Factors of Highly Occupied Airbnbs in New York City

Introduction

Airbnb has quickly become one of the most popular ways to find lodging for travels around the world. With its vast array of options and competitive pricing, it has revolutionized the hospitality industry and acquired market share from existing options (Borrison, 2015). Compared to traditional hotels, Airbnbs can offer more amenities, more privacy, and a unique vacation experience. For example, Airbnbs offer residential-style amenities like living rooms, kitchens, dining rooms, laundry facilities, and, in some locations, a private place to park your car. Globally, the average cost of an Airbnb is $137 per night while the global average cost of a hotel is $167 per night (Pacaso, 2023). A hotel room may be more expensive because it is a fully managed accommodation that includes staff members who assist you during your stay.

Airbnb is a platform that allows users, “guests”, to book homestays at properties owned by others. The owner of the property is known as the “host” and they use the platform to advertise their property to potential guests from around the world. Users can filter for properties based on the price, number of guests, type of property, and many other features. Based on their experience, guests can rate and review properties they have stayed at.

New York is Airbnb’s 3rd largest market, with around 35,000 active listings (Bae, 2019). Despite having relatively stringent requirements for listings, the New York market has retained this global position quite well, due to its strong tourism rates. Though it may lose its position to cities less constrained by regulations, New York will most likely continue to be a major market for Airbnb.

Competition to maintain a fully booked property can be fierce. In order to attain more customers and drive sales, it is critical that the content of the posting be optimized in order to be attractive to potential guests. However, behind the success of Airbnb lies a massive amount of data.

Objective and Business Justification

From booking trends to user behavior, the Airbnb dataset from Kaggle.com has the potential to help explore the patterns and trends that underlie this popular platform. In this project, we aim to gain insights into what factors customers value when searching for lodging in New York. By determining the components of an Airbnb that correlates to highly booked properties, a competitive advantage will be gained. With this information, marketing strategies can be improved, ultimately leading to more bookings and revenue.

Research Question

Our primary research question is: What factors in a New York City Airbnb posting leads to a high occupancy ratio? Occupancy ratio is a term used to indicate the duration a space is rented out compared to the total time the space is available.

Hypothesis

We hypothesize that amenities provided, the number of accommodations for a property, the property rating, the price, the presence of a cleaning fee, and the strictness of the cancellation policy, and the word choice in the name and description would affect the occupancy ratio.

There may be some amenities or conveniences that guests prefer, making certain Airbnbs more popular than others. The accommodates and rating of the property are anticipated to be positively correlated with the occupancy. Guests may prefer properties allowing more inhabitants as this typically denotes a larger property. Additionally, highly rated properties will likely be popular, while Airbnbs with low ratings may be unattractive to guests. The price of the property and the presence of a cleaning fee may negatively correlate to the occupancy as too expensive Airbnbs may be outside of the customer's budget. The strictness of the cancellation policy may negatively correlate as well since guests may feel more comfortable not being locked into a booking. Guests may look for certain keywords in the name or description when booking, making postings that contain them are attractive.

Dataset

Data Source: https://www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset

The Kaggle dataset includes 74111 entries and 30 columns of data, listed below. Each entry represents a different Aibnb property. The columns of data are a mix of numerical and categorical variables that represent information about the property.

id, log_price, property_type, room_type, amenities, accommodates, bathrooms, bed_type, cancellation_policy, cleaning_fee, city, description, first_review, host_has_profile_pic, host_identity_verified, host_response_rate, host_since, instant_bookable, last_review, latitude, longitude, name, neighbourhood, number_of_reviews, review_scores_rating, thumbnail_url, zipcode, bedrooms, beds

Missing data in the dataset are represented by null values. 13 columns contain missing data, listed below. Most have missing values comprising less than 1% of the data, but the bolded columns contain 5% or more.

bathrooms, ***first_review****, host_has_profile_pic, host_identity_verified,* ***host_response_rate****, host_since,* ***last_review****,* ***neighbourhood****,* ***review_scores_rating****,* ***thumbnail_url****, zipcode, bedrooms, beds*

From the dataset, 21986 entries meet the requirements for usage for analysis.

