Airbnb price analysis and prediction in Seattle project proposal

Yueyuan He, Chenghao Wu, Tiehao Chen, and Jian Jian
Group 6

Project Overview:
This project is to identify the dominant factors that affect the price of the houses of Airbnb and provide market strategies to increase the revenue for the Airbnb's stakeholder. Airbnb has become a worldwide application that people use broadly during personal or business travels for its convenience and cost-effectiveness. For this project, we have done the research, explored some public datasets on websites and decided to focus on the dataset in Seattle eventually. By analyzing the strength of the correlation between the price and various dependent variables, the insights we gained from the dataset would help us to answer the major business question: How can we help Airbnb hosts to adjust their pricing strategy to obtain the optimal gross margin.

Dataset Description:
The Airbnb dataset contains 106 columns and 9024 rows, which is available at **http://insideairbnb.com/seattle/**. Since we have a comparatively large number of attributes, we performed data cleaning and wrangling in the first place. We categorized the columns by their different property(eg. Host info, Guest info, time, geography), and selected the columns related to the diverse kind of business questions. So the features can be manipulated in different ways regarding various business questions respectively.

There is a large number of features in the dataset, so the below is only a sample list:

FeatureType	Host Info	Reviews	Geo	Time	House	Useless
Feature Name	Host_since	reviews_per_month	latitude	availability_30	bedrooms	host_acceptance_rate
	Host_verifications	number_of_reviews	longitude	availability_60	bathrooms	instant_bookable

Analytic Approach:
Based on business understanding, we selected the features that have a strong correlation to the housing price. We used Google Colab as the cloud computing solution for team members collaboration. By streaming the analyzing process, the filtered dataset will be fed into the data pipeline. For a particular model, we will reduce the dimensions of the dataset, impute the missing value and outliers, scale and visualize the distribution of each feature to check the effectiveness of the data cleaning process. Using proper machine learning models, we can predict the housing price on the test dataset. Depending on the cases when analyzing, we will compare the model performance among Classification, Regression, Deep Learning, Clustering, Association Rules along with other models to generate the eventual conclusion.