# Final Project

## --- Cousera Capstone

## Airbnb Recommendation for Travelers in NYC

### Qingqing Cao

#### 01/08/2020

#### Introduction:

New York is one of the world's major commercial, financial and culatual centers. Its core, Manhattan, is the most densely populated borough. It is known of many major attractions. Every year, many travelers choose New York for different reasons. They may come for sight seeing, shopping, arts and shows, nightlife, or business trips, etc. Travelers usually wish to stay in neiborhoods that close to their places of interests, and they usually have a budget limit for accomendations. This project is to help travelers to choose the best area for choosing airbnb based on their locations and interests.

People who might be interested in this projects are traverlers who plan to choose airbnb in NYC.

#### Data usage:

The purpose of traveling are divided into four categories: outdoors, arts, shopping, and food. Boroughs and Neighborhood information are acquired from the lab data. Locations related to these categories are explored and clustered using Foursquare API. Airbnb data is downloaded from the website: [https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/data (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/data)](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/data).

Dataset example is in Dataset part.

## Dataset

Request neighborhoods in Mahantton

In [8]:

```
manhattan_data = neighborhoods[neighborhoods['Borough'] == 'Manhattan'].reset_index(dr
manhattan_data.head()
```

Out[8]:

| | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| **0** | Manhattan | Marble Hill | 40.876551 | -73.910660 |
| **1** | Manhattan | Chinatown | 40.715618 | -73.994279 |
| **2** | Manhattan | Washington Heights | 40.851903 | -73.936900 |
| **3** | Manhattan | Inwood | 40.867684 | -73.921210 |
| **4** | Manhattan | Hamilton Heights | 40.823604 | -73.949688 |

Read airbnb data from csv file. The csv file contains airbnb name, host_name, borough, neighbourhood_group,position, room_type, price, etc.

In [9]:

```
airbnb=pd.read_csv('AB_NYC_2019.csv')
airbnb.head()
```

Out[9]:

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | latitude |
|---|---|---|---|---|---|---|---|
| **0** | 2539 | Clean & quiet apt home by the park | 2787 | John | Brooklyn | Kensington | 40.64749 |
| **1** | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | 40.75362 |
| **2** | 3647 | THE VILLAGE OF HARLEM....NEW YORK ! | 4632 | Elisabeth | Manhattan | Harlem | 40.80902 |
| **3** | 3831 | Cozy Entire Floor of Brownstone | 4869 | LisaRoxanne | Brooklyn | Clinton Hill | 40.68514 |
| **4** | 5022 | Entire Apt: Spacious Studio/Loft by central park | 7192 | Laura | Manhattan | East Harlem | 40.79851 |

In [10]:

```
airbnb.drop(columns=['id','host_id','minimum_nights','number_of_reviews','last_review'
airbnb.head()
```

Out[10]:

| | name | host_name | neighbourhood_group | neighbourhood | latitude | longitude | ro |
|---|---|---|---|---|---|---|---|
| **0** | Clean & quiet apt home by the park | John | Brooklyn | Kensington | 40.64749 | -73.97237 | |
| **1** | Skylit Midtown Castle | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | I |
| **2** | THE VILLAGE OF HARLEM....NEW YORK ! | Elisabeth | Manhattan | Harlem | 40.80902 | -73.94190 | |
| **3** | Cozy Entire Floor of Brownstone | LisaRoxanne | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | I |
| **4** | Entire Apt: Spacious Studio/Loft by central park | Laura | Manhattan | East Harlem | 40.79851 | -73.94399 | I |

We assume the reviews per month is related to the popularity of the host. The one with higher reviews per month is more popular.

