CAPSTONE PROJECT-THE BATTLE OF THE NEIGHBORHOODS

Applied Data Science Capstone by IBM/Coursera

Table of contents:

1. Introduction: Business Problem

"Prospects of starting a Restaurant-cum-Catering service by inspecting the Zones of Chennai"

Chennai being one of the metropolitan areas, is one of the growing IT hubs of India. With a population of 8.7 million people (86,96,010) in an area of 426 km^2, the city of Chennai also has a lot of leading industries including automobile, textile, petrochemicals, and hardware manufacturing. All of this makes it a potential place to start a new business.

While looking for places to open a business, we need to select the busiest zones in Chennai where a constant crowd is guaranteed. In a city like Chennai there will be a huge competition for businesses. Keeping this in mind, the surrounding of the selected zones should not have a lot of similar businesses as ours. Analyzing the office areas of the zones, it is expected that there will be a lot of restaurants. But a catering service is an idea which is not much explored in these areas. So, opening up a catering service which will also operate as a restaurant will be a brilliant idea to try.

The Business Problem can be stated as:

“What is the best place to open a Restaurant-cum-Catering Service in Chennai?”

Target Audience:

- The primary target audience for this project are definitely the entrepreneurs who want to open up a new business
- Investors who want to invest in good business ideas
- Offices in the locality of the business who will be interested in a contract-based catering service or employees who are interested in placing a catering order
- Students who are exploring Data Science and are trying to learn the art of telling a story by training, analyzing and learning from a data

2. Data: Requirements and collection

To open a business in an area, one needs to analyze the area, based on the average land prices, housing prices, most frequent venues, target audience, the competition and many other factors.

In this project, the data requirements and collection are as below:

Zones Data (along with Coordinates)

- **Requirement:** There are 15 zones in Chennai with a total of 200 wards. The basic data required to start this project is the names of all these Zones along with their coordinates
- **Collection:** Web scape the data of Zones of Chennai using ‘BeautifulSoup’. Use ‘Python Geocoder’ to get the latitude and longitude values of these zones.

Professional Venue Data

- **Requirement:** From these 15 zones we need to find out which zones have the most professional venues like offices, hospitals, industries, factories etc. In other words, we need to know in which zones we will have a constant flow of people (customers).
- **Collection**: Using ‘Foursquare’ by giving a specific category ID we can find the most frequent professional venues in these 15 zones.

Nearby Venues Data

- **Requirement:** We need to have an idea about the competition before we open a business. So, we need data about the most frequent venues nearby each selected zone.
- **Collection**: Explore the zones using ‘Foursquare’

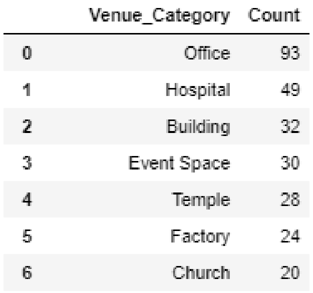
Pricing Data

- **Requirement:** Pricing data will help us in two ways:
  1. By giving us an estimate of the price values if you want to buy the land or rent it for the business.
  2. By giving us an idea about what kind of resident customers we are dealing with
- **Collection:** Websites have pricing data for all zones of Chennai. (It is generally difficult to find accurate pricing data.)

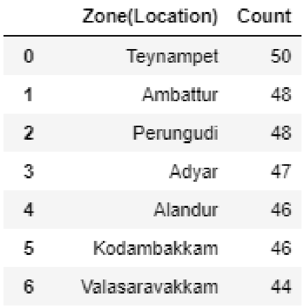
3. Methodology

Exploratory Data Analysis:

From the 494 professional venues obtained from all the 15 zones was processed to find the most frequent venue and the top professional zones.

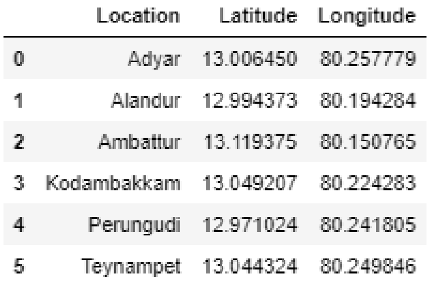


	Venue_Category	Count
0	Office	93
1	Hospital	49
2	Building	32
3	Event Space	30
4	Temple	28
5	Factory	24
6	Church	20



