**Problem Statement**

We are using two datasets:

1. Seattle Airbnb Listings
2. Boston Airbnb Open Data

For the first dataset our main aim would be to predict the Hotel ratings and to analyze how all the features affect the hotel ratings.

For the second dataset we have a timeline available and hence we would like to work on forecasting a hotel/room price for the upcoming year based on the previous years’ data.

Our aim is to give insights into how different features of the hosted homes such as location, space, neighborhood, amenities, interaction and many other factors affect the prices of the households and also the ratings and customer satisfaction. The aim is to study the correlation between these parameters as well as to predict the prices and ratings and other parameters based on different regression techniques and datasets used. We have selected data from two cities for this purpose- Boston and Seattle. The reason behind this is to show how different parameters for these cities compare with each other.

**Collection and Description DataSets**

**Boston Airbnb Open Data:**

<https://www.kaggle.com/airbnb/boston#listings.csv>

This dataset describes the listing activity of homestays in Boston, MA, since 2008.

It has about 95 columns.

**Seattle Listings**

[**https://www.kaggle.com/shanelev/seattle-airbnb-listings**](https://www.kaggle.com/shanelev/seattle-airbnb-listings)

The data was scraped on December 19th, 2018 and contains roughly 8000 listings of current Airbnb listings in Seattle. The data has the price, reviews, latitude, longitude, bedroom, bathroom, number of guests it accommodates, room type, and more.

**Data Preprocessing Techniques Applied**

We will carry out the following steps for Data Preprocessing:

1. Remove identifiers - remove any column that uniquely identifies a tuple
2. Convert string columns to numeric if that is possible and relevant
3. Identify missing values and replace them with relevant values
4. Detect outliers and highly unusual occurrences
5. One Hot Encoding would need to be executed on several of the Boston Attributes
6. Try imputation techniques for missing data (0, mean, fast forward, backward tracking, row removal)

**Visual Examination**

We will show how the predicted features such as Ratings and Prices compare to other parameters. We will perform Exploratory Data Analysis to examine the correlation between these features with other parameters and plot them against each other. The relationship between those features will be examined by computing a correlation matrix. When there are columns that are highly correlated and supply redundant information, only one of them will be used and others removed. In the case of regression techniques, Simple linear regression will be performed with respect to each variable to examine linear dependency of the response variables on those features.

**Modeling Techniques Chosen**

To predict ratings, We will use the following modeling techniques:

1. Null Model
2. Simple Regression
3. MultiLinear Regression
4. Quad Regression
5. Neural Net

To Forecast Prices, we will use the following modeling techniques:

1. NullModel
2. RandomWalk
3. AR(1)
4. ARIMA
5. LSTM

**Explanation of Why Techniques Were Chosen**

We will try all the five modelling techniques and find out the best suitable model for the data.

**Feature Selection**

For feature selection we can use the following techniques;

1. Forward selection
2. Backward elimination
3. Recursive Feature elimination

(We can compare the results obtained by using all three methods)