Data Preprocessing

The data cleaning process for Airbnb listings in New York City began by loading the dataset into the working environment. To ensure the analysis was focused on the intended geographical area, the dataset was filtered to include only listings located in New York City. Unnecessary columns, such as 'id' and 'thumbnail', were removed to simplify the dataset and focus on relevant attributes. Null values were identified and replaced with the string "missing" to maintain consistency and prevent loss of information. Columns containing date information were converted to a date format to facilitate easier manipulation and analysis.

New columns were created to provide additional insights: 'age' represents the time between the last and first review of a listing, 'hostAge' indicates the time between the last review and the host's join date, and 'review ratio' was calculated by dividing the number of reviews by the age of the listing, providing a measure of the listing's popularity and activity over time. The amenities column, which contained strings of amenities, was split into lists of individual words/values to enable analysis of each amenity. The occurrences of each word/value were counted, and the top 10 most common amenities were identified based on these counts and turned into binary columns. To impute the missing values in columns ‘beds’, ‘bedrooms’, ‘bathrooms’, and ‘host_response_rate’, “missing” was used. Using 0 would not be accurate, and we wanted to avoid removing rows for these factors entirely. Rows with infinite values and values above 1 for ‘occupancyRatio’ were removed due to being misrepresentative. These are due to calculations missing values and the same booking receiving multiple reviews respectively. To clean the dataset up further, we harmonized the other binary columns (1=True, 0=False) and removed redundant features like “first_review”. Any additional rows containing missing values were excluded from further analysis.

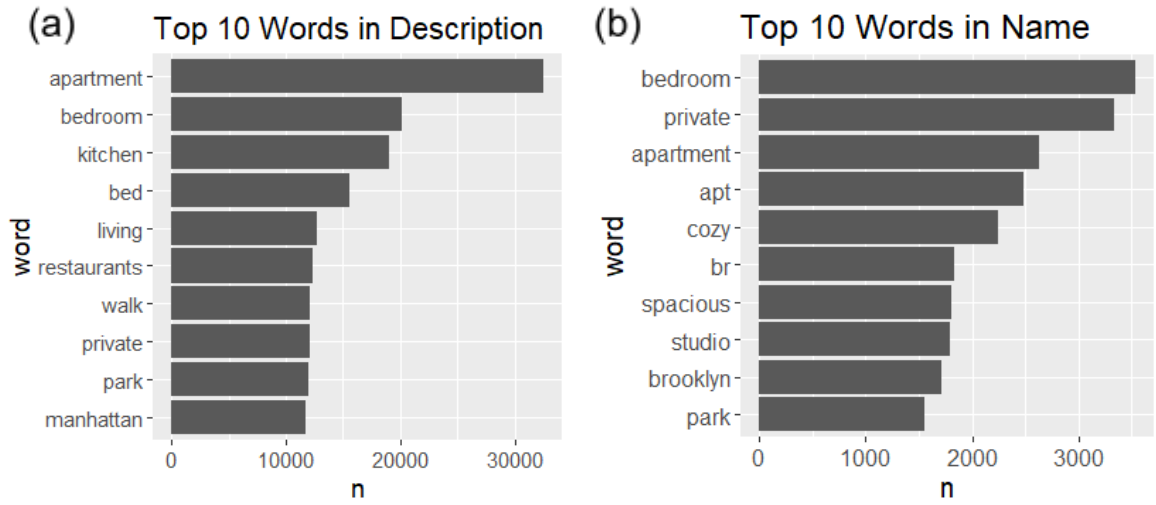
The ‘name’ and ‘description’ columns of the dataset contain unstructured text generated by the host to detail and advertise their Airbnb. The same procedure was used to preprocess the text in both columns to standardize the text data, as follows. To clean the text, non-alphanumeric characters were removed and the text was converted to the same sentence case. Next, tokenization was performed to break down the data into individual words, discrete elements also known as ‘tokens’. Stopwords, common words in the English language which hold little useful information, such as “is”, “are”, “and” were removed.

Feature Selection

<u>Natural Language Processing (NLP)</u>
The top 10 most used words for both the ‘description’ and ‘name’ were determined. Initially, some of the most common words were determined as numbers, such as ‘2’ or ‘1’. This is likely due to the advertising of the number of bedrooms, beds, or bathrooms contained in Airbnb. As this information is conveyed by other columns in the dataset, it was decided that they should be excluded from NLP. Returning to the cleaning step of preprocessing, digits were also removed.

