1st Capstone Project

# Seattle ⌂ airbnb Listing Price Prediction
## Yueh-Tung (Nicole) Chao
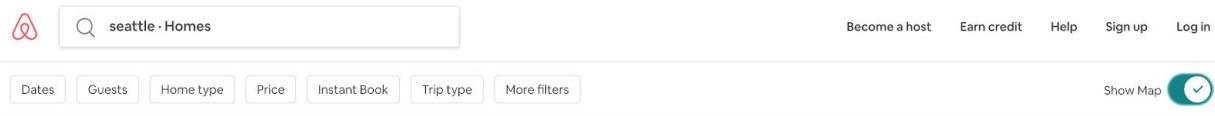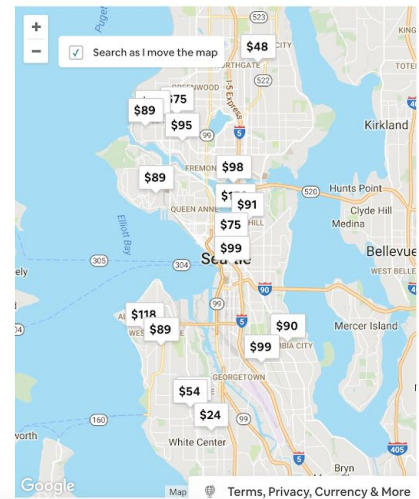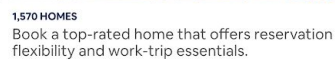




https://www.airbnb.com/s/homes?query=seattle&refinement_paths%5B%5D=%2Fhomes&collection_ids%5B%5D=2&map_toggle=true

## Problem

Airbnb is an American company operating online marketplace and hospitality service for people to lease or rent short-term lodging. The company does not own any real estate. Instead, it is a broker which received percentage service fees in conjunction with every booking.

As from 2008, Airbnb has become one of the most popular hospitality services around the world, which has over 5 million lodging listings in 81,000 cities and 191 countries and has facilitated over 300 million check-ins.

One challenge that Airbnb hosts & Airbnb itself face is determining the optimal nightly rent price. As hosts, if charing too much, then the renter would select more affordable alternatives. On the other hand, if charging too less, then they lose the potential revenue. As for airbnb, if they can recommend optimal listing price to hosts to maximize their revenue, they can receive more service fees as well.

So in this capstone project, we propose to predict airbnb listing price so that both hosts & AIrbnb can optimize their revenue.

## Client

As mentioned in previous session, by predicting listing price, it can be used for both hosts and airbnb for better pricing optimization & maximize the revenue.

## Data Set

- The listing data of Seattle Airbnb can be found on Kaggle: Seattle Airbnb Open Data.
- To visualize data on map, the shapefile of Seattle neighborhood can be found on Zillow: Zillow Neighborhood Boundaries (See figures in next page).
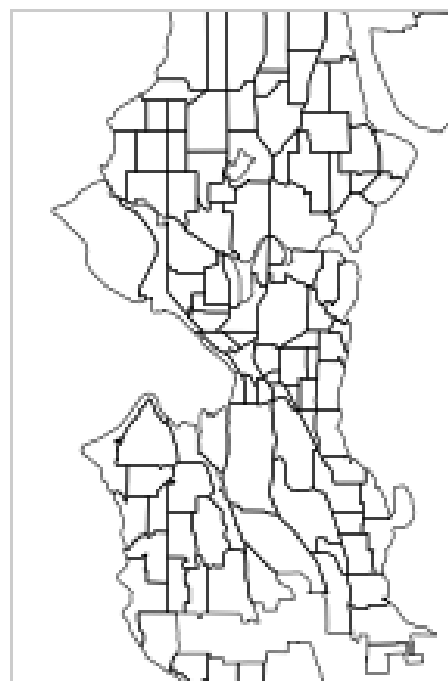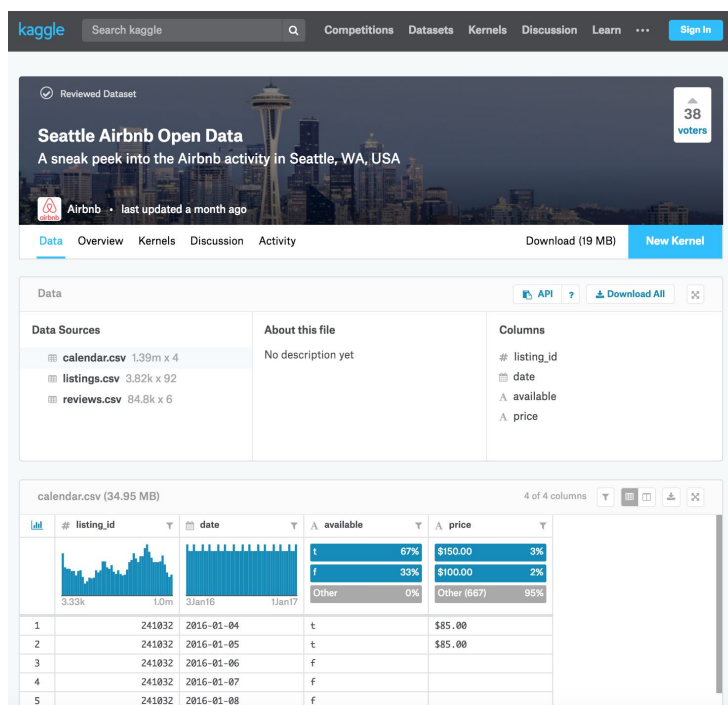
# Data Wrangling

