IBM Data Science Capstone Project

Recommender System for identification of Amenities in the vicinity of Hi-Tech City in Hyderabad using Four Square Application Programming Interface (API) Data

By

Dr.V.Venkat Ramayya

1. Introduction:

According to the Wikipedia, The Hyderabad Information Technology and Engineering Consultancy City, abbreviated as HITEC City, is an Indian Information Technology, Engineering, Health informatics, and Bioinformatics; business district located in Hyderabad, India. HITEC City is spread across 200 acres (81 ha) of land under suburbs of Madhapur, Gachibowli, Kondapur, Manikonda, and Nanakramguda, the technology township is also known as Cyberabad. HITEC City is within two kilometers of the residential and commercial suburb of Jubilee Hills. The place is known for a number of software companies, restaurants, parks, places of cultural importance. Hyderabad Metro Rail (HMR), a prestigious Project of Govt. of Telangaa undertaken by L & T Ltd., is the most recent development and due to which a metro station was also set up in Hitech City catering to the needs of public transportation.

2. Problem Definition & Objectives:

In the Capstone Project of Coursera, it is proposed to develop a recommender system for HITEC City through identification of different types of amenities and categorise them. The objectives of the present study are listed below.

1. To collect required geospatial data to identify amenities in the vicinity of Shilparamam, HITEC city.
2. To collect geoJSON data for the chosen study area.
3. To collect Four square API data thorough general and premium calls to harness the information about the venues and their categories in the vicinity of HITEC City, Hyderabad.
4. To recommend venues to the users based on ratings and other statistics for a chosen venue category.

Target Audience

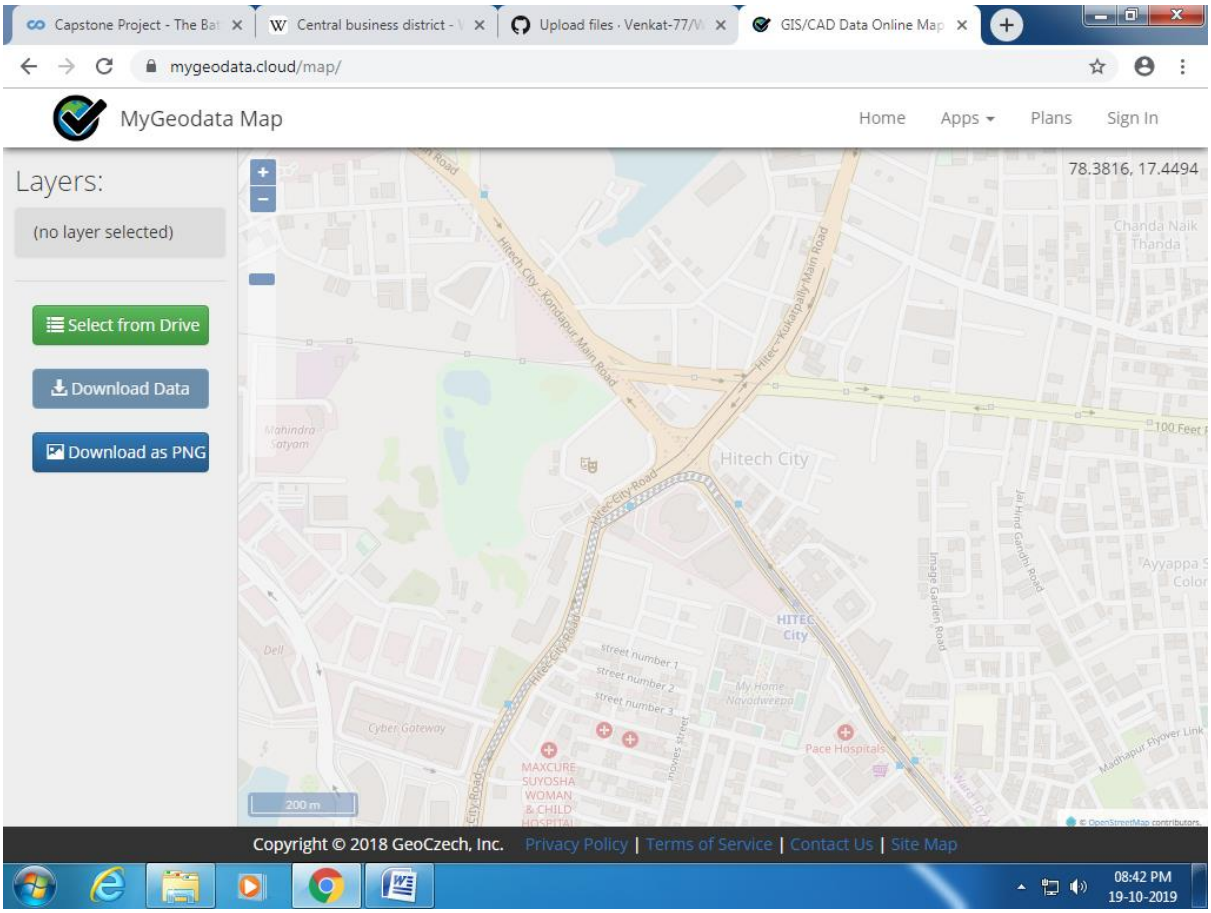
The target audience who will be benefited by this project are:

- People seeking different types of best amenities such as restaurants, coffee shops, furniture malls, textile shops etc in the vicinity of the study area.
- As the recommender system will aim to identify different types venues around a chosen neighbourhood, this data would be useful for business expansion who would like to set up their new centres to promote their business.

3. Data Collection:

The data required for the study is collected in the following sequence.

1. The geospatial data of the HITEC city around Shilparamam is collected using https://mygeodata.cloud/map/ in the form of *.osm* file (open street map). If the file size exceeds 5MB, the service will be a paid one. An image of the map is shown below.



2. In the next step the *.osm* file is converted to geoJson file using https://mygeodata.cloud/converter/ to get the information about the different types of attributes of the study area. This becomes our starting point for the data analysis. The features that we can collect using the service are addresses, amenity points, amenity polygons, bridges, buildings, land cover, power lines railways, roads etc in the study area.
3. The neighbourhoods in the vicinity of the shilparamam are collected from the geoJSON file and then onwards, and data is manipulated to proceed futher using Foursquare API location data.
4. Cluster Analysis will be used to generate clusters of amenities in the study area.

5. In this project, Foursquare API data was alone used to develop the recommender system.

   A sample of geoJSON data is presented below for reference.

```
[5]: hitech_data

[5]: {'type': 'FeatureCollection',
      'name': 'amenity_points',
      'crs': {'type': 'name',
       'properties': {'name': 'urn:ogc:def:crs:OGC:1.3:CRS84'}},
      'features': [{'type': 'Feature',
        'properties': {'osm_id': 6574136929,
         'shop': 'furniture',
         'brand': 'IKEA',
         'name': 'IKEA',
         'man_made': None,
         'amenity': None,
         'highway': None,
         'operator': None,
         'leisure': None,
         'tourism': None,
         'sport': None,
         'addr:housename': None,
         'addr:housenumber': None,
         'place': None,
         'office': None,
         'brand:wikidata': 'Q54078',
         'brand:wikipedia': 'en:IKEA',
         'cuisine': None,
         'name:te': None,
         'payment:debit_cards': None,
         'addr:city': None,
         'payment:coins': None,
         'addr:postcode': None,
```

4. Methodology

The objectives of the study and the data requirements and sources are presented in the earlier section. The methodology adopted in the present study to fulfil the stated objectives is presented in this section.

Steps involved

a. Install required packages

The required packages are to be installed in the Jupyter Notebook environment. The packages and libraries used in the project are:
geopy, folium, numpy, pandas , json, requests, matplotlib, sklearn, kmeans csv, wordcloud, & xlrd

b. Collect *.osm* file of the study location

The data required for developing recommender system is collected from mygeodatabase. As stated in the Section 3, the geospatial data of the