In [11]:

```
airbnb.columns=['name','host_name','borough','neighbourhood','latitude','longtitude','
airbnb_m=airbnb[airbnb['borough']=='Manhattan'].reset_index(drop=True)
airbnb_m.fillna(0,inplace=True)
airbnb_m.head()
```

Out[11]:

| | name | host_name | borough | neighbourhood | latitude | longtitude | room_type | pri |
|---|---|---|---|---|---|---|---|---|
| **0** | Skylit Midtown Castle | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 2 |
| **1** | THE VILLAGE OF HARLEM....NEW YORK ! | Elisabeth | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 1 |
| **2** | Entire Apt: Spacious Studio/Loft by central park | Laura | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | |
| **3** | Large Cozy 1 BR Apartment In Midtown East | Chris | Manhattan | Murray Hill | 40.74767 | -73.97500 | Entire home/apt | 2 |
| **4** | Large Furnished Room Near B'way | Shunichi | Manhattan | Hell's Kitchen | 40.76489 | -73.98493 | Private room | |

## Mathedology

What we do in the mathedology part is 1) use Foursquare API to explore restaurant, arts, shopping center, outdoor activity locations in Manhattan; 2) show these locations on the map and use labels to show their name and category; 3) cluster airbnb in Manhattan by their prices and popularities; 4) show the airbnb locations on the map.

In [18]:

```
manhattan_restaurants.head()
```

Out[18]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Type | Venue | Venue Latitude | Venue Longitude | Ve Categ |
|---|---|---|---|---|---|---|---|---|
| **0** | Marble Hill | 40.876551 | -73.91066 | food | Arturo's | 40.874412 | -73.910271 | P P |
| **1** | Marble Hill | 40.876551 | -73.91066 | food | Tibbett Diner | 40.880404 | -73.908937 | D |
| **2** | Marble Hill | 40.876551 | -73.91066 | food | Dunkin' | 40.877136 | -73.906666 | Dc S |
| **3** | Marble Hill | 40.876551 | -73.91066 | food | Land & Sea Restaurant | 40.877885 | -73.905873 | Seaf Restau |
| **4** | Marble Hill | 40.876551 | -73.91066 | food | Boston Market | 40.877430 | -73.905412 | Amer Restau |

In [20]:

```
manhattan_arts = getNearbyVenues(names=manhattan_data['Neighborhood'],
                                 search_query='arts',
                                 latitudes=manhattan_data['Latitude'],
                                 longitudes=manhattan_data['Longitude'],
                                )
manhattan_arts.head()
```

Out[20]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Type | Venue | Venue Latitude | Venue Longitude | Ve Categ |
|---|---|---|---|---|---|---|---|---|
| **0** | Marble Hill | 40.876551 | -73.910660 | arts | Villa Lobos | 40.875592 | -73.909496 | M Ve |
| **1** | Marble Hill | 40.876551 | -73.910660 | arts | Sonnet Project - Sonnet #152 | 40.880538 | -73.911295 | Perfor Arts Ve |
| **2** | Chinatown | 40.715618 | -73.994279 | arts | Museum at Eldridge Street | 40.714724 | -73.993497 | Mus |
| **3** | Chinatown | 40.715618 | -73.994279 | arts | Metrograph | 40.714999 | -73.991035 | I M The |
| **4** | Chinatown | 40.715618 | -73.994279 | arts | Sofar HQ | 40.713523 | -73.996289 | M Ve |

In [22]:

```
manhattan_shopping = getNearbyVenues(names=manhattan_data['Neighborhood'],
                                   search_query='shops',
                                   latitudes=manhattan_data['Latitude'],
                                   longitudes=manhattan_data['Longitude'],
                                  )
manhattan_shopping.head()
```

Out[22]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Type | Venue | Venue Latitude | Venue Longitude | \ Cat |
|---|---|---|---|---|---|---|---|---|
| 0 | Marble Hill | 40.876551 | -73.91066 | shops | T.J. Maxx | 40.877232 | -73.905042 | Depar |
| 1 | Marble Hill | 40.876551 | -73.91066 | shops | Rite Aid | 40.875467 | -73.908906 | Phai |
| 2 | Marble Hill | 40.876551 | -73.91066 | shops | Vitamin Shoppe | 40.877160 | -73.905632 | Suppl |
| 3 | Marble Hill | 40.876551 | -73.91066 | shops | Lot Less Closeouts | 40.878270 | -73.905265 | Dis |
| 4 | Marble Hill | 40.876551 | -73.91066 | shops | GameStop | 40.874267 | -73.909342 | Game |

In [24]:

