**FINAL PROJECT**

**NEW YORK CITY AIRBNB OPEN DATA**

**BHARGAV DASARI**

1. **Project Description**

Since its launch in 2008, Airbnb has allowed both hosts and visitors to extend their travel options and offer more distinctive, customized ways to see the world. This dataset provides metrics and listing activity information for New York City in 2019.

Airbnb has been utilized by hosts and guests to increase travel options and offer a more distinctive, customized method of seeing the world, still maintaining a happy guest experience.

For this reason, the initiative examines stats and listings on Airbnb in New York City, New York, USA. The "New York City Airbnb Open Data" collection that was chosen contains, among other things, 2019 data on neighborhoods, locations, types of rooms, and reviews.

This project's goal is to examine and learn from the Airbnb dataset for New York City in 2019. We'll look at a variety of topics related to Airbnb listings in New York City, such as price patterns, well-liked locations, and different kinds of rooms, host traits, and elements that affect the minimum number of nights needed for a listing.

1. **Dataset** : **New York City Airbnb data in 2019**

**Data Source:** New York City Airbnb Open Data (kaggle.com)

**Data variables:** name, host\_id, neighbourhood\_group, neighbourhood, latitude, longitude, room\_type, price, minimum\_nights, number\_of\_reviews, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365

Types of variables with respect to information content:

* **Demographic data:** host\_id , neighbourhood\_group, neighbourhood , room\_type
* **Economic data:** price, calculated\_host\_listings\_
* **Behavioral data:** minimum\_nights , number\_of\_reviews , reviews\_per\_month
* **Geographic data:** latitude and longitude, neighbourhood\_group , neighbourhood
* **Availability data:** availability\_365

**Response Variables:** price, availability\_365, number\_of\_reviews or reviews\_per\_month minimum\_nights, calculated\_host\_listings\_count

**Data Inspection:** After inspecting and analysing data I didn’t found any duplicate values present in dataset. I found few null values present in dataset and removed them using R Studio

Number of records: 582638(After removing null values)

1. Exploratory analysis and research question(s)

This study addresses the following research questions:

* How does the number of listings a host manage to impact the price variation across different room types?
* What are the factors influencing the frequency of reviews for Airbnb listings?

4. Data analysis

4.1. Methods and software used.

The exploratory analysis was conducted using R programming language and the tidy verse package. Linear regression and clustering techniques were employed to explore relationships and identify patterns.

4.2. Results

Exploratory Analysis:

I have used different methods like linear regression, poission regression and build a scatter plot to answer the research question “How does the number of listings a host manage to impact the price variation across different room types?”

Scatter Plot:

I used gg plot function to generate a scatter plot which shows the Price Variation by Number of Listings and Room Type

A graph of numbers and a number of listing

Description automatically generated

The key take aways we got from the above scatter plot are:

Entire Homes/Apartments: These listings are more expensive on average, with a wide price range, indicating that hosts with more listings tend to charge higher prices.

Private Rooms: Generally clustered at lower numbers of listings with most prices below $5000, suggesting that private rooms are more affordable and commonly managed by hosts with fewer listings.

Shared Rooms: Represented by the fewest data points and the lowest prices, which implies that shared rooms are the least expensive option and less frequently managed by hosts with multiple listings.

Linear Regression:

Linear regression revealed a significant relationship between the number of listings managed by a host and price variation across different room types. And the following are the insights we got from linear regression model

1. Coefficients:

* Intercept: Think of the intercept as the base price for an Airbnb listing when there are no specific factors considered. It's like the starting point for pricing, which in this case is around $194.65. Imagine this as the base cost for renting out a basic, no-frills accommodation.
* calculated\_host\_listings\_count: For each additional listing managed by a host, the price of their listings tends to increase by approximately $0.23. This reflects how hosts with more properties often incrementally adjust their prices to account for their growing portfolio.
* room\_typePrivate room: Private rooms, compared to entire homes/apartments, are estimated to be around $111.74 cheaper on average. This suggests that guests typically expect to pay less for the privacy and amenities provided in a private room compared to an entire space.
* room\_typeShared room: Shared rooms have an even lower estimated price, around $122.18 less than entire homes/apartments. This indicates that guests are willing to sacrifice privacy for a more budget-friendly option, like a shared room in a hostel or communal living space.

