Coursera - IBM Data Science Capstone Project

The Battle of Neighborhoods

Karthik Srinivasan

The Battle of Neighborhoods

Introduction

Chennai, the capital city of the state of Tamil Nadu, India, is the largest industrial and commercial center of South India. Recent estimates of the economy of the Chennai Metropolitan Area have ranged from US$79 to US$86 billion, ranking it from fourth to sixth most productive metro area of India and the third highest by GDP per capita.

Chennai remains the Chief Retail Industry and Shopping Centre in South India, with some of its suburbs serving as exclusive shopping districts. Since the formation of the city in the seventeenth century, George Town remains one of the chief commercial neighborhoods of the city. However, with the centuries passing, the central business district of the city started shifting towards the south of Fort St. George and moving to its present location at Gemini Circle. The city's retail industry is concentrated chiefly in T. Nagar, which is by far the largest shopping district of India, generating more than twice the revenue of Connaught Place in New Delhi or Linking Road in Mumbai, even by conservative estimates.

In Terms, One-stop destination for all shop solution, people move on to shopping malls in order to cover wide range of products. Retailers and Property Developers eyes for the central location and large crowd at the shopping mall to provide the retail channel to market their products and services. Opening shopping malls is herculean task which requires serious consideration and need an effective modeling and planning to come out of effective solution. Core point of the solution, the location of the shopping mall.

Business Problem

Key objective of the project is to focus on the target audience and help them to bring out better solution for their problems. This project is particularly helpful for property developers, retailers and investors to open a new shopping mall in the neighborhoods of Chennai City. This project is a timely solution for future developers, according to the 2019 Cushman & Wakefield report *Main Streets Across the World*, Khader Nawaz Khan Road at Nungambakkam ranked 10th position in the list of 'Top 10 Global Highest Retail Rental Growth Markets 2019', with 36.7 percent increase in rents.

Using Data Science Methodology and Machine Learning Techniques, it provides business solutions by analyze and cluster the neighborhoods and bring out the central location to build a shopping mall.

Data

The data for this project has been retrieved and processed through multiple sources by carving out exact considerations to the accuracy of the methods used.

Data Acquisition and Cleaning

1. Data Acquisition

The data acquired for this project is a combination of two major sources
First source of the project uses a **Wikipedia** Page
(https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Chennai)
Use of Web scraping techniques to extract data from Wikipedia page along with help of Python requests and package. With help of Python Geocoder packages, the geographical coordinates of Chennai neighborhoods along with latitudes and longitudes.

Second Source, **FOURSQUARE** API to get the venue data of the Chennai City by mainly focusing on the Shopping mall category by data cleaning, wrangling and Machine Learning techniques (K- Means Clustering) and Map Visualization (Folium).

2. Data Cleaning

The first source of data is scraped from **Wikipedia** Page using the **Beautiful Soup Library** in Python. Using this library, the list of neighborhoods data is extracted.

	Neighbourhood
0	Adambakkam
1	Adyar, Chennai
2	Agaram
3	Alandur
4	Alapakkam
5	Alwarpet
6	Alwarthirunagar
7	Aminjikarai
8	Amullaivoyal
9	Andarkuppam
10	Anna Nagar
11	Anna Nagar West
12	Arani, Chennai
13	Ariyalur, Chennai
14	Arumbakkam
15	Ashok Nagar, Chennai
16	Assisi Nagar
17	Athipattu
18	Athipattu Pudunagar
19	Ayanavaram

METHODOLOGY

Data Analyzing and Wrangling

The key point in web scraping the data from Wikipedia Page is get the name of the neighborhoods. But to get the geographical coordinates in the form of latitude and longitude, Foursquare API is in use. To do so, use of **Geocoder Package** that will allow us to convert address into geographical coordinates in the form of Lat and Long. After gathering the data, populate the data into Pandas Data frame.

	Neighbourhood	Latitude	Longitude
0	Adambakkam	12.99192	80.20603
1	Adyar, Chennai	13.00305	80.25193
2	Agaram	13.10953	80.23236
3	Alandur	13.00013	80.20060
4	Alapakkam	13.04610	80.16499

Data Visualization

There are 199 neighborhoods in the city of Chennai region, in broad perspective, it can be visualized in a map using **Folium Package Library**. This allows to perform a sanity check to make sure that the geographical coordinates data returned by geocoder are correctly plotted in the city of Chennai.

