding data to the debate. New York City listings dataset. Retrieved from https://insideairbnb.com/get-the-data/