****

IBM Data Science Capstone Project

Submitted by:

Aveenash J

2020

Table of Contents

[**1.Introduction** 2](#_Toc37078175)

[**1.1 Background** 2](#_Toc37078176)

[**1.2 Problem** 2](#_Toc37078177)

[**1.3 Interest** 2](#_Toc37078178)

[**2. Data Acquisition, Data Cleaning and Data Analysis** 3](#_Toc37078179)

[**2.1 Brief** 3](#_Toc37078180)

[**2.2 K-means clustering** 5](#_Toc37078181)

[**2.3 Links to data sources** 6](#_Toc37078182)

[**3.Findings and Conclusion from the data** 7](#_Toc37078183)

[**3.1 Findings** 7](#_Toc37078184)

[**3.2 Conclusion** 10](#_Toc37078185)

# **1.Introduction**

## **1.1 Background**

Chennai is one of the metropolitan cities in the southern part of India and is the capital of the state of Tamil Nadu. Located on the Coromandel Coast off the Bay of Bengal, it is the biggest cultural, economic and educational centre of south India. According to the 2011 Indian census, it is the sixth-most populous city and fourth-most populous urban agglomeration in India. In such a diverse city, there is always scope for starting a restaurant business.

In Chennai, there are a few neighbourhoods which are already saturated with a lot of restaurants. If a person establishes a restaurant in one of these neighbourhood, he may have an issue in terms of the competition and that too from big players.

## **1.2 Problem**

The big question here is where should one look at starting a restaurant in Chennai? How does one determine which neighbourhood to go for?

In this particular project, I will talk further about the different parameters I would consider before choosing the location and the type of restaurant I would go for.

## **1.3 Interest**

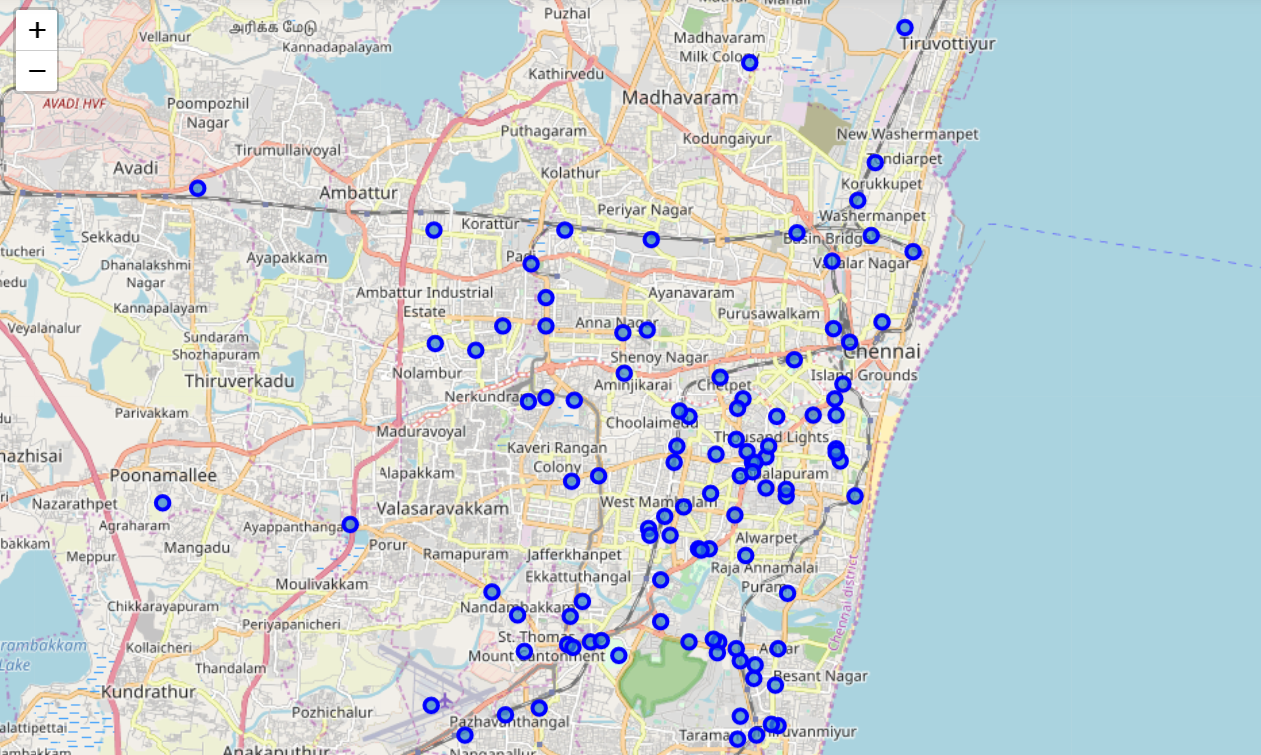
Anyone who is looking to open a restaurant in Chennai may get some insights on the choice of location.