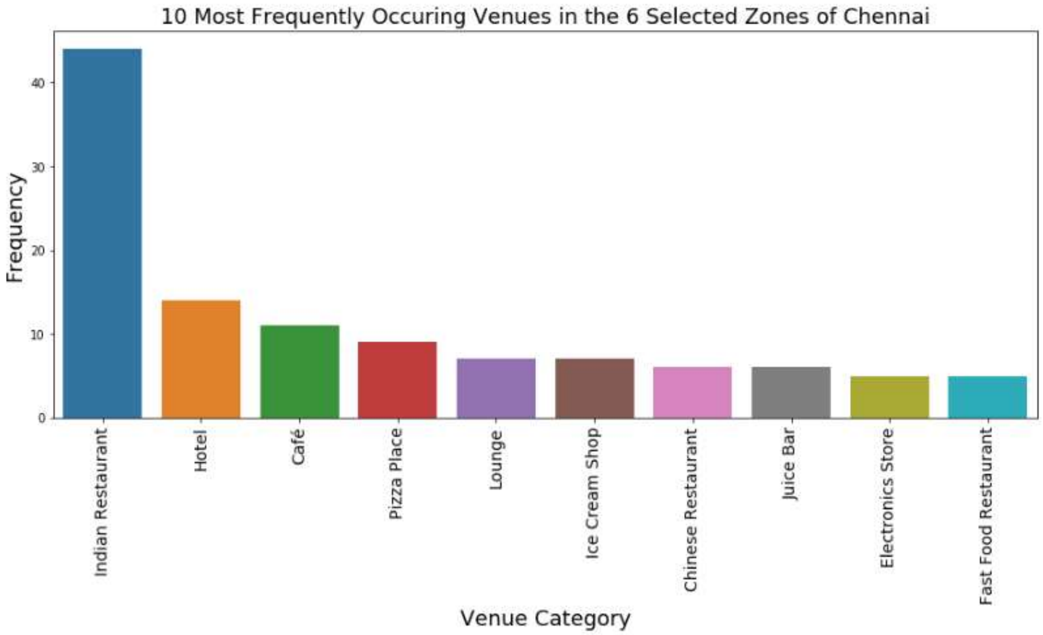
	Zone(Location)	Count
0	Teynampet	50
1	Ambattur	48
2	Perungudi	48
3	Adyar	47
4	Alandur	46
5	Kodambakkam	46
6	Valasaravakkam	44

From this analysis 6 top professional zones are selected for further analysis.

