**Battle of Neighbourhoods of IT SEZ's in and around Chennai**

**A. Ramsrinivas**

1. **Description and Background of the Business Problem**

**Problem Statement: Exploring and Analysing the neighbourhoods of IT SEZ’s located in and around Chennai.**

Chennai is the 2nd largest Exporter of Software in India only next to Bangalore. It has the largest operations for India's top software company TCS, Infosys (has world's largest development center with 25,000 employees in Mahindra world city at Paranur near Chengalpattu, and many centers in IT corridor), and many other CMMI-level 5 companies have their head-offices, regional-offices and development centres here.

All this IT Companies need dedicated spaces to run their operations for which State Government and Private Construction Companies join together and form a zone called “Special Economic Zone” which is in short called SEZ. Examples of some SEZ’s in and around chennai are Mahindra World City, DLF Cyberpark.

This gave me an idea for my Capstone Project, that is to compare the Neighbourhoods of some of the IT SEZ’s to see how similar or dissimilar they are and explore and analyse the Commercial Venues situated around it

**Target Audience:**

1. Business personnel who want to invest in or set up a Commercial Venue around the IT SEZ’s in Chennai region.
2. Real Estate Developers who wants to develop a project near by IT SEZ’s.
3. New Employees to explore the nearby neighbourhood like who want to go to a Restaurant or Multiplex or Malls near to their offices.
4. Budding Data Scientists who do their Capstone Project’s in future using FourSquare Location data
5. **Data Preparation**

Since we need to compare the neighbourhoods of the IT SEZ’s in and around Chennai, first we need to have the list of IT Park’s located in and around Chennai, for which there cannot be a better source than Wikipedia which looks like the below.

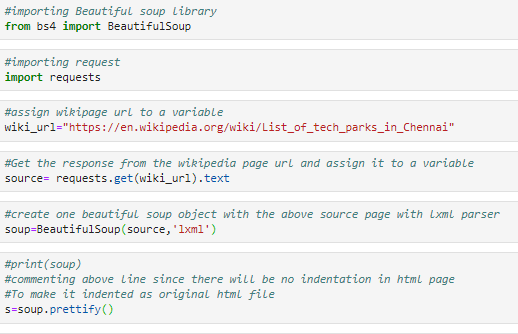


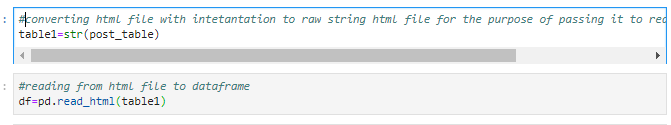
The Above table consists of SEZ name, Area (in Sq. ft),Address, Investment(in crores),Start Date Current Status and Notable Clients who rented out the SEZ space

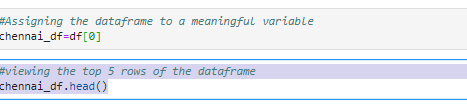
**2.1 Web Scraping:**

Since the above table is the part of the Wikipedia Page, we need to scrap the table alone from the page. For this , Requests library are used to get the XML and HTML source page from the Web and BeautifulSoup4 library object with lxml parser are used to scrap the table we need from the Source page.

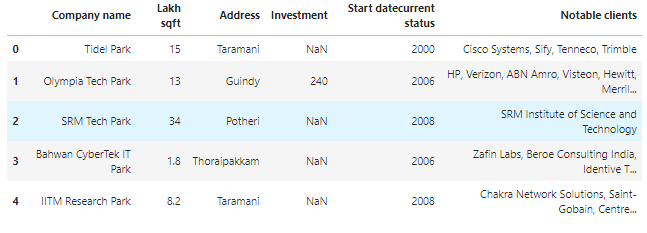
Below is the rough code snippet for it.