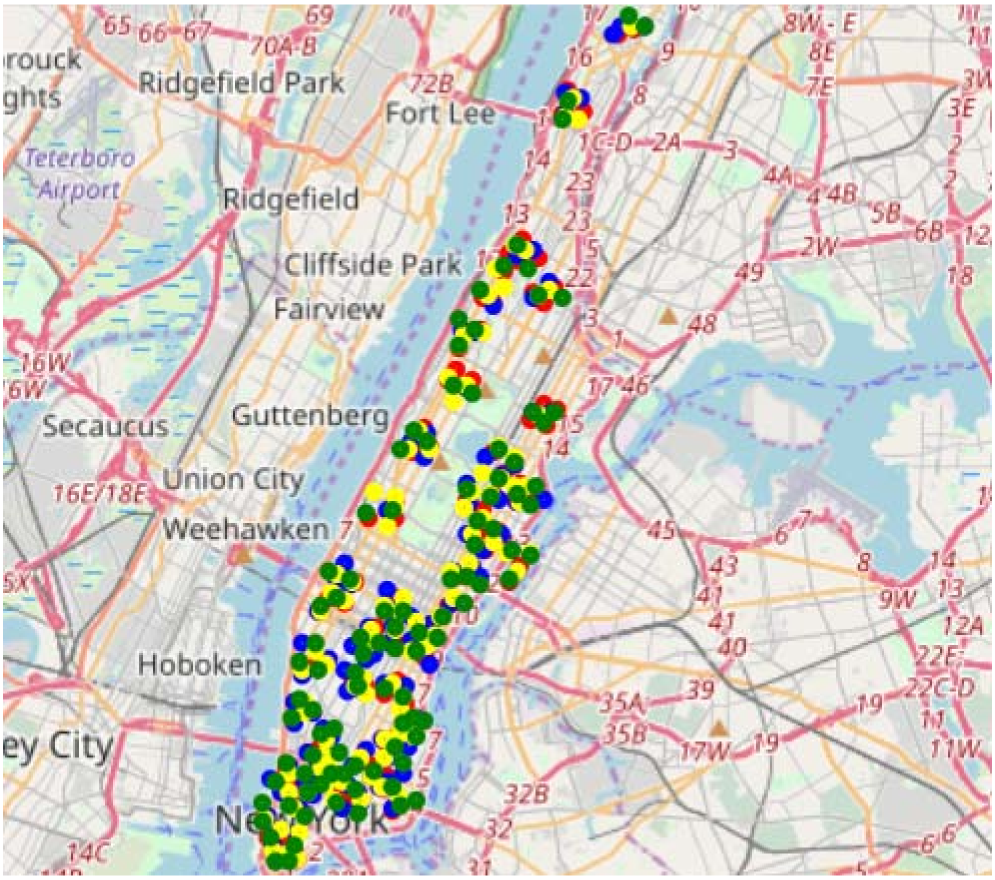
```
manhattan_outdoors = getNearbyVenues(names=manhattan_data['Neighborhood'],
                                   search_query='outdoors',
                                   latitudes=manhattan_data['Latitude'],
                                   longitudes=manhattan_data['Longitude'],
                                  )
manhattan_outdoors.head()
```

Out[24]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Type | Venue | Venue Latitude | Venue Longitude | C |
|---|---|---|---|---|---|---|---|---|
| 0 | Marble Hill | 40.876551 | -73.91066 | outdoors | Bikram Yoga | 40.876844 | -73.906204 | |
| 1 | Marble Hill | 40.876551 | -73.91066 | outdoors | Blink Fitness | 40.877271 | -73.905595 | |
| 2 | Marble Hill | 40.876551 | -73.91066 | outdoors | Planet Fitness | 40.874088 | -73.909137 | |
| 3 | Marble Hill | 40.876551 | -73.91066 | outdoors | Marble Hill Playground | 40.877765 | -73.907994 | Pla |
| 4 | Marble Hill | 40.876551 | -73.91066 | outdoors | Orange Park, Marble Hill, Bronx, NY | 40.877986 | -73.908028 | |

## Results

From the distribution of these clusters in the map, those four colors mix up evenly, which means there is no place that only one or two categories take adavantages.