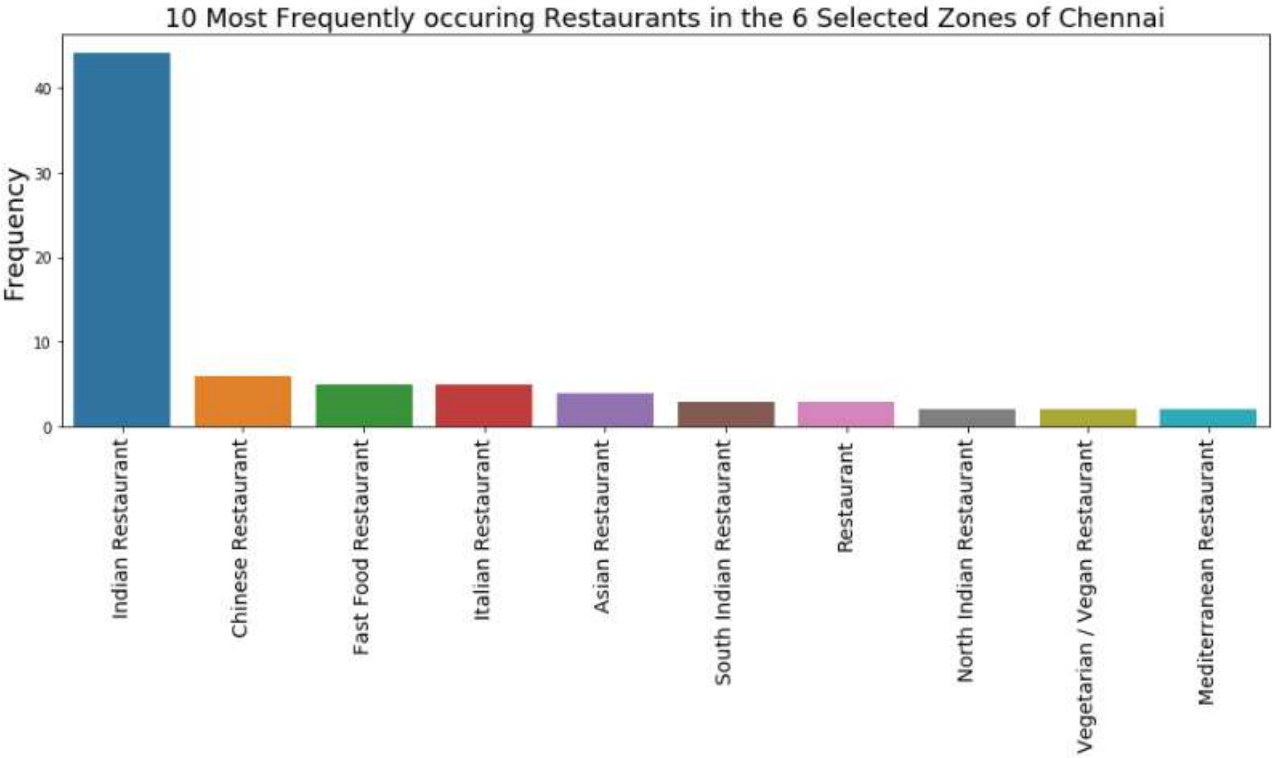


	Location	Latitude	Longitude
0	Adyar	13.006450	80.257779
1	Alandur	12.994373	80.194284
2	Ambattur	13.119375	80.150765
3	Kodambakkam	13.049207	80.224283
4	Perungudi	12.971024	80.241805
5	Teynampet	13.044324	80.249846

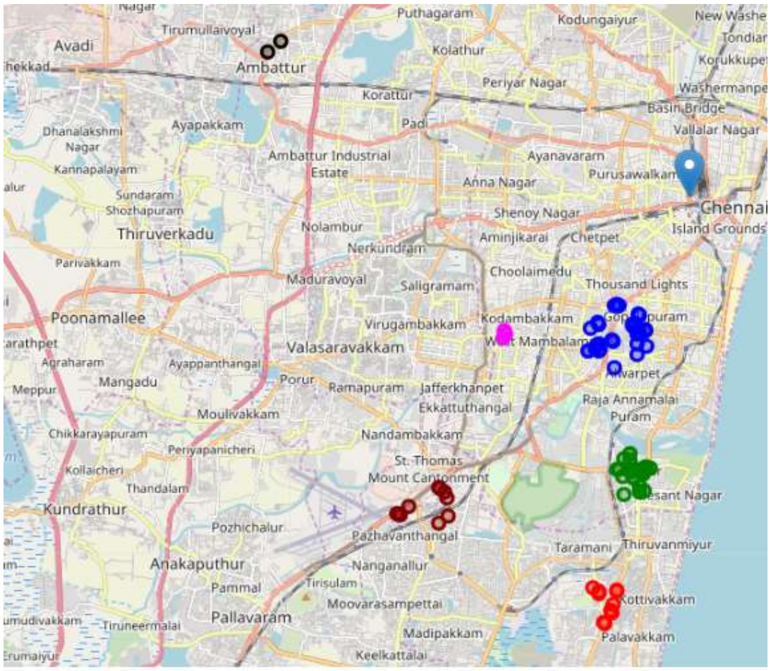
215 venues nearby these selected zones were explored using Foursquare and their details were collected.



Exploring this data, it was found that the most frequent place in these selected zones is "Indian Restaurant". All the restaurants in the selected zones were analyzed to see what the most demanded cuisine is. It was found out to be Indian cuisine.



Maximum restaurants from the collected venues were found to be located in 'Teynampet'. The map shows number of restaurants from the explored venues, per each zone selected.



```
Teynampet      32
Adyar          28
Alandur        11
Perungudi      10
Kodambakkam     3
Ambattur        2
```

This project requires us to find the business or professional zones of Chennai and explore these zones to find out the frequent venues of these zones. All of these is done so that we choose a zone which has more demand for our new business

