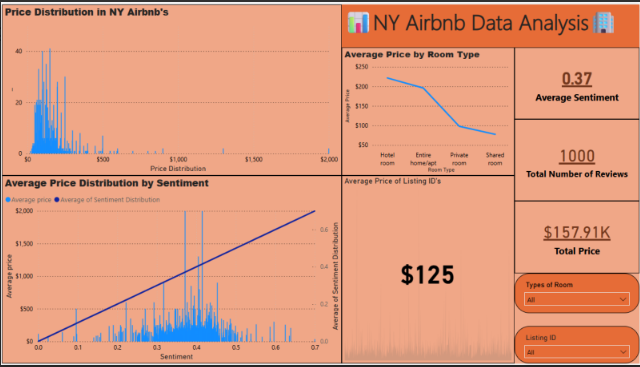
***Unlocking Airbnb Review Dynamics: A Data-Driven Approach***

******

**Executive Summary**

The report uses both insights from the regression analysis and the detailed dashboard to provide and display a thorough understanding of pricing dynamics, review sentiment and key factors that affect Airbnb listings in New York.

The regression model has identified multiple significant variables that influence and play a big role in determining the number of reviews per month in New York Airbnb listings. The key predictors are Availability\_365, review\_scores\_communication, review\_scores\_location and number\_of\_reviews.

Availability\_365 is a variable that measures the listing’s availability throughout the calendar year. It contains a range of 0-365 (inclusive). The availability factor has a negative coefficient meaning it has a negative relationship with reviews per month, this suggests that the higher the availability of a listing the number of reviews tend to be lower.

Review\_scores\_communication is a variable that measures the listing's communication. It contains a range from 7-10 (inclusive). The review score's communication factor has a negative coefficient. Review scores communication has a negative relationship with reviews per month. As the review scores communication increases the number of reviews is likely to decrease

Review\_scores\_location is a variable that measures the listing’s location. It contains a range from 7-10 (inclusive). The review score’s location factor has a negative coefficient. Review scores location has a negative relationship with reviews per month. As the review location increases the number of reviews per month is likely to decrease.

Number of reviews is a variable that measures the number of reviews a listing has. It contains a range from 0 - 590 (inclusive). The number of reviews has a positive coefficient. It has a positive relationship with the number of reviews per month. Ofcourse, as expected when the number of reviews per month increases the number of reviews per month is very likely to increase.

The model’s strength has strong independent power, with an R-Squared of 0.843, meaning that the model explains 84.3% of the variation in the number of reviews per month. The highly significant predictors and the high f statistic further support the strength of the model.

In the dashboard important objectives were investigated. The objectives were the Price distribution and roomtype, Price vs sentiment and the overall market overview.

The price distribution of Airbnb listings in New York are heavily skewed, with the majority of listings priced below $500. There were a few listings that were sitting outside the majority, with a few of them even sitting at $2000, which plays a role in raising the right tail of the distribution. Room type plays a crucial role in being able to determine price. Hotel rooms have the highest average price, followed by entire homes/apt, private rooms, and shared rooms, with shared rooms being the most affordable.

There is a positive correlation between sentiment and price, properties that receive a higher sentiment rating tend to be priced higher. This suggests that guest satisfaction can be reflected and linked by sentiment analysis, clearly showing a significant impact in pricing.

The dataset holds 1000 reviews, reflecting a decently active market with a total listing value of $157,910. The average price per listing is $125 meaning most listings are in the mid range pricing tier. The average sentiment score of 0.37 implies that the overall sentiment is leaning in between neutral and slightly positive (-1 to 1).

The combination of regression analysis and the dashboard insights provide a comprehensive view of the factors that influence both the number of reviews and the pricing of Airbnb listings in New York. By focusing on increasing guest satisfaction, particularly in areas like communication and location, hosts can increase both the frequency of reviews and justify higher prices, ultimately boosting their revenue potential in a competitive market.

**Introduction:**

This report displays a detailed analysis of the key factors that influence and impact New York’s Airbnb listing performance. Using regression analysis, we identify which variables most significantly impact the number of reviews a listing receives each month. These insights help exhibit patterns that may not be immediately apparent.