First, we input data & do data wrangling. We conduct the following steps to clean up data and pick useful features.

1. Select columns that might be useful
   a. Visually inspect data & select columns that might be useful.
2. Analyze further & remove unwanted columns
   a. Originally there are 51 features/columns in the dataset, we further remove some unwanted columns.
   b. For example, we drop city-related columns as we are only predict airbnb listing price for one city now so these columns have same values.
3. Drop duplicates (if applicable)
   a. We also removed duplicate samples (rows) & duplicate columns (like neighborhood & neighborhood_cleansed).
4. Fix data types

a. Then we look through all the data & fixed data types. For example, we identify categorical data & use sklearn LabelEncoder to encode then for further usage.

b. We also combing latitude & longitude data using binning techniques.

c. Replace some data with special characters, like '$' or '%' with numerical data.

d. Replace dates with datetime object and transformed them into numerical data by calculating number of dates between the dates and the last-scraped-date, etc.

5. Fix missing values

a. Plot scatter plot for columns with missing values and inspect the trend. Confirmed that most of the features are either linear or random to target.

b. Fill missing values with mean.

6. Find & fix outliers

a. Remove minimum night outlier by cap it to be 99.5 percentile of the data.

7. Drop unnecessary features as we only have 3818 data points so we might not want to have over 20 features.

a. Plot scatter plot of each feature vs. target, i,e, price and drop some unnecessary columns, i.e., columns that do not seem to affect price.

## Initial Findings

During exploratory data analysis, we ask the following questions:

1. What might be the most important features that affect listing price?
2. Does location/neighborhood affect listing price?
3. If yes, what made these neighborhood special?

and ask follow-up questions if needed.

Initial findings are:

1. As we expected, listing price is correlated to accommodates, bedrooms, bathroom, neighborhood, and property type.

a. Interestingly looks like neighborhoods facing water (either bay or lake) seem to have higher listing prices, like Portage Bay (See the figure in next page).

b. However there might be some exceptions:

    i. There might be extremely high listing price although there's only one bedroom, but rarely happened.

    ii. If the property has a lot more bathrooms, ex. 8, but it's a dorm, the listing price would not increase proportionally.

    iii. If the property has a lot more bathrooms, ex. 8, but can only accommodate 2 people, the listing price would not increase proportionally, either.

    iv. If the property is a big house which can accommodate 14 people but its accommodates vs. bedrooms ratio and/or accommodates vs. bathrooms ratio are higher, that means more traveler need to share the bedrooms / bathrooms, the lisilistingng price would not increase proportionally, either.

2. Property-type-wise, most of the property type of Seattle listings are house and apartment.
   a. Wallingford & First Hill have most number of listings whose property type is house.
   b. First Hill & Belltown have most number of listings whose property type is apartment.
   c. Over 13% of Seattle Airbnb listings are located in First Hill.
3. In terms of bedrooms and bathrooms, 1 bedroom 1 bathroom Airbnbs which can accommodate 1 ~ 4 people are most common in Seattle.
   a. Over 50% of Seattle Airbnb listing are of this type.
   b. About 10% of them are located in First Hill.
   c. For the rest of Airbnbs, not many of them can accommodate more than 10 people. In fact, only 1.3% of listings can accommodate 10 or more than 10 people, and other than First Hill, most of them are located in some suburban area.

## Other Potential Data Sets

- More data can be found at: [Inside Airbnb](#).