# **2. Data Acquisition, Data Cleaning and Data Analysis**

## **2.1 Brief**

The starting point to the project was collecting the geographical data with respect to the neighbourhoods within Chennai. This scrapped this data from one of the websites (link below), although the list of neighbourhoods was not exhaustive, I believed that it was sufficient enough to reach the goals I had in mind.

The below map shows the neighbourhoods that I have select for my analysis.



Map 1: Neighbourhoods selected

**Note:** The only major drawback was that Foursquare had limited data with respect to venues in my city. However, the same model may be used on other neighbourhoods as well where the results will be more accurate.

The latitude and longitude data of neighbourhoods that was obtained by web scrapping was the starting point, this data was then used to obtain details of places within a 500-meter radius of each of the neighbourhoods.

The below table shows details of some of the nearby venues that was obtained from Foursquare.



Table 1: details of some of the nearby venues that was obtained from Foursquare. (This table is not exhaustive; it is only a part of all the data obtained)

Then, the data from Foursqare was sorted and converted into meaningful tables which is when I determined that Foursquare did not have sufficient data with respect to nearby venues for many of the above neighbourhoods. I filtered out the neighbourhoods where I did not have sufficient data.

Finally, I came up with a table which showed details of 10 most common venues in each of the neighbourhoods.



Table 2: 10 most common venues in each of the neighbourhoods.

## **2.2 K-means clustering**

I was working with data relating to more than 45 neighbourhoods, and the quantum of data being manipulated was huge. In order to make the entire process simpler, I decided to apply k-means clustering concept on the dataset.

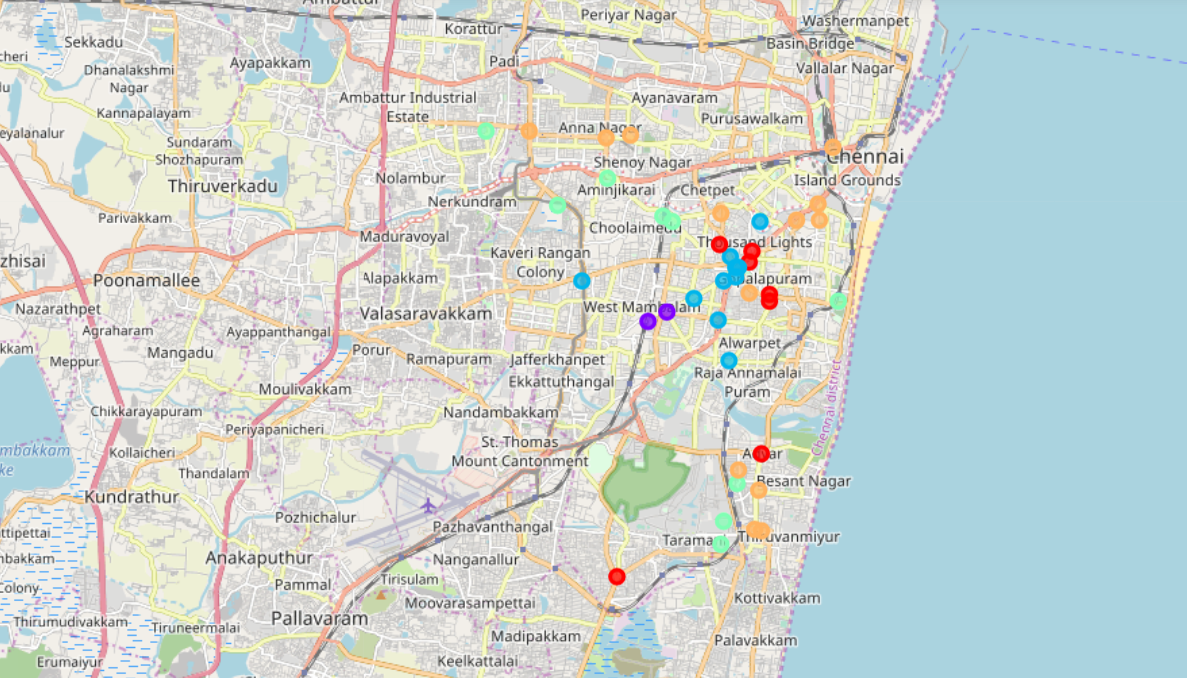
K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data.

