Capstone Project: Battle of the Neighbourhoods – Districts of Hong Kong

Justin Chau

May 21, 2020

1. Introduction

The purpose of this Capstone Project is to identify the venues around the different districts of Hong Kong and determine the venues with the best ratings to assist tourists in making the decision of where to stay.

With a population of over 7.4 million people in a 1,104 square-kilometre area, Hong Kong ranks as one the highest densely populated areas in the world. Nearly 40 million visitors entered Hong Kong last year and with numerous attractions available for tourists, they will want to find a place to stay that suits them the best in terms of surrounding venues.

Using Python and the Foursquare API, we will identify the most common venues in each major district of Hong Kong and use machine learning to cluster these districts to help assist tourists in finding an appropriate location for them to stay.

2. Data

We will require different sets of data in order to complete this project.

1. A list of the different districts of Hong Kong:
   Using a list of Hong Kong districts from Wikipedia, we will use data tools to scrap the district information from Wikipedia.

2. The Longitude and Latitude of these different districts
   Using the Geopy client, we can obtain the coordinates for these districts

3. Venue Data surrounding these districts using the Foursquare API
   Using the Foursquare API, we can obtain venue data surrounding the Hong Kong districts

3. Methodology

3.1

To begin, I obtain the table including all the Hong Kong districts from Wikipedia using the following url: https://en.wikipedia.org/wiki/Districts_of_Hong_Kong
Using the BeautifulSoup library, we pull the data from the Wikipedia table and create a data frame using the pandas library. The resulting data frame is shown below:

	District	Chinese	Population	Area km^2	Density /km^2	Region
0	Central and Western	中西區	244,600	12.44	19,983.92	Hong Kong Island
1	Eastern	東區	574,500	18.56	31,217.67	Hong Kong Island
2	Southern	南區	269,200	38.85	6,962.68	Hong Kong Island
3	Wan Chai	灣仔區	150,900	9.83	15,300.10	Hong Kong Island
4	Sham Shui Po	深水埗區	390,600	9.35	41,529.41	Kowloon
5	Kowloon City	九龍城區	405,400	10.02	40,194.70	Kowloon
6	Kwun Tong	觀塘區	641,100	11.27	56,779.05	Kowloon
7	Wong Tai Sin	黃大仙區	426,200	9.30	45,645.16	Kowloon
8	Yau Tsim Mong	油尖旺區	318,100	6.99	44,864.09	Kowloon
9	Islands	離島區	146,900	175.12	825.14	New Territories
10	Kwai Tsing	葵青區	507,100	23.34	21,503.86	New Territories
11	North	北區	310,800	136.61	2,220.19	New Territories
12	Sai Kung	西貢區	448,600	129.65	3,460.08	New Territories
13	Sha Tin	沙田區	648,200	68.71	9,433.85	New Territories
14	Tai Po	大埔區	307,100	136.15	2,220.35	New Territories
15	Tsuen Wan	荃灣區	303,600	61.71	4,887.38	New Territories
16	Tuen Mun	屯門區	495,900	82.89	5,889.38	New Territories
17	Yuen Long	元朗區	607,200	138.46	4,297.99	New Territories

Figure 1.1

3.2

Next, we obtain the longitude and latitude coordinates for the districts using the Geopy client as shown below:

```
geolocator = Nominatim(user_agent="Hong Kong Districts")

df['Coordinates']=df['Chinese'].apply(geolocator.geocode).apply(lambda x: (x.latitude, x.longitude))
df[['Latitude', 'Longitude']] = df['Coordinates'].apply(pd.Series)

df.drop(['Coordinates'], axis = 1, inplace=True)
df
```

Figure 1.2

3.3

Using the folium library, we can visualize the 17 different districts of Hong Kong. The longitude and latitude coordinates were used to identify each district on the map below:

```
# creating a map of Hong Kong using latitude and longitude values
map_HK = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, district, region in zip(df['Latitude'], df['Longitude'], df['District'], df['Region']):
    label = '{}, {}'.format(district, region)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_HK)

map_HK
```

