AIRBNB REPORT

Vedant Mahendra

This report will go through the process of examining the business team's hypothesis and providing supporting evidence and findings in favor or against the previously formulated hypothesis. It will also provide a brief overview of the procedure, identify other criteria that are relevant predictors of monthly reviews, and lastly assess three prediction models to choose the best one for the current data set.

The data set is called "clean listing" and it is a pandas DataFrame object. We then find columns in the data set that have more than 50% missing values and remove them. This aids in the removal of any columns with insufficient data to be helpful for analysis.

We then fill in any missing values in the "reviews per month" column, as well as other review-related fields, with zeros, allowing the data to be utilized in any computations without creating bias. After filling in the missing values, we remove any remaining rows with missing values from the data set. Many machine learning algorithms cannot handle missing values and will fail if they are present in the data.

Then we iterate through a list of column names, extracting the numeric component of each value in the defined columns. Following that, the values are transformed to floating-point numbers and given to the same columns. The data types of these columns are then shown on the screen. This step is essential since the original values in these columns may not be in a usable format for analysis, such as when they are stored as strings with additional non-numeric characters.

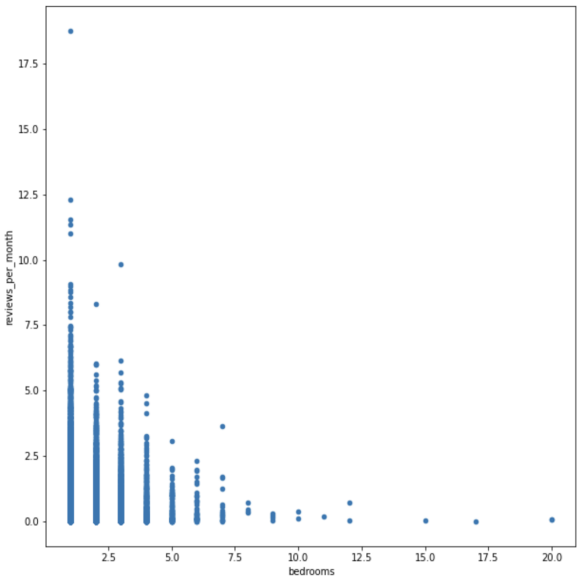
The values are then converted to floating-point numbers after we remove any dollar signs and commas from the "price" column. The "host acceptance rate" and "host response rate" columns are then converted to floating-point values by eliminating the percentage sign at the end of each value and dividing the resultant number by 100. Because these columns were previously saved as strings with a % indicator at the end, this is required.

Next, we build a new DataFrame named "df num" that solely includes the data set's numeric columns. It then provides a lambda function that replaces missing values with the column's mean value and applies it to the "df num" data set, replacing the original values. Furthermore, remove the "latitude" and "longitude" columns because they aren't necessary to compute the reviews per month because the listings are all from the same city, and convert the remaining columns to floating-point integers. This step is required because many machine learning algorithms require all input data to be numeric with no missing values. We guarantee that the data is in a usable format for future analysis by following these procedures.

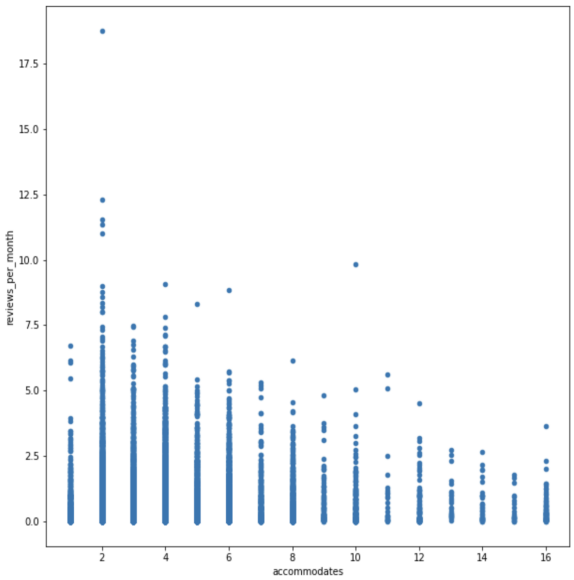
Testing Hypothesis 1:

Claim
The business team believes that larger properties should receive more reviews because larger properties can accommodate more guests and generate more traffic.