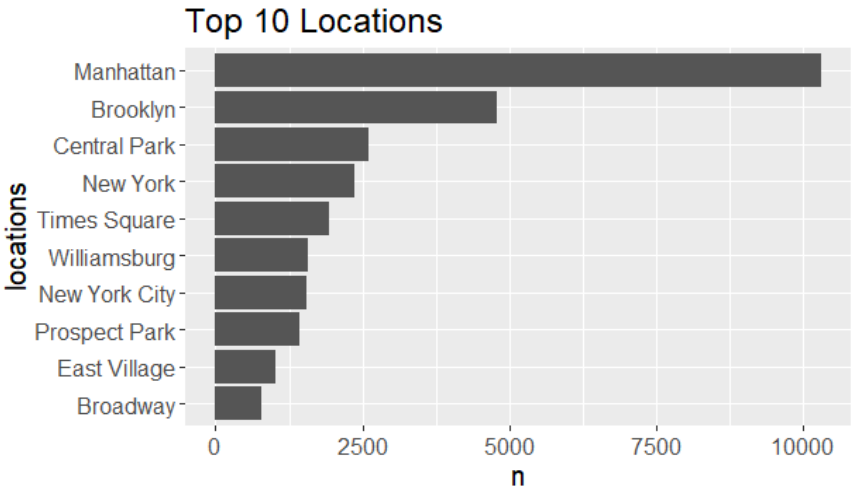
Dummy variables for each of the top words were also added to the dataframe. This will allow for regression analysis to determine significance of the words and any correlation to the occupancy ratio. Figure 1 shows the most commonly used words in the two columns analyzed.

Figure 1. Plot of top 10 most commonly used words in (a) ‘description’ and (b) ‘name’



Named entity recognition (NER) was performed on the preprocessed ‘description’ data. The goal of this step was to determine which specific locations are commonly mentioned. Hosts may want to advertise that their property is in proximity to landmarks or regions in order to attract guests. These locations were added to the data frame as dummy variables. Figure 2 shows the most common locations in the ‘description’ column.

Figure 2. Plot of top 10 most commonly mentioned locations in ‘description’



Sentiment analysis was also performed in order to find common emotional descriptive words. The goal was to see if particular words were correlated with a higher occupancy ratio. However, some of the common sentiment words found may not be valuable. For example, “hell” was a common sentiment word. This is likely due Hell’s Kitchen being a location in New York, and not the hosts using a derivative of “hell” to describe their property. It was decided that this step was superfluous and would add excessive work downstream, so sentiment analysis was excluded.

LASSO

The previously mentioned top 10 names, descriptions, and locations were added to the cleaned up data and LASSO was run from there. After running LASSO, 21 features were removed (out of 62 features) as these were identified to be either redundant or insignificant by the model. An example of a redundant feature being removed is the model keeping “New.York.City” but not “New.York”.

Principal Component Analysis (PCA)

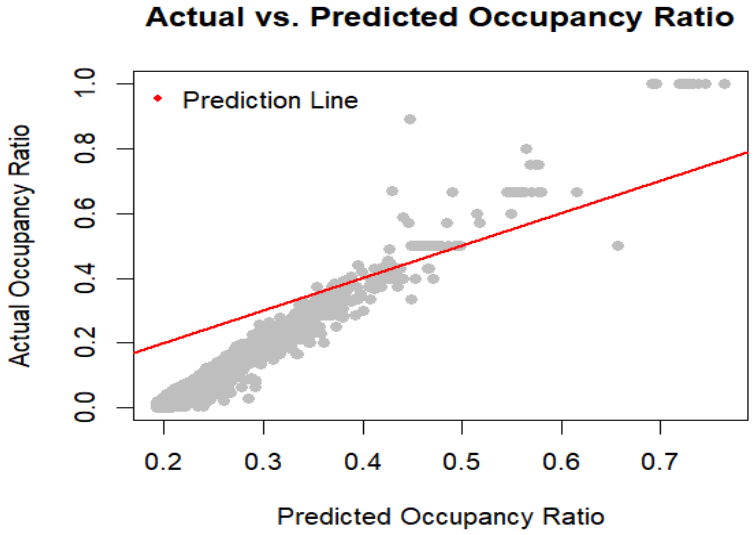
PCA was also run on the data without the NLP columns. However, it does not outright remove features and it has difficulty using categorical and binary columns (we removed these columns entirely when running PCA). This on top of its low adjusted R^2 (-3.003e-05) when running a linear regression model on its top five principal components (accounting for about 70% of the variance), is why we will not be moving forward with PCA.