Similarly, a dashboard was designed to explore New York’s Airbnb listings market trends including price distribution, relationship between sentiment and price, and the different types of rooms in determining listing value. These visual insights lay out a high level view of the market and allow hosts to understand their performance and uncover opportunities for growth.

The combination of these strategies support the report by equipping stakeholders with the knowledge needed to make informed decisions.

**Data Collection & Preparation:**

The development of this report began from 2 datasets. New York.csv which contained the metadata in the form of different variables about Airbnb listings in New York. The second dataset was Aggregate sentiment analysis NY reviews.csv which contained the average sentiment expressed by guests for each of the Airbnb listings. The sentiment values lie between -1 to 1 where -1 is strong negative and 1 is strong positive.

The New York Airbnb dataset had contained different variables ranging from room\_type, host response/acceptance, room details, requirements and cancellation policy. In the variables there were many null values. There was an option to just remove null values completely as there were only around 3-4% (30-40/1000) of the variables were null but because the dataset doesn’t have that much volume I replaced them instead. Through the use of Power BI power query software I was able to remove and replace all null values with the average. The variables that contained null values and were replaced were:

* Bathroom: nullValues → 1
* Bedroom: nullValues → 1
* Beds: nullValues → 2
* Review rating: nullValues → 94.3
* Review Accuracy: nullValues → 9.6
* Review Cleanliness: nullValues → 9.2
* Review Checking: nullValues → 9.8
* Review Communication: nullValues → 9.8
* Review Location: nullValues → 9.5
* Review Value: nullValues → 9.3
* Reviews Per Month: nullValues → 0.88

After multiple validation checks all the variable datatypes were correct, the variables were in their correct data types, all null values were replaced, the expected range of the variables were correct, and all variables were consistent.

A merged\_df was developed by merging both datasets into 1 through their common variable which was listing ID. The final merged\_df consisted of everything from the fixed and correct New York.csv and the sentiment variable in Aggregate sentiment analysis NY reviews.csv. In the final merged\_df the sentiment variable had consisted of a few nulls, the same decision was made and all null values were replaced with the average.

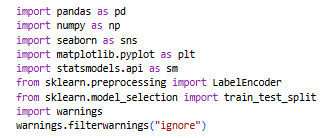
* Sentiment: nullValues → 0.37

**Analytical Methodology:**

When developing a machine learning model it is very important to select your independent and dependent variables. The dependent variable in merged\_df was the number of reviews per month. Selecting the correct algorithm is extremely important, due to the number of reviews being a continuous variable it was very clear to select a Supervised Regression Algorithm.

The goal of a regression model is establish the existence of association between two variables, but not causation. Extremely useful in determining key factors that have significant on their Key Performance Indicators

The first step was importing and loading all the relevant libraries.



*Figure 1: Installing Libraries.*

After connecting my google drive account and linking the merged\_df into the Jupyter Notebook, the next objective is to determine which variables are most suitable in predicting the number of reviews per month. There was no need to do another data validation check or data preparation check.

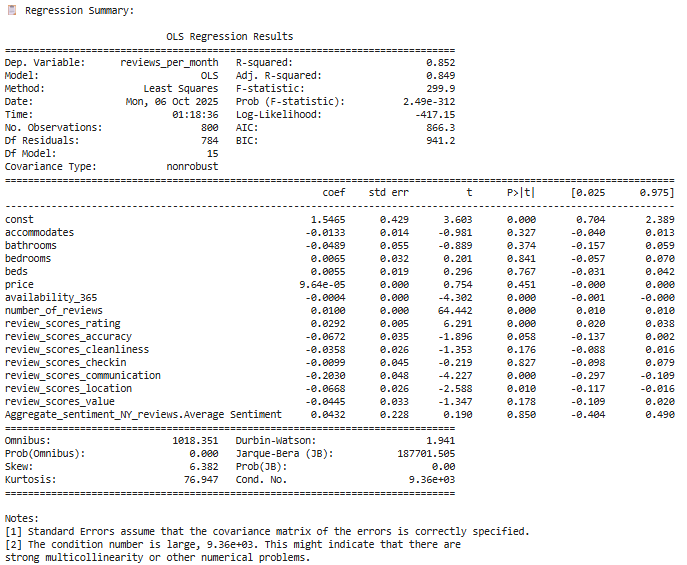
Before running a check through all variables and determining which are most valuable. The dataset was split into 2 factors. The factors are Listing Details & Review Details. A thorough review was done and the variables selected were:

**Listing Details:**

* Accommodates
* Bathrooms
* Bedrooms
* Beds
* Price
* Availability 365

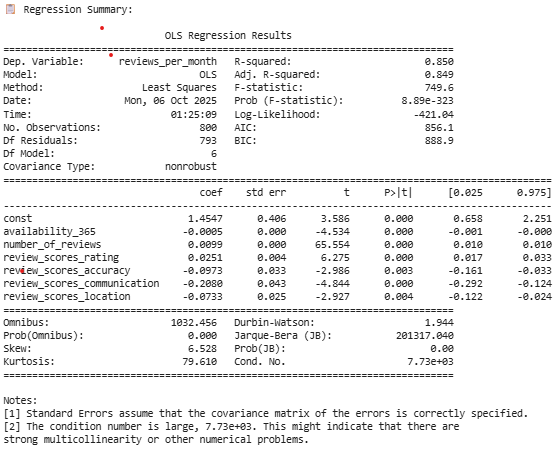
**Review Details:**

* Number of Reviews
* Review Scores Rating
* Review Scores Accuracy
* Review Scores Cleanliness
* Review Scores CheckIn
* Review Scores Communication
* Review Scores Location
* Review Scores Value
* Aggregate Sentiment



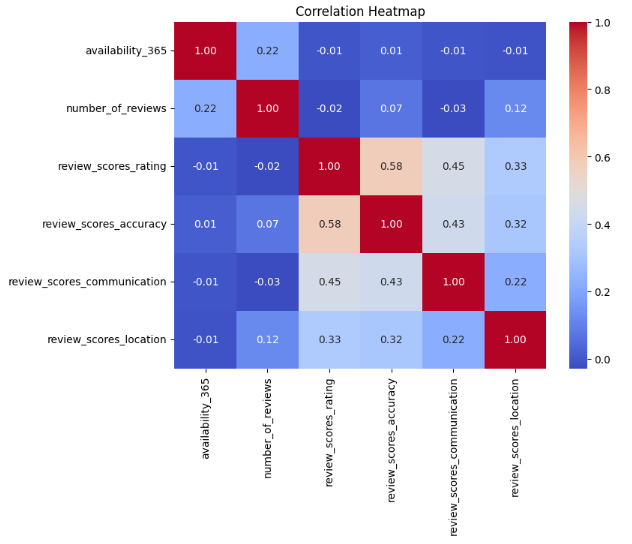
*Figure 2: Model 1 Summary*

As shown in the figure above the R-Squared is acceptable but there are many variables that are insignificant (p\_value > 0.05). A simplified process was taken by removing those variables and keeping the significant variables.

