IBM Data Science – Final Report

MAY 3

Authored by: Eswar Sunder

The Battle of the Neighborhoods

This report is a part of the final IBM capstone project.

This project is designed such a way it doesn't take lot of time from my current work life. So, this going to be a simple project without much assumptions or pushing the limits.

Problem statement:

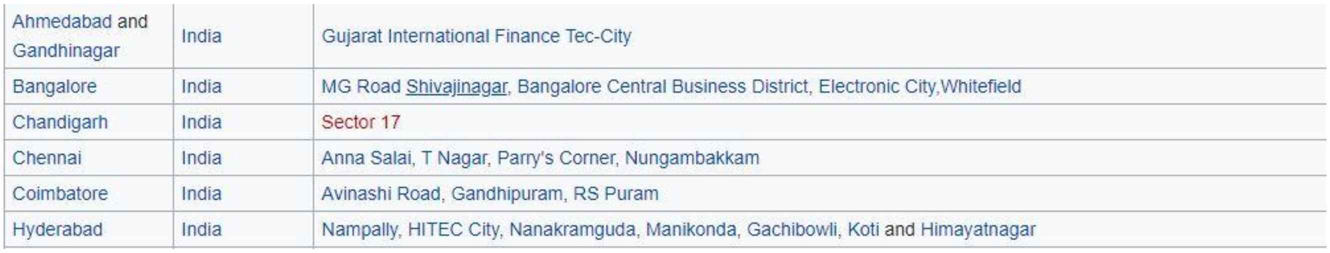
A Multinational Company X wants to setup its first branch in India. The choice of the place where the company is going to setup will depend on the venues in that area. In order to identify the ideal place, the company is interested to do analysis on the Central Business Districts of India. Central Business districts are the commercial and business center of the city. In other words, this is the "happening" places in the city. In India, there are around 12 top tier cities and each of them have their own one or many districts. The Company has many businesses and wants to keep the identity confidential. The analyst doesn't know for which business the company are picking the place to remove any bias from the analysis. So, the company profile or type of business or any other information is unknown.

The outline of the objectives:

- Identify the CBDs in India
- Collect the venue data of the CBDs.
- Perform a clustering analysis among the CBD to identify the similar groups
- Profile the clusters based on the analysis

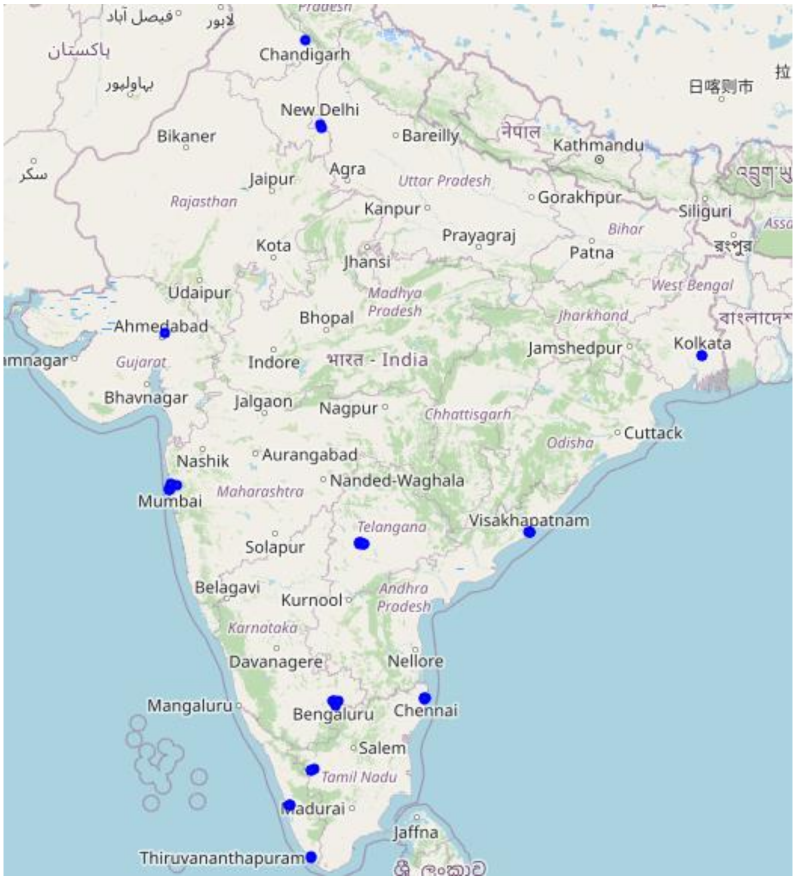
Data Sources and Description:

In order to identify the CBDs, we can scrape the Wikipedia website A snippet of the table that should be scrapped is shown below.