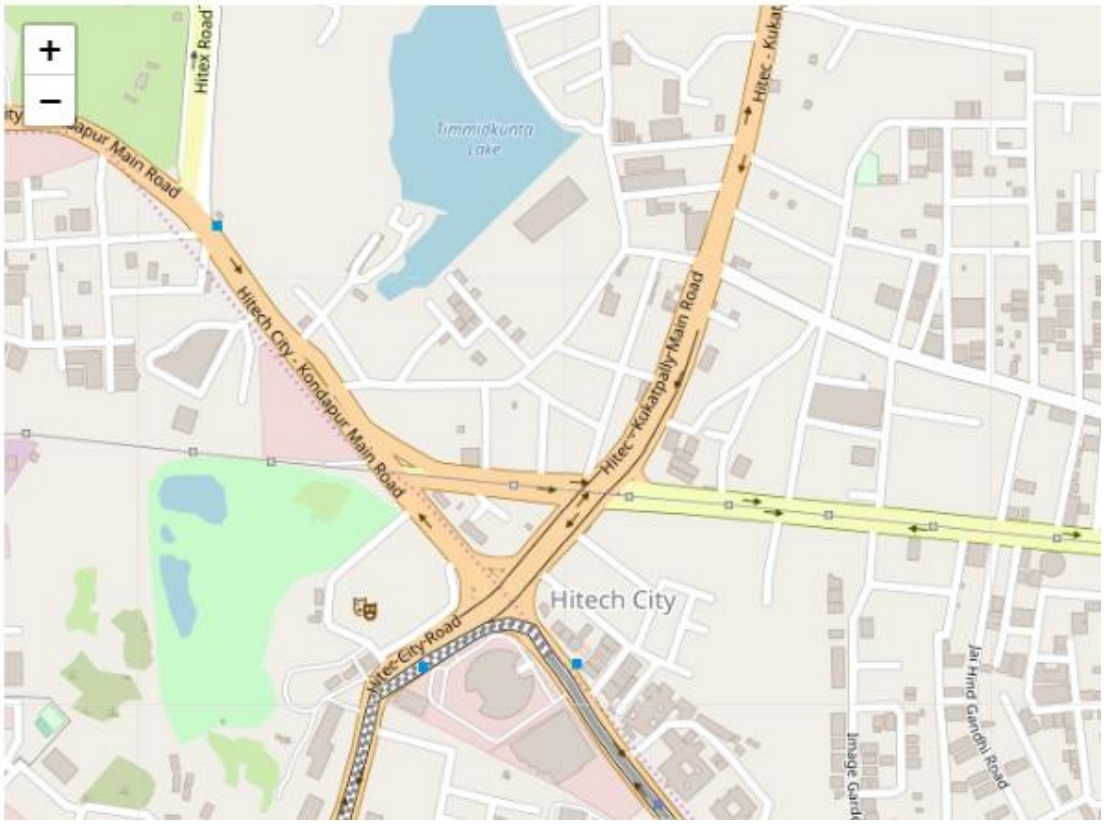
HITEC city around Shilparamam is collected using https://mygeodata.cloud/map/ in the form of *.osm* file (open street map).

c. Convert the *.osm* File to geoJSON file

In this step, the *.osm* file is converted to geoJson file using https://mygeodata.cloud/converter/ to get the information about the different types of attributes of the study area. This becomes starting point for the data analysis. The features that we can collect using the service are addresses, amenity points, amenity polygons, bridges, buildings, land cover, power lines railways, roads etc in the study area.

```
[25]: map_hitech = folium.Map(location=[latitude, longitude], zoom_start=20)
      map_hitech
```

[25]:



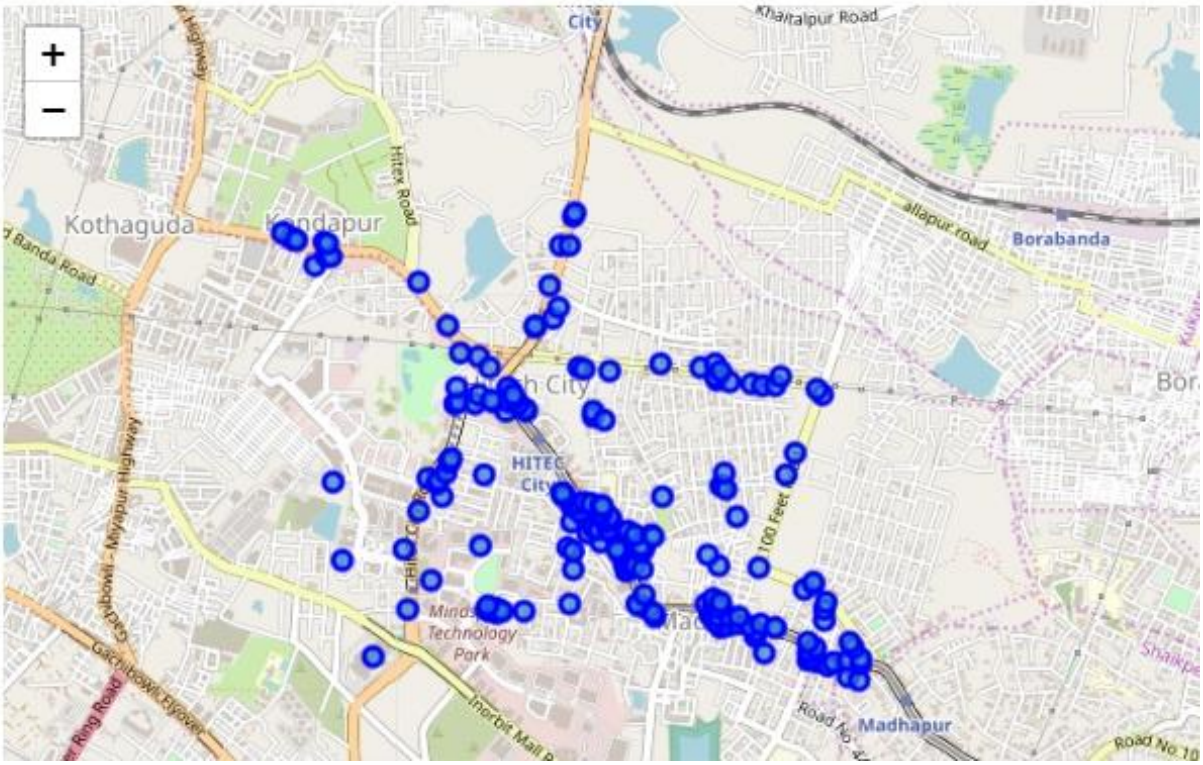
d. Extract Amenity Points in the study area from the geoJSON file.