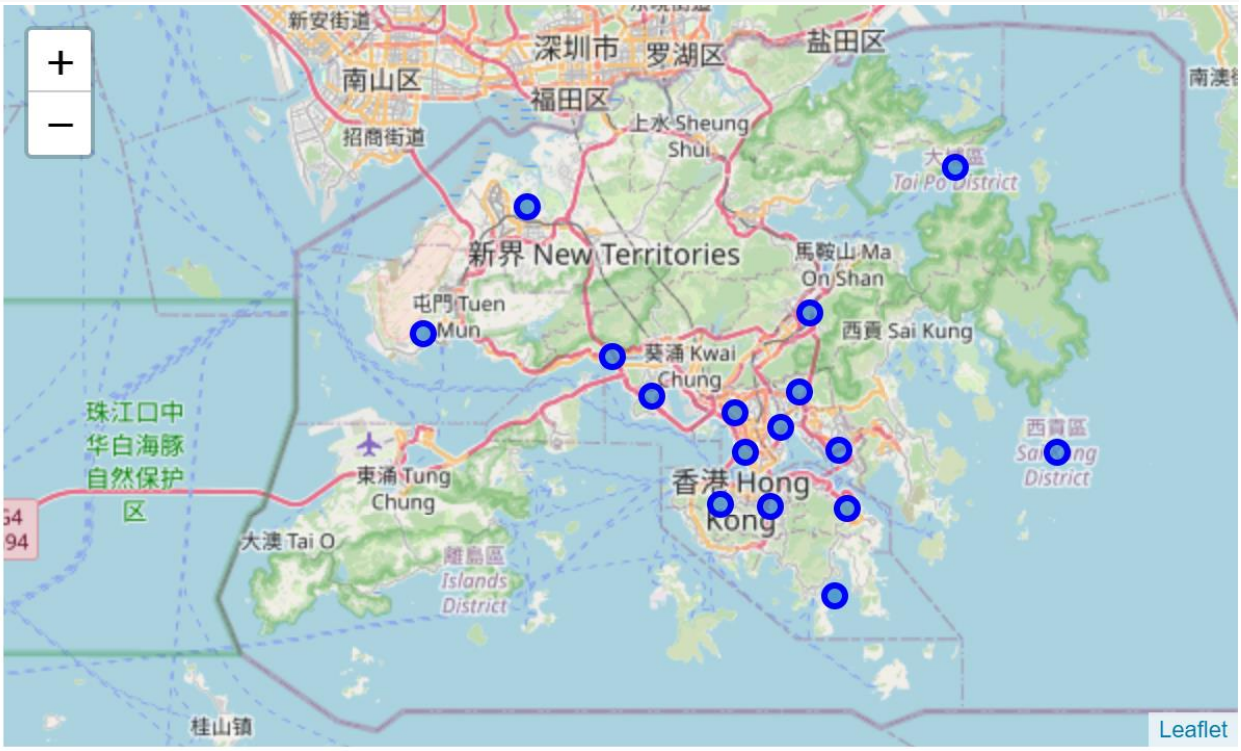

Figure 1.3

3.4

Next, using the Foursquare API, we can get the top 100 venues that are located within a 500-metre radius of the major districts in Hong Kong.

```
HK_venues = getNearbyVenues(names=df['District'],
                                    latitudes=df['Latitude'],
                                    longitudes=df['Longitude']
                                   )
```

```
Central and Western
Eastern
Southern
Wan Chai
Sham Shui Po
Kowloon City
Kwun Tong
Wong Tai Sin
Yau Tsim Mong
Islands
Kwai Tsing
North
Sai Kung
Sha Tin
Tai Po
Tsuen Wan
Tuen Mun
Yuen Long
```

Figure 1.4

After determining there are 115 different venue categories, we display the top 20 most frequently occurring venues on the chart below:

```
import seaborn as sns
from matplotlib import pyplot as plt

s=sns.barplot(x="Venue Category", y="Frequency", data=HK_venue_frequency)
s.set_xticklabels(s.get_xticklabels(), rotation=45, horizontalalignment='right')

plt.title('Top 20 Most Frequent Venues in Hong Kong Districts', fontsize=15)
plt.xlabel("Venue Category", fontsize=15)
plt.ylabel ("Frequency", fontsize=15)
plt.savefig("Most_Freq_Venues1.png", dpi=300)
fig = plt.figure(figsize=(18,7))
plt.show()
```

