Proposal for Airbnb dataset

Introduction and Problem statement

Airbnb has become a major player in the lodging industry taking 20% of US lodging market share in just 11 years since the company began in 2008. This phenomenon is owed not only to their range of economic lodging options but also the ease of accessing lodging information from all over the world in one single platform. For many tourists, Airbnb also provides an immersive experience into the culture and people of the vacation destination by staying at a local venue instead of the similar hotel/motel experience. For these reasons, Airbnb have experienced a meteoric rise in popularity among youth travellers. On the other hand, hosts also enjoy the passive income that comes with renting out their home/space and potentially befriending world travellers that decides to stay at their venue. However, putting up their homes on airbnb can be quite a bit of tasks and can be difficult to predict how their home compete with neighboring airbnb homes in attracting customers into selecting their homes. Therefore, in this project, I will examine the number of reviews and how it correlates with description of the home, the geographical location of the home and other home characteristics. Number of reviews was selected because the airbnb dataset does not have numbers of previous occupants as an obtainable feature and numbers of reviews, in the general sense, typically correlates well with number of consumers that previously used the service.

Audiences

Two main audiences would have an interest in the project: the hosts and Airbnb. Hosts can directly benefit from the study to condition their homes to high number of reviews. Based on the results coming out of the algorithm, hosts can improve their homes or language used when posting his home information or even how their home would predictably do in the market given some fixed conditions like bedroom numbers and geographical location. With the information the host can adjust operational cost for maintaining the rooms and area as well as investment in amenities for a long term profit like buying a television for the room.

Airbnb also would be interested in the project because it effectively tells a story on how the market is behaving for consumer interested in Airbnb services. Airbnb can use this and customer's data to hypothesize a wholesome picture of the demand and supply of certain types of homes and use this information to motivate hosts to have certain characteristics to improve their services and drive business growth.

Questions to analyze

What are the most common words or phrases used by host home? Do some of these words or phrases have a correlation to the airbnb with high review rates? What are common amenities in these homes? How does location affect review rates? Do cost play a commanding role in review rates? How do ratings affect review rates? Can we predict review rates reliability with given features?

Data:

The dataset was obtained at http://insideairbnb.com/get-the-data.html. The detailed Los Angeles listing and calendar dataset were used. In the listing dataset contains the general information about the host home like number of bedrooms, description of the home, price, etc. This dataset has over 100 features and 45000 rows. On the other hand, calendar shows all the data in listing and the availability and price in time series format.

Approach to solve the problem

The data will be wrangled by pandas and nltk for the natural language processing part of the dataset. The data will be visually analyzed by matplotlib and seaborn. The statistical analysis will be conducted in combination of numpy and scipy. Finally the machine learning will be conducted with scikit learn and keras.