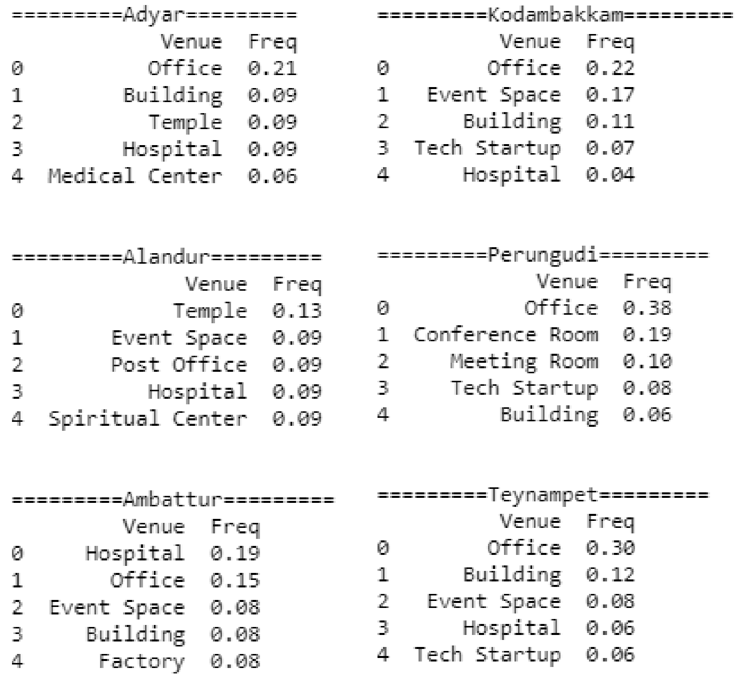
Firstly, we have collected all the required data and have done some exploratory data analysis to find the top 6 professional Zones of Chennai based on the professional venues' frequency in that zone. We found that the most frequent Professional Venue in all the zones combined is an "Office". Frequent venues were explored in these selected zones and it was found that venue category of "Indian Restaurant" is the most frequent venue nearby these selected zones. From this it is clear who our potential customers are and what they prefer.

Secondly, we need to analyze the data a little more to get insights into the venue category. This can be done by using one-hot encoding.

Thirdly. We will use a machine learning method called K-Means Cluster to cluster the zones into groups depending how similar or dissimilar they are.

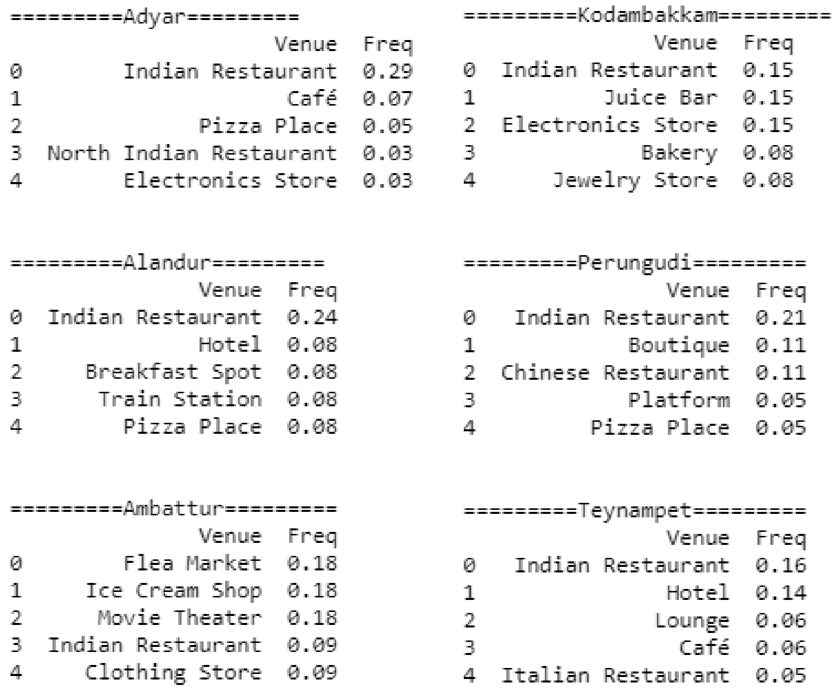
4. Analysis

Using one hot encoding the venue categories are analyzed. This is done for both professional venues and other venues for all the selected zones.



```
=========Adyar=========
            Venue  Freq
0          Office  0.21
1        Building  0.09
2          Temple  0.09
3        Hospital  0.09
4  Medical Center  0.06


=========Alandur=========
              Venue  Freq
0            Temple  0.13
1       Event Space  0.09
2       Post Office  0.09
3          Hospital  0.09
4  Spiritual Center  0.09


=========Ambattur=========
         Venue  Freq
0     Hospital  0.19
1       Office  0.15
2  Event Space  0.08
3     Building  0.08
4      Factory  0.08
```

