Movie Theater Analysis

IBM Coursera Data Science- Capstone Project

1.Introduction

1.1 Background

Chennai is one of the major cities in the Southern part of India. It is also called the ‘Cultural Capital of India’. Thus it is one of the most important cities in India which preserves the rich cultural diversity and tradition. Rightly so it is one of the places where movies are considered to be a part of life. So that’s why me being a Chennaite wanted to analyse on Movie theatres.

1.2 Problem

The main purpose of this project is to analyze the existing theatres in and around the neighborhoods of Chennai and to find a scope for a new theater and to find the places where these new theaters can be constructed.

1.3 Interest

Obviously anyone who is a fan of Cinema, or anyone who wants to enter the theatre business or anyone who wants to extend their number of theatres would be interested in the analysis.

2. Data Acquisition and Cleaning

2.1 Data sources

There are 3 main sources of data for this project.

a. The List of Neighborhood of Chennai data
b. The latitude and longitudes of the neighborhoods
c. The venue data from FourSquare API

The data acquisition and processing stage is as follows:

The data is first scraped from the web using the BeautySoap Parser to parse the neighborhood information from the wikipedia page 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai'.

For the list of neighborhoods, we use the geocoder API to get the latitudes and longitudes. The data is preprocessed, to eliminate some mismatches on the latitude and longitude as there was no correct links/datasets to obtain the same.

The dataset would look as below.

Out[5]:

	Borough	Neighborhood	Latitude	Longitude
0	North	Red Hills	13.10862	80.20615
1	North	Royapuram	13.10728	80.29295
2	North	Korukkupet	13.11805	80.27774
3	North	Vyasarpadi	13.11911	80.25750
4	North	Tondiarpet	13.13145	80.27997
...	...	...	...	...
95	South & East	Kolappakkam	13.07209	80.20186
96	South & East	Mambakkam	12.90110	80.18384
97	South & East	Palavakkam	12.95832	80.25608
98	South & East	Varadharajapuram	13.04658	80.07157
99	South & East	Medavakkam	12.92084	80.18565

Using the latitude and longitudes obtained we then use the Fousquare API to get the list of venues in and around the latitude and then summarize them as below.

Out[21]:

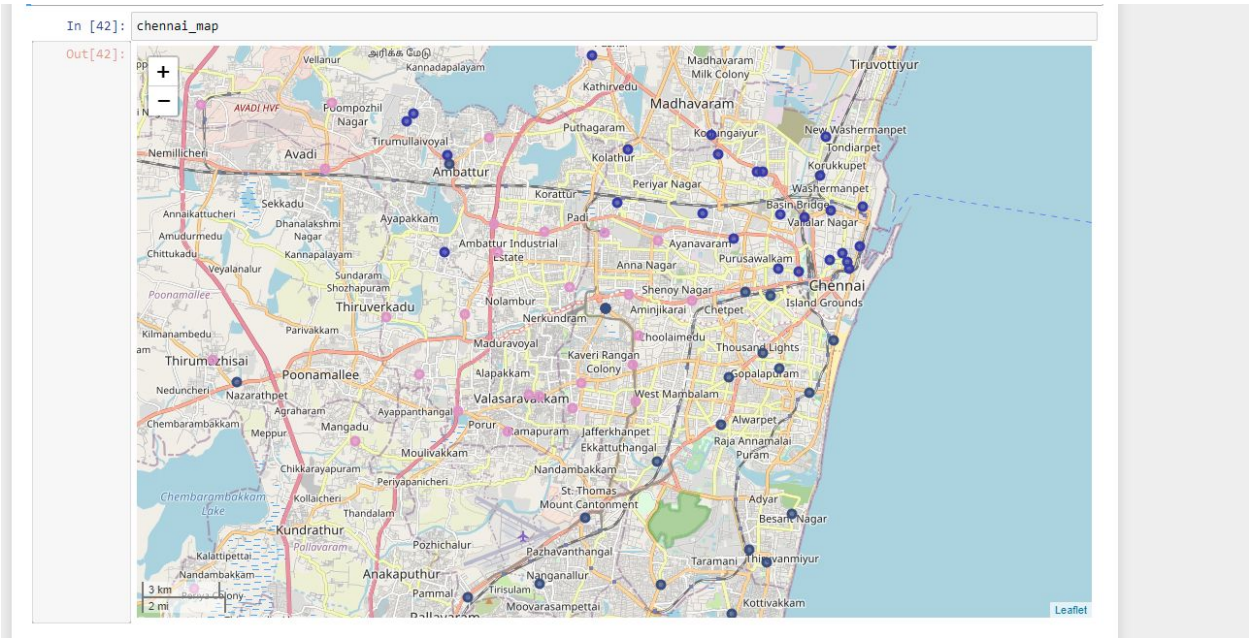
	Borough	Neighborhood	Latitude	Longitude	Total Venue	Total Movie Venue
0	North	Red Hills	13.10862	80.20615	8.0	1.0
1	North	Royapuram	13.10728	80.29295	4.0	0.0
2	North	Korukkupet	13.11805	80.27774	4.0	0.0
3	North	Vyasarpadi	13.11911	80.25750	4.0	0.0
4	North	Tondiarpet	13.13145	80.27997	3.0	0.0
...	...	...	...	...	...	...
95	South & East	Kolappakkam	13.07209	80.20186	15.0	2.0
96	South & East	Mambakkam	12.90110	80.18384	3.0	0.0
97	South & East	Palavakkam	12.95832	80.25608	15.0	1.0
98	South & East	Varadharajapuram	13.04658	80.07157	7.0	0.0
99	South & East	Medavakkam	12.92084	80.18565	11.0	0.0