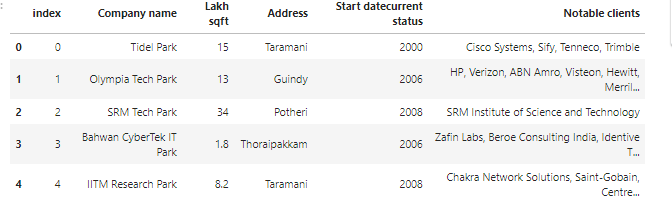
Below is the first version of the dataframe we get after the above code



**2.2 Feature Selection and Engineering:**

Firstly, I drop the Investment column, since many rows does not have it. Then I chose only the rows which are either already completed or just now completed. For this I have replaced the Under Construction and Planning rows with Null values and then dropping the corresponding rows.

After this I have resetted the row index, since I have dropped some of the rows, we get the below dataframe.

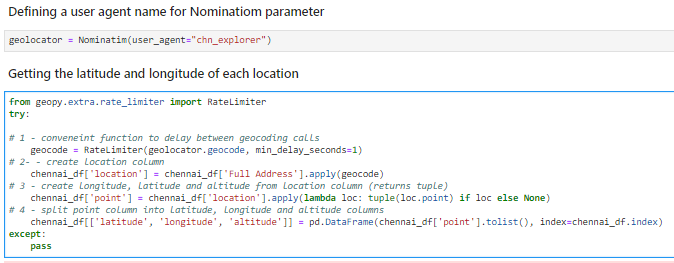


* 1. **Getting the co-ordinates of the IT Parks.**

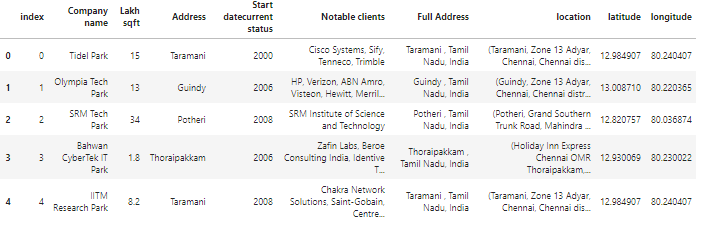
Then we need to get the co-ordinates of each of the IT SEZ in our dataframe. For this, we use Geopy library.

Then we add ‘Tamil Nadu, India’ to each values of the Address column to get the latitude and longitude of each places and store it in a separate column

Below is the rough code snippet.



After some manipulation, we get the below dataframe.



* 1. **Using FourSquare Location data**

Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these IT parks within the radius of 500m.

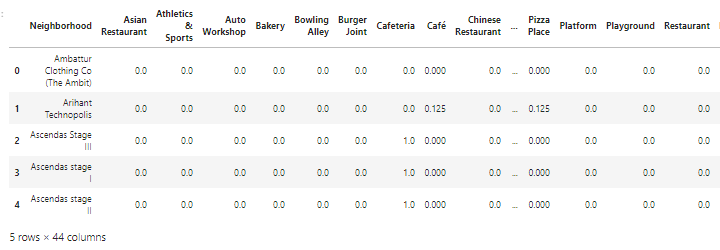
The popular spots returned depends on the highest foot traffic and thus it depends on the time when the call is made. So we may get different popular venues depending upon different time of the day. The call returns a JSON file and we need to turn that into a data-frame which looks like the below.



1. **Data Visualization and Exploratory Data Analysis**

Here we will do some basic Data Visualization and Exploratory Data Analysis. We start by applying one hot encoding technique for the venue categories to convert Venue Categories variable to Numerical Variable, then we add Neighbourhood column from the original dataframe to this encoded dataframe and move neighbourhood column to the first column

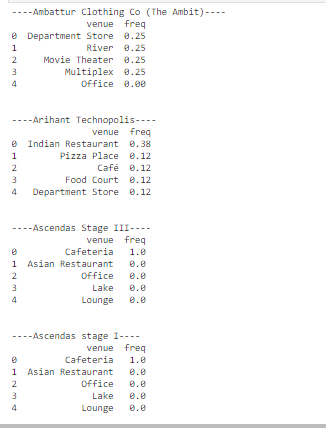
Next, let's group rows by neighbourhood and by taking the mean of the frequency of occurrence of each category. Then we get the below dataframe.