XGBoost

XGBoost model for linear regression was performed to analyze the fit of the features on the target variable of occupancyRatio as well as identify the critical features of the airbnb dataset. The train and test data was split 70-30 and parameters were selected by trials. The results were measured for Root Mean Square Error which came at 2.5% and Residuals vs Actuals, Feature Importance and distribution of residuals were analyzed. The Gini Impurity Index for the XGBoost parameters was found to be at 0.41, indicating a spread of importance across the multiple features of the models

In the actual vs predicted occupancy ratio (Figure 3), most predicted values stayed below and around the perfect prediction line, which dictates the model can be evaluated for further insights.

Figure 3. Plot of actual vs predicted occupancy ratio.



Next, we measure the distribution of the Residuals against their frequency. For illustration purposes, values <-0.3 are ignored due to their smaller size. Most residuals fall within 10% - 20% which is inline with the MSE calculated earlier. We then measure Gain of the features to

the Importance Table to identify the dominant features. Only top 5 features with their highest to lowest contribution are measured.

Random Forest

By running the Random Forest algorithm on the target variable, we validate the model fit of features and target variable, we observe an MSE 0.19% and the model summary explains 82% of the variation. The top 5 features are again tested to see how they align with the findings of the XgBoost Model. The gini impurity index for Random Forest was 0.68 indicating a moderate feature importance mix. Note that both models identify the following features as most significant features: *age, number_of_reviews, hostAge, log_price, instant_bookable*.

Multiple Linear Regression on XGBoost Driven Critical Features

Multiple Linear Regression was used to estimate how strong the relationship was between *OccupancyRatio* (dependent variable) and the other variables (independent variables).

Dependent variable: *OccupancyRatio*
Independent variables: *log_price, accommodates, cleaning_fee, host_has_profile_pic, host_identity_verified, instant_bookable, latitude, longitude, number_of_reviews, age, hostAge, Wireless.Internet, Kitchen, Heating, Essentials, Air.conditioning, Smoke.detector, TV, Hangers, Shampoo, Manhattan, Central.Park, Times.Square, New.York.City, Prospect.Park, East.Village, Broadway, private, apartment, apt, cozy, br, spacious, studio, brooklyn, apartment.1, bedroom.1, kitchen, bed, walk, private.1, park.1*

Prior to modeling, the data was split into training and testing sets in the ratio of 70:30 so that we could use the testing set to validate the final selected model. We also omitted objects with log_price equal to 0 since it could not be 0.

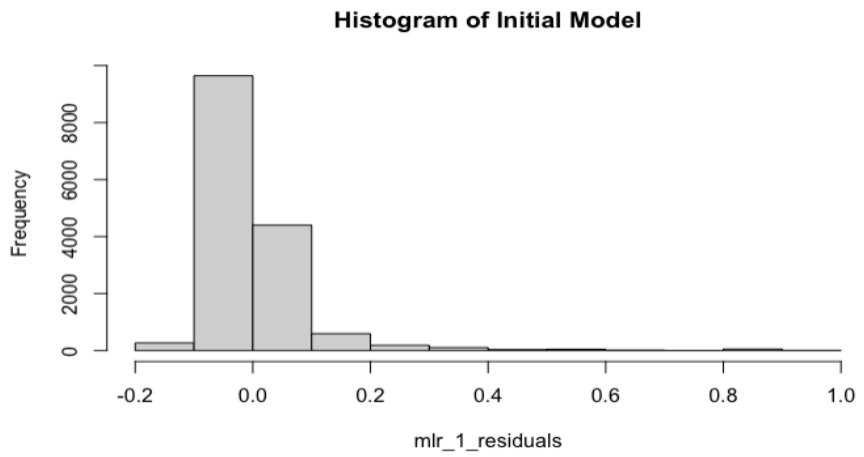
We used all the variables to build our initial multiple linear regression model.