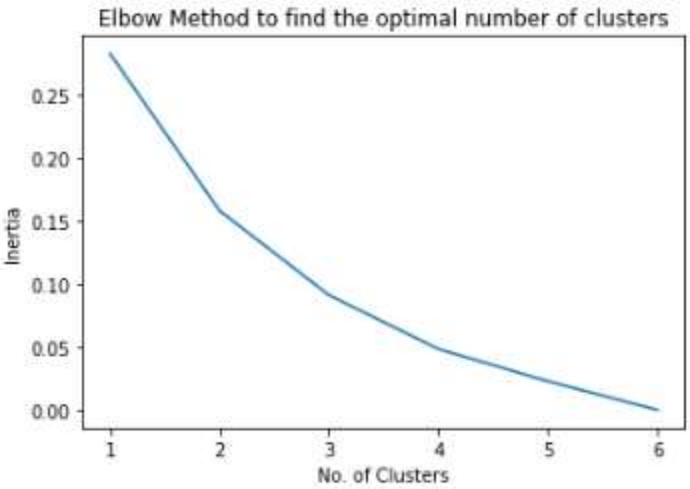
```
=========Kodambakkam=========
          Venue  Freq
0        Office  0.22
1   Event Space  0.17
2      Building  0.11
3  Tech Startup  0.07
4      Hospital  0.04


=========Perungudi=========
             Venue  Freq
0           Office  0.38
1  Conference Room  0.19
2     Meeting Room  0.10
3     Tech Startup  0.08
4         Building  0.06


=========Teynampet=========
          Venue  Freq
0        Office  0.30
1      Building  0.12
2   Event Space  0.08
3      Hospital  0.06
4  Tech Startup  0.06
```



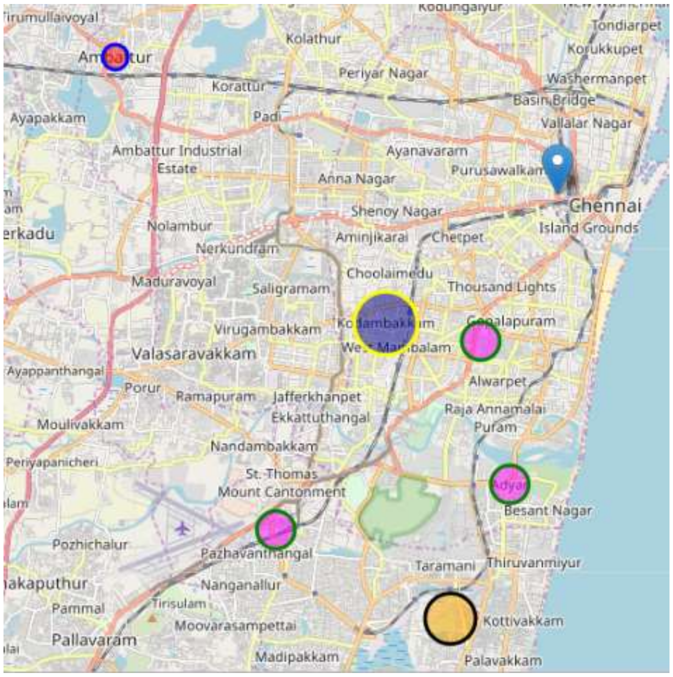
```
=========Adyar=========
                    Venue  Freq
0        Indian Restaurant  0.29
1                     Café  0.07
2              Pizza Place  0.05
3  North Indian Restaurant  0.03
4        Electronics Store  0.03


=========Alandur=========
               Venue  Freq
0  Indian Restaurant  0.24
1              Hotel  0.08
2     Breakfast Spot  0.08
3      Train Station  0.08
4        Pizza Place  0.08


=========Ambattur=========
               Venue  Freq
0        Flea Market  0.18
1      Ice Cream Shop  0.18
2      Movie Theater  0.18
3  Indian Restaurant  0.09
4     Clothing Store  0.09
```

```
=========Kodambakkam=========
               Venue  Freq
0  Indian Restaurant  0.15
1          Juice Bar  0.15
2  Electronics Store  0.15
3             Bakery  0.08
4      Jewelry Store  0.08


=========Perungudi=========
                Venue  Freq
0   Indian Restaurant  0.21
1            Boutique  0.11
2  Chinese Restaurant  0.11
3            Platform  0.05
4         Pizza Place  0.05


=========Teynampet=========
                Venue  Freq
0   Indian Restaurant  0.16
1               Hotel  0.14
2              Lounge  0.06
3                Café  0.06
4  Italian Restaurant  0.05
```