I decided to divide the neighbourhoods into 5 different clusters based on their characteristics. The table below is a consolidated table with the cluster labels.



Table 3: consolidated table with the cluster labels

Below is the map showing how the neighbourhoods have been clustered.



Map 2: Neighbourhoods after clustering

## **2.3 Links to data sources**

1. <https://chennaiiq.com/chennai/latitude_longitude_areas.asp>
2. <https://foursquare.com/>

# **3.Findings and Conclusion from the data**

## **3.1 Findings**

Based on the analysis, I determined that clusters 0,2 and 4 had a good number of restaurants and fast food joints and most of these neighbourhoods were in the heart of the city.

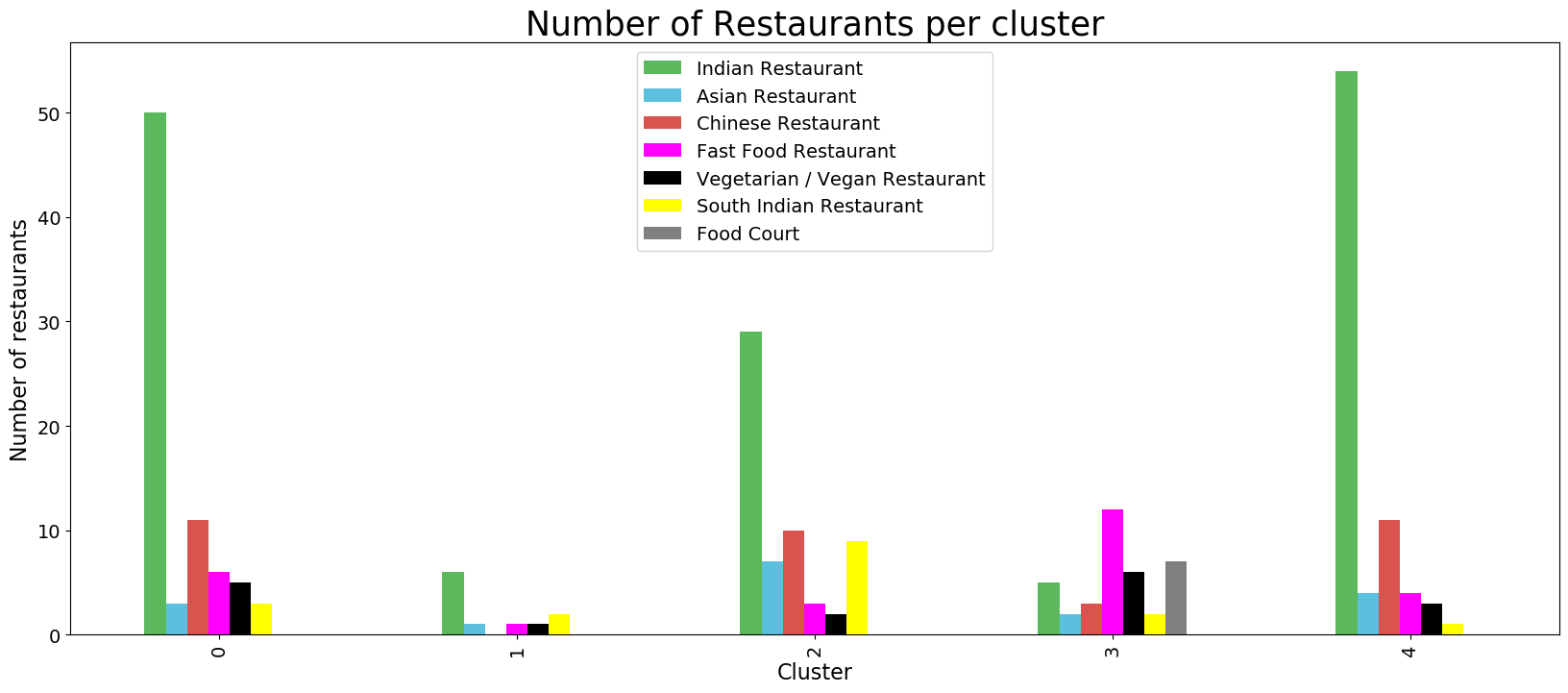


Table 4: Breakdown of restaurants per cluster

Now let us look at the neighbourhoods in these clusters and the top 10 most common venues in these neighbourhoods.