Let's print each neighbourhood along with the top 5 most common venues.

Below is the code snippet used.

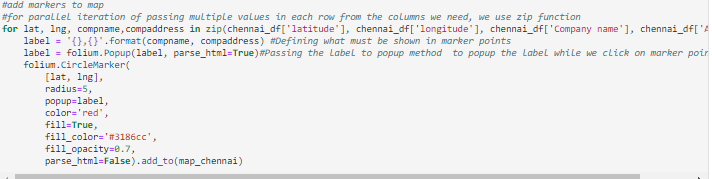




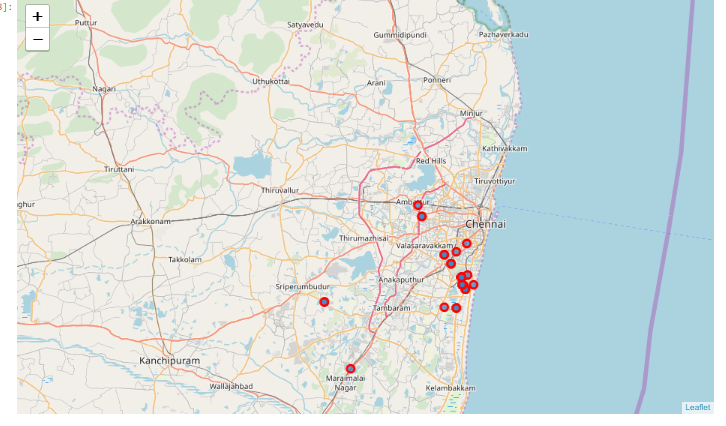
Then we Create a map of Chennai with neighbourhoods superimposed on top using Folium Library

Below is the code snippet used





**Map without clustering:**



Then we will sort the venues in the descending order and let's create the new dataframe and display the top 10 venues for each neighbourhood.

Below is the code snippet used.



Below is the dataframe we get after all this processing is done.



1. **Clustering the IT Parks**

Finally, we try to cluster the IT parks using K-means Clustering. Clustering works on the principle that similar neighbourhoods are grouped together.

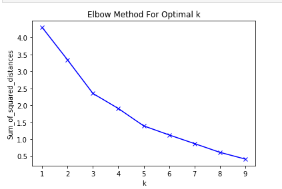
K-means clustering is one of the examples of the Non-supervised learning technique, so there is no need for separate training dataset.

First, we need to find optimal value of K by Minimizing Sum of Squared distances within the clusters.

Below is the code snippet used for it.

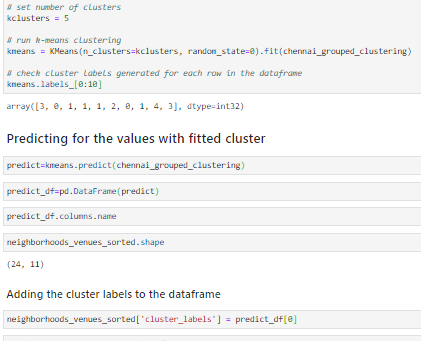


Below is the Elbow Diagram we get after we plot.



By observing the above plot, we observe that after K=5 rate of Intra Cluster Distances decreases at a small rate. There we chose K as 5

Then after choosing K=5, we fit the data and predict the data and add the cluster labels back to the dataframe. Below is the code snippet for it.

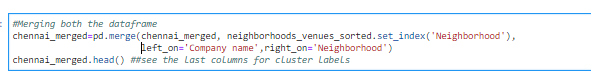


Below is the dataframe after adding the cluster labels.

