Analyzing Neighborhoods of Chennai for starting a New Restaurant.

NIVEDHA SUDHAKAR

09/09/20202

Contents

Introduction

CHENNAI is the capital of Tamil Nadu. It is one of the most important metros in India. The population of Chennai is around 71 lakhs! The region around Chennai has served as an important administrative, military, and economic center for many centuries. Chennai is located on the south–eastern coast of India in the north–eastern part of Tamil Nadu on a flat coastal plain known as the Eastern Coastal Plains attracts many visitors either as tourists or as part of its large workforce. The vast majority claim Chennai is one of the best cities in India. The city is full of restaurants serving thousands of hungry customers, every day. The diversity in the population of the city has brought in a vast diversity in food habits of people. In this project we will study the neighborhoods and make recommendations accordingly.

Business Problem

Our client is an investor who is interested in investing in a restaurant in Chennai. They have approached us to study the market and suggest them a location in one of the neighborhoods which would be in best interest of the business. Our main objectives of this project would be to extract and analyze right data about various neighborhoods of Chennai using various data science techniques and suggest our client a fitting location for their restaurant.

Data

In order to achieve our final goal, we will need the following data:

- Neighbourhoods of Chennai.
- Geographical coordinates of the neighbourhoods.
- Venue data from FourSquare.

Neighbourhoods Data

This data was extracted from Areas of Chennai Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai) using web scraping with BeautifulSoup library in Python. This will give us a detailed list of neighbourhoods present in Chennai.

Geographical Coordinates

Later, the geographical coordinates of various naighbourhoods were extracted using GeoPy library in Python. Geographical coordinates are necessary for plotting maps during the project for visualizing our data. After using GeoPy we added two columns to our dataframe with latitude and longitude information of each neighbourhood as shown below:

	Neighbourhood	Latitude	Longitude
0	Chitlapakkam	12.93277	80.14387
1	Chromepet	12.95234	80.14411
2	Cowl Bazaar	12.98861	80.15100
3	Egattur (Kanchipuram District)	12.82725	80.22866
4	Guduvancheri	12.83790	80.05327

Venue Data from FourSquare

Later we extracted venue data using FourSquare API. This venue data was used to study the venues in various neighbourhoods in Chennai. This data provided important details of various restaurants in the area and helped us understand the competition. This data was very important because it helped us draw the main conclusion of the project.

Methodology

Feature Extraction

Feature extraction was carried out through One Hot Encoding. In this method, each feature is a category that belongs to a venue which is then converted into binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighbourhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

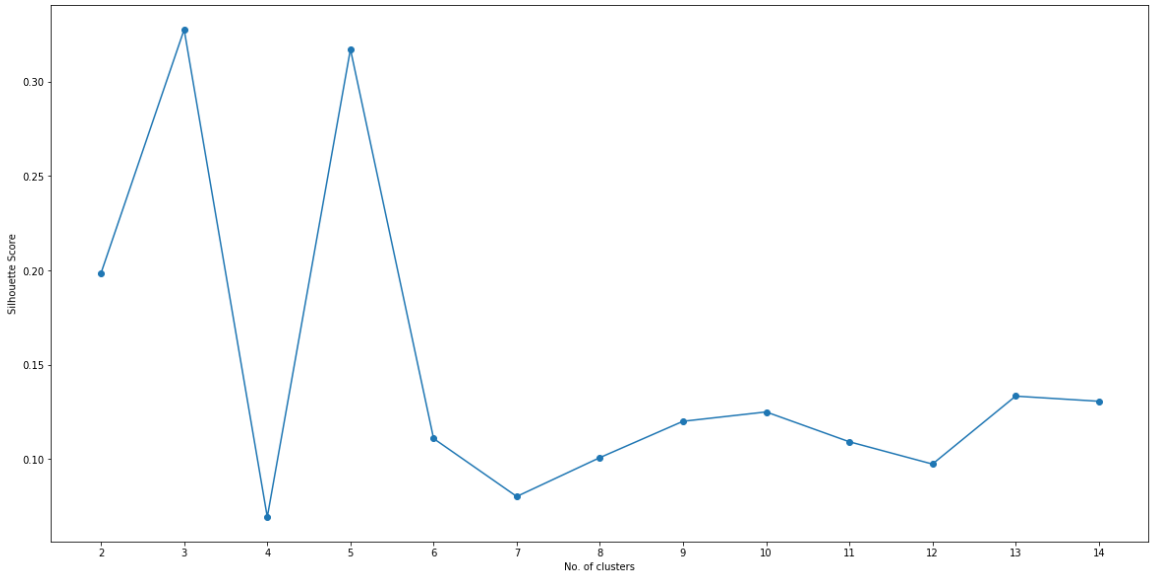
```
man_1hot = pd.get_dummies(explore_man[['Venue Category']], prefix="", prefix_sep="")


# Add neighbourhood column back to dataframe
man_1hot['Neighbourhood'] = explore_man['Neighbourhood']


# Move neighbourhood column to the first column
fixed_columns = [man_1hot.columns[-1]] + man_1hot.columns[:-1].values.tolist()
man_1hot = man_1hot[fixed_columns]

man_1hot.head()
```

Unsupervised Learning

Unsupervised learning was carried out in order to find out the similarities between found similarities between neighbourhoods. K-Means, a clustering algorithm, was implemented. In this case K-Means is used due to its simplicity and its similarity approach to find patterns.