MODELING

Using the final dataset containing the neighborhoods in Chennai along with latitude and longitude, we can find all the venues within a 2000-meter radius of each neighborhood by connecting the Foursquare API. This returns a JSON file containing the venues in each neighborhood which is converted to a Pandas dataframe. This dataframe contains all the venues along with their coordinates and category.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Adambakkam	12.99192	80.20603	more for you	12.996285	80.207074	Department Store
1	Adyar, Chennai	13.00305	80.25193	Nalli, Adayar	13.002910	80.251950	Women's Store
2	Adyar, Chennai	13.00305	80.25193	That Madras Place	13.005848	80.250726	Café
3	Adyar, Chennai	13.00305	80.25193	ibaco	13.005864	80.251764	Ice Cream Shop
4	Adyar, Chennai	13.00305	80.25193	Wonton	13.005047	80.251690	Chinese Restaurant

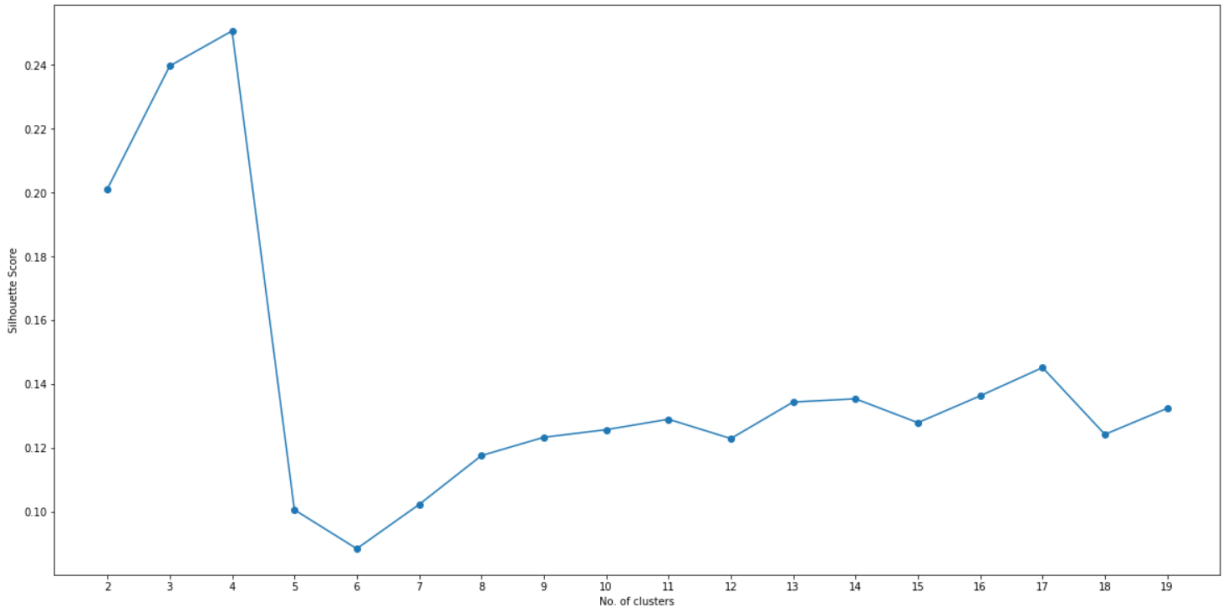
Next process begins with analyzing each neighborhood by grouping the rows and taking the mean frequency occurrence of each venue category. By doing so, the data can be used for clustering. Since we are analyzing the "SHOPPING MALL" data, filter out that as venue category for the neighborhoods.

	Neighbourhood	Shopping Mall
0	Adyar, Chennai	0.000000
1	Agaram	0.000000
2	Alandur	0.000000
3	Alapakkam	0.000000
4	Alwarpet	0.000000
5	Alwarthirunagar	0.000000
6	Aminjikarai	0.071429
7	Amullaivoyal	0.000000
8	Anna Nagar	0.000000
9	Anna Nagar West	0.000000
10	Arani, Chennai	0.000000
11	Ariyalur, Chennai	0.000000
12	Arumbakkam	0.000000
13	Ashok Nagar, Chennai	0.000000

Optimal Number of Clusters

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.

The silhouette ranges from -1 to +1 where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Based on the Silhouette Score of various clusters, the optimal cluster size is determined.



K – Means Clustering

k-means clustering is a method of vector quantization, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

It is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will cluster the neighborhoods based on the optimal score of clusters and their frequency of occurrence for shopping mall.