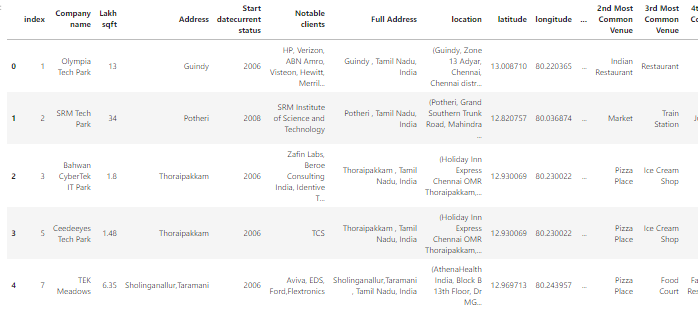


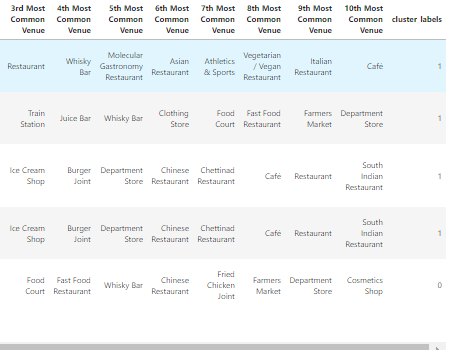
Now, Let's create a new dataframe that includes the cluster as well as the top 10 venues for each neighbourhood.

Below is the code snippet.



Below is the dataframe





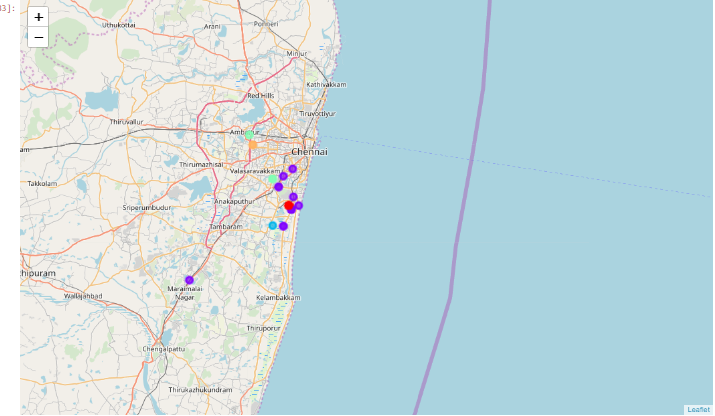
Since the dataframe is big, we split and kept it on top and bottom.

Then we will rename index column as company\_id and then we will visualize the clusters in the Map.

**Code Snippet:**



**Map with Clusters:**



1. **Results and Discussion**

**5.1 Results**

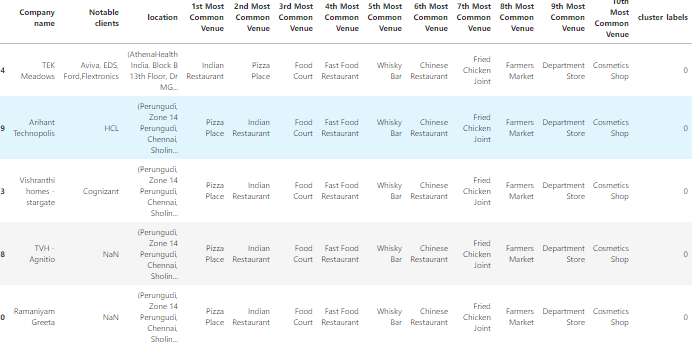
We will see the code snippet for each clusters and corresponding dataframe.

