Choosing the most appropriate district for starting a Multiplex in The Most Densely Populated District of Tamilnadu, India

Meenakshi S Arya

07-06-2020

1. Introduction

1.1Background

TamilNadu is one of the most popular states in the South India as it is the homeground of the Tamil movie industry, popularly known as "Tollywood". People in Tamilnadu are extremely fond of watching movies and dining out. Thus, it is an opportune candidate for setting up a Multiplex with movie theatres and eating joints. These multiplexes are great business ideas as they provide a one-stop solution to all the needs of customers, ranging from household groceries, to beauty products to dining options to options for entertainment.

1.2Problem Statement

Tamilnadu majorly comprises of 37 districts however, the population of a district will determine the feasibility of starting a new multiplex at a given location. The project aims to predict which district will be most suitable for the same considering the overall population and the number of eating joints and multiplexes in a given district.

The battle of these neighborhoods will be unrevealed in a detailed manner in the subsequent parts of the implementation.

Fig 1: Scenic Shore Temple at Mahabalipuram, Tamil Nadu

1.3Target Audience

The project will be of significant interest to the following people.

i. Chain of multiplex owners who are exploring the options of tapping a potential market where not many multiplexes are available and people are willing to spend money and utilize the services provided in these multiplexes.

ii. People interested in uncovering the interesting facts hidden in the data and utilize it to propose a new solution to an existing problem or suggest a solution to a new problem.

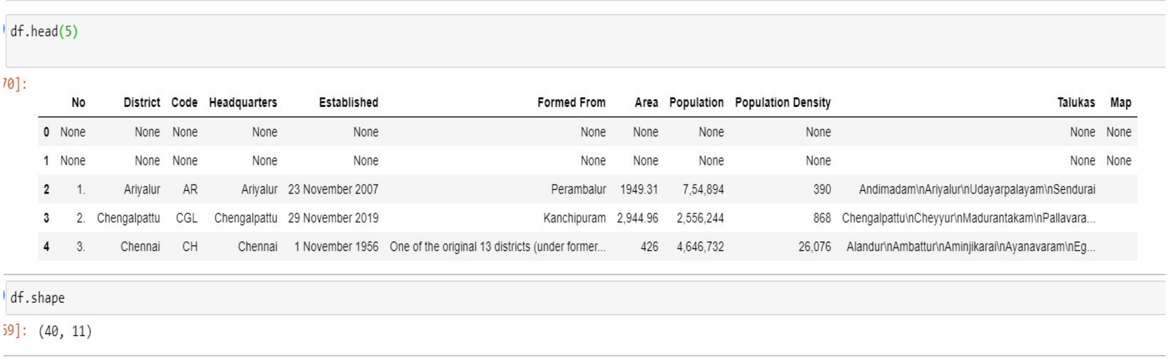
2 Data description & Preprocessing

2.1Dataset Description

The dataset available in wikipedia for districts of Tamilnadu will be used for obtaining the relevant information about the various districts in Tamilnadu, their area, population and talukas within each district. This data along with Foursquare location data will be used for obtaining the information about the number of multiplexes in the most populous districts of Tamilnadu.

2.2Scrapping the relevant data from wikipedia

The table was scrapped from Wikipedia using Beautifulsoup and pandas library to create the initial data-frame of 40 districts alongwith 11 attributes for each district.



```
df.head(5)
```

70]:

	No	District	Code	Headquarters	Established	Formed From	Area	Population	Population Density	Talukas	Map
0	None	None	None	None	None	None	None	None	None	None	None
1	None	None	None	None	None	None	None	None	None	None	None
2	1.	Ariyalur	AR	Ariyalur	23 November 2007	Perambalur	1949.31	7,54,894	390	Andimadam\nAriyalur\nUdayarpalayam\nSendurai	
3	2.	Chengalpattu	CGL	Chengalpattu	29 November 2019	Kanchipuram	2,944.96	2,556,244	868	Chengalpattu\nCheyyur\nMadurantakam\nPallavara...	
4	3.	Chennai	CH	Chennai	1 November 1956	One of the original 13 districts (under former...	426	4,646,732	26,076	Alandur\nAmbattur\nAminjikarai\nAyanavaram\nEg...	

```
df.shape
```

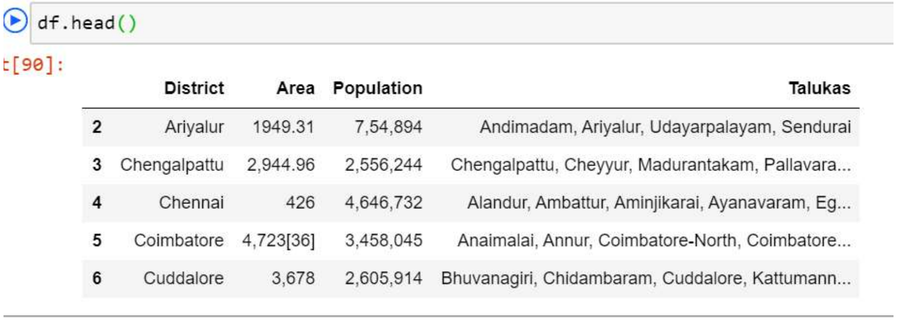
59]: (40, 11)

2.3Data Cleaning and Formatting

Data Cleaning process began with replacing all the missing values and deleting the rows where there were more than 3 missing values in a single row.

