Final Project Memo

Sofia Spasibenko

An Overview of the Dataset

I will be conducting my final project on an Airbnb data set. The data set includes: "250,000+ listings in 10 major cities, including information about hosts, pricing, location, and room type, along with over 5 million historical reviews." It includes variables such as: "Percentage of times the Host responds", "Total listings the Host has in Airbnb", "Binary field to determine if the Host has a profile picture", "review_scores_checkin", "review_scores_communication" and 28 others. The data is available through Kaggle (https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews), and I want to primarily work with the Listings.csv since it also includes data about reviews.

There are over 250,000 observations and 33 predictors. For about 46% of the data there are missing values on the host data such as response time and rate, as well as acceptance. I would ideally not include those reviews, since it would still leave me with an ample amount of data to work with. For most entries, if one variable is absent, so are others.

An Overview of My Research Question

I'm most interested in predicting "Listing's overall rating (out of 100)" and I'll probably use the subsequent subdivisions of the score ("Listing's cleanliness", "Listing's communication with the Host score", etc.) to help guide my analysis. As of right now, I'm interested in seeing how host activities can affect the rating of the property, but I'm also interested in doing a general review of the data and seeing where the best-priced properties are and which predictors affect the ratings the most. The listing's overall rating will be my response variable: it is a quantitative (rated out of 100) and seems to skew towards more positive reviews. Because my outcome variable is numerical, I want to use a regression approach for my project. Out of the predictors, I think the pricing, location, and how long the host has been on Airbnb will affect the ratings the most. In general, I want my model to be more inferential since I want to explore the causal relationships between the predictors and the outcome and determine which ones affect the rating of the Airbnb the most. I'm not as interested in predicting the ratings as much as I am interested in discovering which variables can help hosts get better reviews.

My Proposed Project Timeline

- October 8
  - Have had the chance to load and explore my data and begin the cleaning process
- October 15
  - Create a correlation plot and other plots for qualitative variables to begin centering on the key points of analysis
- October 22
  - Have my individual charts for each predictor made
- November 15
  - Create predictive models and compare their accuracy
- November 29
  - Though I will try my best to write my report as I go along, by this date I would want to have written down a majority of my report
- December 3
  - Edit and have my project peer reviewed and ready to submit

Questions and Concerns

I'm a little worried for the size of my project since there are many entries to analyze as well as many predictors. I think the scope of the project will be fairly big which is why I wanted to narrow my goal for it slightly to make it a little more manageable. Additionally, my computer is over seven years old and I'm afraid that it might crash working with large amounts of data.

Hopefully, I'm able to figure out how to operate Github a little bit more since the way we tried to connect it to R in section did not seem to work. For now, I'll be pushing to the repositories manually.