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Plotting of the neighbourhoods

We first plot the neighborhoods of Chennai in the map, to check if the plots are spread around the map are if the neighborhoods are confined to a single location. This is to check if the data scarped from web, their latitudes and longitudes are correct.



The above plot shows that the data is widespread. So the data obtained from the Wikipedia page and the corresponding latitude and longitudes are correct.

3.1.2 The spread of theatres over the Borough

We have three boroughs, North, West and ‘South & East’. Now we look at the distribution of theatres among the three regions. This is to check if theaters are confined to a particular region or it is spread across various regions.

```
In [18]: chennai_movie_venues['Borough'].value_counts()
Out[18]: South & East    39
         West            20
         North           15
         Name: Borough, dtype: int64
```

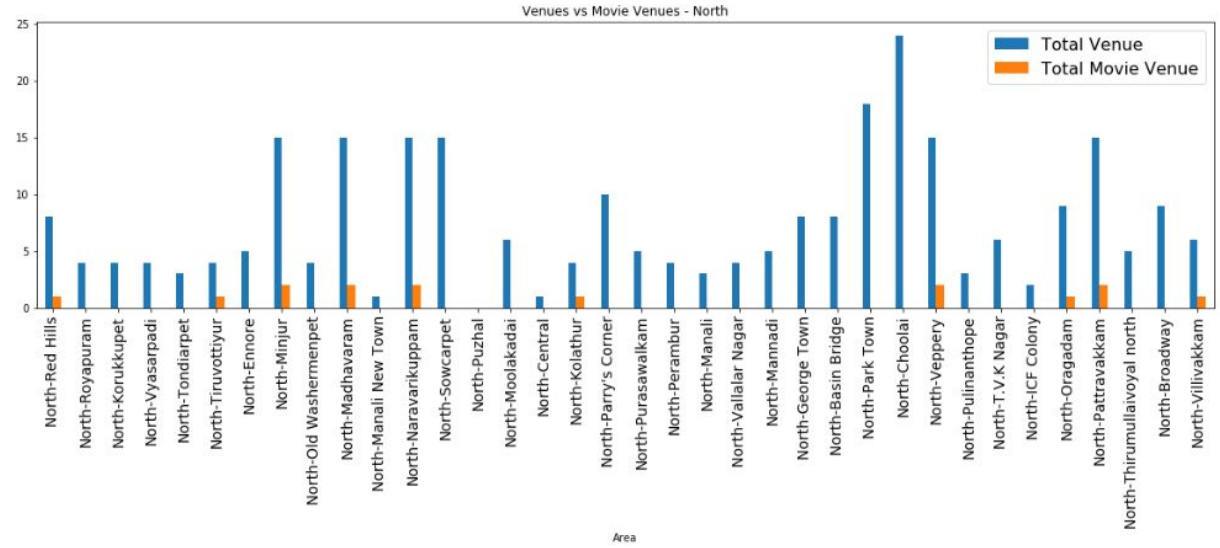
From the above data we can observe that the North neighbourhood of Chennai has a fairly low share of theatres compared to the West and South & east zone.
So we may keep it as a potential candidate for our theater construction.