The neighbourhoods in the vicinity of the shilparamam are collected from the geoJSON file and then onwards, and data is manipulated to proceed futher using Foursquare API location data.

```
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_hitech)

map_hitech
```

[26]:



e. Read Amenities into a Pandas Data Frame and Edit the data for missing values.

The amenities in the study area are read into a pandas data frame. The data frame is edited for missing values and make it ready for further analysis.

[14]:
```
locations.head()
```

[14]:

	osm_id	name	amenity	Latitude	Longitude
0	6574136929	IKEA	None	17.439238	78.375390
1	2713944084	over head water tank	None	17.442588	78.371619
2	4931853973	Hexagon	place_of_worship	17.443600	78.373899
3	655159152	Mindspace	None	17.441420	78.376985
4	893955535	TCS parking	parking	17.442682	78.378129

f. Get the Neighbourhood data by Four Square Call (26 Neighbourhoods)

```
[36]: venues = results['response']['groups'][0]['items']

nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
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()
nearby_venues.shape
```

```
[36]: (26, 4)
```

```
[37]: nearby_venues.to_csv('nearby_venues.csv')
```

g. At this point, the Shilpakala vedika has been chosen to identify and Explore the nearby venues around these 26 Neighbourhoods (5175 nearby venues are returned by Four Square) and prepare data frame.

```
Petrol bunk
Butta Convention Center
Prince DriveInn
Indian Oil Petrol
```

```
[36]: print(neighbourhood_venues.shape)
neighbourhood_venues.head()
```

```
(5175, 8)
```

[36]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category
0	IKEA	17.439238	78.37539	Republic of Noodles	4f89a0b3e4b039b1453f557f	17.442831	78.376914	Asian Restaurant
1	IKEA	17.439238	78.37539	IKEA Hyderabad	57ac118c38fa8f9b5d45395f	17.438940	78.375823	Furniture / Home Store
2	IKEA	17.439238	78.37539	Red Fox Hotel	4e63f252091afcffd9d9de02	17.443244	78.376338	Asian Restaurant
3	IKEA	17.439238	78.37539	Lemon Tree Hotel	4d6dcd43619a236ad8606a8f	17.443275	78.376898	Hotel
4	IKEA	17.439238	78.37539	Paradise	5a6c262cc5309373244cbdcc	17.441355	78.376380	Indian Restaurant

```
[37]: neighbourhood_venues.to_csv('neighbourhood_venues.csv')
```

h. Group the above data based on Neighbourhood and count unique categories. (73 No's)

```
[43]: neighbourhood_venues.groupby('Neighborhood').count()
```

[43]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category
32 Happy Teeth Dental clinic	35	35	35	35	35	35	35
4 Seasons	20	20	20	20	20	20	20
@home	29	29	29	29	29	29	29
Aadi Raghavendra Udupi Veg	35	35	35	35	35	35	35
Abhiruchi Restaurant	30	30	30	30	30	30	30
Absolute Barbecues (ABs)	29	29	29	29	29	29	29
Ahilobhilam Foods	11	11	11	11	11	11	11
Andhra Bank	23	23	23	23	23	23	23
Anjaneya swami temple	29	29	29	29	29	29	29

```
[58]: print('There are {} unique categories.'.format(len(neighbourhood_venues['Venue Category'].unique())))

There are 73 unique categories.
```

i. The frequency of occurrence of top 5 venues listed by Foursquare in each of those Neighbourhoods is calculated.