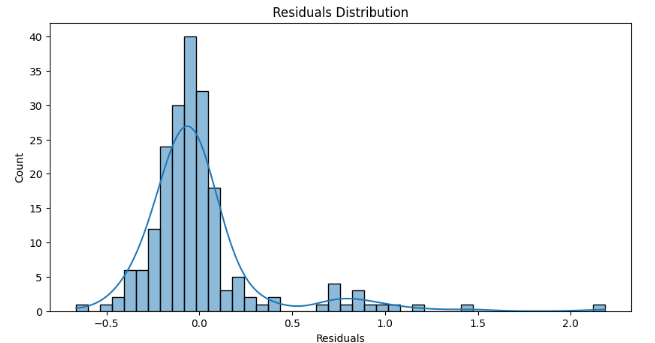


*Figure 3: Model 2 Summary (Simplified Model 1)*

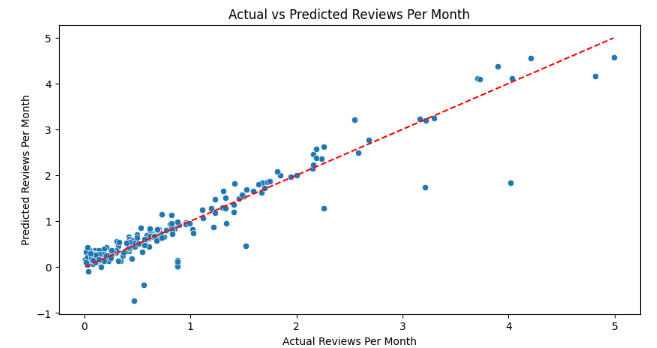
There was a slight decrease at the R-Squared value but regardless it is still an acceptable value. In order to ensure the model was robust is by running a multicollinearity check by checking the correlation between the selected variables, checking the residuals distribution and comparing the Actual vs Predicted Reviews Per Month.



*Figure 4: Model Variables Heatmap*



*Figure 5: Model Residuals Distribution*

**

*Figure 6: Model Actual Vs. Predicted Reviews Per Month.*

**Interpretation of Results**

When interpreting the results there are 4 topics to discuss.These are the heatmap, Regression Summary, Residuals Distribution and Actual Vs Predicted Reviews Per Month.

The Heatmap is a clear display of the variables selected and their correlation between each other. The higher the correlation the higher the odds of multicollinearity. All the variables don’t have an outstanding correlation among each other except for the 2 variables Review Scores Accuracy and Review Scores Rating, sitting at 0.58. I’ve decided to not to remove either one because there wouldn’t be a huge change in model difference.

The regression summary displays an R-Squared value of 0.850. This means that the model explains for 85% of the variance in Reviews Per Month. This indicates a very strong fit. The F-statistic is sitting at 749.6. THe overall model is statistically significant, meaning that at least one predictor variable significantly affects Reviews Per Month. All variables are significant predictors, they all have a p value less than 0.05. Review Scores/ Communication/ Location and availability 365 have negative coefficients. When they increase the Reviews Per Month is likely to decrease. The number of reviews and review scores rating both have a positive coefficient, as they increase the Reviews Per Month is likely to increase.

The residual distribution is clearly skewed to the right. The ideal residual normally distributed around zero, the skewness suggests that some non-normality in errors. However, since the model fit is strong due to the R-Squared, it still has the ability to perform well for prediction.

The Actual vs Predicted plot displays the actuals reviews per month and the predicted reviews per month. The red dashed line is closely aligned with the data points. Overall this demonstrates that predicted values track with the actual data, reinforcing that the model generalizes strongly.

Overall the model explains 85% of the variation in reviews per month, indicating a strong relationship between listing details and review details. Key insights are that if listings have a higher number of Reviews they’re going to tend to have a higher number of Reviews Per Month. Review Scores Rating and Accuracy scores are positively associated with Number of Reviews, highlighting the importance of maintaining high quality listings. However some listings consistently outperform model expectations due to the skewness in the residual distribution.

**Recommendations & Actionability**

Overall, the model is good but there is still room for improvements. Strong Predictors of review frequency per month are Review Scores Accuracy and Review Scores Rating. A focus is definitely maintaining the accuracy and rating scores as well as looking at previous guest feedback and improving on certain aspects. Another way of improving and taking action is to promote Airbnb’s with high reviews as this is a key indicator in receiving a higher Number of Reviews Per Month.

The availability of listings should be managed and controlled seasonally/periodically rather than being open/available 365 days of the year. The limited availability will increase the value and develop scarcity, this could eventually lead to a higher demand and Reviews Per Month.

This analysis shows that listing quality, availability management and social proof are key drivers of monthly reviews. Maintaining high accuracy and rating cores can significantly boost engagement and attract more listing bookings, encouraging reviews for new listings which helps build momentum, while strategically managing availability can enhance demand. Overall, focusing on these factors will improve visibility, increase frequency and increase Reviews Per Month.