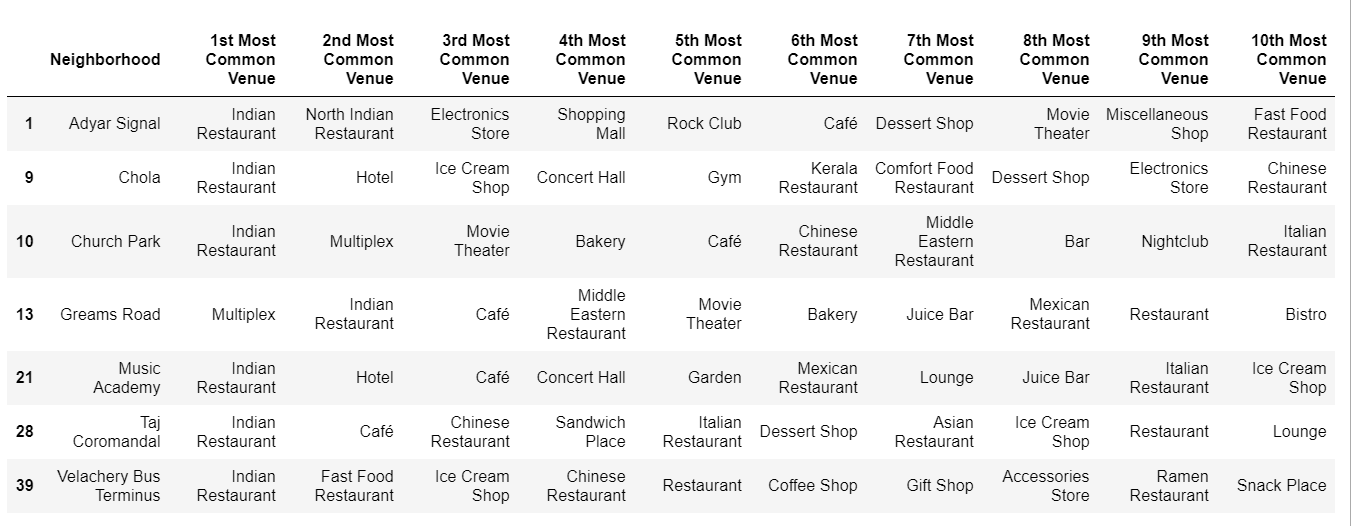


Table 5: Cluster 0 neighbourhoods and common venues



Table 6: Cluster 2 neighbourhoods and common venues

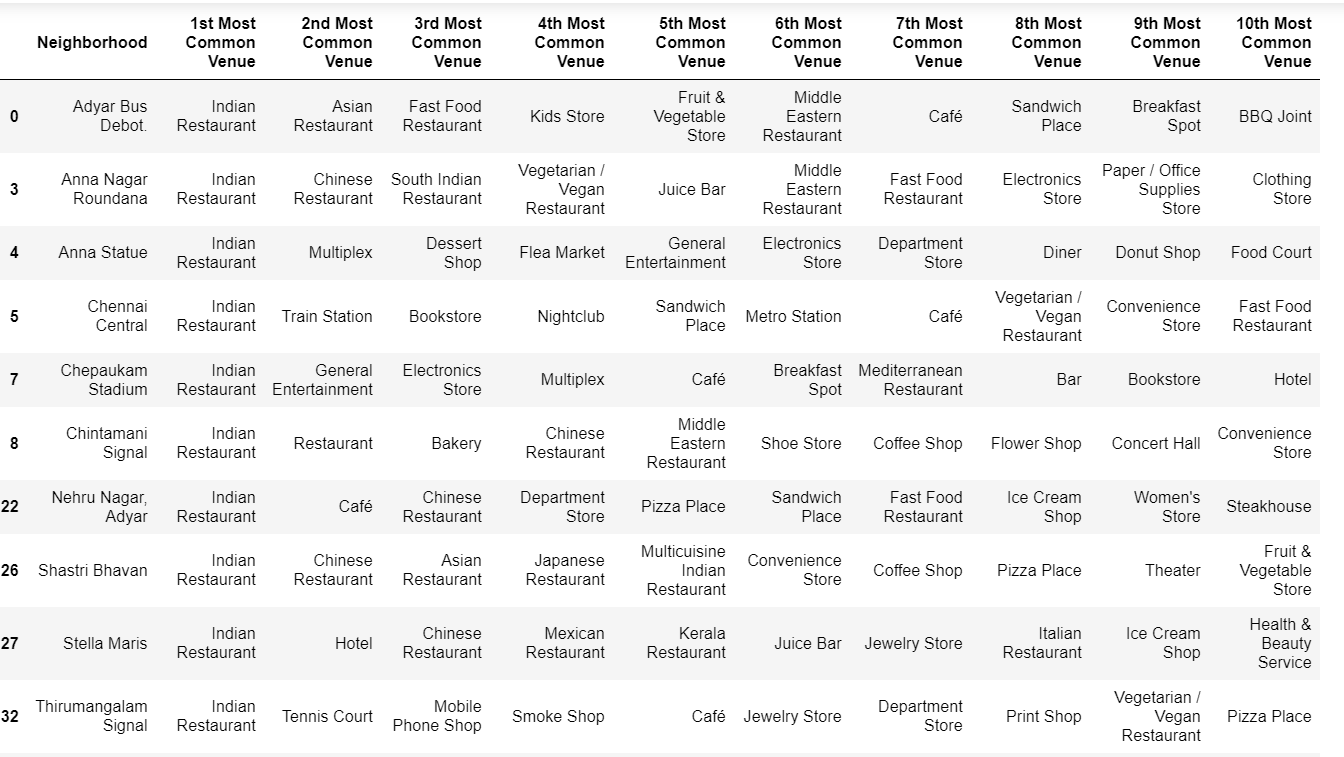


Table 8: Cluster 4 neighbourhoods and common venues (Not exhaustive)

Looking at the data, it can be inferred that the neighbourhoods in the above clusters have sufficient number of restaurants which may be capable enough to meet the demand. Therefore, it may not be a wise decision to establish a restaurant in one of these neighbourhoods.

If we look at the average restaurants per neighbourhood in each of the clusters, the data would support the above inference.