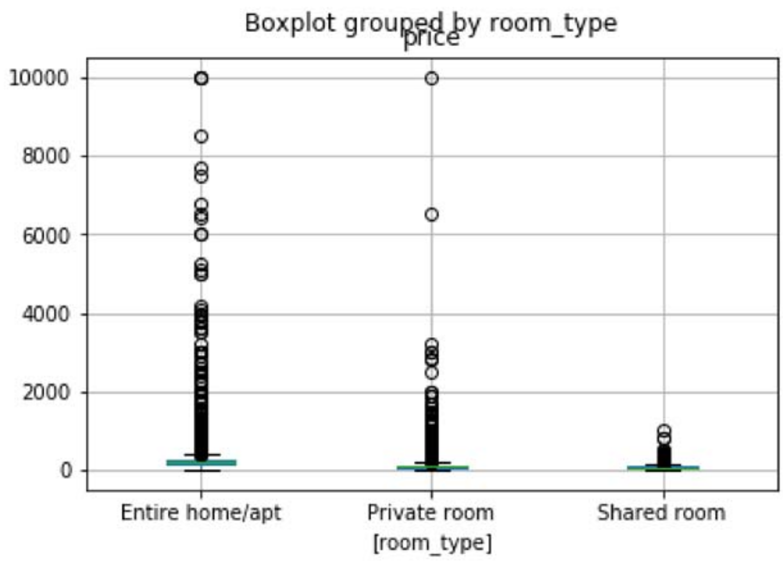

Next, we will analyze airbnb data. Because airbnb dataset is really large, and cannot be fully shown in folium map. Similar clustering operation is conducted to airbnb in Manhattan. First, let's look at how the prices of airbnb affected by locations and room type.

In [38]:

```
airbnb_m.boxplot(['price'],by=['room_type'])
```

Out[38]:

<matplotlib.axes._subplots.AxesSubplot at 0x1b5186e5eb8>



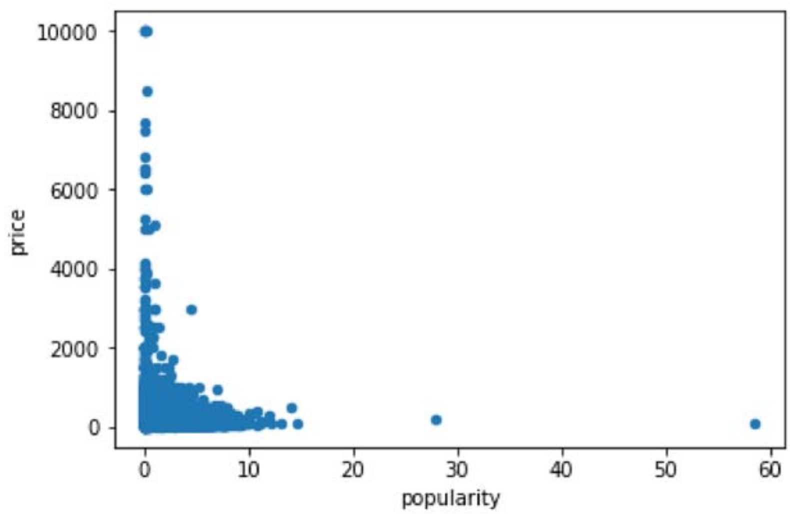

The price is related to the room type. Entire room/apt has the highest mean price compared with private room or shared room. More outliers which prices are higher than the maximum (third quarter + 1.5* interquartiel range) show in the type of entire room/apt, and followed by private room, and shared room.