To validate the assertion that larger houses receive more reviews per month as a result of more traffic, linear regression was used to model the connection between the number of bedrooms, the number of persons accommodated, and the number of reviews per month. The model was trained using a set of input data, and the projected model's mean squared error (MSE) was determined using a loss function. The resulting MSE of 1.03 indicates that the model predicts the number of reviews every month based on the number of beds and accommodations reasonably well. This might imply that the data does not support the assumption that larger properties receive more reviews.



Reviews per month v/s bedrooms



Reviews per month v/s Guests accommodated

Further, to confirm our findings we conducted an analysis using two graphs. The first graph plotted the distribution of reviews per month against the number of bedrooms in a property, with the assumption that the number of bedrooms is correlated with the size of the property. The second graph plotted reviews per month against the number of guests a property accommodates, providing a measure of property size using these two parameters.

Upon examining the resulting graphs, we observed that properties with fewer bedrooms and fewer guests accommodated had more reviews per month. This outcome contradicts our initial hypothesis that larger properties would have more reviews per month, as results show that properties with fewer bedrooms and guests accommodated have more reviews per month.

Testing Hypothesis 2:
Claim: listings that are priced higher than listings of similar sizes and/or locations will receive fewer reviews than those that are priced lower.

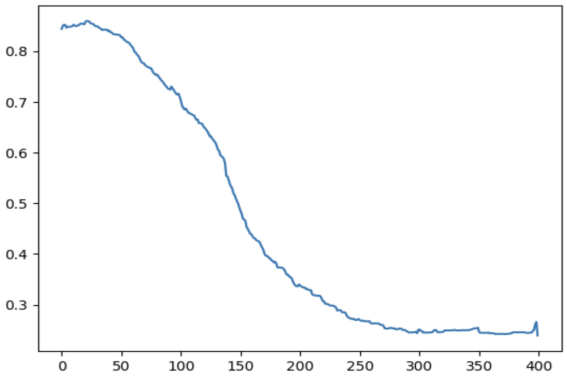
Part 1: Considering Location AND size (together)

In order to analyze the relationship between property pricing and the number of reviews per month, the properties were grouped based on location and size. The median price for properties of similar size in the same neighborhood was calculated and used as a benchmark to determine whether each individual property was priced appropriately. Properties that were within $50 of the mean price were considered to be priced appropriately, while those that were more than $50 above or below the median price were considered to be overpriced or underpriced, respectively. The mean number of reviews per month was then calculated for both underpriced and overpriced properties and compared. The results showed that properties that were priced lower than the mean price received a lower mean number of reviews per month compared to properties that were priced higher than the mean price, contradicting the initial hypothesis.

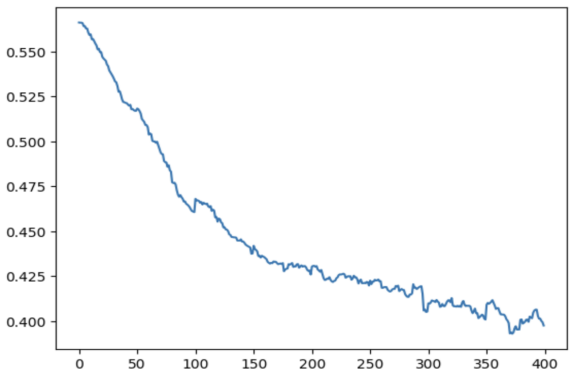
Buffer : $50
Mean Reviews for underpriced listings: 0.51
Mean Reviews for overpriced listings: 0.82

Graphs showing how variarion of buffer size impacts the reviews per month for underpriced and overpriced listings.

