Coursera Capstone
Battle of Neighborhoods in Ha Noi Old Quater

By: Đỗ Quang Anh

September/2019

1. Introduction and Objective

Hanoi is the capital of [Vietnam](#). It covers an area of 3,328.9 square kilometres (1,285 sq mi). With an estimated population of 7.7 million as of 2018, it is the [second largest city](#) in Vietnam. The [metropolitan area](#), encompassing nine additional neighbouring provinces, has an estimated population of 16 million. Located in the central area of the [Red River Delta](#), Hanoi is the commercial, cultural, and educational centre of [Northern Vietnam](#). Having an estimated nominal GDP of US$32.8 billion, it is the second most productive economic centre of Vietnam, following [Ho Chi Minh City](#).

The Old Quarter ([Vietnamese](#): *Phố cổ Hà Nội*) is the name commonly given to the historical civic urban core of Hanoi, located outside the [Imperial Citadel of Thăng Long](#). This quarter used to be the residential, manufacturing and commercial center, where each street was specialized in one specific type of manufacturing or commerce.

Another common name referring to approximately the same area is the 36 streets ([Vietnamese](#): *Hà Nội 36 phố phường*), after the 36 streets or guilds that used to make up the urban area of the city.

Business problem

My friend wanted to open a restaurant or a cafe in **Old Quarter**, but he didn't know where to open with little competition. This data analysis article will clarify and may help him with some useful information for his decision.

Target Audience of this project

This project is particularly useful to property developers and investor looking to open or invest in new business like restaurant, cafe, food, hotel in Ha Noi Old quarter

2. Data acquisition and cleaning

Ha Noi Neighborhoods Data

List of districts, wards of Ha Noi from the following URL: https://www.gso.gov.vn/dmhc2015/Default.aspx

Google map API

This project would use Google Map API Geocoder to get the Latitude and Longitude of each area

Foursquare API

This project would use Four-square API as its prime data gathering source. This API provides the ability to perform location search, location sharing and details about a business.

Step by Step

Downloaded data is excel file includes all districts and wards of Vietnam, filtering the data only for wards of Hanoi Old Quarter.

Next, I used Google API to get the longitude and latitude of each ward, to serve as input to the nearby Foursquare API search.

Then combine the longitude and latitude obtained with the districts and wards data.

```python
# define a function to get coordinates
def get_latlng(neighborhood):
    # initialize your variable to None
    lat_lng_coords = None
    # Loop until you get the coordinates
    while(lat_lng_coords is None):
        g = geocoder.arcgis('{}, Malaysia'.format(neighborhood))
        lat_lng_coords = g.latlng
    return lat_lng_coords

coords = [ get_latlng(neighborhood) for neighborhood in dfoldtown["area"].tolist() ]

# create temporary dataframe to populate the coordinates into Latitude and Longitude
df_oldtown_coords = pd.DataFrame(coords, columns=['Latitude', 'Longitude'])

df_oldtown_coords.head()
```

I then used the Foursquare API to search and continued to merge it into district and ward data. I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account to obtain account ID and API key to pull the data. From Foursquare, I can pull the names, categories, latitude, and longitude of the venues.

The Foursquare API returns all the venues, so to get only food and beverage service venues (bar, coffee, restaurant, food, wine, ice cream ...). I used the Pandas function to filter the result.

I used Pandas to make the horizontal bar chart of the top 20 Venues in Ha Noi Old Quarter:
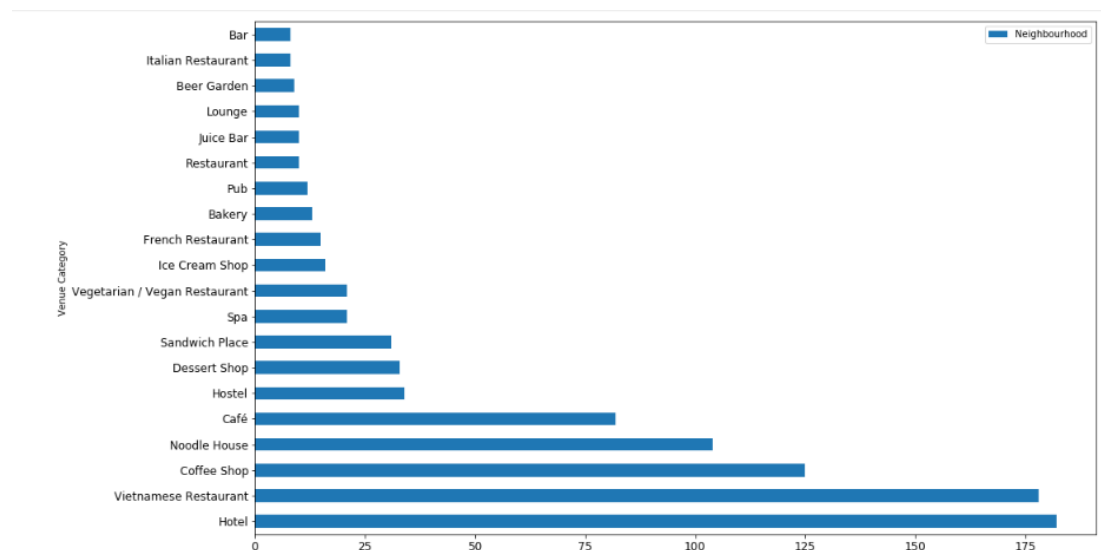
3. Methodology

