<u>Predicting the Best Areas to Start an Italian Restaurant in Pune</u>

<u>Jeet Shah</u>

August 27,2020.

1 Introduction

1.1 Background

It's quite a difficult and time consuming task to analyze "Where to start my restaurant" as this "Where" part is most difficult and is an important factor in determining the success of your hefty investment. A Good restaurant with excellent interior and Food Quality and competent rates can also fail if it's established on a wrong area. For E.g. if the area is on industrial zone or human presence is far away then in spite of maintaining everything it will cause problem. So a Good Area with <u>good demand and supply</u> with all above mentioned points related to restaurant will help the restaurant to succeed.

1.2 Problem

Apart from above mentioned points it is more difficult to analyze for a Specific Cuisine related restaurant which is our case of "<u>Italian Restaurant</u>". Besides this a specific cuisine restaurant requires to analyze the most important <u>"Where" </u>question. A good Italian restaurant can run only and only on areas where there is a <u>Good Demand & Good Supply</u> apart from other facilities and services that a restaurant provides.

1.3 Interest

As this is a complex problem , So this basically attracts a newbie to analyze how this actually works and extract exact data and output.

2 Data Acquisition & Cleaning

2.1 Data Sources

This project will use data from :

- Geopy - For getting the co-ordinated of different locations.
- Foursquare API - To get the list of venues and their details around a given location.
- Geocoder – To extract latitude and longitude from Areas and the city itself.

3 Methodology

Below are the step-by-step points that consists of Methodology steps.

1. Getting the co-ordinates of the target city.

```
data = requests.get('https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Pune').text

soup = BeautifulSoup(data,'html.parser')

neighbourhood_list =[]

#find_all to get whole data
#for i in soup.find_all('div',class_='mw-category')[0].find_all('a'):
#for store in soup.find_all('div',class_='mw_category')[0].find_all('a'):
    #neigbourhood_list.append(store.text)

for i in soup.find_all('div',class_='mw-category')[0].find_all('a'):
    neighbourhood_list.append(i.text)


neighbourhood_df = pd.DataFrame(neighbourhood_list,columns=['Locality'])
neighbourhood_df.head()
```

2. Getting the list of neighborhoods and their co-ordinates.

In [9]:
```
def calculate_latitude_longitude_all_areas(localities):
    locate = geocoder.arcgis('{},Pune,India'.format(localities))
    getlatlong = locate.latlng
    return getlatlong
```

In [11]:
```
# Getting Latitudes and Longitudes for All Areas of Pune

store_localities =[]

for i in neighbourhood_df['Locality'].tolist():
    store_localities.append(calculate_latitude_longitude_all_areas(i))
```