2.4Feature Selection

The table contained a lot of features such as 'No', 'Code', 'Headquarters', 'Established', 'Formed From', 'Population Density' and 'Map' which is not relevant to project. The features were selected on the basis of the requirement of the project and after dropping the non-relevant attributes, the table looked like this



```
df.head()
```

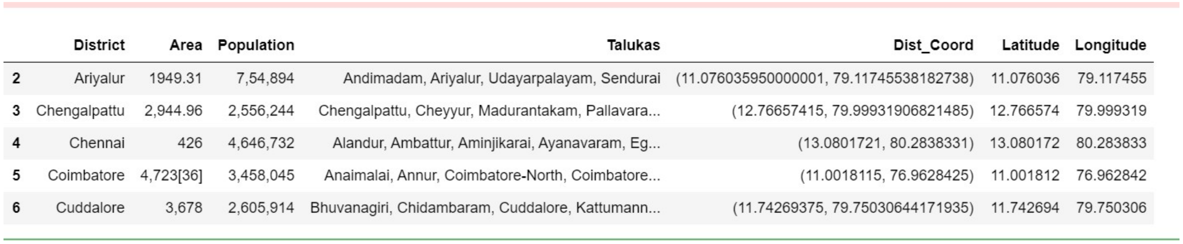
:[90]:

	District	Area	Population	Talukas
2	Ariyalur	1949.31	7,54,894	Andimadam, Ariyalur, Udayarpalayam, Sendurai
3	Chengalpattu	2,944.96	2,556,244	Chengalpattu, Cheyyur, Madurantakam, Pallavara...
4	Chennai	426	4,646,732	Alandur, Ambattur, Aminjikarai, Ayanavaram, Eg...
5	Coimbatore	4,723[36]	3,458,045	Anaimalai, Annur, Coimbatore-North, Coimbatore...
6	Cuddalore	3,678	2,605,914	Bhuvanagiri, Chidambaram, Cuddalore, Kattumann...

The above view of the dataset provides all the relevant information required for exploring the neighborhoods and provides a detailed picture of the corresponding district, as later on most venues within 1-kilometer radius of the these districts will be considered.

2.5Obtaining the coordinates of Tamil Nadu and subsequently that of its districts

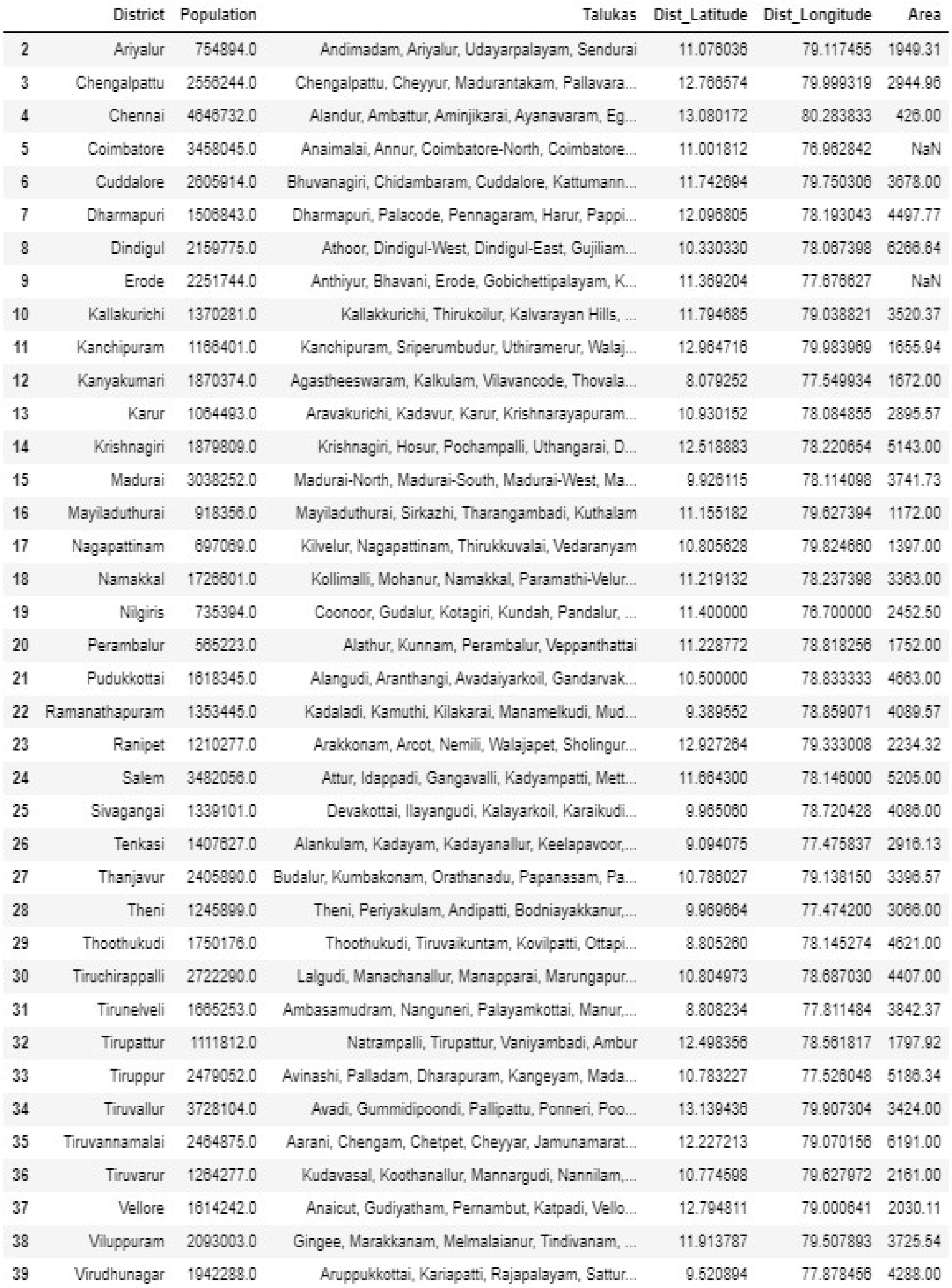
The coordinates of the state are obtained using Nominatim package from Geopy library and subsequently the coordinates of each district are obtained and used for construction of a new data frame with additional information about the latitude and longitude information about each district.



