**Introduction**

Develop a model to predict Airbnb price in NYC with a dataset of 49,076 listings. The dataset with 16 potential predictor variables is provided to help build a linear regression model. Additionally, 30% of the data will be used as a test set to evaluate the model accuracy. The model incorporates supervised machine learning techniques such as classifications and unsupervised machine learning such as clustering to predict the price.

**Conceptual Model**

The objective of this model is to use relevant and statistically significant variables to predict the Airbnb price. The information regarding these variables of the listing will be gathered through an onboarding process of Airbnb as the host enters the information related to their property and their preference; the inputs feed into the prediction model which gives the host a recommended price for the listing.

To start off, the model needs to be trained with cleaned data where there is no empty. The listing with $0 in price is replaced with the mean of all listings since they are likely unlisted by the hosts. Variables including all the id and host name are not predictive variables, therefore, taken out from the model.

Additionally, a few attributes are highly skewed and can skew the prediction since it is a linear regression model. Therefore, attributes such as *last review* are categorized by the review recency, *calculated host listings count* and *minimum nights* are categorized based on volume. The categorical including *room type, neighbourhood* and *neighbourhood group* are converted into dummy variables. Note that a large number of variables are created in this process, however, not all variables will be statistically significant in predicting the price and might even lower the model accuracy. After adjusting the categorical data, the first linear model is created to identify variables with very little significance. Only the variables that have a p-value of lower than 0.05 were reserved to rebuild the linear regression model.

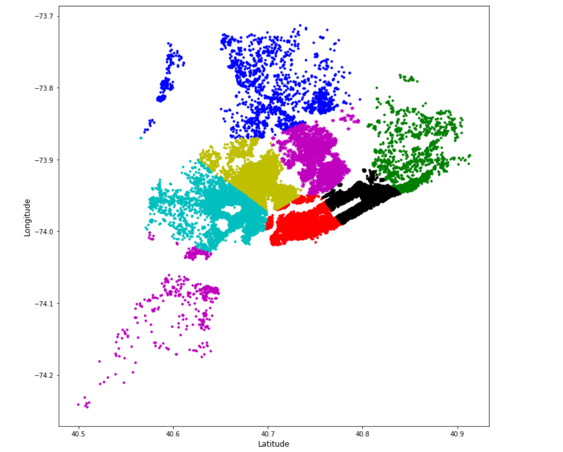
With the *longitude* and *latitude* data, the model aims to incorporate clustering using k mean to identify the geo-cluster of the Airbnb locations.

With the name column, sentiment analysis is conducted on the to analyze the valence score of the listing title using VADER in Python. The valence score of -1 indicates extreme negative tone in the listing title and +1 indicates extreme positive.

The model uses the linear regression function from scikit-learn to build a prediction model based on 70% of the training set data. The remaining 30% of the data is used to test the accuracy of the prediction from the trained model. To validate the accuracy of this Linear Regression model, there are a few statistical metrics used: adjusted R^2, mean absolute error, mean squared error, root mean squared error.

**Result Analysis**

Before incorporating cluster analysis, the model achieves an adjusted R^2 of 0.369, mean absolute error of 73.869, mean squared error of 54298.71, and root mean squared error of 233.02.

When incorporating clustering, the model chooses a cluster of 8 (k=8) from the elbow method to cluster the listing based on the latitude and longitude coordinates. However, the model does not see a significant improvement in the model accuracy. Therefore, the number of clusters and the incorporation of k mean to a linear model need to be optimized if given more time.

Interestingly upon reviewing all the significant variables, variables such as the number of reviews and review recency have a negative coefficient. This implies that recent and large amount of reviews do not necessarily mean a higher competitive price, possibly due to the age of the property and affects the attractiveness of the listing to travelers. Room type of entire home/apt seems to have a large positive coefficient, which makes sense to charge extra when the host has a spacious place. Small minimum A close up of a newspaper

Description automatically generatednights (1 night) has a larger coefficient, indicating that prices can go up more as less the restriction in the booking. In terms of neighborhood locations, prime locations such as Battery Park City and group such as Manhattan have a greater coefficient than the others; this observation is intuitive since location usually plays an important role in Airbnb booking decisions.

**Recommendation to Airbnb**

It is recommended that Airbnb launches a pricing recommendation project which assists new host onboarding by suggesting a competitive pricing based on the linear regression model. The model should include the variables exhibiting on the left to suggest a price for new hosts. This pricing recommendation project can give the new hosts guidance on how to price their listings and reduce their effort and time in research on what a competitive price should look like. To maintain a platform business model, Airbnb needs to attract high volume on both host and traveller sides of the platform. Making the onboarding user experience seamless adds values to the hosts and attracts more people to list their properties on Airbnb. Ultimately, the model enables an increased revenue for Airbnb since the company takes a percentage from every booking. To implement the model, a web-based questionnaire should be developed carefully with questions containing the variables required for this linear regression model.

**Recommendation to Airbnb Hosts**

It is recommended that Airbnb host analyze their current listing and alter the controllable factors in order to achieve the price that they want. Although a host cannot change the location of their listings, factors such as minimum nights can be changed to encourage a higher price. A host can initially encourage some review, however, the listings with too many reviews need to be managed since too many could repulse the travellers away.

**Summary & Improvement**

In conclusion, a linear regression model is developed through statistical analysis using both Python. It incorporates relevant predictor variables of Airbnb price prediction using categorization and clustering and text analysis, ultimately using to guide the new hosts’ onboarding and current host’s pricing strategy.

The model can be improved if given more time. For example, text mining can be helpful in determining which keywords in the name help drive up the price and those keywords can be suggested to the new hosts. Lastly, more unsupervised machine learning techniques such as random forest and KNN can be used in analyzing the geodata because the location is probably a crucial factor for the listing price.