In [12]:
```
store_localities[:5]
```

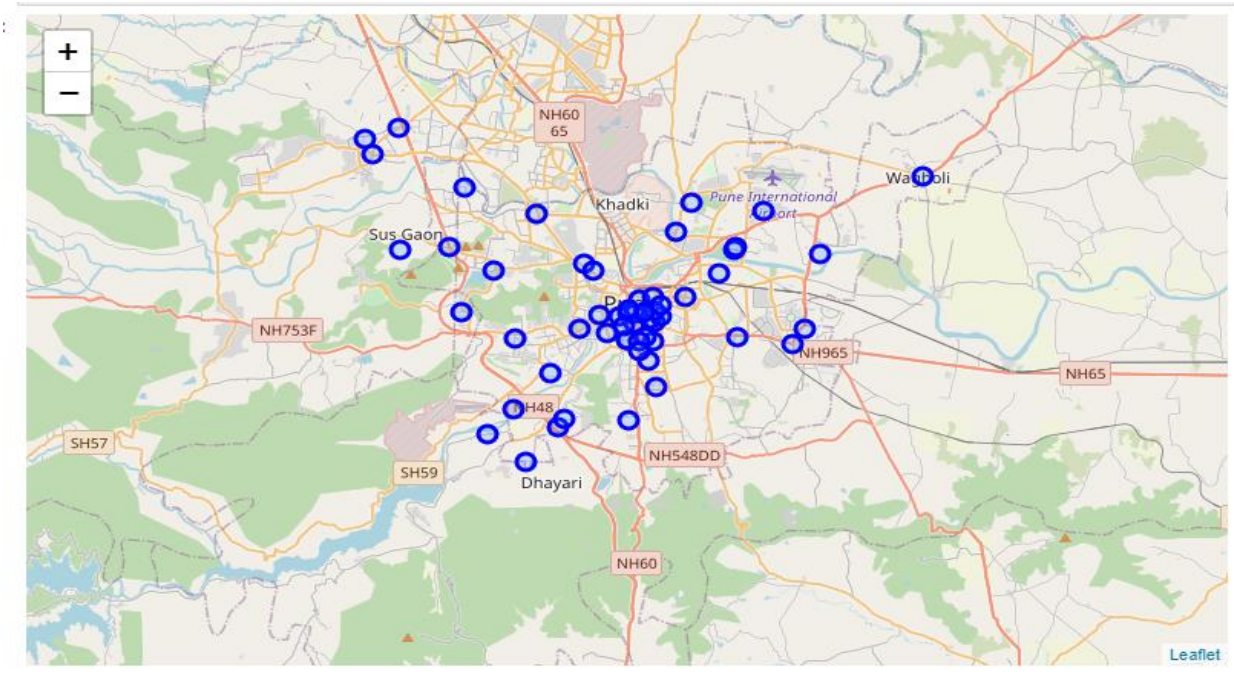
Out[12]:
```
[[18.516483671884753, 73.85387026191101],
 [18.563450000000046, 73.81227000000007],
 [18.576020000000028, 73.77983000000006],
 [18.548200000000065, 73.77316000000008],
 [18.517544858465925, 73.77853184068661]]
```

]:
```
coordinates = pd.DataFrame(store_localities,columns=['Latitudes','Longitudes'])

neighbourhood_df['Latitudes']=coordinates['Latitudes']
neighbourhood_df['Longitudes']=coordinates['Longitudes']

neighbourhood_df.head()
```

]:

	Locality	Latitudes	Longitudes
0	Appa Balwant Chowk	18.516484	73.853870
1	Aundh, Pune	18.563450	73.812270
2	Balewadi	18.576020	73.779830
3	Baner	18.548200	73.773160
4	Bavdhan	18.517545	73.778532

3. Mapping them on Map Using Folium



4. Using Foursquare API to extract nearby Venues from all areas

In [32]:
```
#foursquare
CLIENT_ID = 'M1NURV4RJYRINESBG1AZJ2LLFM0VN4K4FIDUHYQRK5GO0RBL'
CLIENT_SECRET = 'SRW3IY5XJ5G3M3CSYKVABCW2B3DDDR5QEELMBNJX1UX5CDVI'
VERSION = '20180605' # Foursquare API version
```

In [36]:
```
Limit = 10
radius = 2000

venues =[]

for lat,long,locality in zip(neighbourhood_df['Latitudes'],neighbourhood_df['Longitudes'],neighbourhooc
    url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&rac
    results = requests.get(url).json()['response']['groups'][0]['items']


    for venue in results:
        venues.append((locality, lat, long, venue['venue']['name'], venue['venue']['location']['lat'],

    #for venue in results:
        #venues.append((locality,lat,long,venue['venue']['name'],venue['venue']['location']['lat'],venu
```

In [37]:
```
venues[0]
```

Out[37]:
```
('Appa Balwant Chowk',
 18.516483671884753,
 73.85387026191101,
 'Sujata Mastani',
 18.511792754341577,
 73.85214493967393,
```

```
In [38]: venues_df = pd.DataFrame(venues)
         venues_df.columns = ['Locality', 'Latitude', 'Longitude', 'Venue name', 'Venue Lat', 'Venue Lng', 'Venu
         venues_df.head()
```

Out[38]:

	Locality	Latitude	Longitude	Venue name	Venue Lat	Venue Lng	Venue Category	Venue ID
0	Appa Balwant Chowk	18.516484	73.85387	Sujata Mastani	18.511793	73.852145	Ice Cream Shop	4bd12ba141b9ef3b12a4fbe5
1	Appa Balwant Chowk	18.516484	73.85387	Bhagat Tarachand	18.514332	73.851317	Indian Restaurant	4c41785da5c5ef3bb73eb06f
2	Appa Balwant Chowk	18.516484	73.85387	Hotel Madhuban	18.519248	73.848688	Tea Room	50f6c177e4b0e9762504f426
3	Appa Balwant Chowk	18.516484	73.85387	Raja Dinkar Kelkar museum	18.510744	73.854389	History Museum	4d96d24fc910d7ce1b454755
4	Appa Balwant Chowk	18.516484	73.85387	Mad Over Donuts	18.519335	73.845320	Donut Shop	4feebcafe4b0da11fdbe582b

5. Categorizing the Data to get total count venue category wise.

```
In [42]: demo1_df = pd.DataFrame({'Venue Category':complex_df.index[:50]})
         category_strength=[]
         for i in range(50):
             category_strength.append(complex_df['Strength'][i])
         demo2_df = pd.DataFrame(category_strength, columns=['Strength'])
         demo_df = pd.DataFrame({'Venue Category': demo1_df['Venue Category'], 'Strength': demo2_df['Strength']}
         demo_df.head()
```

Out[42]:

	Venue Category	Strength
0	Indian Restaurant	61
1	Ice Cream Shop	35
2	Snack Place	26
3	Café	23
4	Vegetarian / Vegan Restaurant	23

6. Using Word Cloud to visualize frequency of Categories of Venues

Here we come to know Indian Restaurants are having most frequency in Pune.

7. Using OneHotEncoding and Transpose method to convert data into more visualizable form to clearly see details. E.g. if there is Indian Restaurant in Appa Balwant Chowk then it will be displayed with 1.
   Encoding is used to make Textual Data more lenient and flexible for statistical modeling. It can be analyzed with ease.

```
62]: blr_onehot = pd.get_dummies(venues_df[['Venue Category']], prefix="", prefix_sep="")

     blr_onehot['Locality'] = venues_df['Locality']

     #moving the locality column to the front
     blr_onehot = blr_onehot[ [ 'Locality' ] + [ col for col in blr_onehot.columns if col!='Locality' ] ]
     blr_onehot.head(10)
```

62]:

	Locality	ATM	American Restaurant	Arcade	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Garden	Bistro	...	South Indian Restaurant	Southern / Soul Food Restaurant	Sporting Goods Shop
0	Appa Balwant Chowk	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Appa Balwant Chowk	0	0	0	0	0	0	0	0	0	...	0	0	0
2	Appa Balwant Chowk	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Appa Balwant Chowk	0	0	0	0	0	0	0	0	0	...	0	0	0
	Appa													

8. To Group By areas to get the mean of Italian Restaurant in Pune .
   For better analyzing and clustering areas this step is implemented.