	District	Area	Population	Talukas	Dist_Coord	Latitude	Longitude
2	Ariyalur	1949.31	7,54,894	Andimadam, Ariyalur, Udayarpalayam, Sendurai	(11.076035950000001, 79.11745538182738)	11.076036	79.117455
3	Chengalpattu	2,944.96	2,556,244	Chengalpattu, Cheyyur, Madurantakam, Pallavara...	(12.76657415, 79.99931906821485)	12.766574	79.999319
4	Chennai	426	4,646,732	Alandur, Ambattur, Aminjikarai, Ayanavaram, Eg...	(13.0801721, 80.2838331)	13.080172	80.283833
5	Coimbatore	4,723[36]	3,458,045	Anaimalai, Annur, Coimbatore-North, Coimbatore...	(11.0018115, 76.9628425)	11.001812	76.962842
6	Cuddalore	3,678	2,605,914	Bhuvanagiri, Chidambaram, Cuddalore, Kattumann...	(11.74269375, 79.75030644171935)	11.742694	79.750306

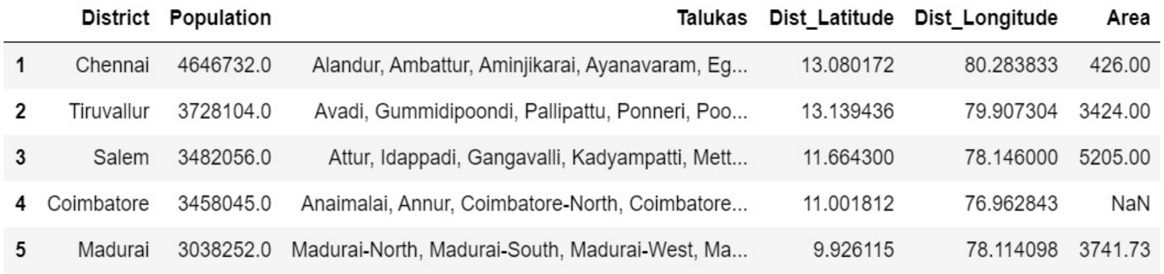
2.5.1 Updating incorrect Latitudes and Longitudes

Some of the coo-ordinates of the major districts returned by Geopy are incorrect as the name of the district probably matches the other more popular place with the same name. e.g. the coordinates of District “Salem” were provided as {44.9391565, -123.033121} whereas the actual coordinates are {11.6643, 78.1460}. The correct coordinates were looked up in google and the replacement was done. Similarly, several other columns were cosmetically modified so that the data looks in a proper and consistent format. Post processing, the data frame had the looked as follows:

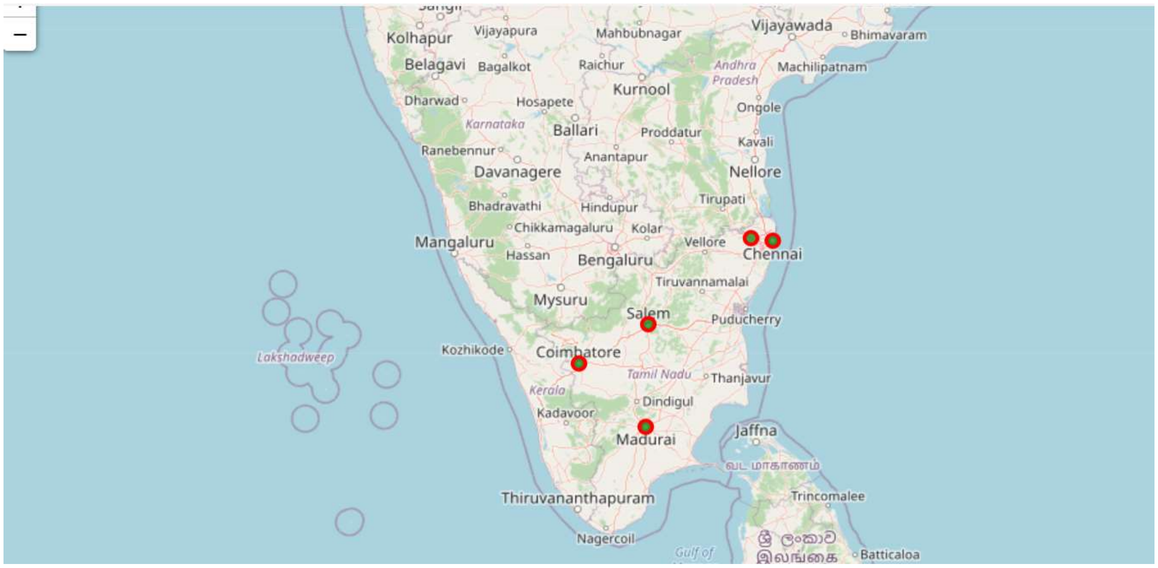


	District	Population	Talukas	Dist_Latitude	Dist_Longitude	Area
2	Ariyalur	754894.0	Andimadam, Ariyalur, Udayarpalayam, Sendurai	11.076036	79.117455	1949.31
3	Chengalpattu	2556244.0	Chengalpattu, Cheyyur, Madurantakam, Pallavara...	12.766574	79.999319	2944.96
4	Chennai	4646732.0	Alandur, Ambattur, Aminjikarai, Ayanavaram, Eg...	13.080172	80.283833	426.00
5	Coimbatore	3458045.0	Anaimalai, Annur, Coimbatore-North, Coimbatore...	11.001812	76.962842	NaN
6	Cuddalore	2605914.0	Bhuvanagiri, Chidambaram, Cuddalore, Kattumann...	11.742694	79.750306	3678.00
7	Dharmapuri	1506843.0	Dharmapuri, Palacode, Pennagaram, Harur, Pappi...	12.096805	78.193043	4497.77
8	Dindigul	2159775.0	Athoor, Dindigul-West, Dindigul-East, Gujiliam...	10.330330	78.067398	6266.64
9	Erode	2251744.0	Anthiyur, Bhavani, Erode, Gobichettipalayam, K...	11.369204	77.676627	NaN
10	Kallakurichi	1370281.0	Kallakkurichi, Thirukoilur, Kalvarayan Hills, ...	11.794685	79.038821	3520.37
11	Kanchipuram	1166401.0	Kanchipuram, Sriperumbudur, Uthiramerur, Walaj...	12.964716	79.983969	1655.94
12	Kanyakumari	1870374.0	Agastheeswaram, Kalkulam, Vilavancode, Thovala...	8.079252	77.549934	1672.00
13	Karur	1064493.0	Aravakurichi, Kadavur, Karur, Krishnarayapuram...	10.930152	78.084855	2895.57
14	Krishnagiri	1879809.0	Krishnagiri, Hosur, Pochampalli, Uthangarai, D...	12.518883	78.220654	5143.00
15	Madurai	3038252.0	Madurai-North, Madurai-South, Madurai-West, Ma...	9.926115	78.114098	3741.73
16	Mayiladuthurai	918356.0	Mayiladuthurai, Sirkazhi, Tharangambadi, Kuthalam	11.155182	79.627394	1172.00
17	Nagapattinam	697069.0	Kilvelur, Nagapattinam, Thirukkuvalai, Vedaranyam	10.805628	79.824660	1397.00
18	Namakkal	1726601.0	Kollimalli, Mohanur, Namakkal, Paramathi-Velur...	11.219132	78.237398	3363.00
19	Nilgiris	735394.0	Coonoor, Gudalur, Kotagiri, Kundah, Pandalur, ...	11.400000	76.700000	2452.50
20	Perambalur	565223.0	Alathur, Kunnam, Perambalur, Veppanthattai	11.228772	78.818256	1752.00
21	Pudukkottai	1618345.0	Alangudi, Aranthangi, Avadaiyarkoil, Gandarvak...	10.500000	78.833333	4663.00
22	Ramanathapuram	1353445.0	Kadaladi, Kamuthi, Kilakarai, Manamelkudi, Mud...	9.389552	78.859071	4089.57
23	Ranipet	1210277.0	Arakkonam, Arcot, Nemili, Walajapet, Sholingur...	12.927264	79.333008	2234.32
24	Salem	3482056.0	Attur, Idappadi, Gangavalli, Kadyampatti, Mett...	11.664300	78.146000	5205.00
25	Sivagangai	1339101.0	Devakottai, Ilayangudi, Kalayarkoil, Karaikudi...	9.965060	78.720428	4086.00
26	Tenkasi	1407627.0	Alankulam, Kadayam, Kadayanallur, Keelapavoor,...	9.094075	77.475837	2916.13
27	Thanjavur	2405890.0	Budalur, Kumbakonam, Orathanadu, Papanasam, Pa...	10.786027	79.138150	3396.57
28	Theni	1245899.0	Theni, Periyakulam, Andipatti, Bodniayakkanur,...	9.969664	77.474200	3066.00
29	Thoothukudi	1750176.0	Thoothukudi, Tiruvaikuntam, Kovilpatti, Ottapi...	8.805260	78.145274	4621.00
30	Tiruchirappalli	2722290.0	Lalgudi, Manachanallur, Manapparai, Marungapur...	10.804973	78.687030	4407.00
31	Tirunelveli	1665253.0	Ambasamudram, Nanguneri, Palayamkottai, Manur,...	8.808234	77.811484	3842.37
32	Tirupattur	1111812.0	Natrampalli, Tirupattur, Vaniyambadi, Ambur	12.498356	78.561817	1797.92
33	Tiruppur	2479052.0	Avinashi, Palladam, Dharapuram, Kangeyam, Mada...	10.783227	77.526048	5186.34
34	Tiruvallur	3728104.0	Avadi, Gummidipoondi, Pallipattu, Ponneri, Poo...	13.139436	79.907304	3424.00
35	Tiruvannamalai	2464875.0	Aarani, Chengam, Chetpet, Cheyyar, Jamunamarat...	12.227213	79.070156	6191.00
36	Tiruvarur	1264277.0	Kudavasal, Koothanallur, Mannargudi, Nannilam,...	10.774598	79.627972	2161.00
37	Vellore	1614242.0	Anaicut, Gudiyatham, Pernambut, Katpadi, Vello...	12.794811	79.000641	2030.11
38	Viluppuram	2093003.0	Gingee, Marakkanam, Melmalaianur, Tindivanam, ...	11.913787	79.507893	3725.54
39	Virudhunagar	1942288.0	Aruppukkottai, Kariapatti, Rajapalayam, Sattur...	9.520894	77.878456	4288.00