3.1.3 Comparison between total number of venues vs theaters.

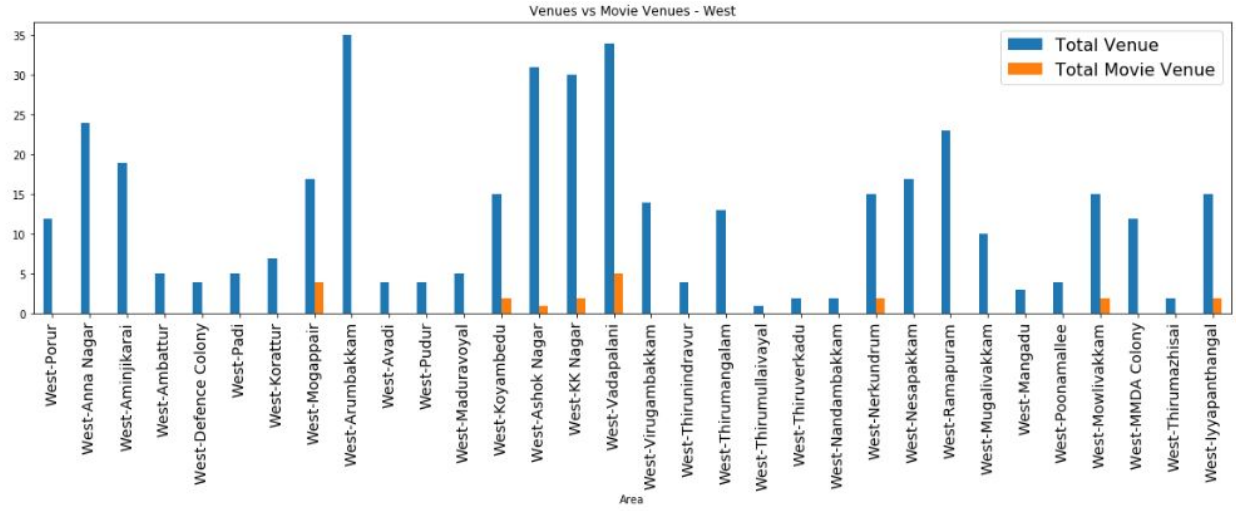
In the previous step we observed that the North neighborhood had low share of theaters. We further analyse to find out if it was due to the availability of venue details for the region or is it actually because of the unavailability of theaters.

The below plots are plotted using the data.