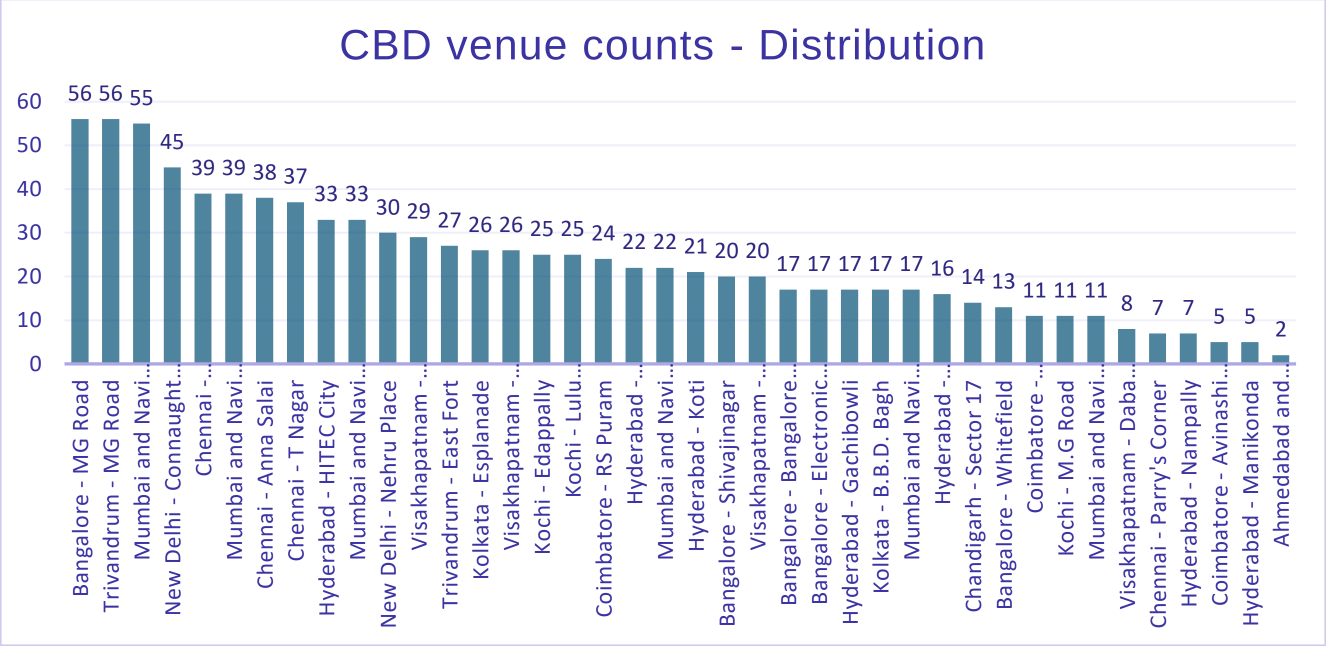


Ahmedabad and Gandhinagar	India	Gujarat International Finance Tec-City
Bangalore	India	MG Road Shivajinagar, Bangalore Central Business District, Electronic City,Whitefield
Chandigarh	India	Sector 17
Chennai	India	Anna Salai, T Nagar, Parry's Corner, Nungambakkam
Coimbatore	India	Avinashi Road, Gandhipuram, RS Puram
Hyderabad	India	Nampally, HITEC City, Nanakramguda, Manikonda, Gachibowli, Koti and Himayatnagar

In the table each place separated by comma is a CBD. One missing key information is the latitudes and longitudes of the CBD in the wiki page. Considering there are only 40 CBDs, we can manually find them. There are going to be 2 levels of data. One is a city level data, and another is CBD level data. Any information that is common for the city for example crime rate, climate, temperature etc. will not be used in the clustering exercise so that the final cluster won't have inherent bias based on the city information. If we use this information, then CBDs in the same city is more likely to be clubbed together. In the map of India, the CBD locations are like the following—