For further processing, the 5 most populous Districts of Tamil Nadu were extracted and a new dataframe consisting of “Chennai”, “Coimbatore”, “Tiruvallur”, “Salem” and “Madurai” was created. The same was plotted by importing Folium.



	District	Population	Talukas	Dist_Latitude	Dist_Longitude	Area
1	Chennai	4646732.0	Alandur, Ambattur, Aminjikarai, Ayanavaram, Eg...	13.080172	80.283833	426.00
2	Tiruvallur	3728104.0	Avadi, Gummidipoondi, Pallipattu, Ponneri, Poo...	13.139436	79.907304	3424.00
3	Salem	3482056.0	Attur, Idappadi, Gangavalli, Kadyampatti, Mett...	11.664300	78.146000	5205.00
4	Coimbatore	3458045.0	Anaimalai, Annur, Coimbatore-North, Coimbatore...	11.001812	76.962843	NaN
5	Madurai	3038252.0	Madurai-North, Madurai-South, Madurai-West, Ma...	9.926115	78.114098	3741.73

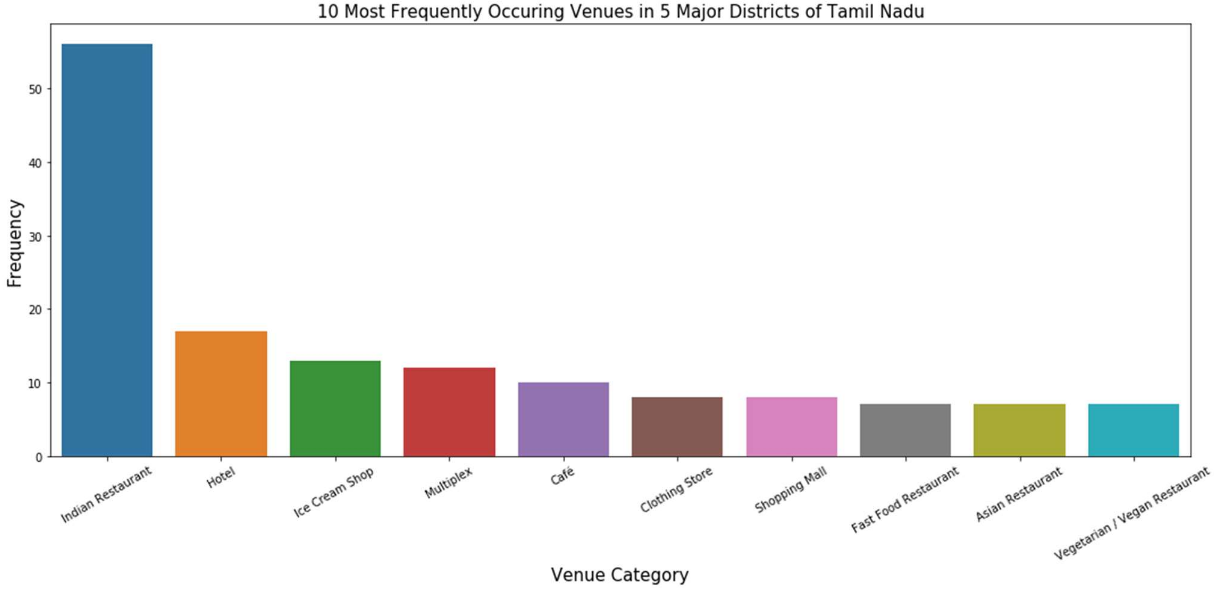


3 Using Foursquare API to explore Tamil Nadu for the number of multiplexes in the top 5 most populated districts

Using personal Foursquare credentials, The Foursquare API to obtain the 100 most common venues within 1 kilometer of each major district.

