Coursera Data Science Capstone Project :

Priyank Shah

1   Introduction

This project is to analyze the districts in Houston and find one for a young couple in their 20's. They enjoy going to wine bars and yoga studios and would prefer a location where these are in plenty. The objective of this project is to analyze the districts in Houston and find one that fits this couples needs. This is a great project to use for the Foursquare API. This dataset can be used by anyone looking for different venues in the Houston Districts, especially someone looking at areas with a higher concentration of wine bars and yoga studios.

2   Data

There are multiple datasets that are needed to analyze this data. These datasets will be used to analyze the different districts and then select one based on the couples needs.

2.1 District Data Set:

This dataset is extracted from the Wikipedia page and we use Beautiful Soup to scrape the webpage. Below is the dataset

Official district name ⇕	Encompassed neighborhoods ⇕
5 Corners District	Almeda, Fondren Gardens, Winchester
Baybrook Management District	Baybrook Mall
Downtown District	Downtown
East Downtown Management District	East Downtown
Generation Park Management District	Undeveloped area
Greater East End Management District	East End, Magnolia Park, and Harrisburg
Greater Northside Management District	Near Northside, part of the Heights, Independence Heights, Northline
Houston Southeast	Third Ward, the Texas Medical Center, Riverside Terrace, South Union / OST
International Management District	Alief and Little Saigon
Memorial Management District	Memorial City
Midtown Houston	Midtown[5]
Near Northwest Management District	Greater Inwood
North Houston District	Greenspoint
Southwest Management District	Sharpstown, Mahatma Gandhi District, portions of Chinatown
Spring Branch Management District	Spring Branch
Upper Kirby District	Upper Kirby
Uptown Houston District	Uptown Houston
Westchase District	Westchase

2.2 Geocoding Data:

This dataset is obtained using the geocoding API that are then used in the dataframe. This gives us the latitude and longitude for each district which will help us when we create the maps in Folium.

Using Geocoder API to get geo data

```python
API_KEY='65183f4caad7471ba05770b07bb594a1'
import json

latitudes = [] # Initializing the latitude array
longitudes = [] # Initializing the longitude array

for postal_code in postal_codes :
    place_name = postal_code + " Houston, TX" # Formats the place name
    url = 'https://api.opencagedata.com/geocode/v1/json?q={}&key={}'.format(place_name, API_KEY) # Gets the proper url to make t
    obj = json.loads(requests.get(url).text) # Loads the JSON file in the form of a python dictionary

    results = obj['results'] # Extracts the results information out of the JSON file
    lat = results[0]['geometry']['lat'] # Extracts the latitude value
    lng = results[0]['geometry']['lng'] # Extracts the longitude value

    latitudes.append(lat) # Appending to the list of latitudes
    longitudes.append(lng) # Appending to the list of longitudes
```

0	5 Corners District	29.987904	-95.479542
1	Baybrook Management District	29.576304	-95.128733
2	Downtown District	32.782611	-96.808781
3	East Downtown Management District	29.831391	-95.276863
4	Generation Park Management District	29.907106	-95.179968
5	Greater East End Management District	29.741647	-95.256811
6	Greater Northside Management District	30.004339	-95.412106
7	Houston Southeast	29.767424	-95.366939
8	International Management District	29.941288	-95.327246
9	Memorial Management District	29.935417	-95.458328
10	Midtown Houston	29.741415	-95.353201
11	Near Northwest Management District	29.919558	-95.493324
12	North Houston District	29.944719	-95.416074
13	Southwest Management District	29.937680	-95.392751
14	Spring Branch Management District	29.937680	-95.392751
15	Upper Kirby District	29.755609	-95.416672
16	Uptown Houston District	29.808889	-95.269734
17	Westchase District	29.728413	-95.551102

2.3 Venue:

This dataset is obtained from the Foursquare API. We get information about venues around the Houston districts. This data will then be analyzed and loaded into Folium maps for analysis

Using Foursquare API

```python
CLIENT_ID = ████████████████████████████████████████ # your Foursquare ID
CLIENT_SECRET = '████████████████████████████████' # your Foursquare Secret
VERSION = '20180604'
LIMIT = 30

explore_df_list = []

for i, nbd_name in enumerate(df['Postcode']):

    try :
        ### Getting the data of neighbourhood
        nbd_name = df.loc[i, 'Postcode']
        nbd_lat = df.loc[i, 'Latitude']
        nbd_lng = df.loc[i, 'Longitude']

        radius = 500 # Setting the radius as 500 metres
        LIMIT = 100 # Getting the top 100 venues

        url = 'https://api.foursquare.com/v2/venues/explore?client_id={} \
        &client_secret={}&ll={},{}&v={}&radius={}&limit={}'\
        .format(CLIENT_ID, CLIENT_SECRET, nbd_lat, nbd_lng, VERSION, radius, LIMIT)

        results = json.loads(requests.get(url).text)
        results = results['response']['groups'][0]['items']

        nearby = json_normalize(results) # Flattens JSON

        # Filtering the columns
        filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
        nearby = nearby.loc[:, filtered_columns]

        # Renaming the columns
        columns = ['Name', 'Category', 'Latitude', 'Longitude']
        nearby.columns = columns

        # Gets the categories
        nearby['Category'] = nearby.apply(get_category_type, axis=1)

        # Gets the data required
        for i, name in enumerate(nearby['Name']):
            explore_df_list.append([nbd_name, nbd_lat, nbd_lng] + nearby.loc[i, :].values.tolist())

    except Exception as e:
        pass
```