Mean reviews vs buffer for overpriced properties



Mean reviews vs buffer for underpriced properties



In the overpriced listing we can see that there is that there is a sudden drop in the reviews per month, similarly for underpriced listings we can see that there is steady decrease as the buffer increases.

After all the testing and evidence that we have identified, we can evidently conclude that the results are in contradiction to the initial hypothesis presented when location and size are taken into consideration, and price is not a function of apartment size or its location.

Part 2: Considering Location OR size (independently)

When we consider, the parameters individually in the sense we consider location or size, we come to a different conclusion.

If we consider just the size, we get a review per month we get overpriced reviews per month as 0.512 and underpriced reviews per month are 0.90. Further, if we take location into consideration, the analysis is inconclusive. Therfore it in conjunction with our initial hypothesis is confirmed.

Predictive Model Comparision:

Linear regression is a statistical approach for modeling the connection between one or more independent variables and a dependent variable. It presumes that the connection between the dependent and independent variables is linear and seeks the best-fitting line to characterize this relationship.

Linear regression is utilized in the above case to estimate the number of reviews every month based on other variables in the dataset. The linear regression model's R square value is 0.35, and the linear regression model's mean square error (MSE) is 0.74, indicating the average difference between the predicted and true values. A lower MSE value suggests that the model fits the data better.

Lasso regression is a sort of linear regression that uses regularisation to improve the performance of the model. Regularization is a method that adds a penalty term to the model's loss function, which helps to decrease overfitting and enhance model interpretability. Lasso regression employs the L1 regularisation term, which effectively performs feature selection by reducing the coefficients of less significant characteristics to zero.

A Lasso regression model is trained on the dataset to estimate the number of reviews every month in the given situation. The R square value of the model is 0.23, and the MSE is 0.84, which is greater than the R square value of the linear and random forest regression models.This suggests that the Lasso model may not be as effective at fitting the data and making accurate predictions.

Random forest regression is an ensemble learning approach that integrates the predictions of numerous decision tree models to generate a more accurate and stable prediction. It works by training numerous decision tree models on various subsets of data and then averaging the predictions of each tree.

A random forest regression model is trained on the dataset to estimate the number of reviews every month in the given situation. The Random Forest model has a R square value of 0.52 and an MSE of 0.61, which is lower than the linear regression model and has the greatest R square value of the three models examined. This shows that, among the three models evaluated, the

Random forest model is the greatest match for the data and may produce more accurate forecasts.