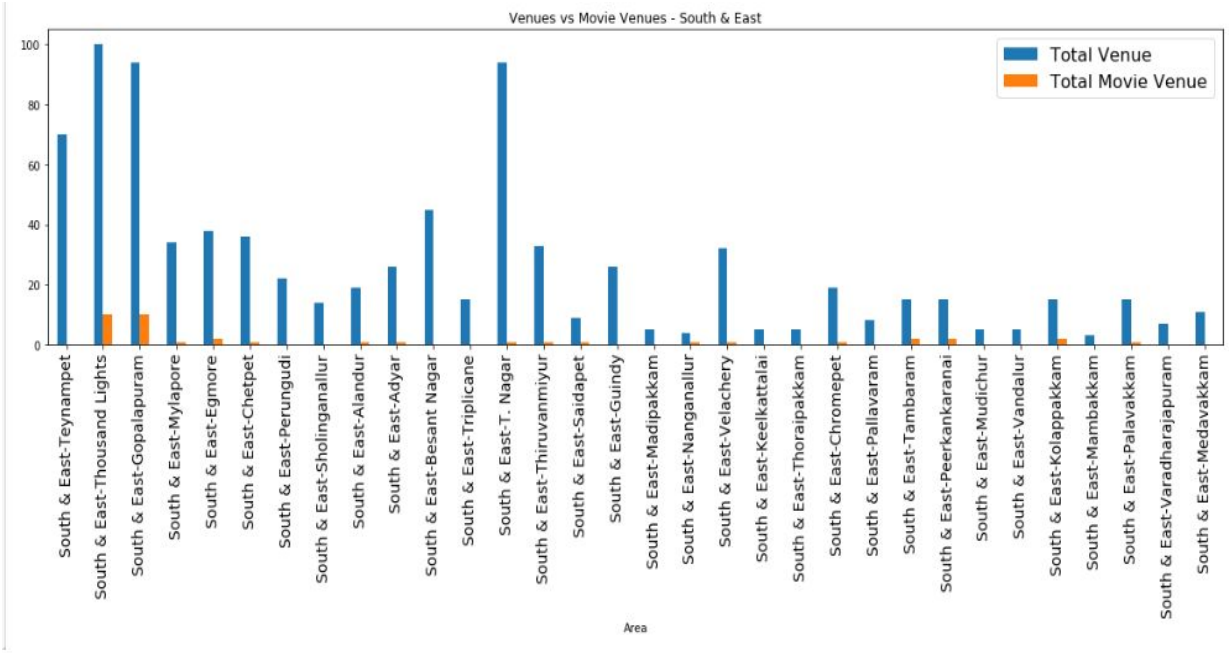
North Region:



West Region:



South and East:



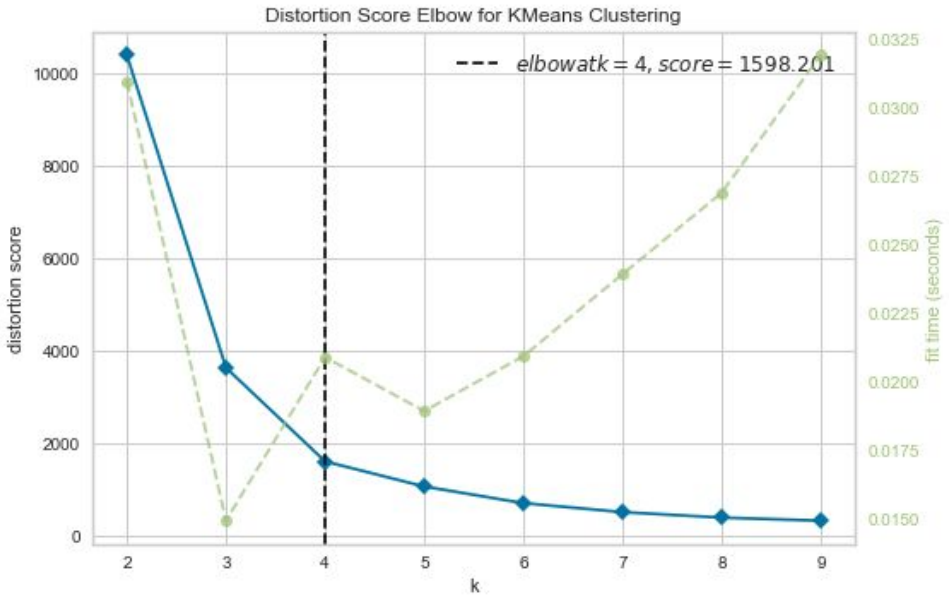
On looking at the data we can see that we have venue details for all of the regions, however in the North data we could see the less number of theaters. So it further strengthens our assumption that the north region has low share of theaters. Even Though West region is having medium number of theaters they are more common in certain places like Vadapalani,

Mogappair etc and the rest of the regions do not have any/ very little theaters. In South & East region we could see that there are atleast 1 thetre in every region, thus the graph shows very little height bars in yellow, but they are spread over the regions. So this makes the West our second candidate for the theater launch, but it must be in area where there are no theaters.

3.2 k-means Clustering:

To identify the clusters having no/very few theaters we use the k-means clustering algorithm to identify the clusters.