2. Statistical Significance:

* The symbols (\*\*, \*\*\*, etc.) next to each coefficient indicate their statistical significance. Essentially, they tell us how confident we can be that the relationship we observe is real and not just due to chance.
* In our case, all coefficients except the interaction terms are statistically significant, meaning there's strong evidence that they indeed impact listing prices.

3. Model Fit:

* The R-squared value (0.08386) tells us how well our model explains the variation in listing prices. Here, it suggests that our predictors (number of listings managed and room type) explain about 8.39% of the variability in prices.
* The F-statistic tests whether our model as a whole is useful in explaining price variability. With a significant p-value, our model seems to be effective in this regard.

In essence, this model gives us valuable insights into how hosts' management practices and the type of accommodation they offer influence Airbnb listing prices.

Poission Regression:

This model helps us understand how the number of reviews received by Airbnb listings is affected by certain factors, like the number of listings a host manages and the type of room they offer.

1.Coefficients:

* Intercept: Imagine this as the baseline number of reviews a listing might get when a host doesn't manage any other listings and offers a standard room type.
* calculated\_host\_listings\_count: For each additional listing a host manages, we expect a slight decrease in the average number of reviews.
* room\_typePrivate room: Listings offering private rooms tend to receive slightly more reviews compared to those offering entire homes/apartments.
* room\_typeShared room: Conversely, listings offering shared rooms typically receive fewer reviews compared to those offering entire homes/apartments.

2.Significance:

* + We look at the p-values to see if these relationships are statistically significant. The "\*\*\*" symbols indicate very strong evidence that the relationships are real and not just due to random chance.

3.Model Fit:

* + The null deviance and residual deviance are like measures of how well the model fits the data. Lower values are better, and the difference between them tells us how much better our model is compared to a simpler one with just an intercept.
  + The AIC helps us compare this model to others. Lower AIC values suggest a better fit.
  + The Fisher Scoring iterations are just part of how the computer figures out the best values for the coefficients.

In simple terms, this model tells us that the number of reviews a listing gets depends on how many other listings the host manages and the type of room they offer

So in conclusion for this exploratory analysis we can say that Both linear regression and Poisson regression models provide valuable insights into the relationship between predictor variables (e.g., number of listings, room type) and response variables (listing prices, number of reviews). Coefficients and their significance levels help understand the magnitude and direction of these relationships.

And the scatter plot generated give a pictorial representation of price when compared to Number of Listings and Room Type

These findings can inform pricing strategies for hosts and help guests make informed decisions when booking accommodations on Airbnb.

Predictive Analysis:

I have used different methods like linear regression build a random forest to answer the research question

“What are the factors influencing the frequency of reviews for Airbnb listings?:

Linear Regression:

The regression model aims to explain the relationship between the number of reviews received by Airbnb listings and several predictor variables: `location\_encoded`, `room\_type\_encoded`, and `price`. Let's break down what this model is telling us:

1. Model Purpose:

* + The goal of this model is to understand how the number of reviews received by Airbnb listings is influenced by their location, room type, and price.
  + It helps us identify which factors contribute to more reviews for a listing and by how much.