Explore top 5 Venues in each neighbourhood

```
[46]: num_top_venues = 5

for hood in neighbourhood_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = neighbourhood_grouped[neighbourhood_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

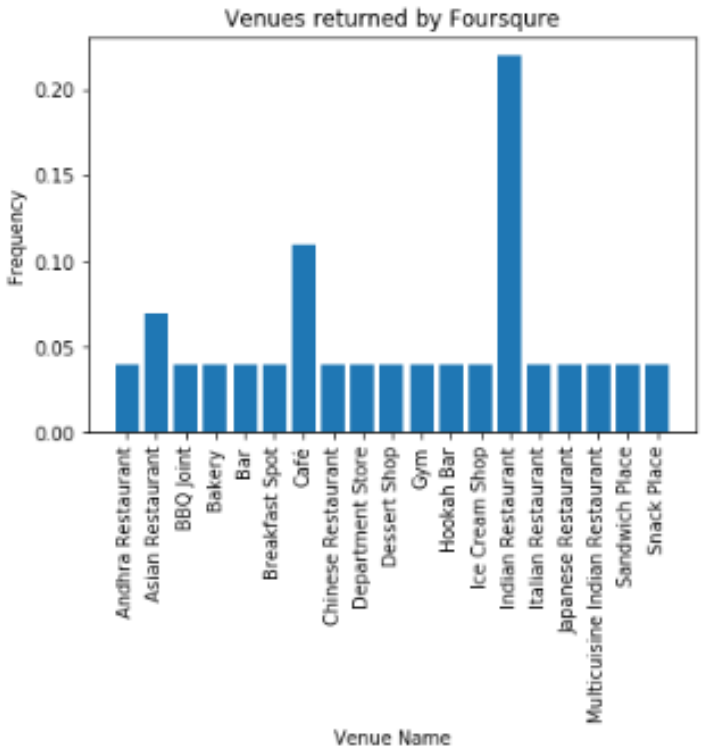
```
----32 Happy Teeth Dental clinic----
                venue  freq
0   Indian Restaurant  0.14
1               Café  0.11
2             Bakery  0.09
3  Chinese Restaurant  0.06
4    Asian Restaurant  0.06


----4 Seasons----
                  venue  freq
0                 Café  0.15
1   Indian Restaurant  0.15
2  Fast Food Restaurant  0.10
3      Bed & Breakfast  0.05
4                  Gym  0.05
```

j. Nth (1st to 10th) Most common Venue for each of these Neighbourhoods is assessed.