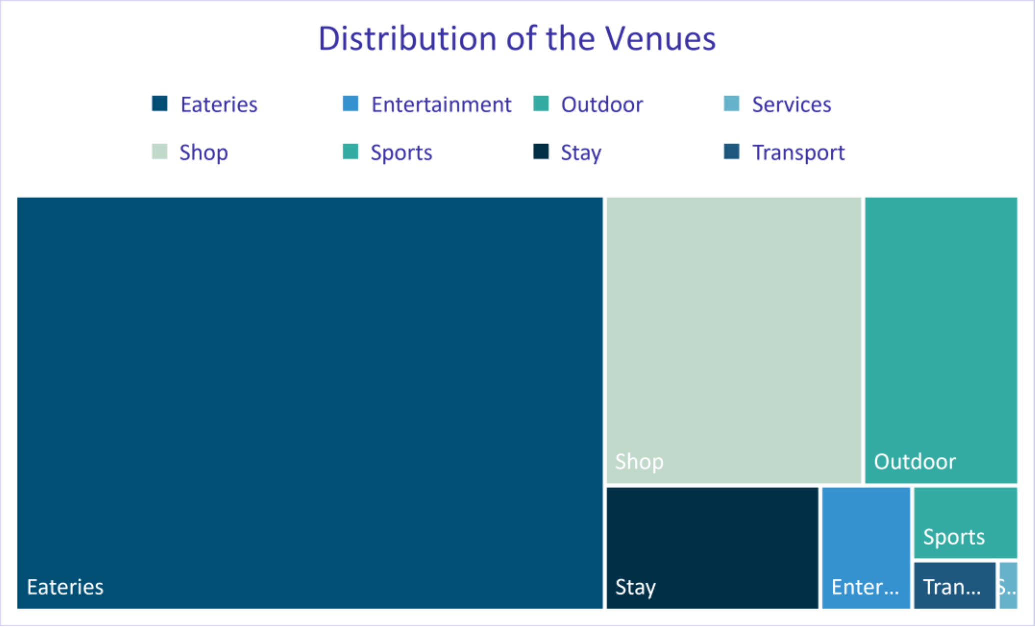
Using Foursquare API and the latitudes and longitudes, we can get the information of the specific Area. Now we have a data with venues and the dataset is ready for any clustering exercise. The final dataset that goes into the clustering algorithm should be in the following format; CBDs as rows and the venue categorical aggregated over the n venues in that CBD as the column. There are around 185 venue categories.



Methodology:

a. **Exploratory analysis**:

Upon the first inspection, it is quite apparent that foursquare data is not rich in quite a few CBDs. Please check the figure below. Nevertheless, I want to send all the data into pipeline and choose an appropriate clustering algorithm to isolate them automatically. Most of the venues we get is predominantly eateries and food outlets. The distribution of the venue categories is shown in the table below. This means the clustering exercise is going to be mostly driven by the type of food outlets in the area.



Where the groups mentioned above consists of the following,

Eateries	Restaurants, Snack bar etc
Entertainment	Move theater, multiplex etc
Outdoor	Beach, Park etc
Services	General Services
Shop	All Stores and shops
Sports	Stadiums and courts
Stay	Hotel, hostel and BnBs
Transport	Bus, train and other transport terminals

b. **Choice of the Clustering algorithm.**

Choosing the right Cluttering algorithm is a bit tricky. We did Agglomerative hierarchical clustering in the data. In Agglomerative or bottom-up clustering method, in the starting stage each observation is assigned to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat the steps until there is only a single cluster left.