Note that a variable will be significant if its p-value is less than 0.05. In addition to providing that information about the model, it also renders the Adjusted R-square, which evaluates the performance of models against each.

P-value < 0.05: *log_price, accommodates, cleaning_fee, host_identity_verified, instant_bookable, number_of_reviews, age, Wireless.Internet, Kitchen, Heating, Smoke.detector, Hangers, Times.Square, New.York.City, bedroom.1*
MSE: *0.0091*
R-squared: *0.2427*
Adjusted R-squared: *0.2407*

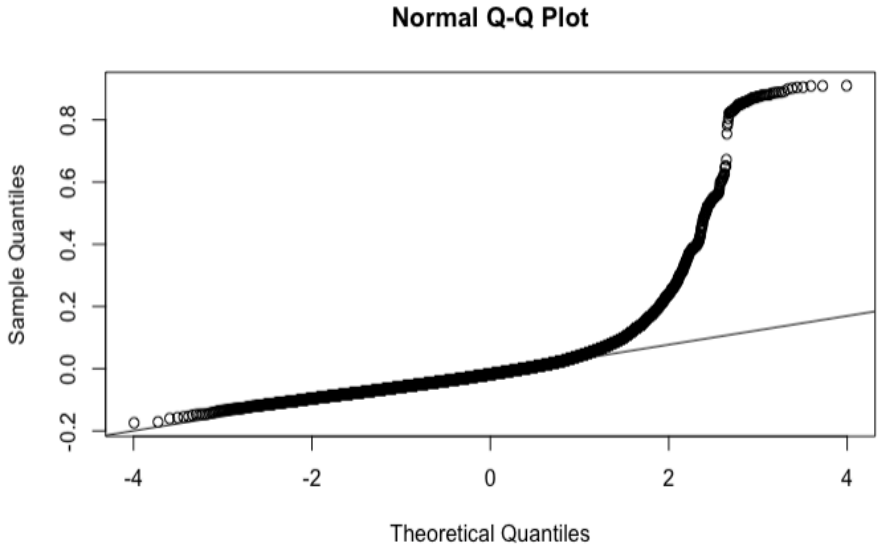
The histogram looks skewed to the right; hence we can't conclude the normality with enough confidence. Instead of the histogram, let's look at the normal Q-Q plot. If there is normality, then the values should follow a straight line.

Figure 4. Histogram for the original linear model with all the variables



From the plot, we observe that only a few portions of the residuals lie in a straight line. Thus, we assume that the residuals of the model do not follow a normal distribution.

Figure 5. Normal Q-Q plot for the original model with all the variables



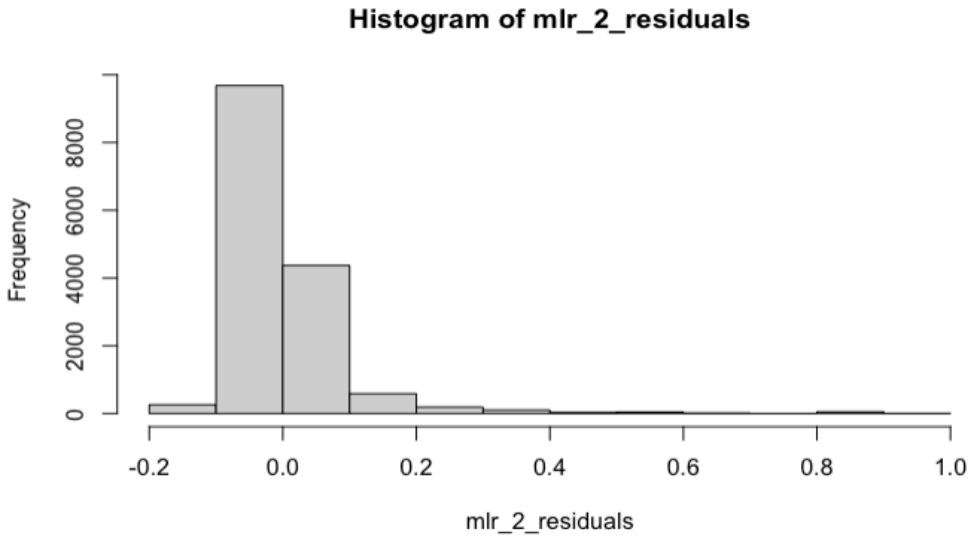
To build our second multiple linear regression model, we only used the selected variables that are considered to be significant by the initial model.