After data acquisition and cleaning, this project applies K-mean clustering unsupervised machine learning algorithm to cluster the venues based on a list of locations for different types of food and beverage service points such as bars, cafes, Chinese restaurants, Vietnamese restaurants, Seafood restaurants, etc. This would give a better understanding of the similarities and dissimilarities between the chosen neighborhoods to retrieve more insights.

Analyze Each Neighborhood, group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Next, create the new data frame and display the top 10 venues for each neighborhood.

Then use the Kmean algorithm from the sklearn library to divide it into 5 groups with similar properties. Next, assign labels from Kmean result to each neighborhood using the Pandas merge function.

```
# set number of clusters
kclusters = 5

hn_grouped_clustering = HN_grouped.drop('Neighbourhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(hn_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
# to change use .astype()
```

The next step, I used folium to draw a map and folium to draw circle marker to visualize the resulting clusters.

```
# create map
map_clusters = folium.Map(location=[lat_HN, long_HN], zoom_start=11)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(HN_merged['Latitude'], HN_merged['Longitude'], HN_merged['area'], HN_merged['Cluster_Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```
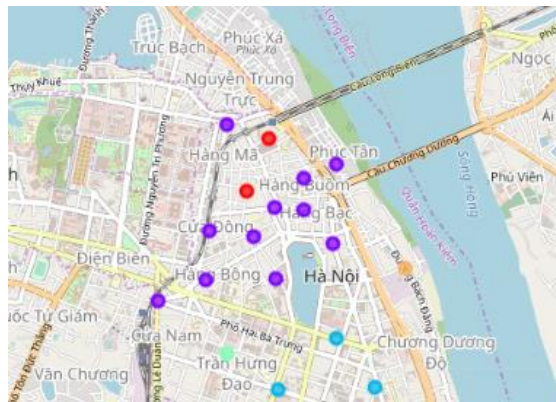
The next step is to display and view a list of each cluster, here to see what the clusters have in common, from which to draw useful information.

4. Results and Discussion

Cluster 1

```
HN_merged.loc[HN_merged['Cluster_Labels'] == 0, HN_merged.columns[[0] + list(range(5, HN_merged.shape[1]))]]
```

	ward	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Phường Đồng Xuân	0	Noodle House	Vietnamese Restaurant	Hostel	Hotel	Dessert Shop	Sandwich Place	Coffee Shop	Pub	Food	Tea Room
2	Phường Hàng Mã	0	Coffee Shop	Noodle House	Dessert Shop	Vietnamese Restaurant	Food	Cocktail Bar	Salad Place	Sandwich Place	Seafood Restaurant	Hotel
5	Phường Hàng Bồ	0	Hotel	Noodle House	Vietnamese Restaurant	Hostel	Sandwich Place	Coffee Shop	Dessert Shop	Café	Food	Massage Studio

Cluster 2