3.1Exploring the Data and Major Districts of Tamilnadu

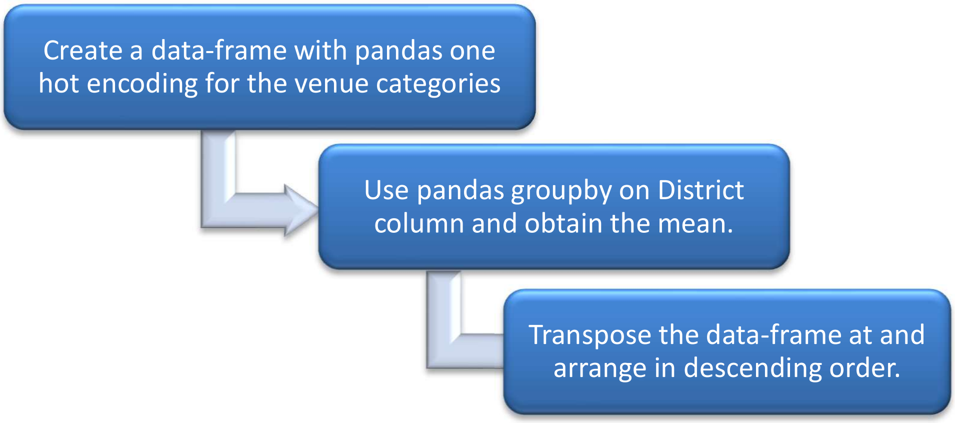
From the Foursquare data, the number of unique venue categories obtained were 74, however as the focus of the study is multiplexes, concentration was shifted to the multiplex category. As the focus is on 5 most populous districts, it was seen that there are only 12 multiplexes among the 500 top venues in these 5 districts. A plot of the ten most frequent venues in these 5 districts are as below



As Indian restaurants are the most frequently visited venues, here's a glimpse of what a typical south Indian Thali looks like.

Fig 2: A typical south Indian thali

The information about the top 5 venues of each district is obtained as follows.