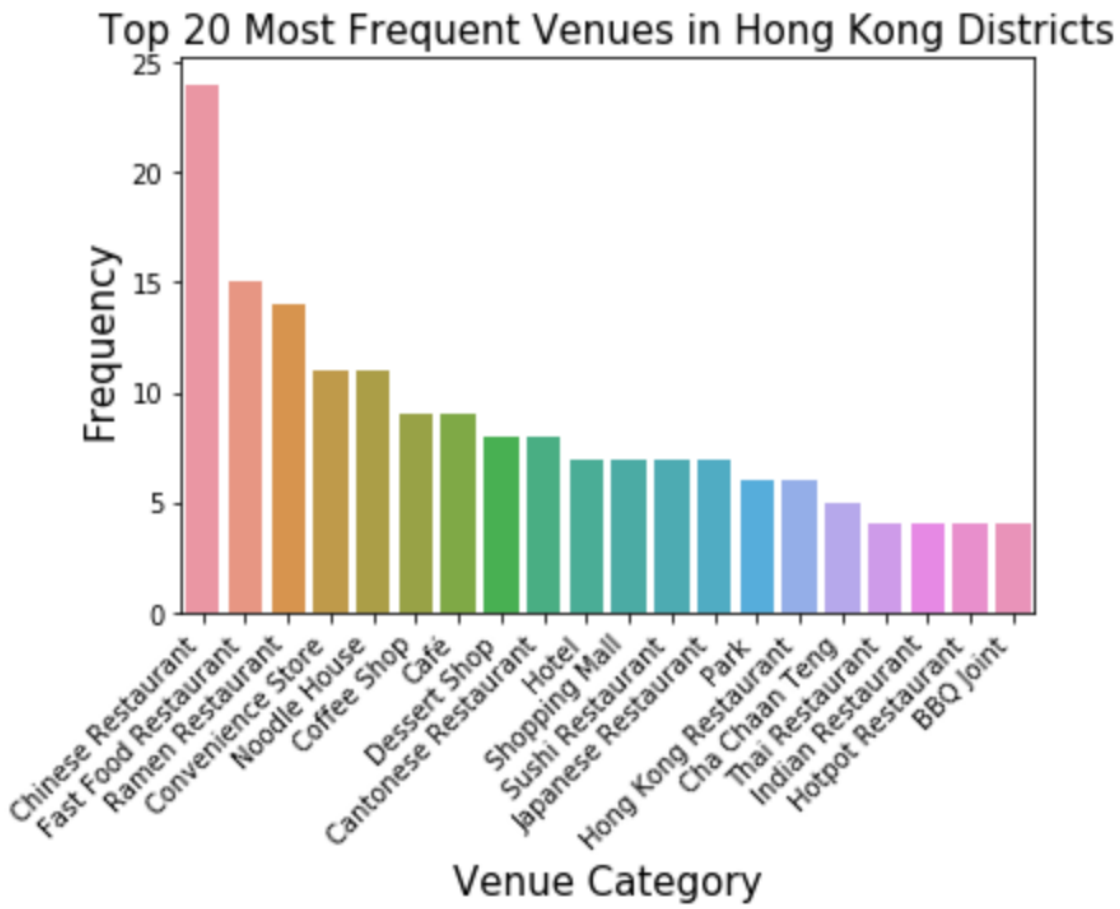

Figure 1.5

First, we create a new pandas dataframe using all the venue categories with one hot encoding:

```
# one hot encoding
HK_onehot = pd.get_dummies(HK_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
HK_onehot['District'] = HK_venues['District']

# move neighborhood column to the first column
fixed_columns = [HK_onehot.columns[-1]] + list(HK_onehot.columns[:-1])
HK_onehot = HK_onehot[fixed_columns]

HK_onehot.head(15)
```

