Final Project Data Memo

Noe Arambula

April 9, 2022

Contents

Data-set Overview

I intend to use a data set from insideairbnb.com which sources the data from publicly available information on the official Airbnb site. The data-set includes information about the hosts of the living space being offered, the living space/listing itself, and reviews. The data set I will use will be strictly from the Los Angeles area. There are 33,330 observations and 74 predictors. The dataset has both numeric and character/string data variables as well categorical and continuous variables. Most of the data is continuous and I will be able to change the data I need into continuous variables. For example there is a variable that lists the amenities a particular listing has and for my purposes I plan on converting that into a number, so instead of listing amenities I will convert it into the total number of amenities per listing. There is some missing data although only for a about 3-5 variables/predictors from what I can see and I plan to just leave those columns out as I have a lot of data to work with and I am not sure how useful they would even be for the variable I am trying to predict.

Research Questions Overview

I am interested in predicting the price of a listing and the review rating. I am interested in answering the question of how different aspects affect a listings price, what has the most influence on a listings price and how does that correlate with a listings review score. Is there any correlation between a listings review score and other aspects of an airbnb. So, my response variable will be price which is the price of booking an air bnb listing for a night and will depend on things such ass the number of amenities, bathrooms, guests allowed, review score, etc. Ideally I believe I would like to bin the responses into different categories for bedrooms or guests as that would clearly have a big affect on the listing price.

Additionally, each listing also has the general area/neighborhood of Los Angeles that the listing is in such as Santa Monica, Hollywood, silver lake etc. and I believe it would be interesting to see what area generally has the most expensive air bnbs. I believe my questions will best be answered using regression although there are a few predictors that are categorical which I intend to assign values to/ can make into dummy variables. I also am not too familiar with categorical approaches at the moment so I will be looking forward to seeing different models which we are taught in class that might be better than regression. I think the predictors that will be most useful will be the typical housing variables like number of rooms, bathrooms, and maybe the neighborhood that it is in. Although, I am very interested in seeing how the reviews score and number of reviews impact price, there is also a variable for date that the host of the airbnb joined the app which I am excited to see if that has an affect on the listing price.

I believe the goal of my model is mostly inferential as I want to state relationships between outcomes and predictors although I would also like to accurately predict the price I believe I am more focused on testing theories about how everything affects the price.

Proposed Project Timeline

- **First**: I plan to have the data loaded and all the variables ready to go within the next two weeks by **April 25th**
- **Second:** I plan to spend another two weeks on the descriptive analysis and have it done by **April 9th**
- **Third:** I plan to write data models, get results, graphs and translate that into a draft of the paper for the rest of may and have that done by the end of may on **May 31st**
- **Fourth:** Lastly I will make the final edits and the final draft within the last week and a half to turn in on time

Questions/Concerns

My main question is what models are best suited for an inferential type model. I would also like to ask if there is a function or a way to convert a string into the number of words in that string like a count. I am interested in this because I would like to convert amenities into number of amenities and the description of the listing into the number of words used in the listing description to see if having a longer description affects the price or reviews. My main concern is whether or not converting some of the predictors will be possible/ too difficult and if this project will be enough to be sufficient for the final project.