**Cluster1:**

**Code Snippet:**



**Dataframe:**

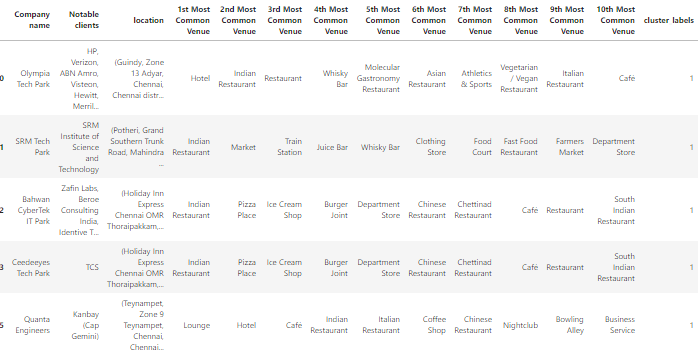


**Cluster2:**

**Code Snippet:**



**Dataframe:**

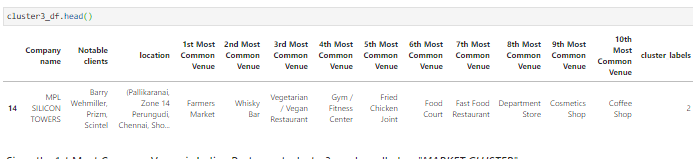


**Cluster3:**

**Code Snippet:**



**Dataframe:**



**Cluster4:**

**Code Snippet:**



**Dataframe:**

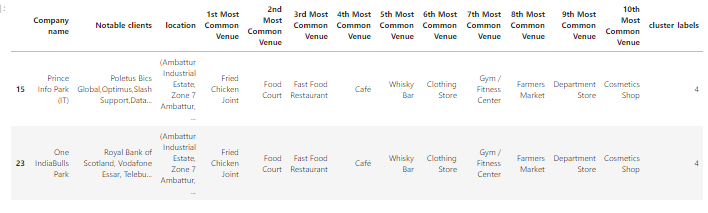


**Cluster5:**

**Code Snippet:**



**Dataframe:**



We will name the Clusters based on the 1st Most Common Venue.

Cluster1 - PIZZA CLUSTER

Cluster2 - INDIAN RESTAURANT CLUSTER

Cluster3 - MARKET CLUSTER

Cluster4 - ASIAN RESTAURANT CLUSTER

Cluster5 - FRIED CHICKEN CLUSTER

**5.2 Discussion**

We will discuss the above results based on mainly 2 person’s point of view: ENTERPRENEUR AND EMPLOYEE

**5.2.1 Entrepreneur point of View**

All the Businesspersons or entrepreneur think on the line that they can set up their shop where there is less competition for the type of shops in that area. We will choose less competition category from the least common venue

We will see just one area from any one cluster for explanation purpose.

*Arihant Technopolis* in *Perungudi* has less no of **Cosmetics shops**, so **Cosmetics Shop vendor** can set up his shop there for better revenue.

**5.2.2 Employee point of View**

Employee thinks on the line that they can explore more varieties and buy products or services at a less price, if the same type of shops is more, we will determine that by 1st Most Common venue.

As seen before,we will see just one area from any one cluster for explanation purpose.

*Prince Info park* in *Ambattur* has more no of **Fried Chicken Shops** , so a person can try Chicken on all the shops and chose his favourite shop depends on the taste and price.

1. **Conclusion**

I would like to conclude my analysis by emphasizing on few points.

* We got a glimpse on how real-world data science projects looks like.
* We got the hands on Geopy, Folium Library and Foursquare API.
* Also tried K-means clustering technique to group the neighbourhoods of IT parks.
* There are few limitations like I have chosen radius as 500m, results may vary if the radius increases or decreases.
* Also, I have used K -means clustering technique, results and accuracy may vary if you chose some other clustering techniques.

Finally, I would like to thank all my tutors from **IBM - Coursera** for teaching me many things on Data Science. Hopefully, this project provides a BIG PICTURE for budding data scientist in future who enrol this beautiful IBM-Coursera Certificate Course.

1. **References**
2. <https://en.wikipedia.org/wiki/List_of_tech_parks_in_Chennai>
3. <https://en.wikipedia.org/wiki/Software_industry_in_Chennai>
4. <https://towardsdatascience.com/geocode-with-python-161ec1e62b89>