Figure 1.6

With pandas' groupby function, we can calculate the mean and frequency of each venue category and display the top 5 most common venues for each major district:

```
num_top_venues = 5

for hood in HK_grouped['District']:
    print("----"+hood+"----")
    temp = HK_grouped[HK_grouped['District'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

Figure 1.7

Last, we will group the major districts using K-means clustering. The districts will be clustered based on the similarity of venue categories. Using the Folium library we can display the major districts below:

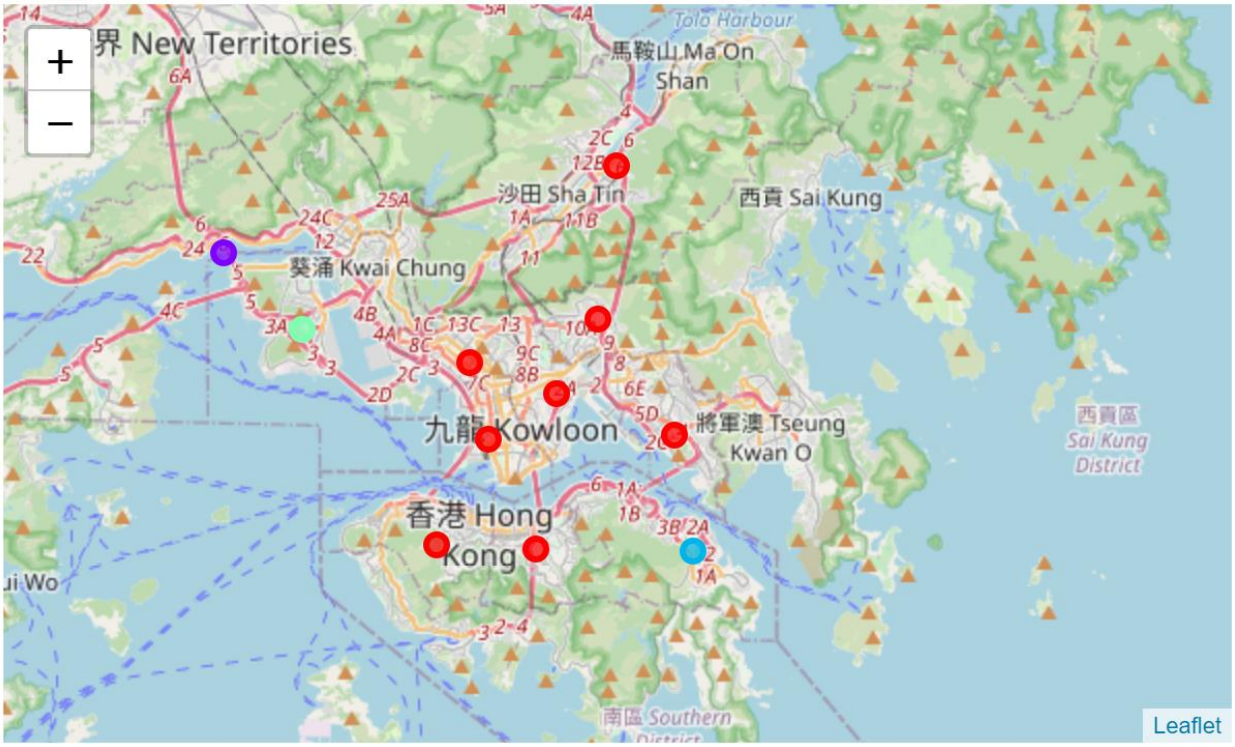

Figure 1.8

4. Results and Discussion

Using various data analysis techniques, we will able to discover information that is potentially useful for tourists travelling to Hong Kong if they are deciding where to spend their time during their stay.

To no surprise, we determine that the top venue category among these districts is Chinese restaurants.

Tourists should travel to Central and Western if they want to experience scenic lookouts.

The Islands, Kowloon City, Kwai Tsing, and Wan Chai districts are most known for the number of restaurants.

Kwai Tsing has the most amount of Korean restaurants located nearby, while Sha Tin has the most Chinese restaurants nearby.

Tsuen Wan should be visited if tourists intend to visit a beach.

Using k-means clustering, we can determine there are 4 different clusters of districts, with 8 districts being part of one cluster. This shows that these 8 districts have similar venues among them. It is however, important to note that our analysis is only based on occurrence of venues among the districts. Price and reviews have not been taken into consideration.

5. Conclusion

When visiting a city for the first time, choosing the right place to stay can be overwhelming. Using data analysis, we were able to analyze the 17 major districts of Hong Kong and cluster them accordingly based on venue availability. Hopefully, these results can assist tourists to stay at a district that is most suitable to them.

Code can be found on GitHub.