```
HN_merged.loc[HN_merged['Cluster_Labels'] == 1, HN_merged.columns[[0] + list(range(5, HN_merged.shape[1]))]]
```

	ward	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Phường Phúc Tân	1	Vietnamese Restaurant	Hotel	Noodle House	Coffee Shop	Café	Vegetarian / Vegan Restaurant	Hostel	Sandwich Place	Ice Cream Shop	Dessert Shop
3	Phường Hàng Buồm	1	Vietnamese Restaurant	Hotel	Noodle House	Coffee Shop	Café	Hostel	Sandwich Place	Dessert Shop	French Restaurant	Vegetarian / Vegan Restaurant
4	Phường Hàng Đào	1	Vietnamese Restaurant	Hotel	Noodle House	Coffee Shop	Hostel	Café	Sandwich Place	Vegetarian / Vegan Restaurant	Dessert Shop	Pub
6	Phường Cửa Đông	1	Hotel	Vietnamese Restaurant	Noodle House	Coffee Shop	Café	Dessert Shop	Spa	Hostel	Juice Bar	Bed & Breakfast
7	Phường Lý Thái Tổ	1	Vietnamese Restaurant	Hotel	Coffee Shop	Café	Noodle House	Spa	Sandwich Place	French Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant
8	Phường Hàng Bạc	1	Hotel	Vietnamese Restaurant	Coffee Shop	Noodle House	Café	Hostel	Sandwich Place	Dessert Shop	Vegetarian / Vegan Restaurant	Pub
9	Phường Hàng Gai	1	Hotel	Vietnamese Restaurant	Noodle House	Coffee Shop	Café	Hostel	Spa	Dessert Shop	Ice Cream Shop	Beer Garden
11	Phường Hàng Trống	1	Hotel	Café	Coffee Shop	Noodle House	Vietnamese Restaurant	Spa	Hostel	Market	Park	Mobile Phone Shop
12	Phường Cửa Nam	1	Vietnamese Restaurant	Hotel	Coffee Shop	Café	Bakery	Sandwich Place	Bar	Asian Restaurant	Korean Restaurant	Restaurant
13	Phường	1	Vietnamese	Hotel	Coffee Shop	Café	Noodle House	Restaurant	Ice Cream Shop	Sandwich Place	Hotpot	Japanese

Cluster 3

```
HN_merged.loc[HN_merged['Cluster_Labels'] == 2, HN_merged.columns[[0] + list(range(5, HN_merged.shape[1]))]]
```

	ward	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	Phường Tràng Tiền	2	Coffee Shop	Café	Vietnamese Restaurant	Ice Cream Shop	Restaurant	Hotel Bar	Hotel	History Museum	Italian Restaurant	Cultural Center
16	Phường Phan Chu Trinh	2	Vietnamese Restaurant	Restaurant	Café	Coffee Shop	Hotel	Ice Cream Shop	Cultural Center	Noodle House	Hotel Bar	Hotpot Restaurant
17	Phường Hàng Bài	2	Coffee Shop	Vietnamese Restaurant	Dessert Shop	Noodle House	Café	Hotel	Steakhouse	Wine Bar	Cultural Center	French Restaurant

Cluster 4

```
HN_merged.loc[HN_merged['Cluster_Labels'] == 3, HN_merged.columns[[0] + list(range(5, HN_merged.shape[1]))]]
```

	ward	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
15	Phường Trần Hưng Đạo	3	Hotel	Chocolate Shop	Supermarket	Vietnamese Restaurant	French Restaurant	Lounge	Cultural Center	Noodle House	Coffee Shop	Restaurant

Cluster 5

```
HN_merged.loc[HN_merged['Cluster_Labels'] == 4, HN_merged.columns[[0] + list(range(5, HN_merged.shape[1]))]]
```

	ward	Cluster_Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10	Phường Chương Dương	4	Café	Coffee Shop	Italian Restaurant	Seafood Restaurant	Smoothie Shop	Noodle House	Supermarket	Juice Bar	Nightclub	Vietnamese Restaurant

After reviewing the data of each cluster, I have some discussions:

- At Cluster 1, 2 ,3 focus mainly on Vietnamese restaurants, Cafe, Hotels. So need to be careful when you intend to open a Vietnamese restaurant or cafe
- At Cluster 4, mainly hotels and supermarkets. Cafe is only ranked 9 out of 10 most popular places, so it is possible to open a cafe in Cluster4
- Cluster 5, the most popular is cafe and Italian restaurant. However, the location here is quite limited in sight. Need to consider carefully when intending to open a new business in this area

5. Conclusion

Finally, I have got a small glimpse of how real-life data-science projects look like. I used various types of APIs to collect data, used the Pandas library to eliminate redundant data, used it, and used Python libraries to draw graphs, using unsupervised machine learning algorithms to group data into similar characteristics. From that it is possible to discover the information that is hidden in it, making it easier to make decisions such as where to open a restaurant or a cafe is appropriate and less competitive.

Final Notes

This is my assignment: a part of the IBM Data Science Course on Coursera.

The full project Jupiter Notebook from data scraping to preprocessing to results here: https://github.com/lego2003/My-Assignment/blob/master/battle-of-neighborhoods-in-hochiminh-city.ipynb