P-value < 0.05: *log_price, accommodates, cleaning_fee, host_identity_verified, instant_bookable, number_of_reviews, age, Wireless.Internet, Kitchen, Heating, Smoke.detector, Hangers, Times.Square, New.York.City*
MSE: *0.0091*

R-squared: *0.2405*
Adjusted R-squared: *0.2398*

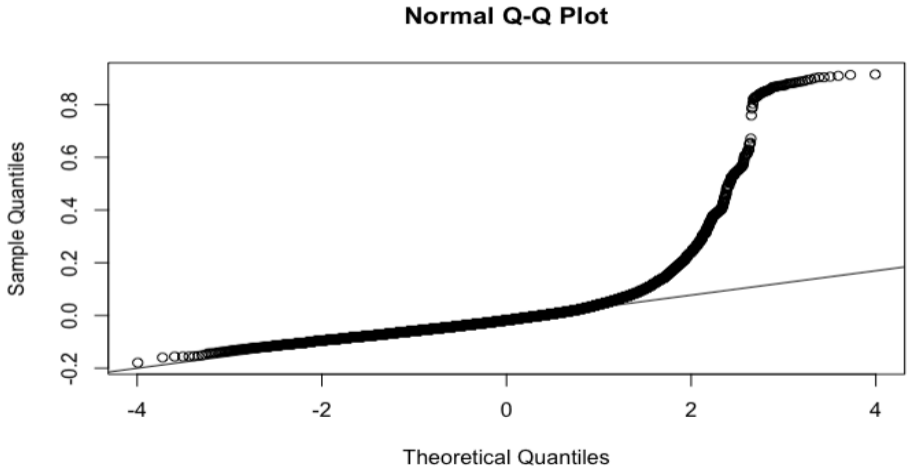
The histogram for the second model also looks skewed to the right; hence we can't conclude the normality with enough confidence.

Figure 6. Histogram for the second linear model with all the variables



From the Q-Q plot for the second model, we don't see a significant change on the straight line.

Figure 7. Normal Q-Q plot for the original model with all the variables



The initial multiple linear regression model has an adjusted R-squared of 0.2407, which is higher than the second model's adjusted R-squared (0.2398). This means that the initial model with all the predictors is better than the second model.

The logical next step of this analysis is to remove the non-significant variables and fit the model

to see if the performance improves. Another strategy for efficiently choosing relevant predictors is through the Akaike Information Criteria (AIC), which does not assume normality.

AIC

We also ran the AIC model on the training set for variable selection to see how the models differ from each other. The smaller the AIC score, the better the model.

AIC score: *-72159.25*
Selected variables: *age, number_of_reviews, instant_bookable, cleaning_fee, host_identity_verified, log_price, Wireless.Internet, Heating, Hangers, Kitchen, accommodates, Smoke.detector, New.York.City, Times.Square, latitude, private, Prospect.Park, apt, hostAge, TV, brooklyn, Broadway*

MSE: *0.0091*
R-squared: *0.2416*
Adjusted R-squared: *0.2405*

With the AIC model, we get a slightly lower adjusted R-squared value (0.2405) compared to the initial multiple linear regression model (0.2407).

Best Subset Regression

The final model we used for variable selection is Best Subset Regression. For our model, we fit up to a 15-variable model. Then, we built a model using the 15 selected co-variables.

MSE: *0.0091*
R-squared: *0.2407*
Adjusted R-squared: *0.2399*

In order to pick a model from the ones that we had built, we did a comparison MSE and Adjusted R-squared of all the models and chose the model with the best combination of both.

The initial multiple linear regression model has the best combination of MSE (0.0091) and Adjusted R-Squared (0.2407).