- K-Means: K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features. It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of cluster is done and the elbow is selected. Then, further analysis of each cluster is done.

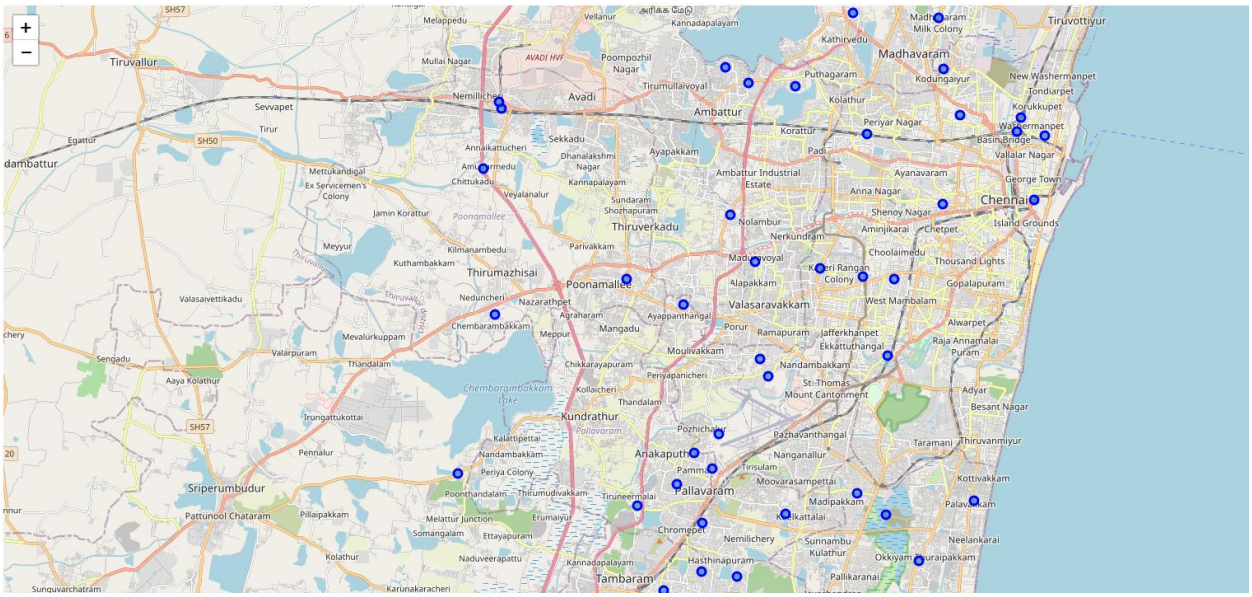
```
max_range = 15 #Max range 15 (number of clusters)

from sklearn.metrics import silhouette_samples, silhouette_score


indices = []
scores = []

for man_clusters in range(2, max_range) :

    # Run k-means clustering
    man_gc = man_grouped_clustering
    kmeans = KMeans(n_clusters = man_clusters, init = 'k-means++', random_state = 0).fit_predict(man_gc)

    # Gets the score for the clustering operation performed
    score = silhouette_score(man_gc, kmeans)

    # Appending the index and score to the respective lists
    indices.append(man_clusters)
    scores.append(score)
```



Plotting

Various plotting techniques we used as well in order to visualize the data. Visualizing data often gives a clear understanding of the data as it is easier to spot patterns in a visualized data as compares to quantitative data.

- Folium: Folium library was used to plot maps of Chennai city as well as neighbourhoods. Folium was also used to visualize the cluster data.



Results

The above mentioned, K-Means clustering method was applied to the dataframe of neighbourhoods of Chennai city. As mentioned earlier the number of clusters that was derived from elbow method was 4. The code as well as plotting of clusters can be seen below:

After visualizing the clusters, the individual clusters were studied, and some important conclusions were derived. The neighbourhood that had the more number of restaurants was cluster number 3.

Discussion

As mentioned earlier the most suitable neighbourhoods for starting the restaurant business are present in the cluster number 3. Our K-Means model worked perfectly and successfully clustered similar neighbourhoods together. After studying all four clusters, it is recommended to the client that neighbourhoods such as Chromepet, Kilpauk and Cowl Bazaar that fall in cluster 4 look like good locations for starting their restaurant business. The client can go ahead and make a decision depending on other factors like availability and legal requirements that are out of scope of this project.

Conclusion

Data analysis and machine learning techniques used in this project can be very helpful in determining solutions of certain business problems. Python's inbuilt libraries such as GeoPy, Folium and BeautifulSoup make it very easy and effective for a data scientist to analyze a geographical location because these libraries make it very easy to extract data that is easily available online. In this project we studied the neighbourhoods of Chennai city and came up with a recommendation of neighbourhoods where our client can start their restaurant business.