**MA 678 Midterm Project Proposal**

**Project description:**

According to data from Airbnb, I plan to analyze the rating of each score (overall satisfaction, out of 5, from tenants) and evaluate how the ratings be affected by factors such as the borough of the Airbnb host, the number of guests a listing can accommodate, the number of bedroom(s) and bathroom(s), room type (“Entire home/apt”, “Private room”, or “Shared room”) or the reviews, than adjusting the rating of each host, exclude the effect of those factors that are not related to price, so that the adjusted rating is able to reflect the rental price.

**Why this project:**

Two reasons that I am interested in choosing this project, one is I noticed the Airbnb hosts whose average price is high tend to have low ratings but the actual experience is good, at the same time some rooms whose ratings are high but the price is cheap, with comparatively bad living experience, so we can say that the ratings of each host are biased, which cannot be referred as the selection standard. The other reason is that if I can figure out those factors that do significantly affect the popularity of a host, it will provide some guidance on how to successfully run an Airbnb host.

**Project dataset:**

Airbnb dataset of Boston from 2014 to 2017, from Airbnb Data Collection.

Retrieved from <http://tomslee.net/airbnb-data-collection-get-the-data>

**Questions to answer:**

What are the critical factors that do influence the rating of popularity of an Airbnb host?

How those factors affect the rating of the popularity of an Airbnb host.

**Potential problem I foresee:**

The entire dataset from Airbnb is too large to process and my device may cannot process such a large data file, so I plan to narrow the scope and analyze the greater Boston area.

The original data table that contains Airbnb records has a large number of missing values, which may cost some time on data cleaning.

**The timeline of working on this project:**

First draft by Nov 27, 2019.

Data cleaning: around 1 week;

Data analysis: building multi-level model and selecting control variables, 1-2 week(s);

Result interpretation: depends on whether the result is consistent as my expectation, roughly 1 week.

**Future direction that I might take this project to:**

The data is part of the Airbnb data competition, so I may continue analyzing datasets and try participating in it; also, I may extent the analyzing scope to states and countries with providing a more general result that is more applicable to wider range of Airbnb hosts and provide a rating system that is free from the bias related to factors which are not related to living experience.