2. Coefficients:

* + Intercept: The intercept represents the expected number of reviews when all predictor variables are zero. Here, it's approximately 33.46. This means that, on average, a listing can expect around 33.46 reviews even without considering other factors.
  + Location Encoded: These coefficients represent the effect of each location (Brooklyn, Manhattan, Queens, Staten Island) on the number of reviews, compared to a reference location (presumably omitted from the output). For example:
  + Brooklyn and Manhattan have negative coefficients, indicating that, on average, listings in these locations tend to receive fewer reviews compared to the reference location.
  + Queens and Staten Island have positive coefficients, suggesting that listings in these locations receive more reviews on average.
  + Room Type Encoded: These coefficients represent the effect of each room type (Private room, Shared room) on the number of reviews, compared to a reference room type (presumably Entire home/apt). For example:
  + Private rooms have a positive coefficient, indicating that, on average, they receive more reviews compared to entire homes/apartments.
  + Shared rooms have a negative coefficient, suggesting that, on average, they receive fewer reviews compared to entire homes/apartments.
  + Price: The coefficient for price is negative (-0.007853), indicating that, on average, as the price increases, the number of reviews tends to decrease slightly.

3. Significance:

* + The p-values associated with each coefficient indicate whether they are statistically significant.
  + Coefficients with p-values less than 0.05 are typically considered statistically significant. In this case, all coefficients except for `location\_encodedBrooklyn`, `room\_type\_encodedPrivate room`, and `location\_encodedQueens` have p-values less than 0.05, suggesting they are statistically significant.

4. Model Fit:

* + The R-squared value (0.003292) is quite low, indicating that the model explains only a small proportion of the variability in the number of reviews.
  + The Adjusted R-squared value (0.003067) accounts for the number of predictors in the model.
  + The F-statistic tests whether the overall model is useful in explaining the variability in the number of reviews.

In summary, this regression model helps us understand how location, room type, and price influence the number of reviews received by Airbnb listings.

Random Forest:

The random forest model aims to predict the number of reviews received by Airbnb listings based on their location, room type, and price. Here's what the output tells us:

1. Model Summary:

* Type of Random Forest: This model is a regression random forest, indicating that it's used for predicting continuous outcomes.
* Number of Trees: The model consists of 500 decision trees.
* Number of Variables Tried at Each Split: At each split in the decision trees, only one predictor variable is considered.

2. Mean of Squared Residuals: This represents the mean squared error (MSE) of the model's predictions on the training data. A lower MSE indicates better model performance in terms of how close the predicted values are to the actual values.

3. Variance Explained: This indicates the proportion of variance in the number of reviews that is explained by the model. In this case, the model explains approximately 1.67% of the variability in the number of reviews.

4. Predictions and Model Performance:

* The model makes predictions on the test dataset, and then the root mean squared error (RMSE) and mean absolute error (MAE) are calculated to evaluate its performance.
* The RMSE (46.50) and MAE (30.33) provide measures of how well the model's predictions match the actual values. Lower values indicate better performance.

5. Feature Importance:

- The `varImpPlot` function generates a plot showing the importance of each predictor variable in the random forest model.Here is the plot generatedA graph with numbers and a dot

Description automatically generated

- The importance values indicate how much each predictor contributes to the overall prediction accuracy of the model. Higher values suggest more influential predictors.

So, the random forest model attempts to predict the number of reviews for Airbnb listings based on their location, room type, and price. While the model explains a small portion of the variability in review frequency, its performance could potentially be improved with further feature engineering or parameter tuning.

In summary, both models contribute to understanding and predicting the frequency of reviews for Airbnb listings based on their location, price, and room type. While the regression model offers insights into the directional relationships between predictors and review counts, the random forest model provides predictive capability. Together, they underscore the importance of location, room type, and price in shaping review frequency, albeit with modest explanatory power. Further refinement and feature engineering may enhance the models' predictive accuracy and explanatory ability.

5.Conclusions / Challenges / Further Analysis:

The regression model sheds light on how location, room type, and price influence Airbnb review counts, despite explaining a small portion of variability. Meanwhile, the random forest model offers predictive capability, albeit with modest accuracy. Challenges include low explained variability and the complexity of factors affecting review frequency. Further analysis should focus on feature engineering, model refinement, temporal analysis, qualitative research, and benchmarking to enhance predictive accuracy and understand the nuanced dynamics of review behavior.

# References

DGOMONOV. (2019). *Kaggle*. Retrieved from New York City Airbnb Open Data: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?resource=download

OpenAI. ( 2024). ChatGPT. San Franciso.