In [33]:

```
airbnb_m.plot(kind='scatter',x='popularity',y='price')
```

Out[33]:

<matplotlib.axes._subplots.AxesSubplot at 0x1b518694d68>



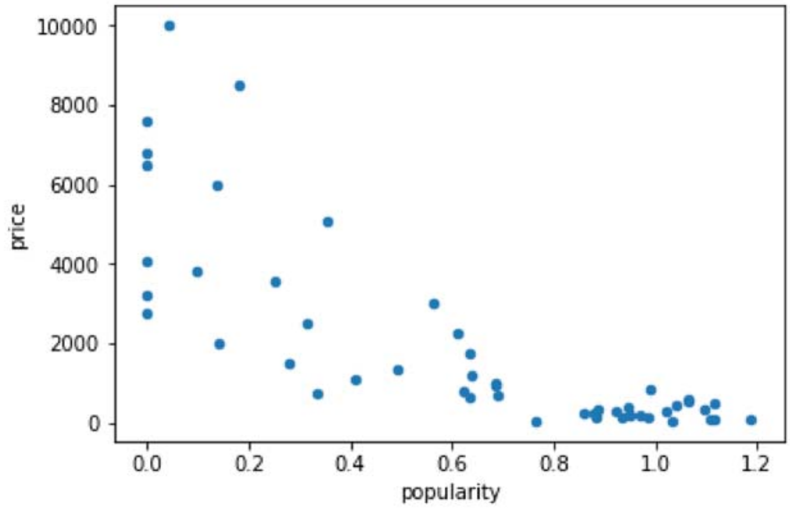

No obvious trend is found in the relationship between the price and popularites if we consider the whole dataset.

In [36]:

```
airbnb_mmean.plot(kind='scatter',x='popularity',y='price')
```

Out[36]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b518707828>
```




From the clustered data, generally cheaper airbnb has higher popularities.

The clusters of airbnb and attractions are shown in the map below


## Discussion

In four categories(restaurants, arts, shopping and outdoors) in Manhattan, arts has the the least number of items, which is 835. Other three categories all has over 1000 items.

All four categories are clustered into 100 clusters and shown in the map separately. Red markers show restaurant clusters; yellow markers show shopping places; blue markers how arts places; green markers show locations for outdoors. From the distribution of these clusters in the map, those four colors mix up evenly, which means there is no place that only one or two categories take adavantages. The more we reach the south of Manhattan, more attractions(include all four categories) show up. Most attractions accumulated on the south of the Broadway-Lafayette Street. And east side of the Manhattan has more attractions than the west side.

There are more than 21000 airbnb in Manhattan, and the price ranges from 0 to 10000. The price is related to the room type. Entire room/apt has the highest mean price compared with private room or shared room. More outliers which prices are higher than the maximum (third quarter + 1.5* interquartiel range) show in the type of entire room/apt, and followed by private room, and shared room. We assume the reviews per month is related to the popularity of the host. The one with higher reviews per month is more popular. No obvious trend is found in the relationship between the price and popularites if we consider the whole dataset.

Then we divided airbnb based on their locations, prices and popularities and sorted the data by descending popularites. From the clustered data, generally cheaper airbnb has higher popularities. This is not always the truth because the popularity of the location is also decided by other reasons such as the hosts' attitudes, the cleaness of the room, etc. The price of the airbnb is generally higher on the south of Manhattan. The two clusters on the north of the Manhattan have relatively low price, which are 39 and 54 dollars per night. Travelers with low budget can consider this area. The mean price of the clusters on the south of the Manhattan is much higher than the mean price of the airbnb. This might be because they are close to many attractions. However, because of the high price, they do not have a high popularity. The majorities of the airbnb apartments are between 23rd and 50th street. The price and popularity are very diverse in this area. As there are many attractions in this area, and the mean prices in this area are generally acceptable. This area is recommended for most travelers in NYC.

## Conclusion

Based on the analysis and discussion above, main ideas are concluded here:

(1) There is no differences among the distribution of restaurants, arts, shopping and outdoors location. Travelers can always enjoy them in the same area.

(2) More attractions lay on the south part of Manhattan, and east side has more attractions than west side of Manhattan.

(3) Airbnb price is related to the room type. Usually private house/apt has higher price than private room or shared room. Generally cheaper airbnb will attract more travelers.

(4) The price of the airbnb is higher on the south of Manhattan and lower on the north tof Manhattan, similar trend as the attractions number.

(5) For most travelers, the area between 23rd and 50th street is most recommended for them to choose airbnb. The price is very diverse in this area and there are many attractions in the area.