pstat231-final

2022-10-05

The dataset include information for AirBnB listings in New York City(Listing name, neighborhood, lat/long, room type, price, etc). It will be directly downloaded from the following link via Kaggle: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data (https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data) There are about 48,000 unique observations and 8 out of the 16 columns may serve as predictors. I'll mainly be working with numerical and categorical data. However, I hope to incorporate text analysis and geographical data during the EDA step. There are no missing data recorded with the metadata. However, I will use mean/median values to fill in missing data w.r.t its neighborhood and mode values for categorical data w.r.t the neighborhood.

I hope to predict a price for a listing with given predictors. This will allow hosts to come up with an optimal price for new listings. During the EDA step, I plan on incorporating text analysis to see what trends in listing titles lead to higher prices. Additionally, I want use geographical data to see which areas get most customer reviews and which locations have higher listing prices. I hope to be able to use clustering algorithms to find trends as a bonus. The response variable is the price for the listing. The model will be predictive. If I incorporate clustering models for analysis, it will also include descriptive traits.

I'd like to spend week 2-3 implementing best practices for text processing/analysis. weeks 3-6 will be spent doing EDA/cleaning weeks 7-8 model building weeks 9-11 final touch ups/ honors contract fulfillment

I anticipate difficulties in text processing for reaching my goals. This will be a new technology learned, so many bugs may come along with it. Visualizations for explaining trends may become tricky for geographical/text data.

I plan on completing the rest of this project in Python using Scikit-learn, numpy, pandas, matplotlib, Keras and Spacy.