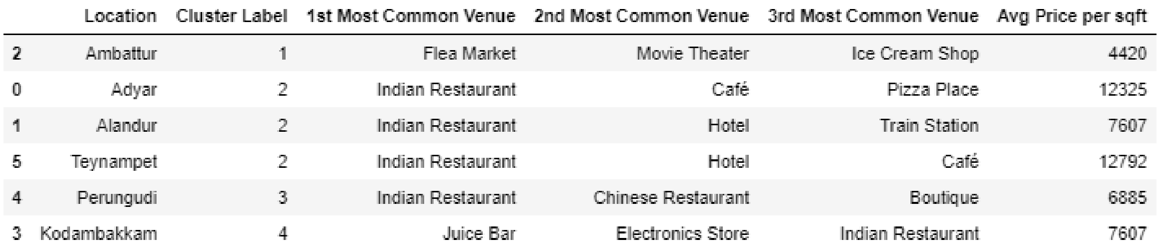
After choosing the number of clusters using Elbow method, the zones were clustered into 4 clusters using K-Means Clustering.



The clustered zones are represented on a folium map as shown in the figure below. It can clearly be seen how the zones are grouped based on the frequent places around these areas.



A final dataframe including the cluster label and average price per sqft of each zone was made.



	Location	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Avg Price per sqft
2	Ambattur	1	Flea Market	Movie Theater	Ice Cream Shop	4420
0	Adyar	2	Indian Restaurant	Café	Pizza Place	12325
1	Alandur	2	Indian Restaurant	Hotel	Train Station	7607
5	Teynampet	2	Indian Restaurant	Hotel	Café	12792
4	Perungudi	3	Indian Restaurant	Chinese Restaurant	Boutique	6885
3	Kodambakkam	4	Juice Bar	Electronics Store	Indian Restaurant	7607

5. Results and Discussion

Results:

- Cluster Results Analysis
  - Cluster 1 contains zones whose common venues are not restaurants: Entertainment and shopping areas
  - Cluster 2 contains zones whose 1st most frequent venue is a restaurant
  - Cluster 3 contains zones with top 2 most common venues being restaurants
  - Cluster 4 again contains zones whose 1st common venue is a restaurant but with the same frequency as the 2nd and 3rd common venues.
- From one hot encoding we found that ‘Adyar’, ’Kodambakkam’, ’Perungudi’ and ’Teynampet’ zones have ‘Office’ as the 1st frequent professional venue with ‘Perungudi’ having the highest frequency among them. This tells us that these 4 zones, out of the 6 selected zones, will be good for our business as our potential customers are employees.
- Also, Indian Restaurants are the most frequent venues near the selected zones, suggesting the type of cuisine customers in that area prefer. Teynampet has the most restaurants, out of all the selected zones, based on the venues explored.
- 'Ambattur' has the least average price per sqft, followed by 'Perungudi', among the selected zones.

Discussions:

Based on the clustering and exploratory data analysis it can be seen that with maximum frequency of offices and moderate restaurants in the area **‘Perungudi’** seems like a potential zone to open up our Restaurant-cum-Catering service. The pricing data also seems favorable to this. Clustering also shows these venues in cluster 3 which represents the cluster with restaurants as the frequent venues.

Although the results seem promising as Perungudi is an area with a lot of offices in the city of Chennai, further analysis needs to be done based on the wards in these zones to get a more accurate location to open up the business.

Since the clustering is done based on only the common venues obtained from Foursquare the results will need more refining. But this preliminary analysis will be of great help in the beginning stages of the business plan.

6. Conclusion

The main objective of this project was to understand how to deal with real life data science projects using some of the popular Python packages such as seaborn, folium, BeautifulSoup and geocoders. I have also got a glimpse of how web scraping is done and how FourSquare can be used to acquire data of frequent venues in a selected area.

The idea of opening a “Restaurant-cum-Catering service in an area which has a huge pool of office workers (‘Perungudi’) is an interesting and a potential idea to try in Chennai where Catering Services are not very well established. Although the analysis is very preliminary and requires a lot of refining based on the data used (refined ward data per each zone, pricing data), this analysis helped me understand Chennai more than I did in the 6 years that I stayed here, and, as mentioned earlier, will be of great help in the beginning stages of the business plan.