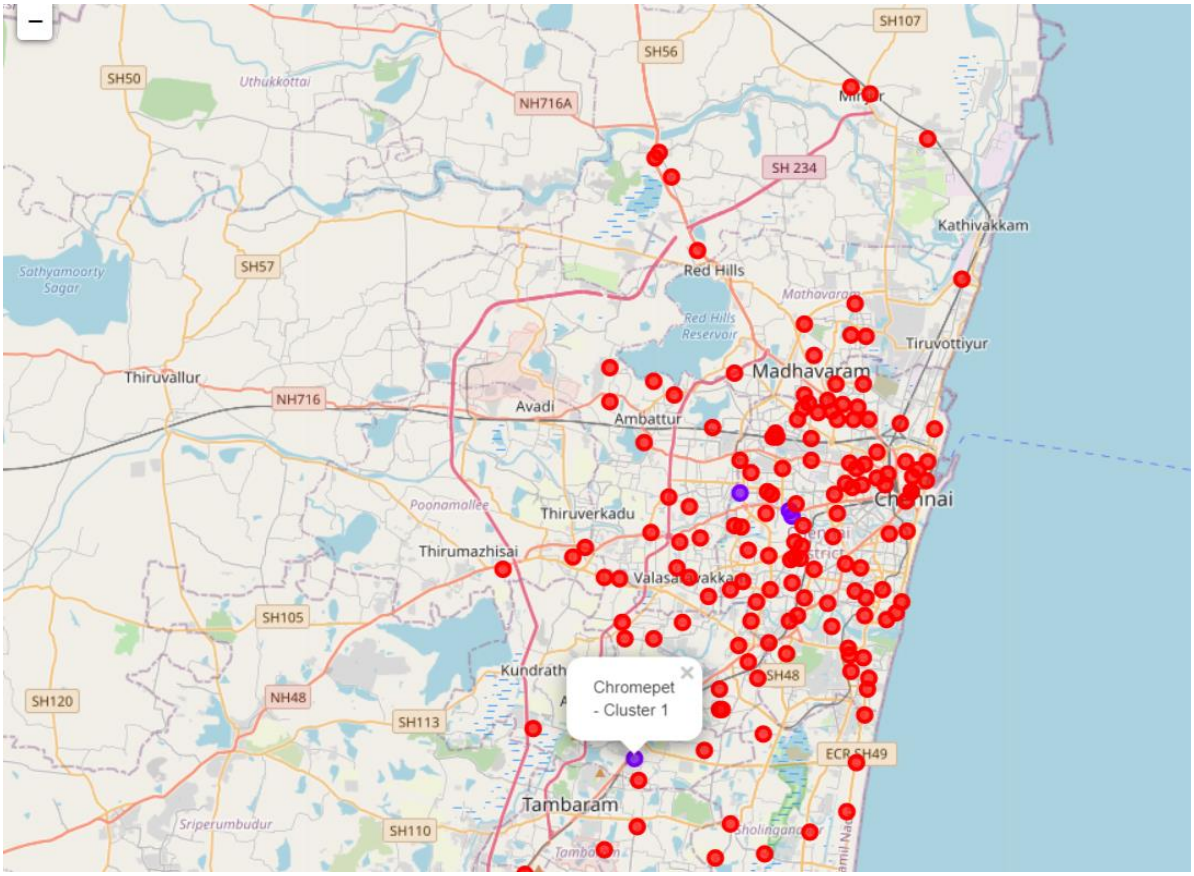
The Clustered neighborhoods are visualized using different colors so as to make them distinguish each other. The results will allow us to identify the neighborhoods have higher concentration of shopping malls to fewer ones. Based on the occurrence of shopping malls in different

neighborhoods, it helps us to answer for the question for which neighborhood are most suitable to open a new shopping mall.

RESULTS

The results from the K means clustering show that we can distinguish the neighborhoods into 3 clusters based on the frequency of occurrence for venue category “Shopping Mall”

- Cluster 0: Neighborhoods with no existence of Shopping Malls.
- Cluster 1: Neighborhoods with low number of Shopping Malls.
- Cluster 2: Neighborhoods with 2 or more number of Shopping Malls.



Each Cluster is color coded for the ease of presentation; we can see that majority of the neighborhood falls in the red cluster with is Cluster 0. Further by, blue Cluster is Cluster 1 and Green Cluster is Cluster 2.

DISCUSSION

The aim of the project is to help the proprietor to find a location to set up a Shopping mall in the neighborhoods of Chennai. From results, we can carve out that in Cluster 0 represents no existence of Shopping mall but as for my suggestion Cluster 1 which represents a great opportunity and high potential areas to open a shopping mall because it falls under the central location of Chennai city and property developers or Rental operators will find a unique selling propositions. From the findings, Cluster 2 is not advisable for the proprietor to open a new Shopping Mall.

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
138	Thirumangalam, Chennai	0.111111	1	13.08285	80.19699
25	Chromepet	0.071429	1	12.95234	80.14411
6	Aminjikarai	0.083333	1	13.07139	80.22256
20	Chinnakudal	0.076923	1	13.07444	80.22167

CONCLUSION

This project helps a proprietor to get a better understanding of the neighborhoods with respect to frequency of occurrence of shopping malls. Future Research of this project will be focusing on the most common venues in each neighborhood of Chennai City and filtering out many factors for a perfect living.