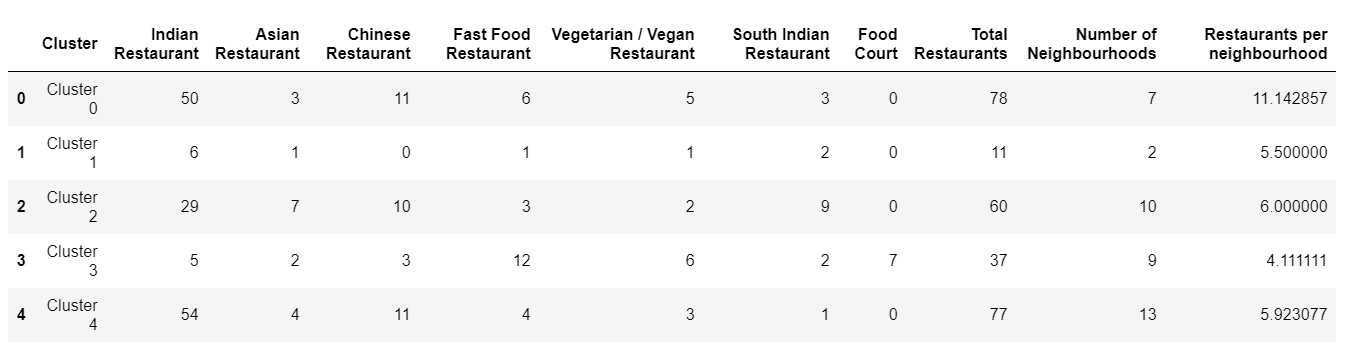
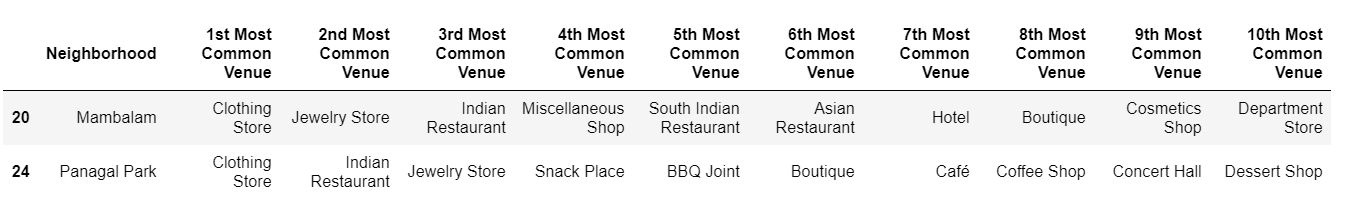


Table 9: Restaurants per neighbourhood

## **3.2 Conclusion**

Having looked at the average number of restaurants per neighbourhood in each of the clusters, it is clear that the neighbourhoods in clusters 1 and 3 are underserved.

Before making our decision, let us look at the neighbourhoods in cluster 1 and 3 to understand if there are any noticeable trends.