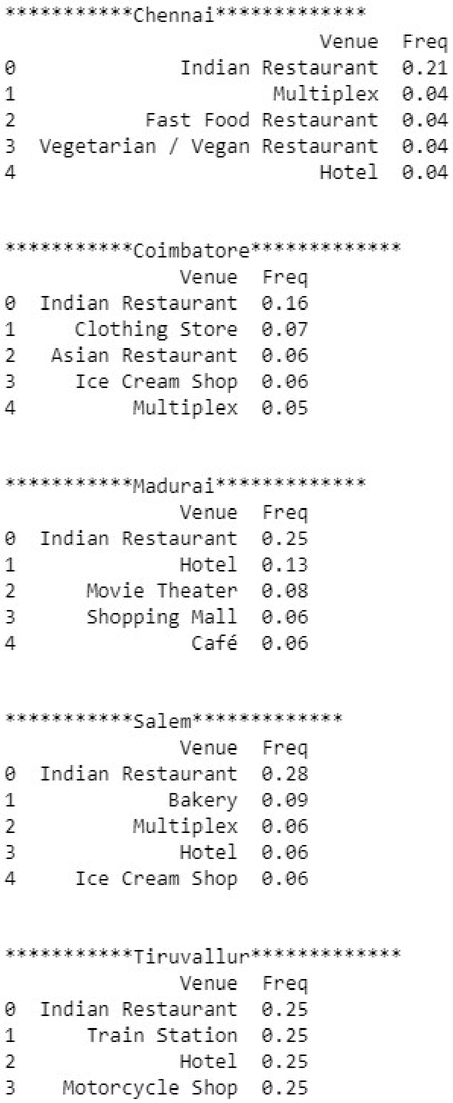


Create a data-frame with pandas one hot encoding for the venue categories

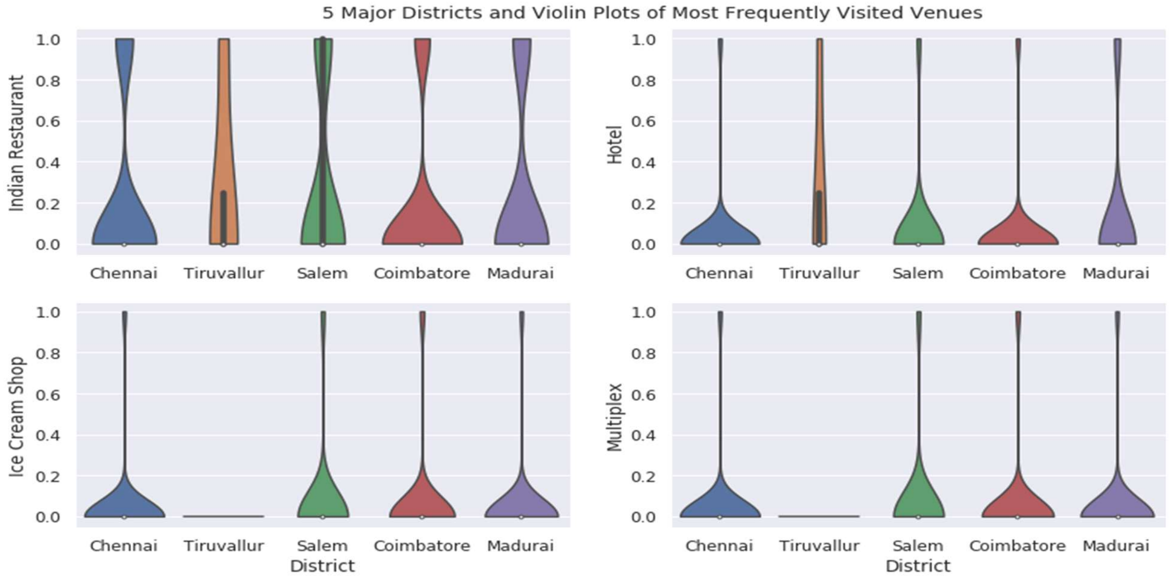
Use pandas groupby on District column and obtain the mean.

Transpose the data-frame at and arrange in descending order.

Implementing them in Pandas outputs the following--

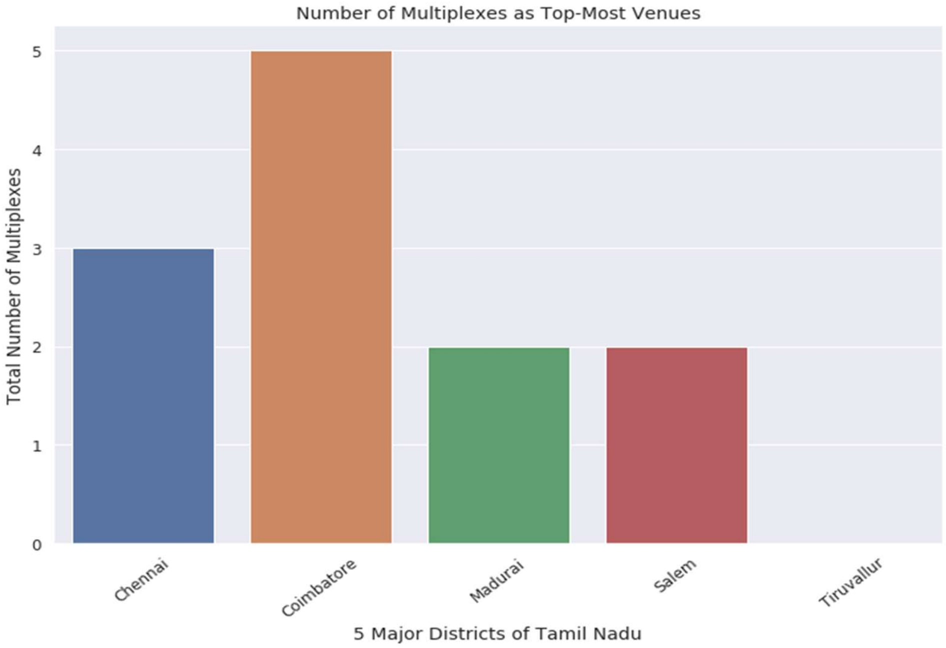


```
***********Chennai*************
                            Venue  Freq
0               Indian Restaurant  0.21
1                       Multiplex  0.04
2            Fast Food Restaurant  0.04
3  Vegetarian / Vegan Restaurant  0.04
4                           Hotel  0.04


***********Coimbatore*************
               Venue  Freq
0  Indian Restaurant  0.16
1     Clothing Store  0.07
2   Asian Restaurant  0.06
3     Ice Cream Shop  0.06
4          Multiplex  0.05


***********Madurai*************
               Venue  Freq
0  Indian Restaurant  0.25
1              Hotel  0.13
2      Movie Theater  0.08
3      Shopping Mall  0.06
4               Café  0.06


***********Salem*************
               Venue  Freq
0  Indian Restaurant  0.28
1             Bakery  0.09
2          Multiplex  0.06
3              Hotel  0.06
4     Ice Cream Shop  0.06


***********Tiruvallur*************
               Venue  Freq
0  Indian Restaurant  0.25
1      Train Station  0.25
2              Hotel  0.25
3    Motorcycle Shop  0.25
```