5 rows × 81 columns

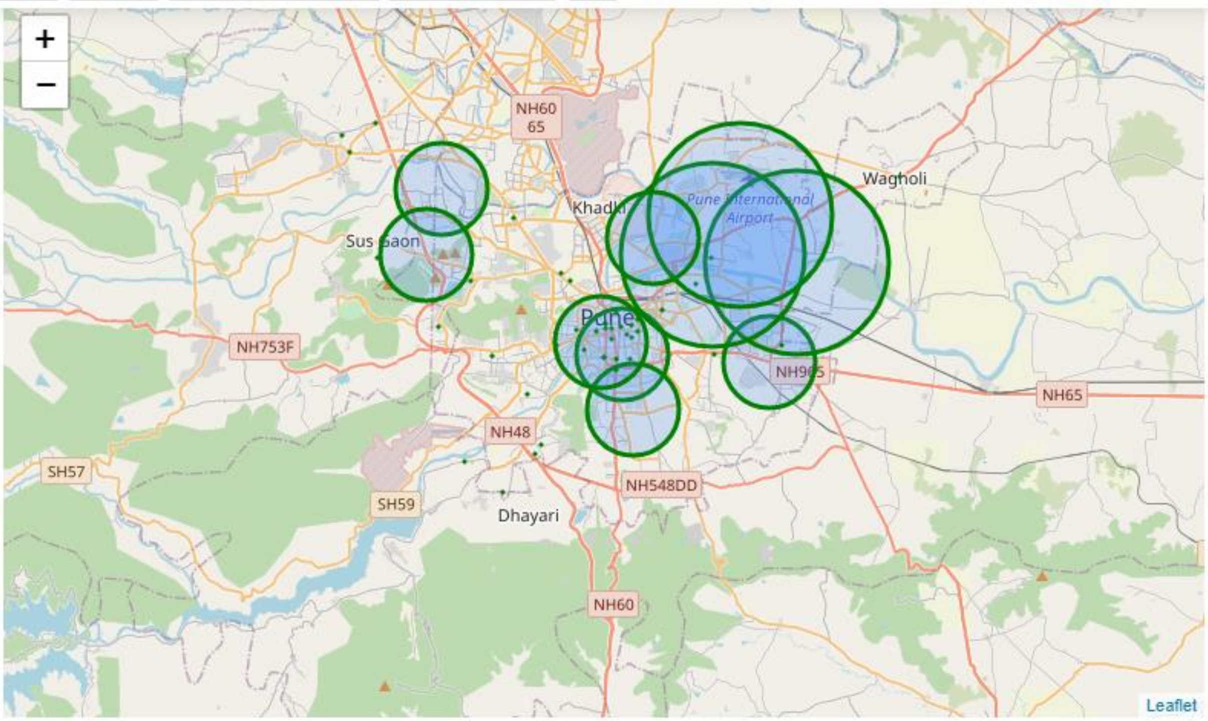
```
In [64]: len(blr_grouped[blr_grouped['Italian Restaurant'] > 0])
Out[64]: 10

In [65]: blr_italian = blr_grouped[['Locality', 'Italian Restaurant']]
         blr_italian.head()
```

Out[65]:

	Locality	Italian Restaurant
0	Appa Balwant Chowk	0.0
1	Aundh, Pune	0.0
2	Balewadi	0.1
3	Baner	0.1
4	Bavdhan	0.0

9. To Display a Map where an Area with most Restaurants is displayed in bigger circles.

```
In [70]: blr_map = folium.Map(location=[18.504220000000032,73.85302000000007 ],zoom_start=11)


#markers for localities
for latitude,longitude,name,strength in zip(neighbourhood_df["Latitudes"], neighbourhood_df["Longitudes
    folium.CircleMarker(
        [latitude, longitude],
        radius=strength*300,
        color='green',
        popup=name,
        fill=True,
        fill_color='#3186ff'
    ).add_to(blr_map)

blr_map
```



10 . Clustering Areas based on Mean of Total restaurants in that particular area.

Out[76]:

	Locality	Italian Restaurant	Cluster Label	Latitudes	Longitudes
26	Magarpatta	0.2	0	18.50927	73.93251
52	Vishrantwadi	0.2	0	18.55533	73.87492
22	Koregaon Park	0.2	0	18.53533	73.89382
28	Manjri	0.0	1	18.48194	73.86562
30	Megapolis Pune	0.0	1	18.54016	73.83355

In [77]:

```
#Cleaning the dataframe for mapping the localities according to their cluster labels
blr_only_labels = blr_labels.drop(columns=['Italian Restaurant','Latitudes','Longitudes'])
blr_only_labels.head()
```

