Data Memo

Ziyi Fang

10/2/2022

An overview of my dataset

It includes 28,580 housing listed on Airbnb in Hawaii, United States. There are 18 variables in the dataset which included the host_id, host_name, neighborhood_group, neighborhood, latitude, longtitude, room type, price, minimum_nights, number of reviews, last_reviews, reviews per month, calculated_host_listing_count, aviliable 365, number_of_review_itm, and license.

The dataset is from Inside Airbnb and it offers different data sets related to Airbnb listings in dozens of cities around the world. Here is the link of the website:http://insideairbnb.com/get-the-data/. And the dataset is the summary information and metrics for listings in Hawaii (good for visualisations) till 12 September, 2022.

I will choose a ramdom sample that contains 1,000 observations in order to get a unbiased but runable result. And I will work with both numerical variables and character variables. There are missing data, like the license column. About 10% of data is missing in this column, but I am thinking about using this as one of my questions. for example, does missing license affect the reviews of an airbnb?

The variable that I am interested in predicting is the review. What are the elements that affect the review of an Airbnb houseing and how to improve the review by changing those elements. Does the review affect the popularity and the price of a housing? I think these questions might be best answered with a regression approach .

The goal of my model is predictive as I am predicting some variables may have positive correlations with my variable and some may have negative correlations.

Proposed project timeline

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc? Provide a general timeline for the rest of the quarter.

At week 3, I will have my dataset loaded so I can begin my exploratory data analysis.

At week 4, I will try to clean up my data and using ramdom samples, and I will refine my demo and add more research questions to build a more detailed outline.

At week 5, I will start to run a few test and predictions to see if my data is fit or not.

At week 6, I will repeat the same process of week 5, and I will go to OH to debug my code.

At week 7, I will finalizing my code and started to work on my design and formatting of the project.

At week 8, I will go to OH again to check if my project need more data or predictions.

At week 9, I will generate my report and preparing for any improvements.

At week 10, finally time to submit!

Questions or concerns

Are we allowed to use multiple different datasets for the report or we can only use one data set for all predictions? Thank you!