DISTRICT 2 – HCM CITY NEIGBORHOOD

BY DINH BAO CHAU

Business Problem

My sister – in - law wanted to open a restaurant or a cafe in District 2, Ho Chi Minh, but she didn't know where to open with little competition. This data analysis article will clarify and may help him with some useful information for her decision.

In this project we will try to find an optimal location for a restaurant or cafe. Specifically, this report will be targeted to stakeholders interested in opening an **Restaurant or Cafe** in **District 2, Ha Noi**, **Viet nam**.

We will use our data science powers to generate a few most promissing neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

DATA

Detail information of neighborhoods in District 2, list of districts, wards of District 2, Ho Chi Minh from the following URL: http://www.pso.hochiminhcity.gov.vn/web/guest/danhmucthongke-danhmuctinhthanhpho http://www.pso.hochiminhcity.gov.vn/web/guest/danhmucthongke-danhmucphuongxa or file data xls from the following: https://github.com/chaudb39/Capstone_Cousera/blob/e4b872054271da617fcb10566faa3ea8966df29a/HCM_DISTRICT2.xlsx

DATA

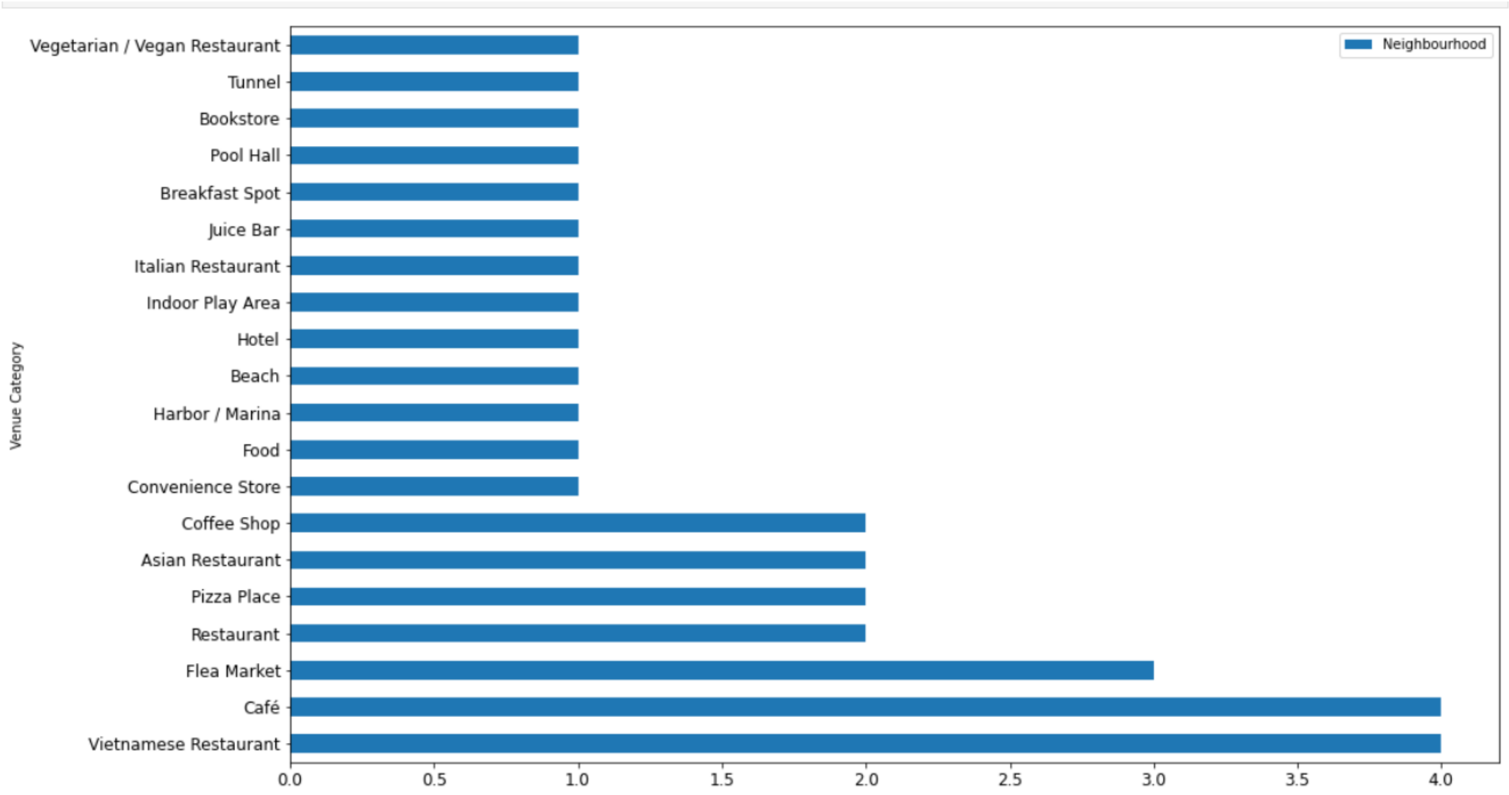
Google map API

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

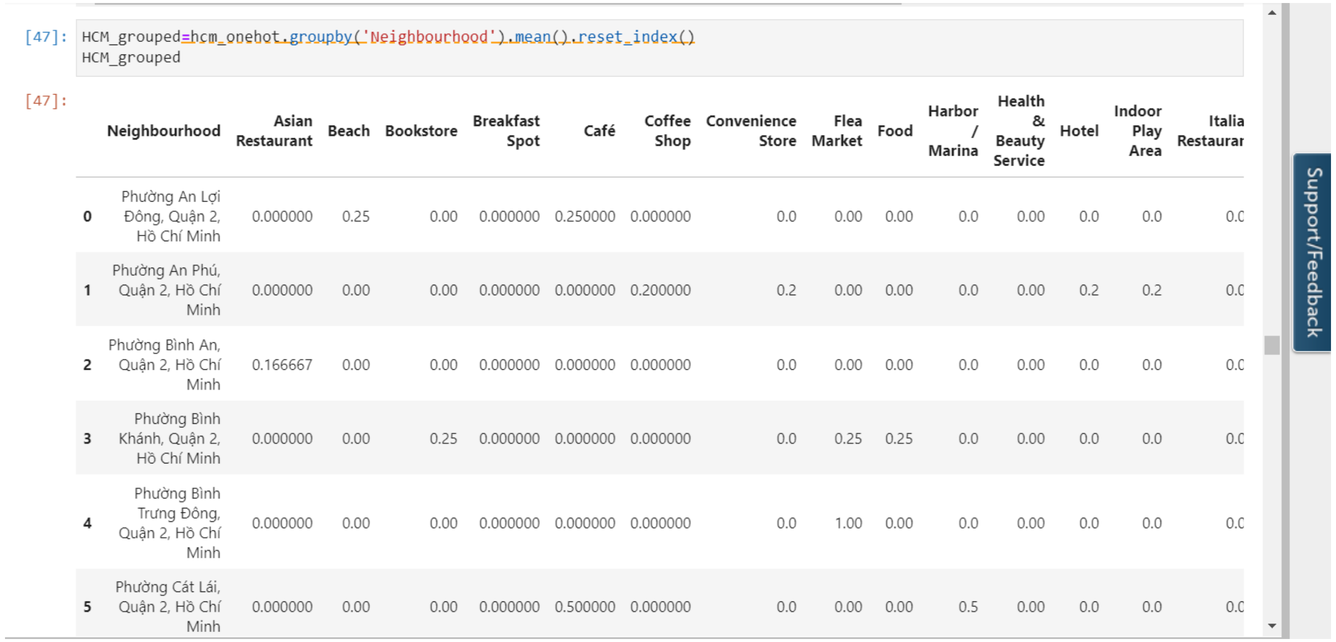
Foursquare API

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.

Char top Venue Category common



The frequency of occurrence of each category:

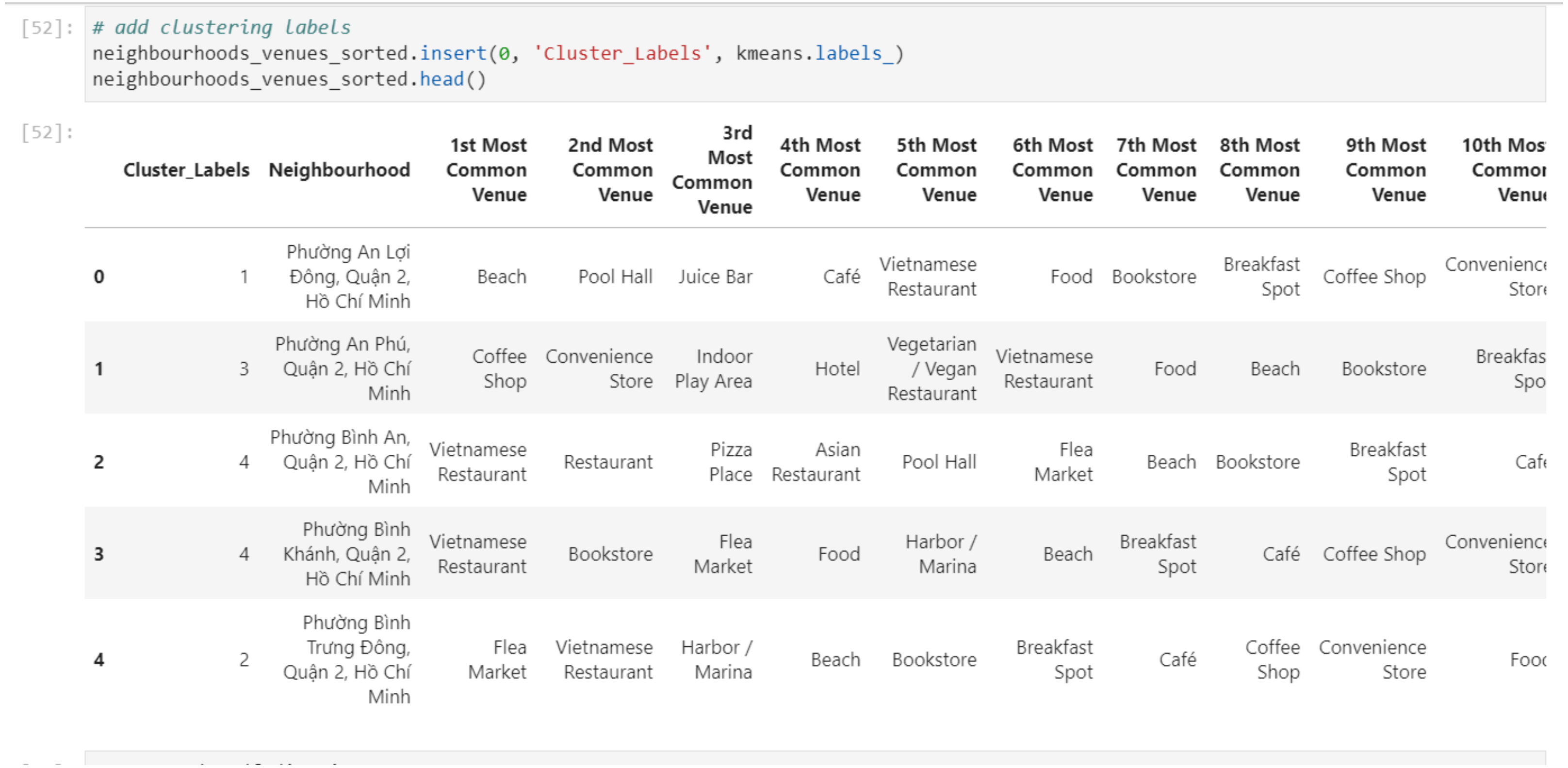


```
[47]: HCM_grouped=hcm_onehot.groupby('Neighbourhood').mean().reset_index()
      HCM_grouped
```

[47]:

	Neighbourhood	Asian Restaurant	Beach	Bookstore	Breakfast Spot	Café	Coffee Shop	Convenience Store	Flea Market	Food	Harbor / Marina	Health & Beauty Service	Hotel	Indoor Play Area	Italia Restaurar
0	Phường An Lợi Đông, Quận 2, Hồ Chí Minh	0.000000	0.25	0.00	0.000000	0.250000	0.000000	0.0	0.00	0.00	0.0	0.00	0.0	0.0	0.C
1	Phường An Phú, Quận 2, Hồ Chí Minh	0.000000	0.00	0.00	0.000000	0.000000	0.200000	0.2	0.00	0.00	0.0	0.00	0.2	0.2	0.C
2	Phường Bình An, Quận 2, Hồ Chí Minh	0.166667	0.00	0.00	0.000000	0.000000	0.000000	0.0	0.00	0.00	0.0	0.00	0.0	0.0	0.C
3	Phường Bình Khánh, Quận 2, Hồ Chí Minh	0.000000	0.00	0.25	0.000000	0.000000	0.000000	0.0	0.25	0.25	0.0	0.00	0.0	0.0	0.C
4	Phường Bình Trưng Đông, Quận 2, Hồ Chí Minh	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.0	1.00	0.00	0.0	0.00	0.0	0.0	0.C
5	Phường Cát Lái, Quận 2, Hồ Chí Minh	0.000000	0.00	0.00	0.000000	0.500000	0.000000	0.0	0.00	0.00	0.5	0.00	0.0	0.0	0.C

The top 10 venues for each neighborhood (CLUSTER):

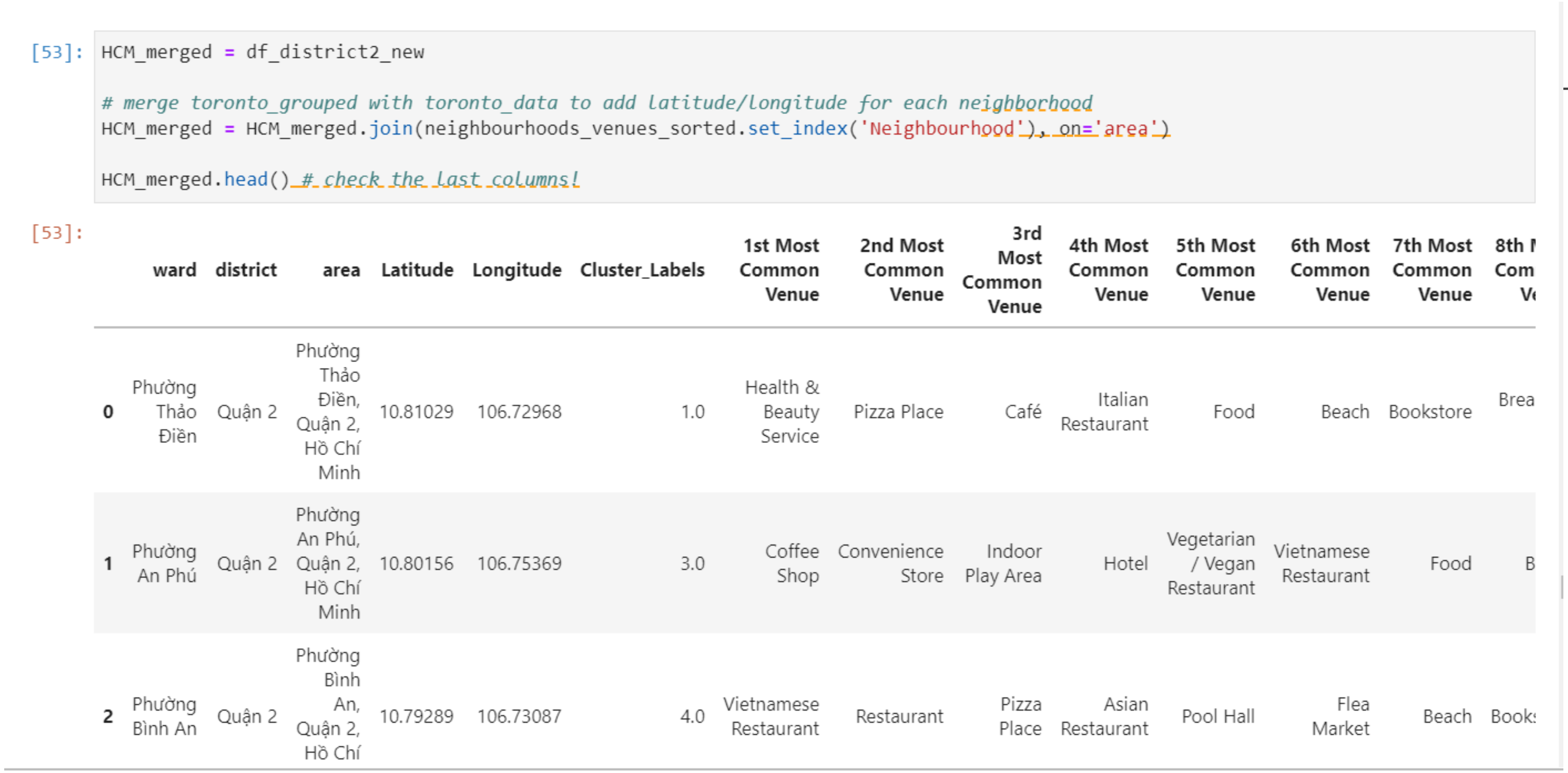


```
[52]: # add clustering labels
      neighbourhoods_venues_sorted.insert(0, 'Cluster_Labels', kmeans.labels_)
      neighbourhoods_venues_sorted.head()
```

[52]:

	Cluster_Labels	Neighbourhood	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 Mos Commo Venu
0	1	Phường An Lợi Đông, Quận 2, Hồ Chí Minh	Beach	Pool Hall	Juice Bar	Café	Vietnamese Restaurant	Food	Bookstore	Breakfast Spot	Coffee Shop	Convenienc Stor
1	3	Phường An Phú, Quận 2, Hồ Chí Minh	Coffee Shop	Convenience Store	Indoor Play Area	Hotel	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Food	Beach	Bookstore	Breakfas Spo
2	4	Phường Bình An, Quận 2, Hồ Chí Minh	Vietnamese Restaurant	Restaurant	Pizza Place	Asian Restaurant	Pool Hall	Flea Market	Beach	Bookstore	Breakfast Spot	Caf
3	4	Phường Bình Khánh, Quận 2, Hồ Chí Minh	Vietnamese Restaurant	Bookstore	Flea Market	Food	Harbor / Marina	Beach	Breakfast Spot	Café	Coffee Shop	Convenienc Stor
4	2	Phường Bình Trưng Đông, Quận 2, Hồ Chí Minh	Flea Market	Vietnamese Restaurant	Harbor / Marina	Beach	Bookstore	Breakfast Spot	Café	Coffee Shop	Convenience Store	Foo

The top 10 venues for each neighborhood (CLUSTER):



[53]:

```
HCM_merged = df_district2_new

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
HCM_merged = HCM_merged.join(neighbourhoods_venues_sorted.set_index('Neighbourhood'), on='area')

HCM_merged.head() # check the last columns!
```

[53]:

	ward	district	area	Latitude	Longitude	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 I Com Ve
0	Phường Thảo Điền	Quận 2	Phường Thảo Điền, Quận 2, Hồ Chí Minh	10.81029	106.72968	1.0	Health & Beauty Service	Pizza Place	Café	Italian Restaurant	Food	Beach	Bookstore	Brea
1	Phường An Phú	Quận 2	Phường An Phú, Quận 2, Hồ Chí Minh	10.80156	106.75369	3.0	Coffee Shop	Convenience Store	Indoor Play Area	Hotel	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Food	B
2	Phường Bình An	Quận 2	Phường Bình An, Quận 2, Hồ Chí	10.79289	106.73087	4.0	Vietnamese Restaurant	Restaurant	Pizza Place	Asian Restaurant	Pool Hall	Flea Market	Beach	Book

METHOGOGY

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

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

LINK NOTE BOOK:

https://github.com/chaudb39/Capstone_Cousera/blob/e4b872054271da617fcb10566faa3ea8966df29a/CourseraCapstone(Week2).ipynb

THANKS FOR YOUR WATCHING!