```
[52]: neighborhoods_venues_sorted.head()
```

	Neighborhood	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	32 Happy Teeth Dental clinic	Indian Restaurant	Café	Bakery	Asian Restaurant	Chinese Restaurant	Afghan Restaurant	Outdoors & Recreation	Fast Food Restaurant	Electronics Store	Department Store
1	4 Seasons	Café	Indian Restaurant	Fast Food Restaurant	Department Store	Nightclub	Dumpling Restaurant	Bed & Breakfast	Sandwich Place	Restaurant	Food Court
2	@home	Indian Restaurant	Fast Food Restaurant	Restaurant	Chinese Restaurant	Hotel	Punjabi Restaurant	Italian Restaurant	Market	Mexican Restaurant	Concert Hall



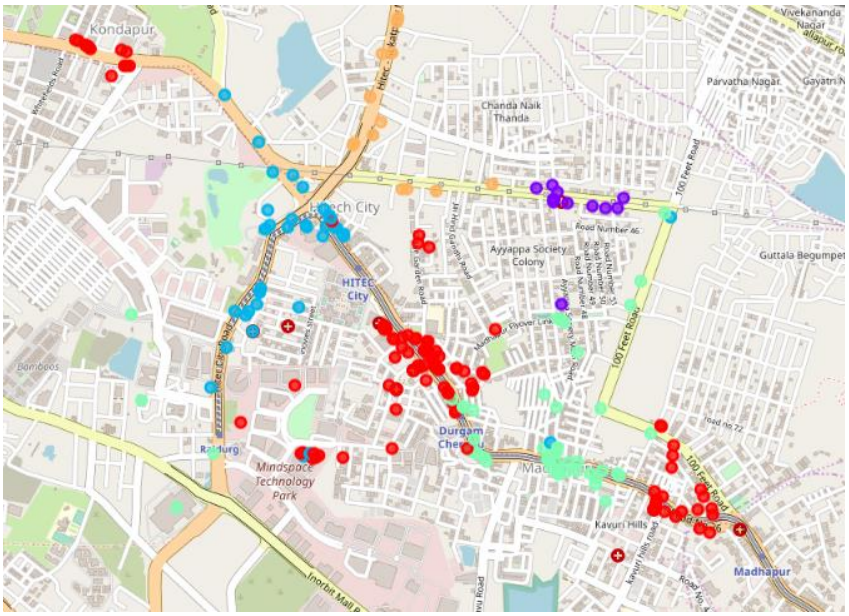
k. Word Clouds are prepared both for 'Venue' and 'Venue Category' to identify the most listed venue for 1st, 2nd and 3rd Common venue.

Word cloud for the Venue Category

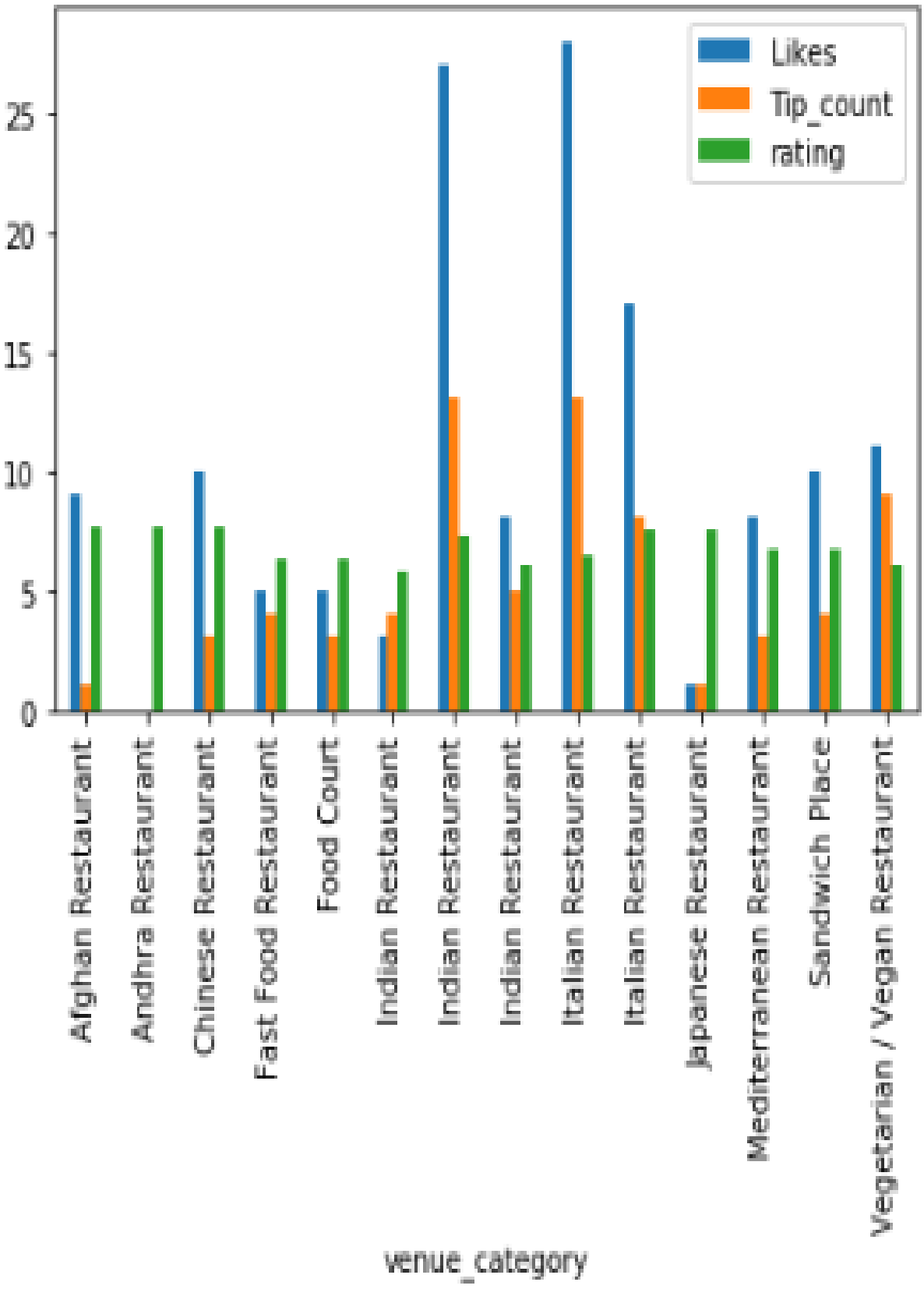
Word Cloud for Venue Name

These word clouds reveal dominant venue categories and venue names data as returned by Four square.

l. Perform the cluster analysis for the Neighbourhoods.



m. Restaurants are explored (as this venue is found to highest frequency) through premium calls for getting information about tip count and other venue statistics. Graphs are prepared for recommending venues for the Category 'Restaurants' based on 'Tip Count', 'Ratings' and 'Likes' .

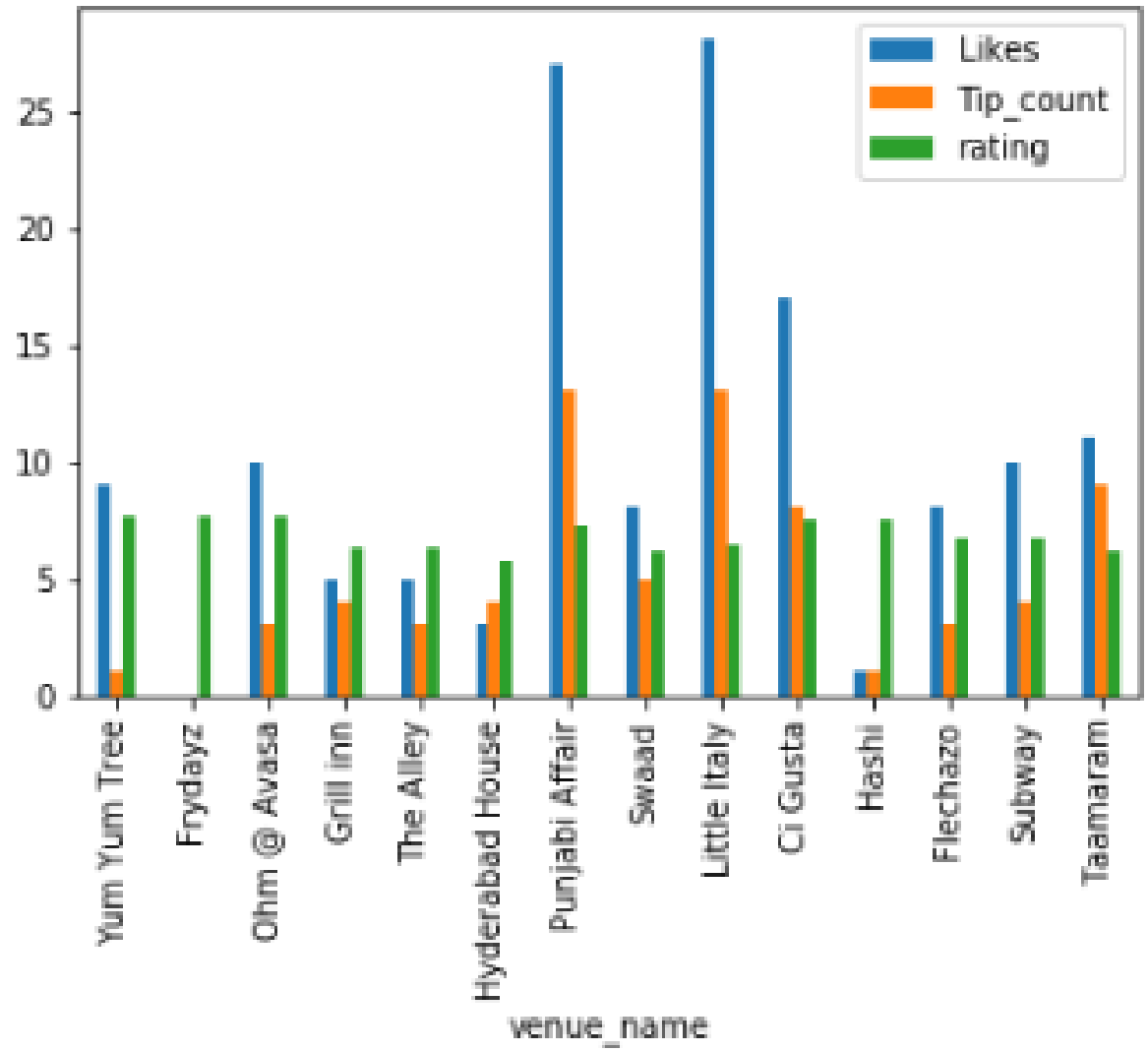

Likes
Tip_count
rating
25
20
15
10
5
0
Afghan Restaurant
Andhra Restaurant
Chinese Restaurant
Fast Food Restaurant
Food Court
Indian Restaurant
Indian Restaurant
Indian Restaurant
Italian Restaurant
Italian Restaurant
Japanese Restaurant
Mediterranean Restaurant
Sandwich Place
Vegetarian / Vegan Restaurant
venue_category

[78]:
```
for i in results_premium:
        venue_name = results_premium[i]['response']['venue']['name']
        venue_cate= results_premium[i]['response']['venue']['categories'][0]['name']  # for vene name
        id = results_premium[i]['response']['venue']['categories'][0]['id']  # for vene id
        tipcount =results_premium[i]['response']['venue']['stats']['tipCount']   # for tip count
        likes = results_premium[i]['response']['venue']['likes']['count'] # venue likes count
        if i!=1 and  i!=4:
            price = results_premium[i]['response']['venue']['price']['message']  # price range not there for 1 4
        if i!=1 and  i!=4 and i!=10:
            rating = results_premium[i]['response']['venue']['rating'] # venue rating   not there for 1 4  10
        df_prem=df_prem.append(pd.Series([venue_name,venue_cate,id,tipcount,likes,price,rating],index=df_prem.columns ), ignore_index=True
```

[79]:
```
df_prem
```

[79]:

	venue_name	venue_category	Venue_id	Tip_count	Likes	Price	rating
0	Yum Yum Tree	Afghan Restaurant	503288ae91d4c4b30a586d67	1	9	Moderate	7.6
1	Frydayz	Andhra Restaurant	54135bf5e4b08f3d2429dfe5	0	0	Moderate	7.6
2	Ohm @ Avasa	Chinese Restaurant	4bf58dd8d48988d145941735	3	10	Cheap	7.8
3	Grill inn	Fast Food Restaurant	4bf58dd8d48988d16e941735	4	5	Cheap	6.3
4	The Alley	Food Court	4bf58dd8d48988d120951735	3	5	Cheap	6.3
5	Hyderabad House	Indian Restaurant	4bf58dd8d48988d10f941735	4	3	Moderate	5.9
6	Punjabi Affair	Indian Restaurant	4bf58dd8d48988d10f941735	13	27	Moderate	7.3
7	Swaad	Indian Restaurant	4bf58dd8d48988d10f941735	5	8	Moderate	6.1
8	Little Italy	Italian Restaurant	4bf58dd8d48988d110941735	13	28	Moderate	6.4
9	Ci Gusta	Italian Restaurant	4bf58dd8d48988d110941735	8	17	Moderate	7.5
10	Hashi	Japanese Restaurant	4bf58dd8d48988d111941735	1	1	Moderate	7.5



