**Midterm project proposal-Haoyu Li**

**Personal statement**

The dataset I choose is about the review rating of the Airbnb in San Francisco area. The reason I choose Airbnb dataset mainly because I want to enter Tech industries as a data analyst or scientist. By analyzing the Airbnb rating dataset is one step further to get into the Tech industries earlier and to see if I can handle such dataset that I may encounter frequently if I really get into the Tech companies. Another reason is that me myself like to travel when vacation comes, and I used Airbnb multiple times. So, I wonder if there is a relationship between rating and different housing criteria, and also if the difference is significant across different city areas, and that interests me and leads me to analyze this dataset.

**Preliminary question I try to answer**

* Are the rating scores significantly different given the Airbnb houses are grouped by city areas or neighbors?
* The relationships between rating scores and different house criteria that hosts list out.

**The data source**: <https://www.kaggle.com/jeploretizo/san-francisco-airbnb-listings>

**Proposed timeline of work**

* EDA: During week 11/9 - 11/13

I will do the EDA based on the rating scoring that I have, to see the data distribution and if there are outliers that will have significant effect on the data distribution.

* Data processing: During week 11/16 - 11/20

Cleaning out unnecessary data points and variable to make the dataset clearer to visualize and prepare the dataset for the modeling step.

* Modeling and model check: During week 11/23 - 11/27

Using multilevel linear model to fit no pooling, partial pooling, complete pooling, and compare models with varying slopes and varying intercept. And check the validity of these models.

* Write/Wrap up: Before 12/1 (If possible)

Write up report and conclusion of the project process, including data cleaning, EDA, modeling and model checking.