# COURSERA

# IBM Applied Data Science Capstone Project

Neighbourhood Analysis of Chennai

By Vaibhav Jain

# INTRODUCTION

Chennai is the capital of Tamil Nadu, an Indian state. It is one of the greatest cultural, commercial, and educational centres in South India. It is one of 4 metropolitan cities of India. Many people have relocated to Chennai in quest of better employment prospects and a better living over the years. Chennai is considered to be one of the most attractive tourist locations in India and due to such factors, the city is home to many youths from all across India and many foreigners. As a result, numerous local businesses, such as restaurants and cafés, are flourishing. The main goal of this project is to explore Chennai's neighbourhoods for potential locations for a new café. This project may be beneficial to entrepreneurs and business owners interested in opening a café. The goal is to look at the locations of existing cafés and determine which site would be best for a new café to open.

# DATA COLLECTION

The data required for this project is mentioned below and the data has been collected primarily from Wikipedia, Foursquare API and Python Libraries.

1) Different Areas of Chennai – Wikipedia

2) Coordinates of Chennai and all of its neighbourhoods – Python Geocoder

3) Venue data of all the neighbourhoods – Foursquare API

The data of the Areas of Chennai was collected from Wikipedia using the following link:

https://en.wikipedia.org/wiki/Areas_of_Chennai#:~:text=The%20city%20of%20Chennai%20is,zones%2C%20consisting%20of%20200%20wards.

Web Scrapping using beautiful soup library in python was used to collect the data. We will then use Python Geocoder library to gather the coordinates of all the areas extracted and use those coordinates to fetch all the Venue information using Foursquare API.

# TARGET AUDIENCE

- This Project targets Local Investors, entrepreneurs looking to start or invest in a new café in a Chennai Locality

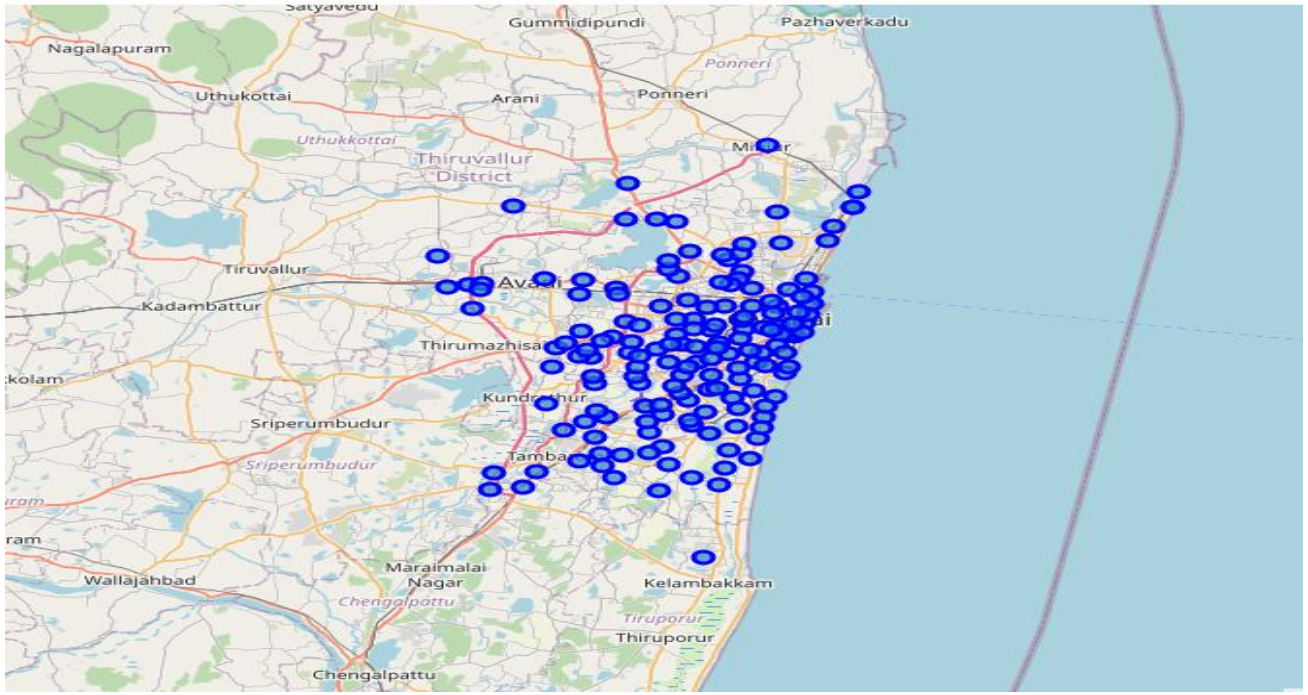


**Fig 1 MAP OF CHENNAI**

# METHODOLOGY

- The neighbourhood data is extracted from Wikipedia, using web scrapping through beautiful soup. The extracted data is then cleaned and saved into the pandas dataframe.
- Then the Geographical Coordinates in the form of longitude and latitude are extracted for the neighbourhoods using the Python Geocoder Library. We then extract the Coordinates of Chennai, India using the same library. The Map of Chennai is visually represented using Folium library.
- Then we use the foursquare API to extract top 100 venue information within a 1000 meter radius. The API calls are made using the Longitudinal and Latitudinal coordinates. The data returned is in the JSON Format which contains the Venue Name, Venue category, Latitude and Longitude. We then check how many venues were returned from each neighbourhood and how many unique categories are there. We then check if Café is present in the categories that are returned. One hot encoding is used on venue categories as they are categorical in nature and then they grouped based on their neighbourhood they belong too.
- We then use Kmeans which is a form of unsupervised learning to create 4 clusters and analyse them. This will allow us to analyse and find out how many café’s are available in a neighbouhood and based on our findings we will be able to determine which neighbourhood could be more suitable for opening a new cafe