In the present work, the scope is limited to restaurants in the HiTech city. Similar kind of exercise can be extended for other category of venues.

5. Results

The methodology presented in the Section 4 and associated results. The data is presented through a number of ways. The neighbourhoods in the vicinity of HiTech city were explored around Shilpa Kala Vedika (26 No's). Further, the venues in the vicinity of the above neighbourhood was explored using ***venues/explore*** call method. Restaurants were found to be very predominant among the other categories. Hence it was decided to explore the Restaurants through premium calls. It was found that the **Little Italy of Italian restaurant category and Punjabi Affair of Indian Restaurant category** were found to have good ratings, tip counts and likes of all venue categories. Both these venues were indexed as '**moderate'** price by the users.

6. Conclusion

It can be concluded that the Four Square API is a good tool for exploring the neighbourhoods for different types of amenities. Premium calls will help us to get the detailed information like venue stats, likes, ratings, images etc. Word Clouds developed from first most to third most common venues under each neighbourhood have helped to understand most frequent venue categories and venues in the HiTech City as listed by Four Square. In the present work, due to non availability of demographic data, the analysis is limited to the results of Four Square API. Better recommender system models can be developed if it is possible to integrate the demographic data.

References:

1 https://mygeodata.cloud/converter

2 https://foursquare.com

3 https://stackoverflow.com

4 https://geeksforgeeks.org