PHÒNG LẬP TRÌNH & MẠNG
TRUNG TÂM TIN HỌC
ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Đồ án:

ỨNG DỤNG CLUSTERING VÀ PHÂN TÍCH DỮ LIỆU VÀO VIỆC CHỌN ĐỊA ĐIỂM MỞ QUÁN ĂN VIỆT NAM TẠI TPHCM

Giáo viên hướng dẫn: Nguyễn Quan Liêm
Học viên thực hiện: Trần Hoàng Long

25/8/2021

Contents

1. Lời giới thiệu

Đồ án này được thực hiện để tìm một địa điểm thích hợp để mở một quán ăn Việt Nam bằng Data Science. Đồ án này hướng đến những người có ý định mở một nhà hàng Việt Nam tại TPHCM nhưng không biết nên mở ở quận nào.

Vì hiện tại số lượng nhà hàng Việt Nam ở TPHCM là nhiều vô kể , vậy nên trong Đồ Án này em sẽ tìm ra những khu vực mà có mật độ cạnh tranh quán ăn Việt Nam thấp nhất để giảm thiểu rủi ro cạnh tranh, ngoài ra em còn xem xét đến những yếu tố khách quan như chi phí bất động sản (nếu thuê, hoặc mua địa điểm kinh doanh), mật độ dân số (chẳng ai đi mở quán ăn kinh doanh ở nơi không có người) và mật độ công ty / nhà xưởng /... (để bán cho công nhân đến ăn sáng / ăn trưa/ ăn tối hoặc ghé quán để " lai rai". Tóm lại là sẽ phân tích và củng cố những findings của mình một cách thật là data driven.

Ngoài ra, đồ án này còn có thể được dùng làm cảm hứng ý tưởng cho nhiều loại hình kinh doanh khác nữa chứ không giới hạn ở việc mở một Quán Ăn Việt Nam.

2. Thu Thập và làm sạch Dữ liệu

2.1 Data sources

Những data cần thu thập là :

- Số quận huyện của TPHCM, cùng với kinh độ vĩ độ (để vẽ lên bản đồ)

- Mật Độ dân số , số công ty , nhà máy , xí nghiệp của từng quận.

- Mật độ những địa điểm của từng quận huyện của TPHCM (sẽ sử dụng G oogle Places API và Foursquare API)

2.2 Data cleaning

Những dữ liệu được thu thập về , ngoại trừ dữ liệu từ API Foursquare tất cả đều cần phải được ETL (đối với dữ liệu dạng JSON của Google Places API) hoặc process và cleansing để có thể đưa về dạng Dataframe.

```
pd.read_html('https://rentapartment.vn/dan-so-dien-tich-quan-tphcm/')
```

```
[          Quận  Dân số (người)  Diện tích (km²)  Số Phường/Xã  \
 0       Quận 1          205180             7.73            10
 1       Quận 2          168680            49.74            11
 2       Quận 3          196433             4.92            14
 3       Quận 4          203060             4.18            15
 4       Quận 5          187510             4.27            15
 5       Quận 6          258945             7.19            14
 6       Quận 7          324620            35.69            10
 7       Quận 8          451290            19.18            18
 8       Quận 9          397000           114.00            13
 9      Quận 10          372450             5.72            15
 10     Quận 11          332536             5.14            16
 11     Quận 12          520175            52.78            11
 12  Bình Thạnh          490618            20.76            20
 13     Thủ Đức          524670            48.00            12
 14      Gò Vấp          663313            19.74            16
 15   Phú Nhuận          182477             4.88            15
 16    Tân Bình          470350            22.38            15
 17     Tân Phú          464493            16.06            11
 18    Bình Tân          702650            51.89            10

    Mật độ dân số (người/km²)
 0                      26543
 1                       3391
 2                      39925
 3                      48578
 4                      43913
 5                      36014
 6                       9095
 7                      23529
 8                       3482
 9                      65113
 10                     64695
 11                      9855
 12                     23632
 13                     10930
 14                     33602
 15                     37392
 16                     21016
 17                     28922
 18                     13541  ,
```

Việc clean data có thể gồm nhiều task và nhiều bước, gồm cả những việc đơn giản như tách khoảng trống trong chuỗi cho tới những việc đặc thù như chuyển đổi tiếng Việt từ có dấu thành không dấu để thư viện Geocoder có thể hiểu được. Phần lớn việc clean data và chuẩn hóa dữ liệu đến từ việc phải lọc lại các dữ liệu trong content được crawl về từ beautiful soup và đưa vào trong các list, từ list chuyển đổi thành dataframe.

```
url = 'https://thongtindoanhnghiep.co/tim-kiem?location=%2Ftp-ho-chi-minh&kwd='
```

```
def get_page_content(url):
   page = requests.get(url,headers={"Accept-Language":"en-US"})
   return BeautifulSoup(page.text,"html.parser")
soup = get_page_content(url)
```

```
soup
```

```
<a href="/tp-ho-chi-minh/quan-7">Quận 7</a>
</li>
<li>
<span class="badge badge-u">13032</span>
<a href="/tp-ho-chi-minh/quan-8">Quận 8</a>
</li>
<li>
<span class="badge badge-u">13027</span>
<a href="/tp-ho-chi-minh/quan-9">Quận 9</a>
</li>
<li>
<span class="badge badge-u">29051</span>
<a href="/tp-ho-chi-minh/quan-binh-tan">Quận Bình Tân</a>
</li>
<li>
<span class="badge badge-u">35668</span>
<a href="/tp-ho-chi-minh/quan-binh-thanh">Quận Bình Thạnh</a>
</li>
<li>
<span class="badge badge-u">31804</span>
```

Tuy nhiên, dữ liệu lấy được từ API Google Places lại là dạng JSON, lists of Dictionary , cần ETL rất nhiều, nhưng hiệu suất Query lại rất thấp do Google chỉ cho query tối đa 20 kết quả 1 lần, 80 lần là tối đa cho 1 địa điểm. Nên nếu cố gắng query nhiều lần để lấy nhiều kết quả có thể dẫn đến bị duplicate hoặc bị miss dữ liệu. Vì vậy nên em không sử dụng Google Places API mà sử dụng Foursquare API

```
In [271]: map_client = googlemaps.Client(API_KEY)
          response2 = map_client.places_nearby(location = '10.780960,106.699110', radius = 2000)

In [272]: business_list = []

In [274]: import time

In [276]: business_list.extend(response2.get('results'))
          next_page_token = response2.get('next_page_token')
          while next_page_token:
            time.sleep(3)
            response = map_client.places_nearby(location = '10.780960,106.699110', radius = 2000, page_token = next_page_token)
            business_list.extend(response2.get('results'))
            next_page_token = response.get('next_page_token')

In [283]: pprint.pprint(business_list)
```

```
  'icon_mask_base_uri': 'https://maps.gstatic.com/mapfiles/place_api/icons/v2/hotel_pinlet',
  'name': 'Novotel Saigon Centre',
  'opening_hours': {'open_now': True},
  'photos': [{'height': 315,
              'html_attributions': ['<a '
                                    'href="https://maps.google.com/maps/contrib/105084937406477605018">Novotel '
                                    'Saigon Centre</a>'],
              'photo_reference': 'Aap_uEDTItGN_kTEvDn1gwA0Go9sP5RNQTIL6rvzwHIXK5fyl6XHrTzeEtbsV-vX3BI02oC8tFFYpqvz1FgEnh_7txx
nRBB8ievnxC8JGQFdt4nFWcD_T8vBHo9813h8FftCUjk8C4e2Ycwwu7ieXNk936Tb41wJ-nehEnJTfePkrgjKURBq',
              'width': 815}],
  'place_id': 'ChIJv3u21EkvdTERd8RlwzZlJMI',
  'plus_code': {'compound_code': 'QMMW+P9 District 3, Ho Chi Minh City, '
                                 'Vietnam',
                'global_code': '7P28QMMW+P9'},
  'rating': 4.4,
  'reference': 'ChIJv3u21EkvdTERd8RlwzZlJMI',
  'scope': 'GOOGLE',
  'types': ['lodging', 'point_of_interest', 'establishment'],
  'user_ratings_total': 2181,
  'vicinity': '167 Hai Bà Trưng , Phường 6'}]
```

Thành quả sau khi đã xử lý dữ liệu

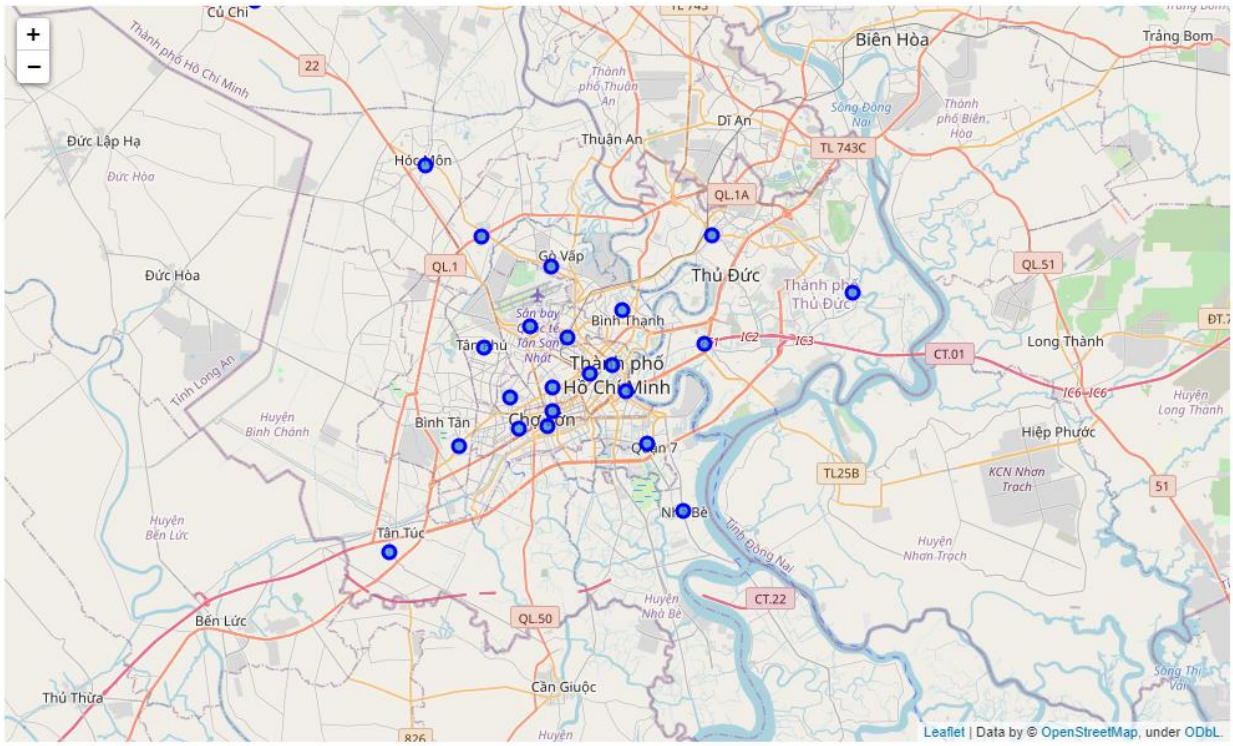
	District	Latitude	Longitude	Population	Square	Ward	Density	Average Housing Price (1M VND)/m2	Total Companies
0	Binh Chanh District	10.679220	106.576540	680000	253.00	0	16	24.1	16227
1	Binh Tan District	10.736840	106.614480	702650	51.89	10	13541	74.5	47222
2	Binh Thanh District	10.810520	106.705050	490618	20.76	20	23632	128	17140
3	Can Gio District	10.415660	106.961300	74960	704.00	0	7	18.4	671
4	Cu Chi District	10.977340	106.502230	403038	435.00	0	21	9.6	5995
5	District 1	10.780960	106.699110	205180	7.73	10	26543	441	10427
6	District 2	10.791990	106.749850	168680	49.74	11	3391	106	27134
7	District 3	10.775650	106.686720	196433	4.92	14	39925	254	12180
8	District 4	10.766700	106.706470	203060	4.18	15	48578	78.8	23411
9	District 5	10.755690	106.666370	187510	4.27	15	43913	241	7453
10	District 6	10.745970	106.647690	258945	7.19	14	36014	121	11739
11	District 7	10.738291	106.718292	324620	35.69	10	9095	98.2	9397
12	District 8	10.747710	106.663340	451290	19.18	18	23529	88.1	22258
13	District 9	10.820050	106.831820	397000	114.00	13	3482	76.4	13032
14	District 10	10.768830	106.665990	372450	5.72	15	65113	211	13027
15	District 11	10.763160	106.643140	332536	5.14	16	64695	161	29051
16	District 12	10.850440	106.627310	520175	52.78	11	9855	54.7	35668
17	Go Vap District	10.833790	106.665560	663313	19.74	16	33602	101	31804
18	Hoc Mon District	10.888360	106.596400	422471	109.00	0	12	25.5	13890
19	Nha Be District	10.701530	106.738180	175360	100.00	0	7	57.5	4625
20	Phu Nhuan District	10.795650	106.674640	182477	4.88	15	37392	186	17996
21	Tan Binh District	10.801484	106.654077	470350	22.38	15	21016	139	41604
22	Tan Phu District	10.790069	106.628524	464493	16.06	11	28922	102	28159
23	Thu Duc District	10.850955	106.753941	524670	48.00	12	10930	76	20378

Dữ liệu hoàn chỉnh từ Foursquare API

```
hcm_venues
```

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Binh Chanh District	10.679220	106.576540	Kedai Sarah	10.688974	106.574965	Women's Store
1	Binh Chanh District	10.679220	106.576540	Lò Bánh Mì Vạn Hoà	10.665982	106.570857	Bakery
2	Binh Chanh District	10.679220	106.576540	Xí Nghiep Sx Hang Thu Cong My Nghe 27-7	10.683414	106.562306	Arts & Crafts Store
3	Binh Chanh District	10.679220	106.576540	National Road 1A	10.683168	106.561552	Bus Station
4	Binh Chanh District	10.679220	106.576540	Ốc chị Lượm	10.663730	106.570333	Seafood Restaurant
...	...	...	...	...	...	...	...
1150	Thu Duc District	10.850955	106.753941	Amy coffee shop	10.844842	106.767268	Coffee Shop
1151	Thu Duc District	10.850955	106.753941	Hội Ngộ quán	10.852376	106.768717	Restaurant
1152	Thu Duc District	10.850955	106.753941	Ti Tach Cafe	10.861972	106.763950	Café
1153	Thu Duc District	10.850955	106.753941	Binh Quoi Ferry	10.837140	106.745524	Boat or Ferry
1154	Thu Duc District	10.850955	106.753941	Bến Đò Bình Quới	10.835886	106.745588	Boat or Ferry

Ngoài ra em cũng sử dụng Folium để marker và plot các quận huyện của TPHCM lên bản đồ, cùng với các cluster market sau đó



3. Exploratory Data Analysis

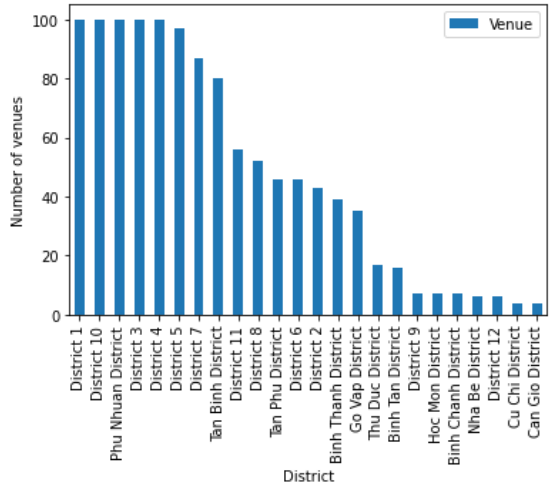
Sau khi tiến hành một vài phương pháp phân tích dữ liệu cơ bản và thống kê , có những thông tin sau.

Những quận có nhiều điểm kinh doanh nhất

Số địa điểm từng quận

```
ax = hcm_venues_group.sort_values(by="Venue", ascending=False).plot(x="District", y="Venue", kind="bar")
ax.set_ylabel("Number of venues")
```

```
Text(0, 0.5, 'Number of venues')
```

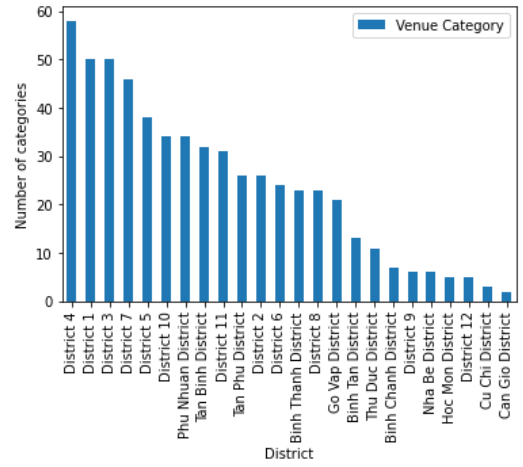


Những Quận có nhiều loại hình kinh doanh nhất

Số loại địa điểm từng Quận

```
hcm_venues_category = (
    hcm_venues.groupby(['District','Venue Category'])
        .count().reset_index()[['District', 'Venue Category']]
            .groupby('District').count().reset_index()
)
# hcm_venues_group_cat
ax = hcm_venues_category.sort_values(by="Venue Category", ascending=False).plot(x="District", y="Venue Category", kind="bar")
ax.set_ylabel("Number of categories")
```

```
Text(0, 0.5, 'Number of categories')
```



Những loại hình kinh doanh được gặp nhiều nhất

Tần số của các loại địa điểm

```
most_venues = hcm_venues.groupby('Venue Category').count().sort_values(by="Venue", ascending=False)
```

```
most_venues.head(10)
```

Venue Category	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude
Café	132	132	132	132	132	132
Vietnamese Restaurant	127	127	127	127	127	127
Coffee Shop	85	85	85	85	85	85
Hotel	49	49	49	49	49	49
Chinese Restaurant	49	49	49	49	49	49
Asian Restaurant	31	31	31	31	31	31
Seafood Restaurant	27	27	27	27	27	27
Vegetarian / Vegan Restaurant	27	27	27	27	27	27
Dessert Shop	25	25	25	25	25	25
Multiplex	25	25	25	25	25	25

Có thể thấy nhiều nhất chính là nhà hàng Việt Nam và quán ăn có kết hợp với uống cà phê (như kiểu cơm trưa văn phòng ,...) , thứ 3 là quán Cà Phê và thứ 4 là Khách sạn.

Lưu ý : vì khái niệm Café và Restaurant đôi khi nó giống nhau, nên ở Việt Nam có khi những nhà hàng Việt Nam cũng bị nhầm lẫn là Café. Tuy nhiên, đồ án này chỉ tập trung vào " Vietnamese Restaurant" để có kết quả khách quan nhất.

Top 10 loại hình dịch vụ theo từng Quận

```
#TOP 10 VENUES CATEGORIES FOR EACH DISTRICT
```

	District	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	Binh Chanh District	Women's Store	Tourist Information Center	Diner	Seafood Restaurant	Bakery	Arts & Crafts Store	Bus Station	Electronics Store	Dumpling Restaurant	Dim Sum Restaurant
1	Binh Tan District	Café	Coffee Shop	Shopping Mall	Food Court	Residential Building (Apartment / Condo)	Sandwich Place	Department Store	Bubble Tea Shop	Multiplex	Fabric Shop
2	Binh Thanh District	Café	Vietnamese Restaurant	Asian Restaurant	Coffee Shop	Multiplex	Seafood Restaurant	French Restaurant	Diner	Gym / Fitness Center	Food Court
3	Can Gio District	Beach	Vietnamese Restaurant	Yoga Studio	Design Studio	Fabric Shop	Electronics Store	Dumpling Restaurant	Diner	Dim Sum Restaurant	Dessert Shop
4	Cu Chi District	Vietnamese Restaurant	Restaurant	Café	Department Store	Electronics Store	Dumpling Restaurant	Diner	Dim Sum Restaurant	Dessert Shop	Design Studio
5	District 1	Vietnamese Restaurant	Hotel	Coffee Shop	Café	Massage Studio	Vegetarian / Vegan Restaurant	Pizza Place	Dessert Shop	French Restaurant	Spa
6	District 10	Vietnamese Restaurant	Café	Coffee Shop	Chinese Restaurant	Vegetarian / Vegan Restaurant	Noodle House	Hotel	Dessert Shop	Seafood Restaurant	Bookstore
7	District 11	Café	Chinese Restaurant	Seafood Restaurant	Vietnamese Restaurant	Dessert Shop	Cantonese Restaurant	Multiplex	Residential Building (Apartment / Condo)	Basketball Stadium	Mobile Phone Shop
8	District 12	Restaurant	Gym	Supermarket	Café	Asian Restaurant	Coffee Shop	Convenience Store	Convention Center	Cupcake Shop	Cocktail Bar
9	District 2	Café	Asian Restaurant	Restaurant	Coffee Shop	Bistro	Juice Bar	Burger Joint	French Restaurant	BBQ Joint	Vegetarian / Vegan Restaurant
10	District 3	Vietnamese Restaurant	Hotel	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Dessert Shop	Pizza Place	Hostel	Beer Bar	Spa
11	District 4	Hotel	Vietnamese Restaurant	Spa	Massage Studio	Café	Coffee Shop	Indian Restaurant	Bar	Noodle House	Burger Joint

Loại hình dịch vụ phổ biến nhất của các quận phần lớn vẫn là Nhà Hàng, quán ăn.

4. Preprocessing

Trong phần đồ án này , task preprocessing chính vẫn là convert categorical data thành categorical interger data (để phục vụ cho Kmean clustering) và convert ordinal data thành categorical data (để phục vụ cho Kmode clustering)

Về task convert categorical data thành interget categorical data, sử dụng onehot encoder để biến những giá trị categorical như “Vietnamese Restaurant” ... thành những giá trị như 0 – tức no , 1 – tức yes. Sau đó sẽ dùng giá trị đó để xây dựng các cluster

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

# add district column back to dataframe
hcm_onehot['District'] = hcm_venues['District']

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

# group the rows by district and by taking the mean of the frequency of occurrence of each category
hcm_grouped = hcm_onehot.groupby('District').mean().reset_index()
hcm_grouped.head()
# group rows by neighborhood order by the mean of the frequency of occurrence of each category
```

	District	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Arts & Crafts Store	Asian Restaurant	...	Turkish Restaurant	Udon Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Wareho St
0	Binh Chanh District	0.0	0.0	0.0	0.0	0.0	0.0000	0.0	0.142857	0.000000	...	0.0	0.0	0.000000	0.000000	
1	Binh Tan District	0.0	0.0	0.0	0.0	0.0	0.0625	0.0	0.000000	0.000000	...	0.0	0.0	0.000000	0.062500	
2	Binh Thanh District	0.0	0.0	0.0	0.0	0.0	0.0000	0.0	0.000000	0.102564	...	0.0	0.0	0.025641	0.102564	
3	Can Gio District	0.0	0.0	0.0	0.0	0.0	0.0000	0.0	0.000000	0.000000	...	0.0	0.0	0.000000	0.250000	
4	Cu Chi District	0.0	0.0	0.0	0.0	0.0	0.0000	0.0	0.000000	0.000000	...	0.0	0.0	0.000000	0.500000	

5 rows × 141 columns

Về task convert ordinal data thành categorical data, sử dụng phương pháp IQR, tính phân bổ hay độ trải để chấm điểm và xếp hạng giá trị thành các categorical data.

Các bước thực hiện :
1. Count số lần xuất hiện của venue
2. Tiến hành thống kê, xem xét phân bổ
3. Tiến hành chấm điểm mức độ

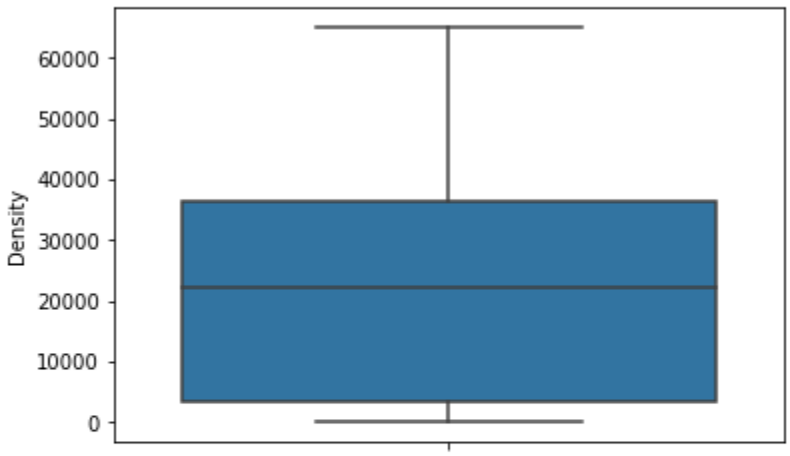
```
df1['Venue_Marks'].describe()
```

```
count    18.000000
mean      7.055556
std       6.310330
min       1.000000
25%       2.250000
50%       4.500000
75%       9.750000
max      21.000000
Name: Venue_Marks, dtype: float64
```

Kết hợp với sử dụng boxplot để convert

```
sns.boxplot(y=AHP_Result['Density'].astype(str).astype(float))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1bba08deb80>
```




```
AHP_Result['Density'] = AHP_Result['Density'].astype(str).astype(float)
```

Range of Average Population Density

- Low : 7 ≤ APD < 3459.25
- Medium : 3459.25 ≤ APD < 22272.5
- High : 22272.5 ≤ APD < 36358
- Very High : 36358 ≤ APD

```
data_kmode['Type'] = data_kmode.apply(lambda x: 'Low' if (x['Venue_Marks'] >= 1 and x['Venue_Marks'] < 3)
                                      else ('Medium' if (x['Venue_Marks'] >= 3 and x['Venue_Marks'] < 5)
                                            else ('High' if (x['Venue_Marks'] >= 5 and x['Venue_Marks'] < 9)
                                                  else 'Very High')), axis = 1)
```

```
data_kmode
```

	District	Venue_Marks	Type
0	Cu Chi District	2.0	Low
1	Binh Chanh District	0.0	Very High
2	Nha Be District	0.0	Very High
3	Hoc Mon District	0.0	Very High
4	District 9	0.0	Very High
5	District 8	4.0	Medium
6	District 7	5.0	High
7	District 6	2.0	Low
8	District 5	9.0	Very High
9	District 4	6.0	High
10	District 12	0.0	Very High
11	District 11	3.0	Medium
12	Binh Tan District	1.0	Low
13	District 2	1.0	Low
14	Thu Duc District	0.0	Very High
15	District 3	13.0	Very High
16	Tan Phu District	6.0	High

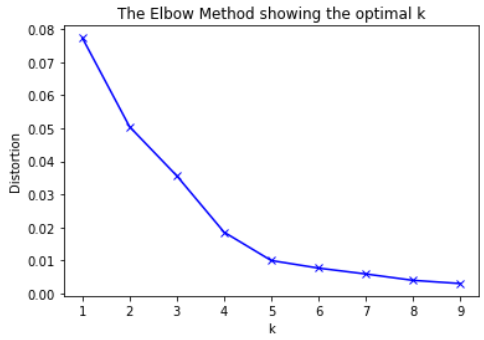
5. Modelling

5.1 Elbow Test

- Vì sử dụng thuật toán Clustering nên phải tìm điểm Elbow
- Thuật Toán Kmeans và Kmode đều cho kết quả Elbow tại k=4

- Kmean

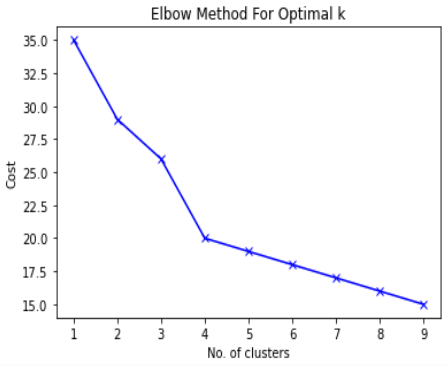
```
# Use the K-Means clustering to do this but first we need to determine how many k we need to use. The *"elbow" method* helps to j
# try with 10 different values of k to find the best one
Ks = 10
distortions = []

hcm_restaurant_clustering = hcm_grouped_restaurant.drop('District', 1)

for k in range(1, Ks):

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

    # find the distortion w.r.t each k
    distortions.append(
        sum(np.min(cdist(hcm_restaurant_clustering, kmeans.cluster_centers_, 'euclidean'), axis=1))
        / hcm_restaurant_clustering.shape[0]
    )

plt.plot(range(1, Ks), distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



- Kmode

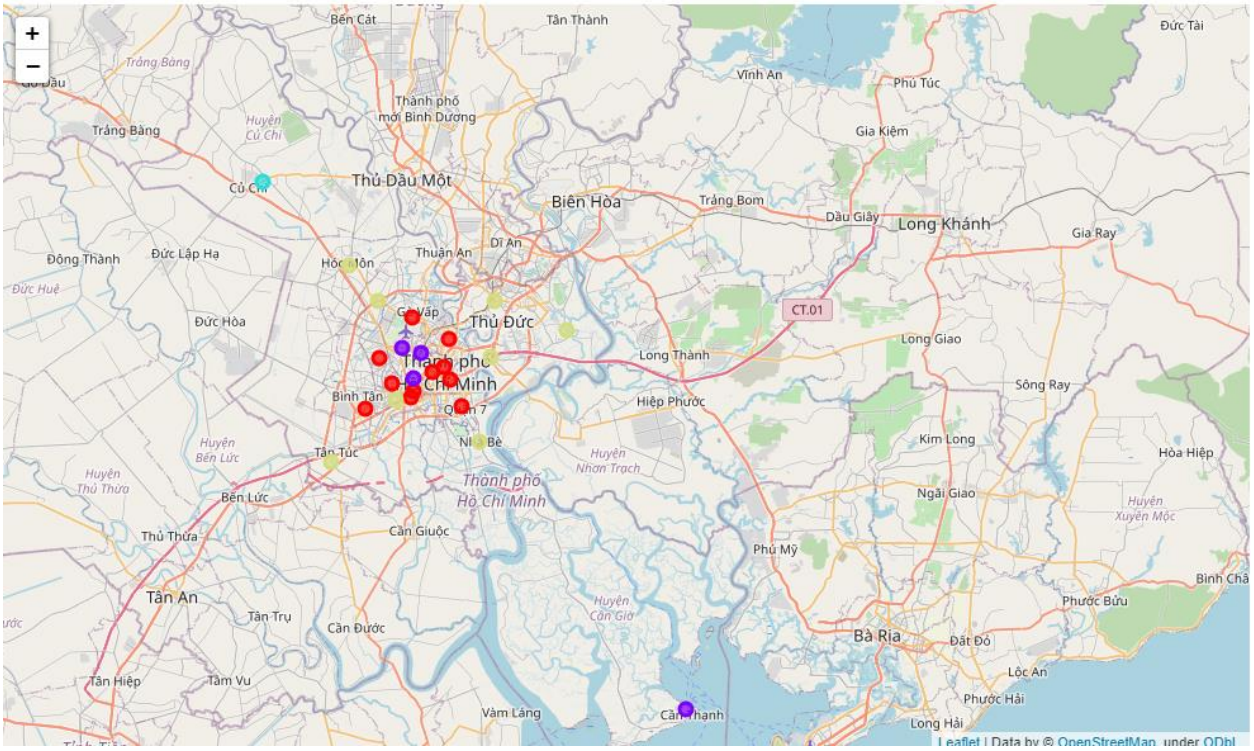
In [244]:
```
# Elbow curve to find optimal K
cost = []
K = range(1,10)
for num_clusters in list(K):
    kmode = KModes(n_clusters=num_clusters, init = "random", n_init = 5, verbose=1)
    kmode.fit_predict(data_kmode)
    cost.append(kmode.cost_)

plt.plot(K, cost, 'bx-')
plt.xlabel('No. of clusters')
plt.ylabel('Cost')
plt.title('Elbow Method For Optimal k')
plt.show()
```

```
Run 5, iteration: 1/100, moves: 0, cost: 16.0
Best run was number 3
```



5.2 Clustering

Kmeans

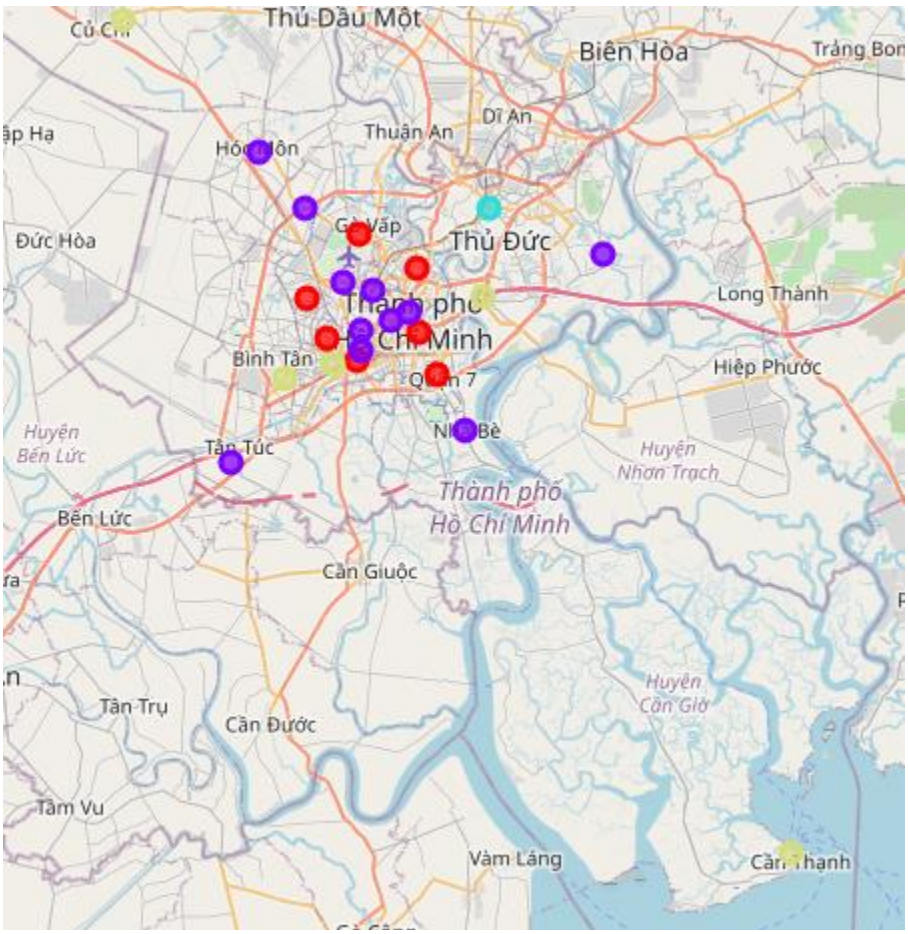


Củ Chi và Cần Giờ là 2 outlier

có 3 cluster rõ rệt :

+ Cluster 0 : những quận phát triển , ở trung tâm
+ Cluster 1 : những quận ở rìa thành phố, giáp với sân bay
+ Cluster 2 là Củ Chi : outliner nên tạm bỏ qua.
+ Cluster 3 là những quận/ huyện ngoại thành

Kmode



Giờ đây Thủ Đức là 1 cluster riêng
Cần Giờ và Củ Chi không còn là outliner

3 cluster còn lại khác biệt rõ rệt :

+ Cluster 0 : những quận phát triển , ở trung tâm

+ Cluster 1 : những quận ở rìa trung tâm, ngoài ra có những huyện ở rìa thành phố nhưng có mật độ dân số cao cũng được vào cluster 1

+ Cluster 2 là Thủ Đức: Thủ Đức đứng 1 cluster riêng cũng hợp lý vì đây là quận mới lên Thành phố. Tương lai sẽ là một khu riêng biệt phát triển độc lập

+ Cluster 3 là những quận/ huyện ngoại thành

5.3 Further Analysis

- Những yếu tố khác :
  - Average Housing Price (chi phí bds)
  - Average Population Density (Mdds)
  - Total Companies (Tổng số công ty, văn phòng cơ quan,..)

District	Average Housing Price (1M VND)/m2	AHP_Level	APD_Level	Cluster Labels	Total Companies
Binh Chanh District	24.1	Low	Low	3	16227
Binh Tan District	74.5	Medium	Medium	0	47222
Binh Thanh District	128.0	High	High	0	17140
Can Gio District	18.4	Low	Low	1	671
Cu Chi District	9.6	Low	Low	2	5995
District 1	441.0	Very High	High	0	10427
District 2	106.0	High	Low	3	27134
District 3	254.0	Very High	Very High	0	12180
District 4	78.8	Medium	Very High	0	23410
District 5	241.0	Very High	Very High	0	7453
District 6	121.0	High	High	3	11739
District 7	98.2	Medium	Medium	0	9397
District 8	88.1	Medium	High	0	22258
District 9	76.4	Medium	Medium	3	13032
District 10	211.0	Very High	Very High	1	13027
District 11	161.0	Very High	Very High	0	29051
District 12	54.7	Low	Medium	3	35668
Go Vap District	101.0	High	High	0	31804
Hoc Mon District	25.5	Low	Low	3	13890
Nha Be District	57.5	Low	Low	3	4625
Phu Nhuan District	186.0	Very High	Very High	1	17996
Tan Binh District	139.0	High	Medium	1	41604
Tan Phu District	102.0	High	High	0	28159

6. Findings

6.2 Discussion

Thực tế, ta chỉ muốn những quận có chi phí bất động sản vừa phải , hoặc "hời hơn" nếu so với mật độ dân số (nghĩa là APD level cao hơn APH Level), chứ không ai muốn những quận có chi phí bất động sản quá cao nhưng mật độ dân số thì lại thấp. Hoặc những quận có mật độ dân số cao , chi phí bất động sản cũng cao không kém thì cũng chẳng có lợi ích gì.

Sau khi phân tích các yếu tố phụ và mức độ cạnh tranh kết hợp với phương pháp Thống kê cơ bản. Em lọc được 3 quận có mối quan hệ giữa chi phí nhà đất và mật độ dân số khá hấp dẫn.

```
AHP_Level  APD_Level
Very High  Very High    5
Low        Low          5
Medium     Medium       4
High       High         4
Very High  High         1
Medium     Very High    1
           High         1
Low        Medium       1
High       Medium       1
           Low          1
dtype: int64
```

AHP Level Low - APD Level High : Quận 12 (chi phí nhà thấp, mật độ dân số cao)
AHP Level Medium - APD Level High : Quận 8 (chi phí nhà tương đối, mật độ dân số cao)
AHP Level Medium - APD Level Very High : Quận 4 (chi phí nhà tương đối , mật độ dân số rất cao)

District	Average Housing Price (1M VND)/m2	AHP_Level	APD_Level	Cluster Labels	Total Companies
District 4	78.8	Medium	Very High	0	23411

District	Average Housing Price (1M VND)/m2	AHP_Level	APD_Level	Cluster Labels	Total Companies
District 8	88.1	Medium	High	0	22258

District	Average Housing Price (1M VND)/m2	AHP_Level	APD_Level	Cluster Labels	Total Companies
District 12	54.7	Low	Medium	3	35668

Nhắc lại ý nghĩa cluster :

* Cluster 0 : Có ít cạnh tranh.
* Cluster 1 : Có sự cạnh tranh tương đối.
* Cluster 2 : Có sự cạnh tranh cao

Mặc dù quận 8 có giá bất động sản ở mức tương đối và mật độ dân cư cao, tuy nhiên vẫn không hấp dẫn do không có nhiều công ty và nhà xưởng hơn quận 4 là mấy, và mức cạnh tranh cũng rất cao

Việc mở quán ăn ở quận 4, mặc dù có lợi thế là giá bất động sản ở mức rẻ hơn so với quận 8 và mật độ dân cư cao hơn, tuy nhiên, mật độ công ty / nhà xưởng trong khu vực này cũng không nổi trội hơn quận 8 và mức cạnh tranh cũng rất cao không khác biệt nên khá rủi ro khi kinh doanh tại đây.

việc mở 1 quán ăn Việt Nam ở quận 12 sẽ là tốt nhất vì có mật độ dân số tương đối, mật độ công ty , nhà xưởng vào mức cao trong khi giá bất động sản lại chỉ bằng 3/5 so với quận 4 và quận 8

Vậy , lựa chọn tốt nhất là nên mở một quán ăn Việt Nam ở quận 12 vì chi phí bất động sản rẻ, mật đô dân cư, công ty & nhà máy xí nghiệp tương đối cao và mức độ canh tranh thấp.

6.2 Conclusion

Mục đích của Đồ Án này là tìm ra quận hợp lý nhất để mở một nhà hàng Việt Nam dựa trên các tiêu chí : Chi Phí cố định vừa phải & kinh tế (chi phí bất động sản), mật độ dân cư đông và số lượng công ty , nhà máy xí nghiệp nhiều để không chỉ kinh doanh cho dân địa phương mà còn cho công nhân của các nhà máy , nhân viên của các công ty do các đối tượng này có xu hướng ra ngoài ăn trưa, ghé ăn sáng ,... hoặc đi làm về ghé ăn nhậu. Sau khi tiến hành phân tích dữ liệu từ nhiều nguồn và xây dựng mô hình phân tích cụm thì quận hợp lý nhất để mở một nhà hàng Việt Nam là quận 12.

6.3 Comments

Ngoài ra, Kết quả này là một sự dự đoán khách quan từ dữ liệu kết hợp với cả suy luận theo tình hình thực tế dựa trên kết quả của dữ liệu. Đương nhiên sẽ không thể tránh khỏi những thiếu sót như mật độ quán ăn chưa phản ánh đúng thực tế (vì tình hình quán ăn của Việt Nam đôi khi không có trong dữ liệu của Foursquare, hoặc dữ liệu từ Foursquare dán nhãn sai, ví dụ như quán ăn cơm trưa văn phòng Việt Nam nhưng lại dán nhãn Cà phê,... Tuy nhiên ban đầu việc collect dữ liệu đã lấy ngẫu nhiên 100 địa điểm ở từng quận, nếu quận nào có mật độ nhà hàng Việt Nam nhiều đương nhiên kết quả sẽ ra nhiều nên cũng hạn chế được vấn đề này.