Table 10: Cluster 1 neighbourhoods and common venues

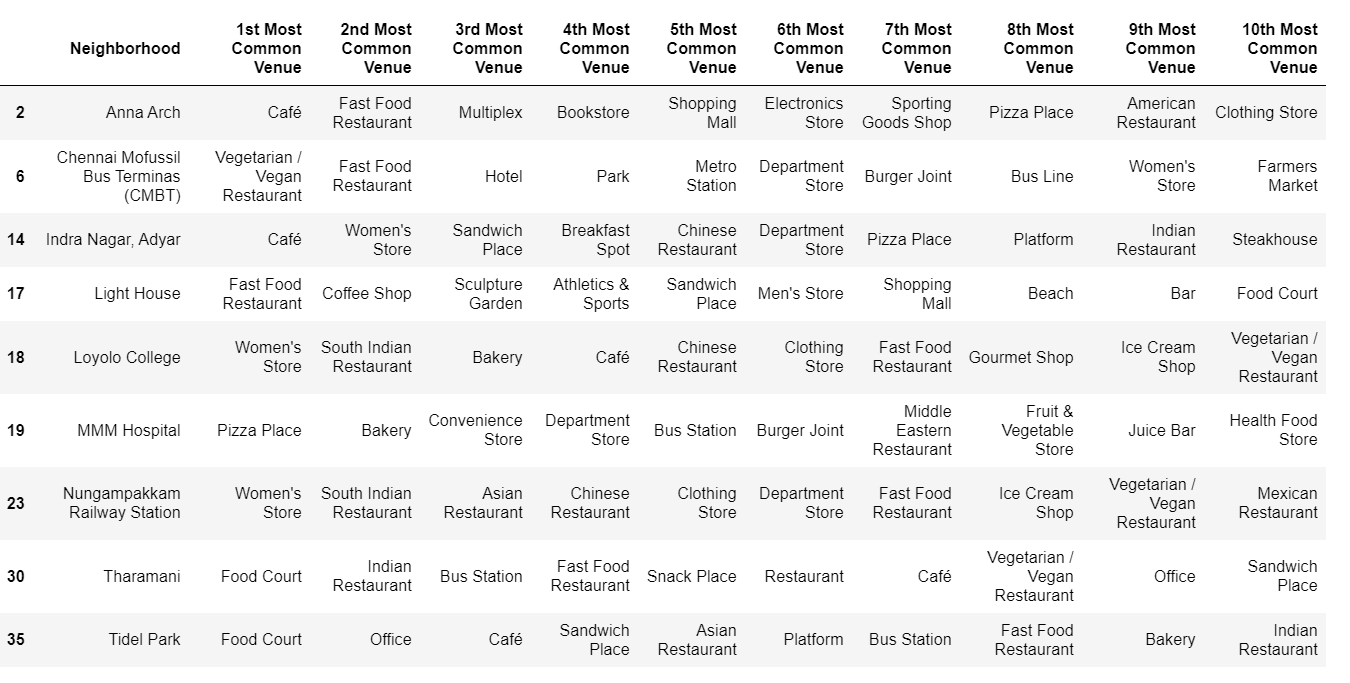


Table 11: Cluster 3 neighbourhoods and common venues

Looking at the average restaurants per cluster table, one may jump to a conclusion that the best choice of location for a restaurant would be a neighbourhood in cluster 3. Upon further analysis, it is clear that one of the neighbourhoods “MMM Hospital” in the cluster is the major reason for the reduced average, and it may not be a wise decision to start a restaurant next to a hospital. Also, it looks as if all the other neighbourhoods in this cluster are sufficiently catered to.

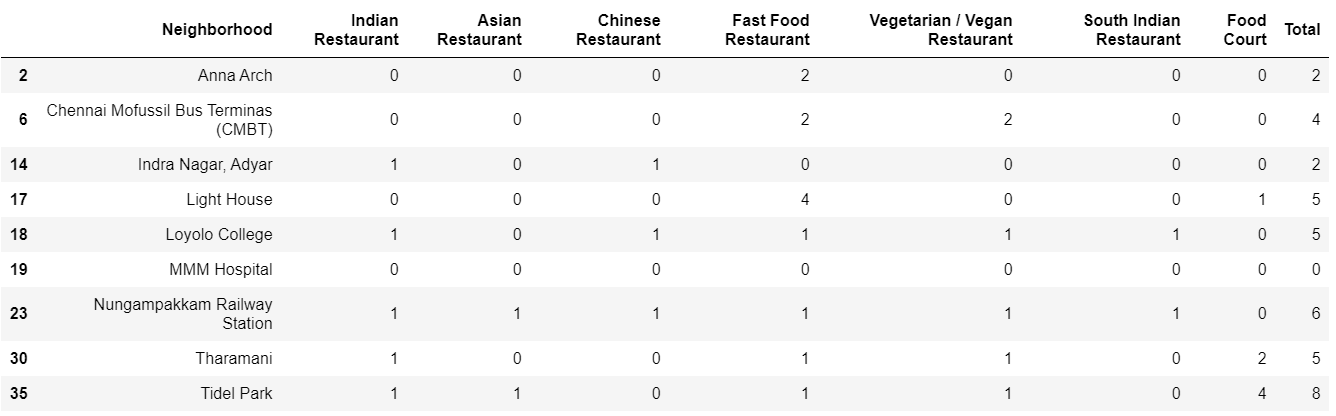


Table 12: Cluster 3 breakdown

Now, it looks like the neighbourhoods in cluster 1 would be a good choice of location to start a new restaurant, keeping in mind the following points:

1. Based on the common venues in cluster 1, these neighbourhoods seem to be the shopping zone of the city which means that the footfall at these neighbourhood is expected to be high.
2. Looking at a breakdown of the number of restaurants in each of these neighbourhoods, it is clear that “Mambalam” is the neighbourhood which is the most underserved.



Table 13: Cluster 1 Breakdown

Based on the above, one may logically conclude that the “Mambalam” is the right neighbourhood to start a restaurant in Chennai.