Figure 9. MSE, R-squared, Adjusted R-squared table

Model <chr>	MSE <dbl>	R_Squared <dbl>	Adjusted_R_Squared <dbl>
Initial Multiple Linear Regression Model	0.0091	0.2427	0.2407
Multiple Linear Regression Model 2	0.0091	0.2405	0.2398
StepWise Regression Using AIC	0.0091	0.2416	0.2405
Best Subset Regression Model	0.0091	0.2407	0.2399

To evaluate how the model performs on future data, we used *predict()* to get the predicted values from the test set. Cross validation was also performed to estimate 3-folds cross-validation prediction error for generalized linear models.

MSE of the final model: *0.0085*

Cross Validation of the final selected model (MSPE): *0.0089*

Comparing the MSE of the test data set which is equal to 0.0085 and the MSPE of the full data is 0.0089, we can see that the values are almost similar. Hence the variables that we have selected for the model are good estimators of the dependent variable.

Conclusions

In this project we aimed to gain insights into what factors customers value when searching for lodging in New York. Despite there being various factors influencing Airbnb occupancy, our initial multiple linear regression model demonstrated modest success, achieving an adjusted R-squared of 0.2407. This model outperformed our other approaches: second multiple linear regression, stepwise regression using AIC, and best subset regression, with adjusted R-squared values of 0.2398, 0.2405, and 0.2399 respectively. However, as the adjusted R-squared values are relatively similar, one of the later models could be used if a simpler model with fewer factors is wanted. These R-squared values are also relatively small, but we will go into how to improve this shortly.

We saw a few different areas of significant predictors at the 0.05 level, with some of them being very expected such as price and the number of people the listing accommodates. There were also some factors more related to the interaction with booking and listing than the stay itself. In this category we saw variables such as the cleaning fee, whether the host is verified, whether the property is instant bookable or not, the age of the property, and also the number of reviews on the listing. The last main category was more experience related with numerous amenities such as internet, heating, smoke detectors, hangers and kitchen availability. On top of this we saw some importance in the mentioning of locations such as Time Square and New York City altogether. These variables suggest that both the intrinsic qualities of the listings and the user experience as a whole significantly affect occupancy rates.

In order to improve our predictive power in the future, our first step would likely be to improve our calculation of the occupancy ratio. In this project we used the ratio of the number of reviews against the number of days since the property was first reviewed. This was certainly a valid option given the data available, but in the future it would be better to have the number of bookings versus the number of days available, since hosts do not need to list the property as available every day after posting.

We also believe it could be advantageous to break the listing down into certain price ranges such as budget listings, regular listings, and luxury listings. These groups likely target different customers with different preferences on what is provided within the stay. It would also be valuable to know the number of other listings also available within that price on the days that a listing is posted, we certainly imagine the supply of properties will factor into the demand and occupancy ratio. Continuing this project with these enhancements could yield better insights and more robust models, ultimately guiding hosts and platform strategies more effectively.

Works Cited `

Bae, J. B. (2019, July 18). *Home Buyer's Guide to Airbnb: NYC*. Medium. https://towardsdatascience.com/home-buyers-guide-to-airbnb-nyc-a240c43ff610

Borrison, R. (2015, Mar 7). *Here's Proof Airbnb Is Shaking Up the Global Hospitality Industry.* The Street. https://www.thestreet.com/technology/heres-proof-airbnb-is-shaking-up-the-global-hospitality-industry-13070686

Clampet, J., Skift. (2016, February 3rd). *Measuring Airbnb's Real Threat to U.S. Hotels Using Industry Metrics.* Skift. https://skift.com/2016/02/03/measuring-airbnbs-real-threat-to-u-s-hotels-using-industry-metrics/

Inside Airbnb. (2024, February 06). *INSIDE AIRBNB: NEW YORK CITY*. http://insideairbnb.com/new-york-city/

McGill University. (2018, JANUARY 31). *The high cost of short-term rentals in New York City*. Phys.org. https://phys.org/news/2018-01-high-short-term-rentals-york-city.html

Montevideo. (2024, March 17th). *What is the Airbnb effect on the global tourism industry?*. MercoPress. https://en.mercopress.com/2019/11/25/what-is-the-airbnb-effect-on-the-global-tourism-industry

PACASO. (2023, November 28). *Airbnb vs. hotel: 10 factors to consider*. Pacaso. https://www.pacaso.com/blog/airbnb-vs-hotel

Rana, S. R. (2024, February). *Airbnb price Dataset*. Kaggle. https://www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset