Coursera Capstone Project : Applied Data Science

Jiajun Zhou
zhouthomas177@gmail.com

1 Introduction

Nowadays, times are exceptionally difficult for our city's restaurants right now, with the spread of the coronavirus, and the unfortunate but necessary social distancing requirements, currently providing many obstacles for Hong Kong's stellar eateries. In light of the current situation, I think that recognition and support of our restaurant scene is particularly important right now. With nearly 7 million inhabitants and the highest metropolitan GDP in the world, an estimated 60 million overseas or mainland visitors are expected to flock to the Hong Kong.

Hong Kong is well-known as one of restaurant capitals of the world. Hong Kong has hit the record high of having 15,000 licensed restaurants, cafes and bars by year end of 2017, with 20.4 restaurants per 10,000 people. It is one of the highest density city of restaurants in the world. Hong Kong is the first batch of Asia's cities to be included in the Michelin's Guide Hong Kong.

I believe it’s difficult for a travelers, especially restaurant-goers, to make a choice from among many options since there is also too much information on the web because everybody’s got their own take of where to go and it’s all so fragmented that you have to assemble it yourself especially if you’re wanting non-touristy recommendations.

2 Business Problem

I have done that by updating the list districts around the city that have been doing great things, even if current events have led to a temporary halt in operations. Read on, then as we present our pick of Hong Kong's 10 most common restaurants and get some inspiration for where your next meal could be. Thus, the main objective of the project will be to find ideal spots in different districts of the city. The Foursquare API will used to get location data and clustering methods to group their restaurant venues information.

3 Data

The data for this project has been retrieved and processed through multiple sources, giving careful considerations to the accuracy of the methods used. For this project we need following data:

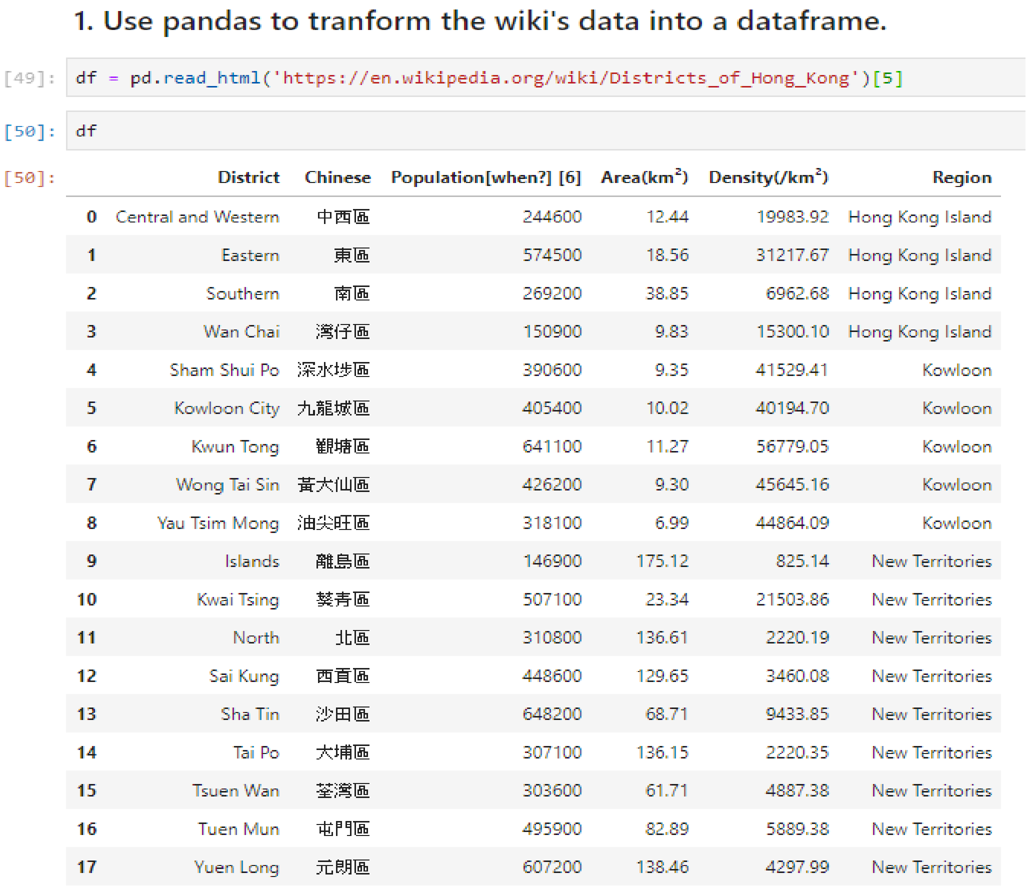
- **Hongkong data that contains list districts (Wards) along with their latitude and longitude.**
  Datasource : https://en.wikipedia.org/wiki/Districts_of_Hong_Kong

  Description: We will Scrap HK districts (Wards) Table from Wikipedia and get the coordinates of these major districts using geocoder class of Geopy client.

- **Restaurants in each neighborhood of Hong Kong:**
  Data source: Foursquare APIs
  Description : By using this API we will get all the venues in each neighborhood. We can filter these venues to get only restaurants.

3.1 Neighborhood

The data of the neighborhoods in HK can be extracted out by web scraping using Pandas for Python. The neighborhood data is scraped from a Wikipedia webpage.



1. Use pandas to tranform the wiki's data into a dataframe.

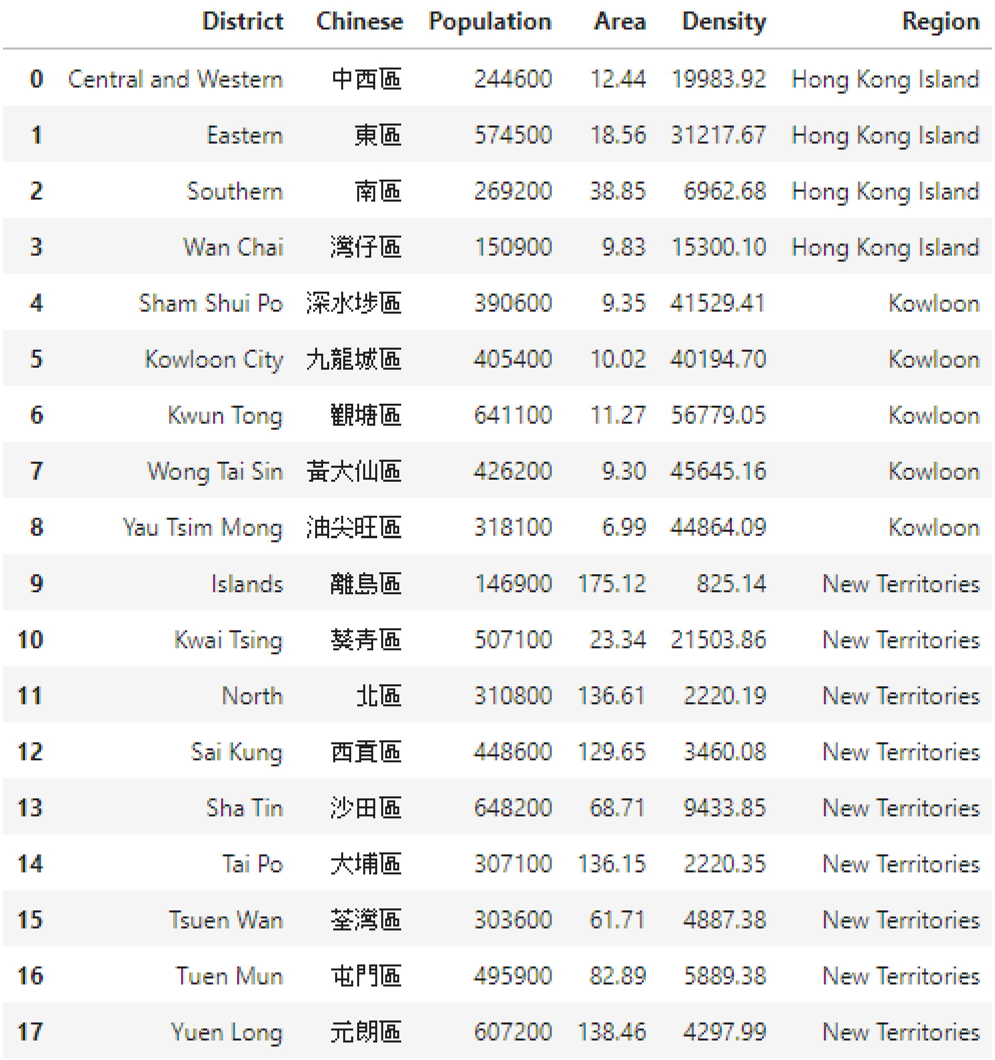
```
[49]: df = pd.read_html('https://en.wikipedia.org/wiki/Districts_of_Hong_Kong')[5]
```

```
[50]: df
```

[50]:

	District	Chinese	Population[when?] [6]	Area(km^2)	Density(/km^2)	Region
0	Central and Western	中西區	244600	12.44	19983.92	Hong Kong Island
1	Eastern	東區	574500	18.56	31217.67	Hong Kong Island
2	Southern	南區	269200	38.85	6962.68	Hong Kong Island
3	Wan Chai	灣仔區	150900	9.83	15300.10	Hong Kong Island
4	Sham Shui Po	深水埗區	390600	9.35	41529.41	Kowloon
5	Kowloon City	九龍城區	405400	10.02	40194.70	Kowloon
6	Kwun Tong	觀塘區	641100	11.27	56779.05	Kowloon
7	Wong Tai Sin	黃大仙區	426200	9.30	45645.16	Kowloon
8	Yau Tsim Mong	油尖旺區	318100	6.99	44864.09	Kowloon
9	Islands	離島區	146900	175.12	825.14	New Territories
10	Kwai Tsing	葵青區	507100	23.34	21503.86	New Territories
11	North	北區	310800	136.61	2220.19	New Territories
12	Sai Kung	西貢區	448600	129.65	3460.08	New Territories
13	Sha Tin	沙田區	648200	68.71	9433.85	New Territories
14	Tai Po	大埔區	307100	136.15	2220.35	New Territories
15	Tsuen Wan	荃灣區	303600	61.71	4887.38	New Territories
16	Tuen Mun	屯門區	495900	82.89	5889.38	New Territories
17	Yuen Long	元朗區	607200	138.46	4297.99	New Territories

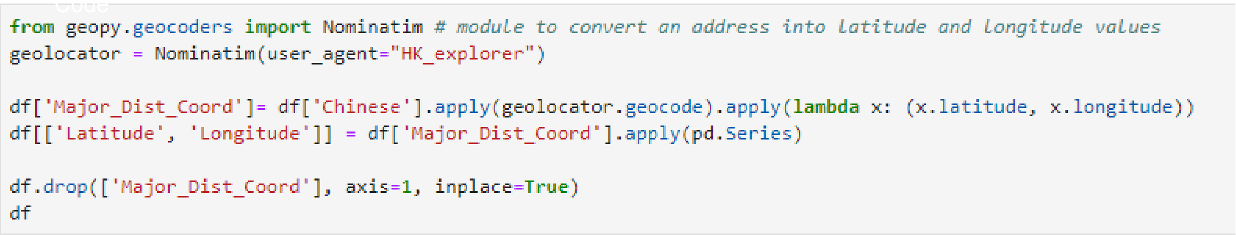
After pre-processing, the data frame is obtained like this:



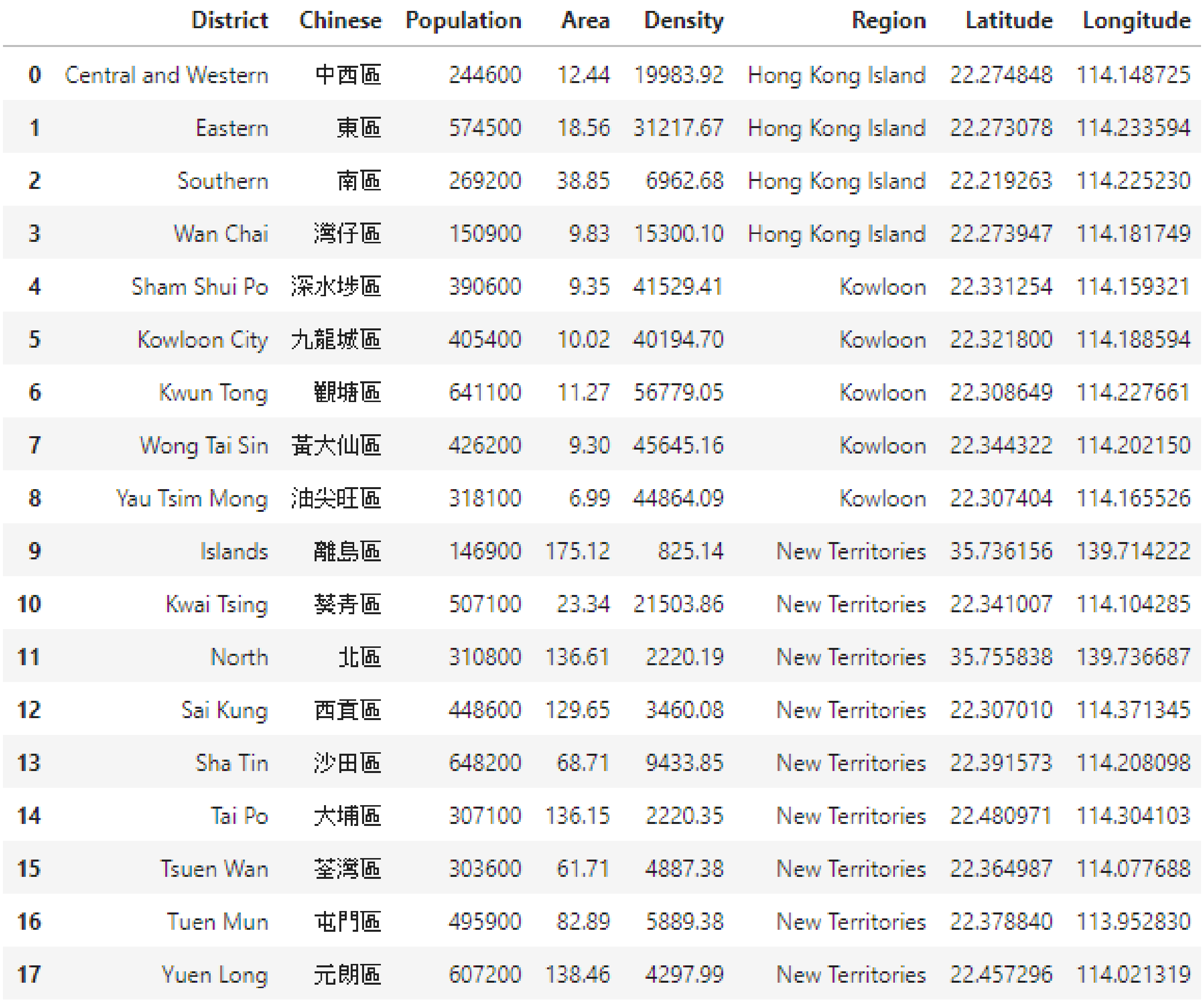
	District	Chinese	Population	Area	Density	Region
0	Central and Western	中西區	244600	12.44	19983.92	Hong Kong Island
1	Eastern	東區	574500	18.56	31217.67	Hong Kong Island
2	Southern	南區	269200	38.85	6962.68	Hong Kong Island
3	Wan Chai	灣仔區	150900	9.83	15300.10	Hong Kong Island
4	Sham Shui Po	深水埗區	390600	9.35	41529.41	Kowloon
5	Kowloon City	九龍城區	405400	10.02	40194.70	Kowloon
6	Kwun Tong	觀塘區	641100	11.27	56779.05	Kowloon
7	Wong Tai Sin	黃大仙區	426200	9.30	45645.16	Kowloon
8	Yau Tsim Mong	油尖旺區	318100	6.99	44864.09	Kowloon
9	Islands	離島區	146900	175.12	825.14	New Territories
10	Kwai Tsing	葵青區	507100	23.34	21503.86	New Territories
11	North	北區	310800	136.61	2220.19	New Territories
12	Sai Kung	西貢區	448600	129.65	3460.08	New Territories
13	Sha Tin	沙田區	648200	68.71	9433.85	New Territories
14	Tai Po	大埔區	307100	136.15	2220.35	New Territories
15	Tsuen Wan	荃灣區	303600	61.71	4887.38	New Territories
16	Tuen Mun	屯門區	495900	82.89	5889.38	New Territories
17	Yuen Long	元朗區	607200	138.46	4297.99	New Territories

3.2 Geocoding

The latitude and longitude of the neighborhoods are retrieved using Google Maps Geocoding API. The geometric location values are then stored into the initial data frame.



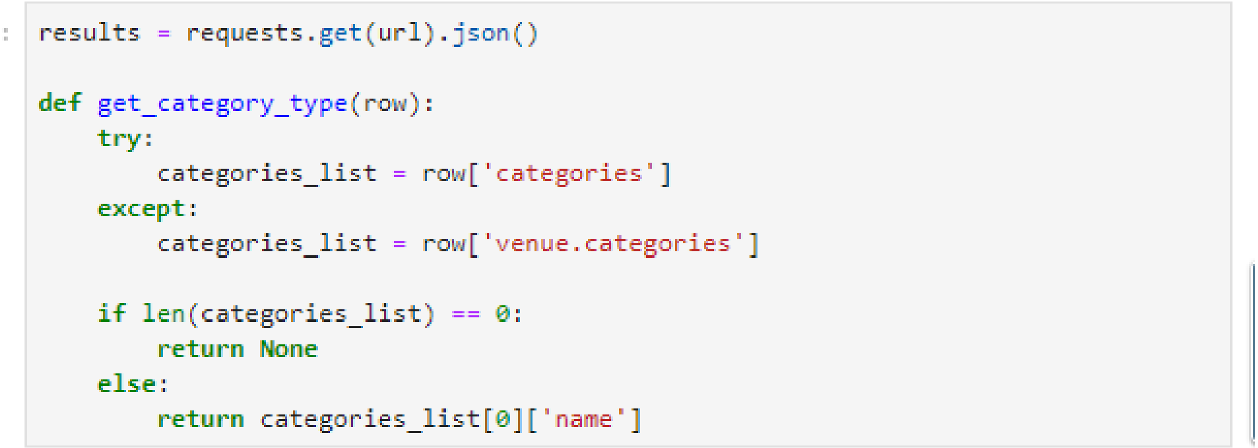
```
from geopy.geocoders import Nominatim # module to convert an address into latitude and longitude values
geolocator = Nominatim(user_agent="HK_explorer")

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

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



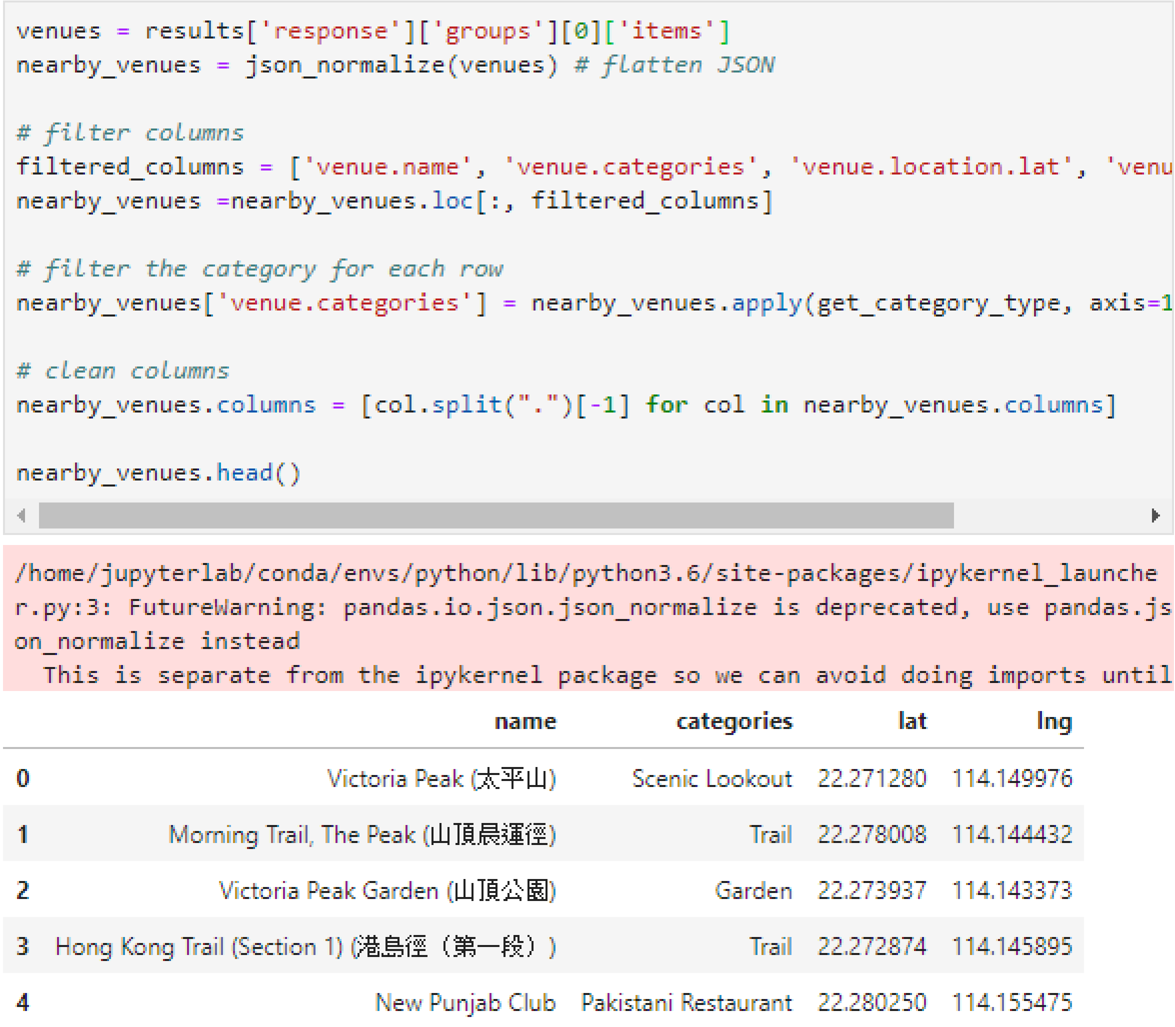
	District	Chinese	Population	Area	Density	Region	Latitude	Longitude
0	Central and Western	中西區	244600	12.44	19983.92	Hong Kong Island	22.274848	114.148725
1	Eastern	東區	574500	18.56	31217.67	Hong Kong Island	22.273078	114.233594
2	Southern	南區	269200	38.85	6962.68	Hong Kong Island	22.219263	114.225230
3	Wan Chai	灣仔區	150900	9.83	15300.10	Hong Kong Island	22.273947	114.181749
4	Sham Shui Po	深水埗區	390600	9.35	41529.41	Kowloon	22.331254	114.159321
5	Kowloon City	九龍城區	405400	10.02	40194.70	Kowloon	22.321800	114.188594
6	Kwun Tong	觀塘區	641100	11.27	56779.05	Kowloon	22.308649	114.227661
7	Wong Tai Sin	黃大仙區	426200	9.30	45645.16	Kowloon	22.344322	114.202150
8	Yau Tsim Mong	油尖旺區	318100	6.99	44864.09	Kowloon	22.307404	114.165526
9	Islands	離島區	146900	175.12	825.14	New Territories	35.736156	139.714222
10	Kwai Tsing	葵青區	507100	23.34	21503.86	New Territories	22.341007	114.104285
11	North	北區	310800	136.61	2220.19	New Territories	35.755838	139.736687
12	Sai Kung	西貢區	448600	129.65	3460.08	New Territories	22.307010	114.371345
13	Sha Tin	沙田區	648200	68.71	9433.85	New Territories	22.391573	114.208098
14	Tai Po	大埔區	307100	136.15	2220.35	New Territories	22.480971	114.304103
15	Tsuen Wan	荃灣區	303600	61.71	4887.38	New Territories	22.364987	114.077688
16	Tuen Mun	屯門區	495900	82.89	5889.38	New Territories	22.378840	113.952830
17	Yuen Long	元朗區	607200	138.46	4297.99	New Territories	22.457296	114.021319

3.3 Venue Data

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another Data Frame to contain all the venue details along with the respective neighborhoods.



```
results = requests.get(url).json()

def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']
```



```
venues = results['response']['groups'][0]['items']
nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venu
nearby_venues =nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues.head()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/ipykernel_launche
r.py:3: FutureWarning: pandas.io.json.json_normalize is deprecated, use pandas.js
on_normalize instead
  This is separate from the ipykernel package so we can avoid doing imports until
```

	name	categories	lat	lng
0	Victoria Peak (太平山)	Scenic Lookout	22.271280	114.149976
1	Morning Trail, The Peak (山頂晨運徑)	Trail	22.278008	114.144432
2	Victoria Peak Garden (山頂公園)	Garden	22.273937	114.143373
3	Hong Kong Trail (Section 1) (港島徑（第一段）)	Trail	22.272874	114.145895
4	New Punjab Club	Pakistani Restaurant	22.280250	114.155475