# RESULTS

**The results are as follows :**

The neighbourhoods are divided into 4 clusters based on the frequency of cafes:

Cluster 0 (Red) : Areas with less or no café’s

Cluster 1 (Purple) : Areas with Decent Amount of Cafes

Cluster 2 (Light Blue) : Areas with high number of cafes

Cluster 3 (Yellow) : Areas with very limited number of Café’s

The results are visualised in Fig 2

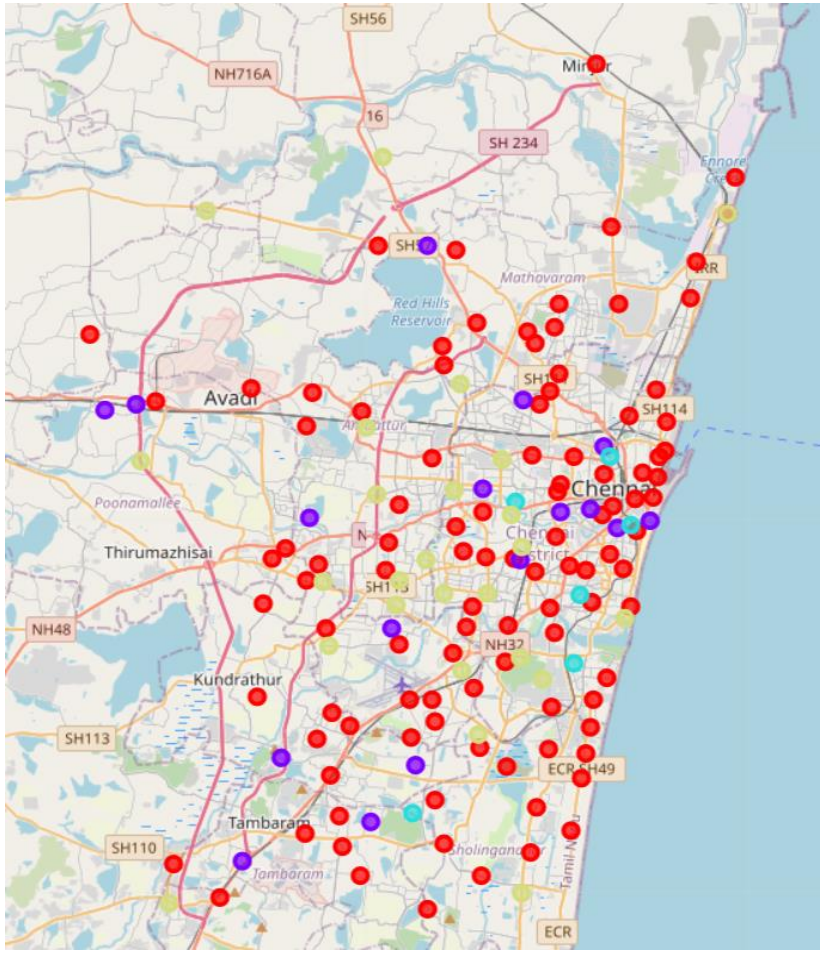


Fig 2

# DISCUSSION

According to the results , a large number of cafes are concentrated in Cluster 2, which has 4 to 5 cafes per Area, while a reasonable number of cafes are located in Cluster 1, which has 2 to 3 cafes each Area. Cluster 3 has only one café each Area, which may be a great opportunity for new investors to open cafes with less competition. Cafes in Clusters 1 and 2 face intense competition from one another, however there are no cafes in Cluster 0 and opening a new cafe could be risky due to unknown circumstances that may have contributed to the resultant 0 cafes. Personally, I would urge investors to avoid areas in clusters 1 and 2, which have a high concentration of cafes and are subject to fierce rivalry, and instead start a cafe in cluster 3.

# CONCLUSION

In this project, We have identified a business problem, specified the data required. We have collected that data using various platforms and techniques and used machine learning to create 4 different clusters of the neighbourhood based on the frequency of the Cafe’s present in an area. We have thus discovered that areas in cluster 1 and 2 will suffer from high competition , cluster 0 has no café’s and cluster 3 which would be an ideal location to open a café given that there are only very limited café’s available which will provide very less competition. This project can further be extended into multiple fields such as discovering and exploring venues such as schools , restaurants, theatres and hospitals. Through this project we hope to provide enough insight to the stakeholders, that they can make smart decision as to which neighbourhood can be the most suitable place for them to invest there money.