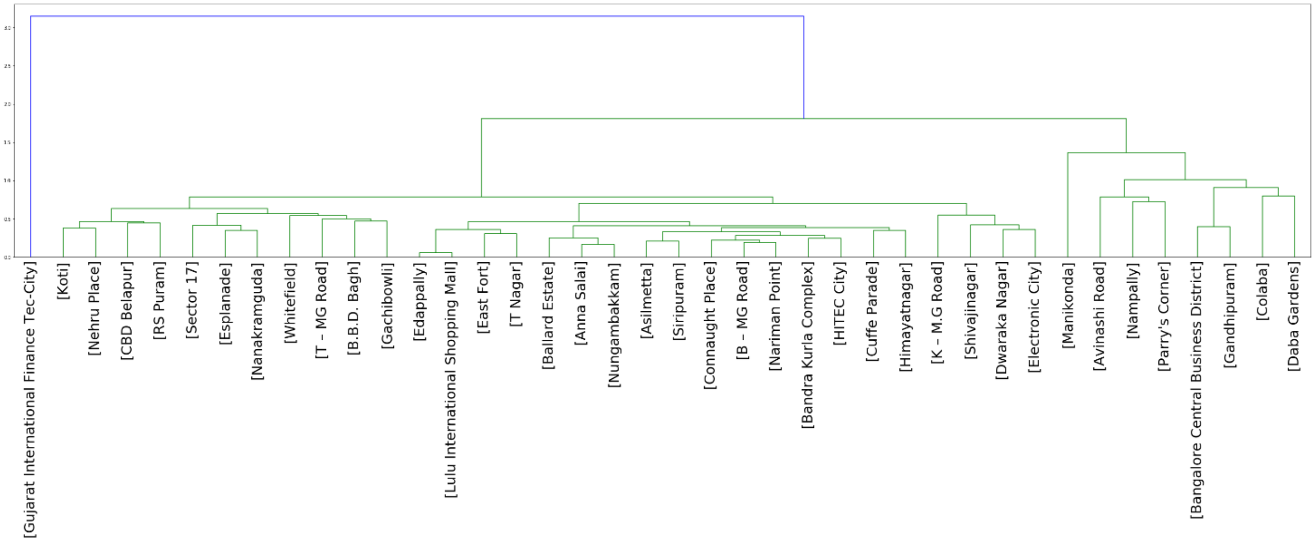
There are two advantages of this algorithm.

1. Euclidean distance is a good measure of the similarity in this case.
2. Client can choose any place that is closest to the ideal branch from the dendrogram.

There is not going to be any new feature engineering or preprocessing. The features as it stands now is as granular it can get. Also, the zeroes in the column is key for clustering. In order to emphasis the point, let's say for example we have two data points place 1 with only 2 venues populated out of the 185 categories and place to with 50 categories. If we use any transformation say min-max transformation of the data, then zero will tend to mean something while calculating Euclidean distance. So, the data is good as it stands now.

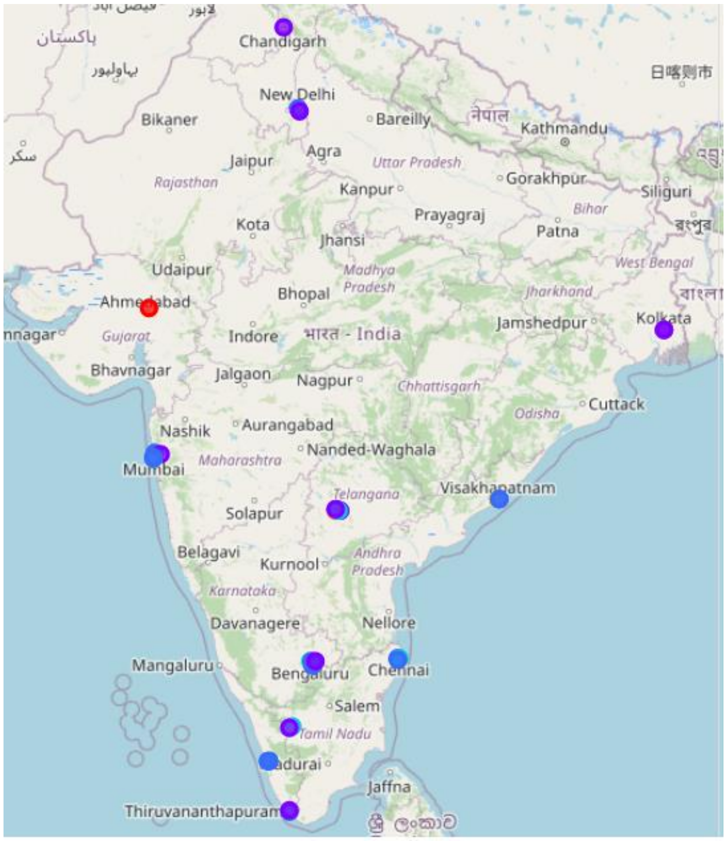
Results:

As the data is predominantly restaurants and eateries, as expected they drive the solution. The results are decent and make intuitive sense. The dendrogram is shown below.