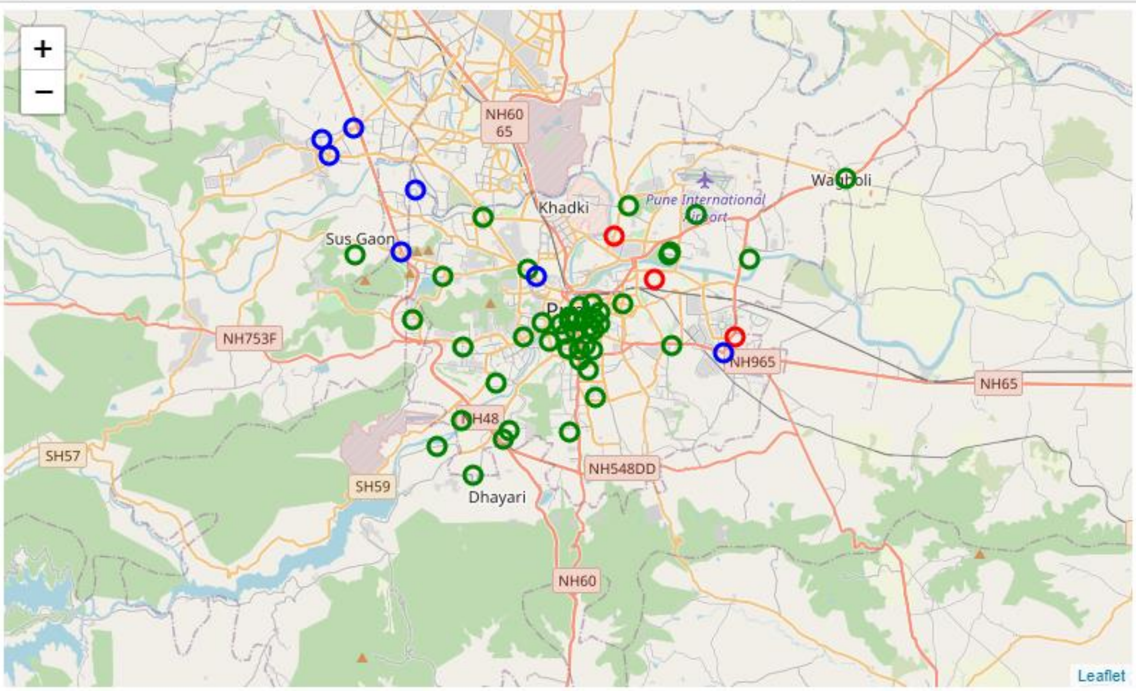
Out[77]:

	Locality	Cluster Label
26	Magarpatta	0
52	Vishrantwadi	0
22	Koregaon Park	0
28	Manjri	1
30	Megapolis Pune	1

In [79]:

```
#Plot the cluster on map
```

Out[79]:



4 Result Section

Here we observed that 3 clusters are formed and we need to analyze that which cluster is to be used for decision making. Ideally 3rd cluster with Shivaji Nagar Area having 0.1 Mean is suitable as it resembles that demand is there but supply is less. Rather than Magarpatta where Demand and supply both are high.

Result:

Cluster 1:

```
In [80]: #Cluster 1
#Dataframe containing localities with cluster label 0, which corresponds to localities with no Italian
cluster_1 = blr_labels[blr_labels['Cluster Label'] == 0]
print("There are {} localities in cluster-1".format(cluster_1.shape[0]))
mean_presence_1 = cluster_1['Italian Restaurant'].mean()
print("The mean occurence of Italian restaurant in cluster-1 is {0:.2f}".format(mean_presence_1))
cluster_1.head()
```

```
There are 3 localities in cluster-1
The mean occurence of Italian restaurant in cluster-1 is 0.20
```

Cluster 2:

```
In [81]: #Cluster 2
#Dataframe containing localities with cluster label 1, which corresponds to localities with high densit
cluster_2 = blr_labels[blr_labels['Cluster Label'] == 1]
print("There are {} localities in cluster-2".format(cluster_2.shape[0]))
mean_presence_2 = cluster_2['Italian Restaurant'].mean()
print("The mean occurence of Italian restaurant in cluster-2 is {0:.2f}".format(mean_presence_2))
cluster_2.head()
```

```
There are 46 localities in cluster-2
The mean occurence of Italian restaurant in cluster-2 is 0.00
```

Out[81]:

	Locality	Italian Restaurant	Cluster Label	Latitudes	Longitudes
28	Manjri	0.0	1	18.48194	73.86562
30	Megapolis Pune	0.0	1	18.54016	73.83355
31	Mukund Nagar	0.0	1	18.49480	73.86229
32	Nana Peth, Pune	0.0	1	18.51510	73.86787
33	Nanded City, Pune	0.0	1	18.45992	73.79015

Cluster 3:

In [82]:
```
#Cluster 3
#Dataframe containing localities with cluster label 2, which corresponds to localities with low density
cluster_3 = blr_labels[blr_labels['Cluster Label'] == 2]
print("There are {} localities in cluster-3".format(cluster_3.shape[0]))
mean_presence_3 = cluster_3['Italian Restaurant'].mean()
print("The mean occurence of Italian restaurant in cluster-3 is {0:.2f}".format(mean_presence_3))
cluster_3.head()
```

```
There are 7 localities in cluster-3
The mean occurence of Italian restaurant in cluster-3 is 0.10
```

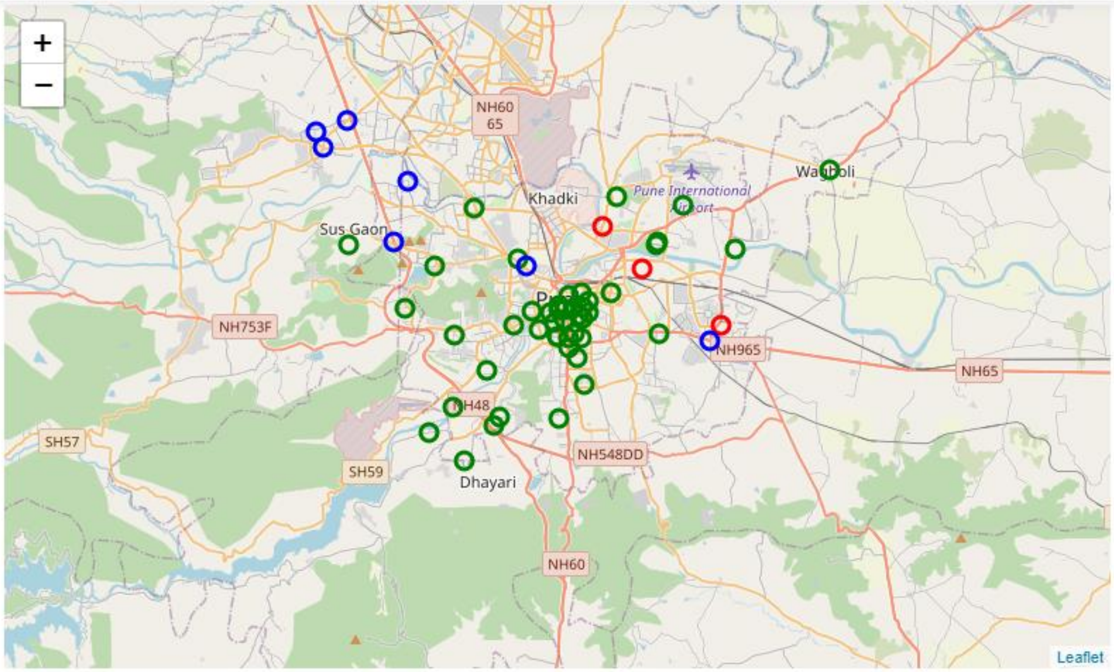
Out[82]:

	Locality	Italian Restaurant	Cluster Label	Latitudes	Longitudes
43	Shivajinagar, Pune	0.1	2	18.53723	73.83808
17	Hadapsar	0.1	2	18.50253	73.92706
18	Hinjawadi	0.1	2	18.59142	73.73895
3	Baner	0.1	2	18.54820	73.77316
2	Balewadi	0.1	2	18.57602	73.77983

5 Discussion Section:

Based on clustering we get to know that clusters give us an extra knowledge to analyze how to perceive things and come up with an exact/near to it decision.

Out[79]:



6. 3 Clusters are formed with **First and Last Most Significant**

1. Thus best Areas to Open **Italian Restaurant** are **Magarpatta,Vishrantwadi,Koregaon Park** on basis of Frequency of Italian Restaurants and also this observation seems to be True as these areas are actually good for Italian Restaurant

2. But for A Startup Italian Restaurant with no brand name and less Risk Cluster 3 Seems to be promising as it has a decent amount of Italian Restaurant and that means **demand** is there but supply is less. So **Shivaji Nagar,Hadapsar,Hinjawadi** seems to be **more promising** for Future Prospective