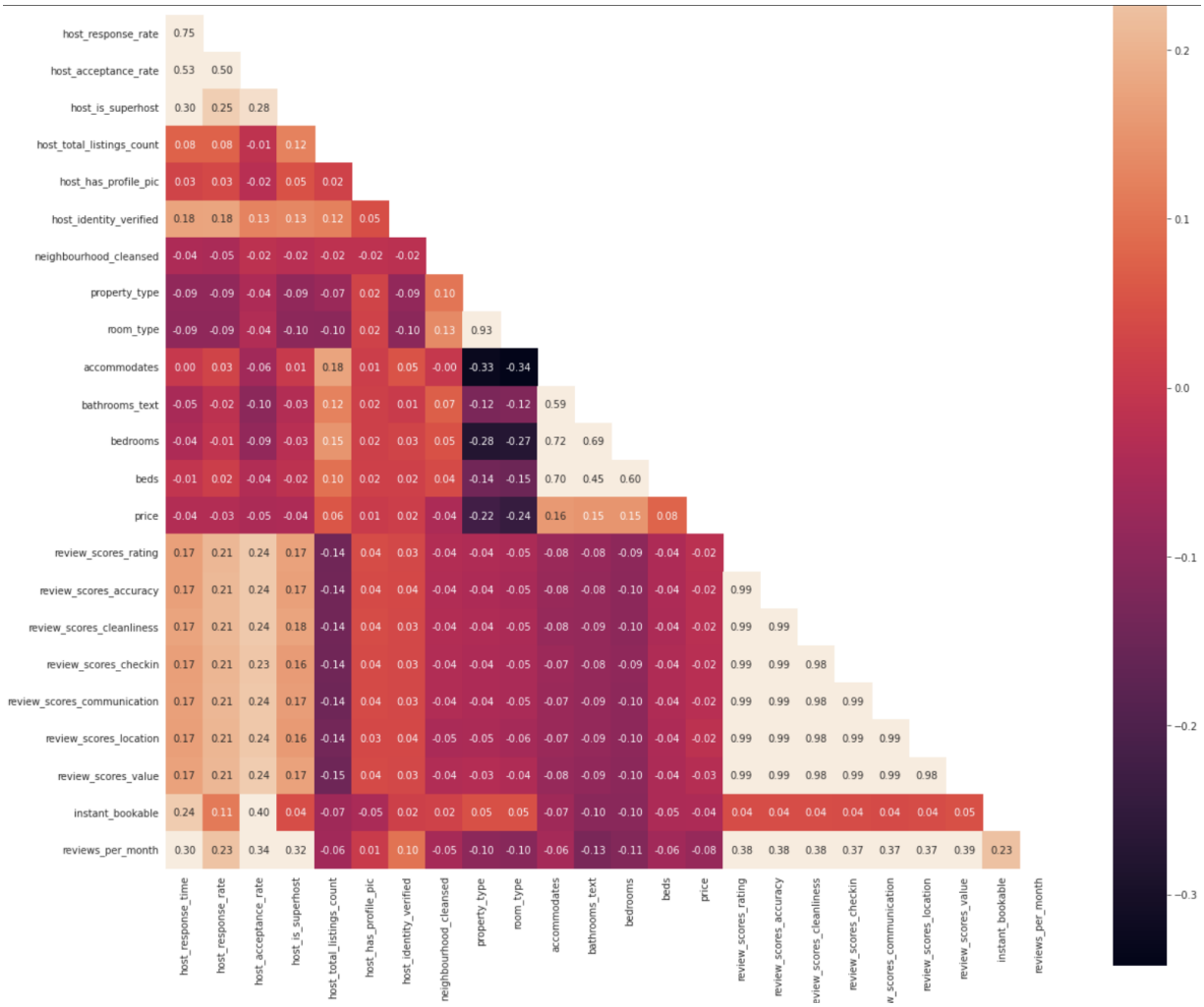
Ordinary least squares (OLS) regression is a kind of linear regression that calculates linear model coefficients by minimizing the sum of squared residuals. It is a subset of generalized least squares and is the most often used approach for fitting linear regression models.

In the current case, OLS regression is used to determine the additional factors in the dataset that are key predictors for the number of reviews every month. The OLS regression model findings can reveal which factors are most closely connected with the number of reviews, which might be valuable for additional research or modeling.

According to the findings of OLS Regression, the following factors are significant predictors of monthly reviews:

- `host_response_time[T.3]` 4.170686e-03
- `host_is_superhost[T.1]` 2.473206e-161
- `host_identity_verified[T.1]` 1.445693e-06
- `property_type[T.13]` 9.837988e-03
- `property_type[T.16]` 3.815671e-02
- `instant_bookable[T.1]` 5.857092e-59
- `bedrooms` 3.454224e-05
- `host_acceptance_rate` 2.574588e-35
- `host_total_listings_count` 6.995643e-11
- `accommodates` 6.588300e-06
- `price` 7.665981e-18
- `review_scores_checkin` 1.750821e-04
- `review_scores_communication` 2.681899e-03
- `review_scores_location` 2.336319e-04
- `review_scores_value` 4.439958e-33

Appendix:



Correlation matrix heat map