Based on the analysis, Ideally, if we use the distance criterion, we will get 2 clusters, one with 39 CBDs, which is not a great solution. As the CBDs have almost similar features based on the Foursquare data it is hard to differentiate them. In the final solution we forced the program to pick 8 clusters which is the least number to separate the 30 CBDs.

The actual results with 8 clusters had 4 one element cluster. They are outliers with respect to the foursquare data. But the 4 individual cluster makes little sense from business standpoint and we cannot omit any places as we don't know much about the company. We rolled the individual groups into 1 cluster called others. Finally we have 5 clusters. Final clusters are attached in the Appendix. The final clusters in Map of India looks like the following.



Discussion:

The interpretation of the clusters are pretty straight forward. We have to see the Venue categories composition with respect to the overall composition. It is called as index numbers. The formula is

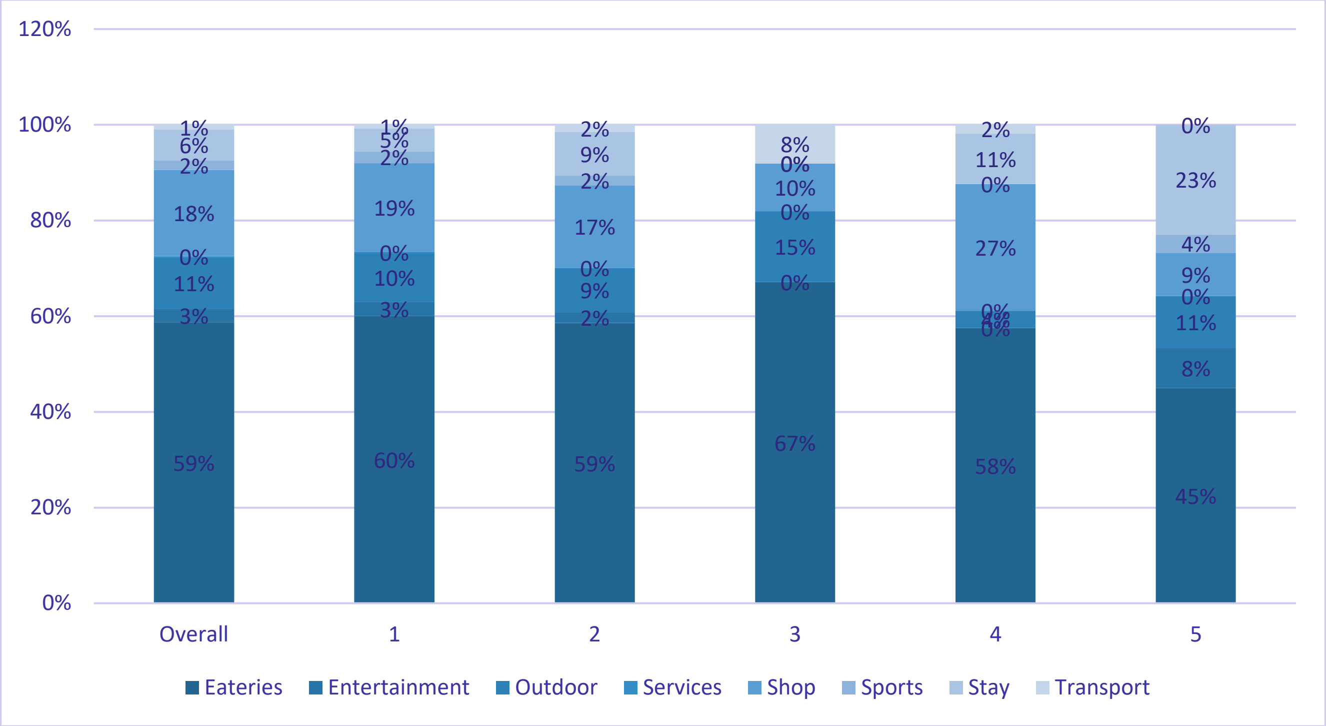
$$Index\ value = \frac{\%\ of\ Venue\ category\ in\ Cluster}{\%\ of\ Venue\ category\ overall} \times 100$$

For example, for Eateries in cluster 3, the index value is

$$Index\ value_{Eateries\ in\ cluster\ 3} = \frac{67\%}{69\%} \times 100 = 113$$

Then we can consider a threshold (which is subjective), in this case I have considered 95 and 105 as thresholds. If an index value is less than 95 then it is skewing low, which means the group has less of it and if it is more than 105; it is skewing high and cluster has more of it.