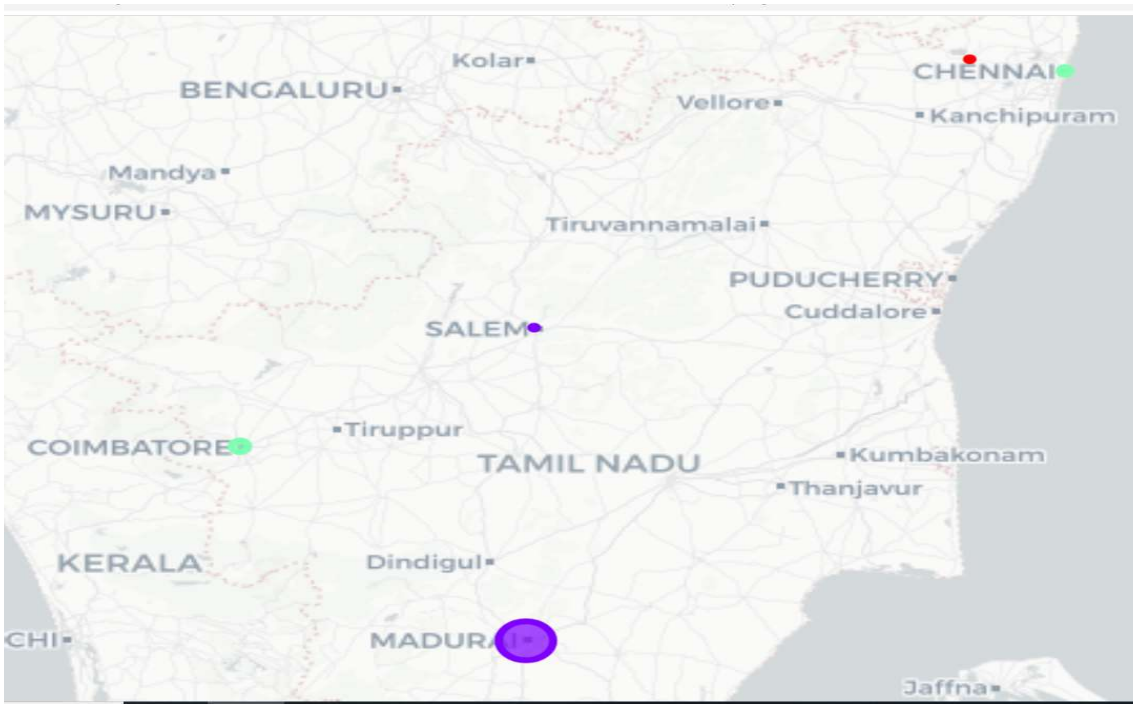
From the above plot, we can infer that every district except for Tiruvallur has some number of multiplexes. However, considering that Tiruvallur is the third most populous district, it lacks any multiplexes. Also, it is important to note here that most common venue in Tiruvallur are Indian Restaurants which is a good indication that people like to venture out to eat.

The same observation is corroborated by the plot given below which shows the number of multiplexes as the top-most venues and Tiruvallur is again lagging behind in the race.



4 Using K-Means Clustering for clustering the Tamil Nadu Districts (Neighborhoods)

Finally, these 5 districts are clustered on the basis of the frequency of venue categories and, K-Means clustering algorithm from Scikitlearn library using 3 cluster centres is used to cluster the districts on the basis of frequency of venue categories. So the entire expectation would be based on the similarities of venue categories.



Here the radius of the circles represents the number of multiplexes as most common venue for the corresponding district and, we have seen before that it is maximum for Madurai district and none for Tiruvallur.

From the most common venues this clustering makes a complete sense as Coimbatore maximum number of multiplexes (green cluster), Chennai (red cluster) followed by Madurai and Salem having same number of multiplexes (purple cluster).

5 Results and Discussion

The results of the exploratory data analysis and clustering are summarized below--

- Indian restaurants top the charts of most common venues in the 5 districts.

- Coimbatore district has the maximum number of multiplexes as the most popular venues.
- Despite being thickly populated, Tiruvallur does not have any multiplex whereas this district has the maximum number of Indian restaurants.

From the analysis, Tiruvallur district is the most appropriate district for opening a new multiplex to provide people with a weekend getaway and an avenue to watch movies as well as eat out.

Some drawbacks of this analysis are-- the clustering is completely based on the most common venues obtained from Foursquare data. Since land price, distance of the venues from closest stations, number of potential customers, benefits and drawbacks of Tiruvallur, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it definitely gives us some very important preliminary information on possibilities of opening restaurants around the major districts of Tokyo.

Also, another pitfall of this analysis could be consideration of the entire district as a whole not taking into consideration various Talukas within each district, taking into account of all the areas under the 5 major wards would give us an even more realistic picture.

6 Conclusion

The project was a real hands-on experience for applying the concepts of software engineering. Right from the conceptualizing of the idea to executing it and then finally drawing conclusions from it for a major business decision has really been informative. Scrapping data from the web, cleaning it and then applying various visualization tools to it so that it can make sense and communicate the information required was really a great learning experience. Finally to conclude this project provided a true glimpse of how real life data-science projects look like. Also it helped in understanding the usage of some frequently used python libraries to scrap web-data, Foursquare API to explore the major districts of Tamil Nadu and visualizing the results of segmentation of districts using Folium leaflet map. Potential for this kind of analysis in a real-life business problem is discussed in great detail. Finally, since the entire analysis was mostly concentrated on the possibilities of opening a multiplex targeting the fact that people are fond of going out and eating out, some of the results obtained are contrary to what the expectations were.