Zhaobin Liu

MA 678 midterm project Proposal

I will choose the data analysis for the midterm project. My dataset will be from Airbnb website: <http://tomslee.net/airbnb-data-collection-get-the-data>. I am still deciding which specific region I am going to use. Since the format of the dataset all look similar each other, it will not affect very much for picking different regions.

The variables in the dataset are: room\_id, host\_id, room\_type, borough, neighborhood, reviews, overall\_satisfaction, accommodates, bedrooms, price, mainstay, latitude and longitude, and last\_modified. The data has already been in the csv file, thus it should straightforward to use in R.

I am going to the study two interesting variables: overall\_satisfaction and the price. The former is the average rating (out of five) that the listing has received from those visitors who left a review. The latter is the price for a night stay.

First, I will study what other variables will affect them individually. For example, for price, does the room\_type will have effects on it? How strong the effect will be? This will be a GLM. Also, do more than one variable will affect price? Do these variables have correlation each other? These could be the multilevel model. I will do the same analysis for overall\_satifaction variable.

Second, I will study the relationship between price and overall\_satifaction to see whether they affect each other strongly or not.

The conclusion/result should be clearly state that which factor/s are the most important for price and the rate and whether they affect each other. What is more, I might be using SQL to import/clean data since the dataset could be very large which R can not deal with.