MA 678 Midterm Proposal

Guangze Yu

Personal Statement:

My career goal is to become a data analyst in a technology company. Except linear regression model and logistic regression model, multilevel model is more realistic and have more replicablity. To be a data analyst, it is very important to deal with data with multiple variables. In real life, there don’t have any regression model which can fit into a relative big dataset. Although it is common to fit simple model into clean dataset and get a relative high accuracy, it is impossible in “dirty” data. How to tell a data-drive story is one of essential skill for a data analyst. Knowing how to separate model into different levels can help to deal with data in which the times of the measurements vary from subject to subject. The Airbnb dataset contains text cleaning, missing-value handling, merging different datasets and so on. It face more challenges than “clean” data as in-class examples. It is a good pratice chance to play with real business data and try to reproduce methodology. Also, multilevel model can extend to more analysis methods like Multivariate responses,Cross-classified data. This project can help me to enrich my knowledge about complicated dataset and prepare for my career skill.

Question:

The relationship between the Price of Boston Airbnb house and other variables like response time, bedroom number and so on. At current stage, the question is board and without any investigate and the question should be more specific after EDA the dataset.

The data source:

Airbnb source of Boston updated 19 September,2021

<http://insideairbnb.com/get-the-data.html>

Including: listings.csv,candelar.csv, review.csv and so on (need to be revised for the further analysis)

Proposed timeline of work:

EDA: 11/15

Data Processing: 11/20

Modeling and Validation: 11/27

Write up: 12/2