The percentages of the venue categories within group and in an overall is shown in the chart below. The table below the chart shows the pattern in this data. Dark blue flags are skewing high and cells with light colors are skewing low.



	Eateries	Entertainment	Outdoor	Services	Shop	Sports	Stay	Transport
Cluster 1			░	▨		▨	░	░
Cluster 2		░	░	░			▨	▨
Cluster 3	▨	░	▨	░	░	░	░	▨
Cluster 4		░	░	░	▨	░	▨	▨
Cluster 5	░	▨		░	░	▨	▨	░

In this case, we can see that, there are business areas that are up for grabs aka opportunity areas and few areas that are ideal to depend on them. All the negative skews, are the opportunity cells. It goes without saying if the company's business involves one of the categories and needs less competition for example, Eateries in Cluster 5 area is an opportunity area. In the above analysis, As we don't know the nature of the company we cannot suggest any cluster as our final solution. In this we will list out the advantages in each cluster and the client and pick the one that fits their profile.

Cluster 1: The CBDs under this cluster are the places with lot of venues. If the company is dependent on all the above groups. Cluster 1 is the ideal place to loot for.

Cluster 2: With high skews in Stay and Transport. If the company wanted a place with high connectivity, this is the cluster to look.

Cluster 3: This is ideal place for companies that requires Eateries, Transport and Outdoor like educational or creative institutions.

Cluster 4: Companies that involves Travels and Tourism. These places are rich in Shops, Stay, Transport. So, it can aid their cause.

Cluster 5: This is probably opportunity CBDs where the company can drive growth.

Conclusion:

The final capstone project was fun. As for as the problem statement and results are concerned, it is decent but there are so many places for improvement. Like we can consider some 3rd party data that involves the crime rates in the area which can be used to make the decision on the safety of the employees, number of universities in the area, which can be used from growth and recruitment perspective, etc. We can do a principle component analysis on the features to get the latent features. One other area which is a problem from my perspective is the sample size. We are measuring 187 odd variables for 40 CBD is very odd. The results as is pretty decent though.

Appendix:

The final tags

City	CBD	Final tags
Chennai	Anna Salai	1
Visakhapatnam	Asilmetta	1
Coimbatore	Avinashi Road	3
Bangalore	MG Road	1
Kolkata	B.B.D. Bagh	2
Mumbai and Navi Mumbai	Ballard Estate	1
Mumbai and Navi Mumbai	Bandra Kurla Complex	1
Bangalore	Bangalore Central Business District	4
Mumbai and Navi Mumbai	CBD Belapur	2
Mumbai and Navi Mumbai	Colaba	5
New Delhi	Connaught Place	1
Mumbai and Navi Mumbai	Cuffe Parade	1
Visakhapatnam	Daba Gardens	5
Visakhapatnam	Dwaraka Nagar	1
Trivandrum	East Fort	1
Kochi	Edappally	1

Bangalore	Electronic City	1
Kolkata	Esplanade	2
Hyderabad	Gachibowli	2
Coimbatore	Gandhipuram	4
Ahmedabad and Gandhinagar	Gujarat International Finance Tec-City	5
Hyderabad	HITEC City	1
Hyderabad	Himayatnagar	1
Kochi	M.G Road	1
Hyderabad	Koti	2
Kochi	Lulu International Shopping Mall	1
Hyderabad	Manikonda	5
Hyderabad	Nampally	3
Hyderabad	Nanakramguda	2
Mumbai and Navi Mumbai	Nariman Point	1
New Delhi	Nehru Place	2
Chennai	Nungambakkam	1
Chennai	Parry's Corner	3
Coimbatore	RS Puram	2
Chandigarh	Sector 17	2
Bangalore	Shivajinagar	1
Visakhapatnam	Siripuram	1
Chennai	T Nagar	1
Trivandrum	MG Road	2
Bangalore	Whitefield	2