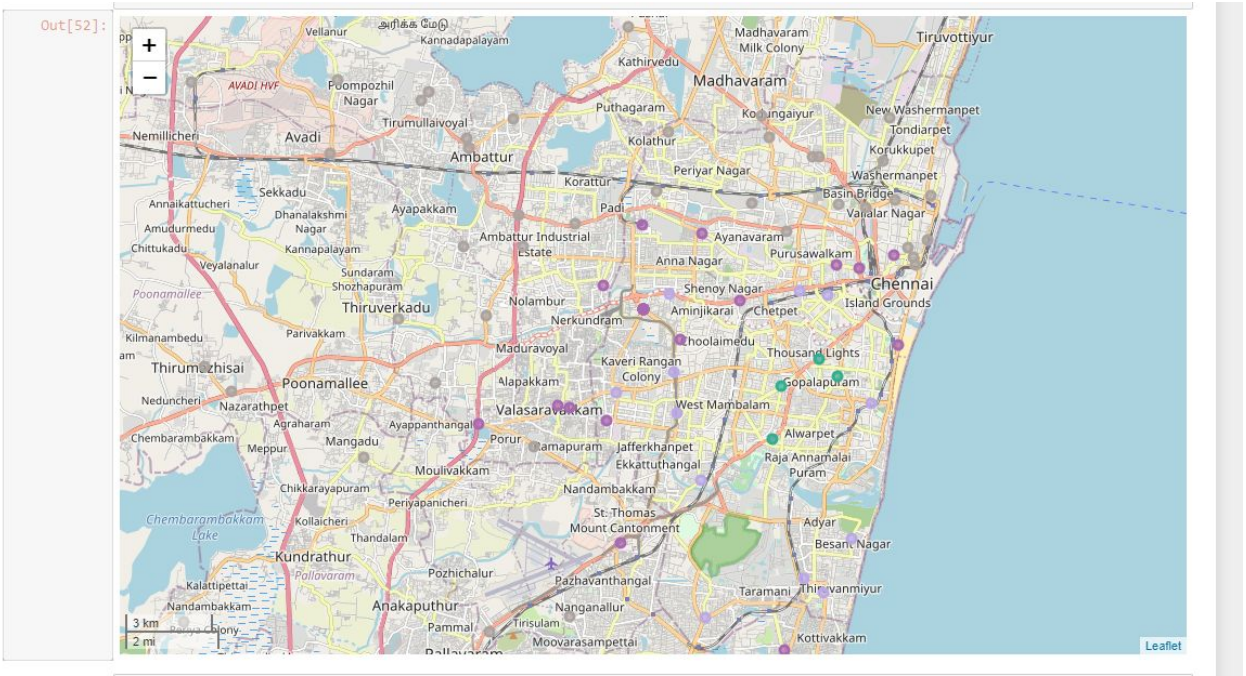
We use the elbow method to identify the optimum level of cluster. The below is the elbow graph obtained.



From the above screenshot we could see that the optimum value is obtained when k equals 4. So the model is fitted with the data to partition the data into 4 clusters.

Results:

As per the model above, the neighborhood was fit into the 4 clusters. Please find the clusters created by the k-means clustering algorithm below.



To identify what each cluster might represent, we calculated the number of theaters in each of the cluster.

```
Cluster Labels
0    31.0
1    21.0
2     7.0
3    15.0
Name: Total Movie Venue, dtype: float64
```

We could find that the cluster 2 (Gray shaded circles on the map) had the least number of theaters. If we locate the grey shaded points on the map we could find that they are predominantly on the northern and western side of the map , while the southern and the eastern side of the map had differently colored clusters.

Discussion:

Both from the EDA (Exploratory Data Analysis) and the K-Means Clustering of the data, we can see that the Northern and western parts of chennai have very low number if theaters and the southern and eastern side had a good number of theaters. This means that both the EDA and the Clustering Algorithm were consistent with each other. If there were any inconsistencies, we might have had to revisit the model.

Conclusion:

Thus based on the report we could come to the conclusion that the Northern and certain Western parts of Chennai can be considered as a correct candidate to launch a new theatre.

Appendix and Acknowledgement:

I hereby am thankful to the sources of data, without which this analysis would not have been performed.

I thank everyone who had reviewed/reviewing my work during the course of the project.

The report could be further strengthened by combining it with various other source of data.