Classification of Delhi Metro stations

Introduction

- Delhi Metro is a rapid transit system serving Delhi and its satellite cities in the National Capital Region of India. As of now, there are a total of 229 metro stations including the Airport Express stations. The first section of the Delhi Metro opened on 25 December 2002 with the Red Line, and has since been expanded to around 347.66 km (216.03 miles) of route length as of 4 October 2019. The network has nine operational lines and is built and operated by the Delhi Metro Rail Corporation Limited (DMRC). The Delhi Metro Rail Corporation makes 2,700 trips per day carrying 1.5 million passengers, who on an average travel a distance of 17 kilometres each.

- For this project, we will try to look at the places surrounding these metro stations and classify them according to the similarity of nearby venues. Almost every one use metro transit to migrate from one place to another for reasons which can be personal of professional. If there are more professional places like companies, offices surrounding a station then it will mostly be used by working professionals. Then there are some stations with many universities or colleges nearby and is used by Students mostly. Stations which have places like amusement parks, malls, monuments are used by people for recreation.

- We can classify stations by primary usage analysing the data that contains the number of nearby venues according to their category. This can help plan further extension of the network and find places for new development.

Data

- In this section we will describe our base data which we will analyse to reach the goal we want.
- We will be requiring some Python libraries and modules like Pandas, Numpy, JSON, requests, geopy, bs4, sklearn, etc.to move further in the project.
- The main source of stations data that will contain the list of all the metro stations under DMRC in Delhi. Is the following Wikipedia page.
- https://en.wikipedia.org/wiki/List_of_Delhi_Metro_stations
- We have to scrape the relevant table data from this url like Station name, Line.
- To get the latitude and longitude values we will use geopy.geocoder module and will geocode each station.

Assumption :-

- There are some stations with more than one line that pass through it. So we have assumed that only the line that is written first on the above url will be the data of our choice to nullify the ambiguity that we may face while plotting on the graph
- To get the nearby venues of all stations we will be using Foursquare API for which we need to create an account on the following url.
- https://developer.foursquare.com/.
- After this we will be required to pre-process data and create a data frame. After the pre-processing our data will look something like in the next slide.

	Station	Line	Latitude	Longitude
0	Adarsh Nagar	1	28.714401	77.167288
46	Ghitorni	1	28.493751	77.149187
81	Jor Bagh	1	28.588239	77.216528
134	New Delhi	1	28.643641	77.221737
68	Jahangirpuri	1	28.725972	77.162658

Data Sample

Methodology

- We have the the data related to the metro station under DMRC, their latitudes and longitudes values with the line of Metro that passed through them.
- For further processing we will be utilizing the **FOURSQUARE API** and explore the venues in a specific radius around that using their coordinates values.
- We will use explore query under venues of the api calls. You can read more about the various Endpoints provided Foursquare api on the following link. **https://developer.foursquare.com/docs/places-api/endpoints/**

Foursquare API Calls

- The advantage of using Foursquare api is that venues are classiied in particular categories. Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)
- Nightlife Spot (4d4b7105d754a06376d81259)
- Outdoors & Recreation (4d4b7105d754a06377d81259)
- Professional & Other Places (4d4b7105d754a06375d81259)
- Residence (4e67e38e036454776db1fb3a)
- Shop & Service (4d4b7105d754a06378d81259)
- Travel & Transport (4d4b7105d754a06379d81259)
- We'll be querying the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance.
- **explore_url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, long, radius, categoryId)**

	Station	Line	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation
0	Adarsh Nagar	1	28.714401	77.167288	1	1	0	8	1	1
1	Ghitorni	1	28.493751	77.149187	2	1	0	5	1	3
2	Jor Bagh	1	28.588239	77.216528	3	2	2	17	2	3
3	New Delhi	1	28.643642	77.221737	2	5	2	36	9	1
4	Jahangirpuri	1	28.725972	77.162658	0	1	0	2	0	1

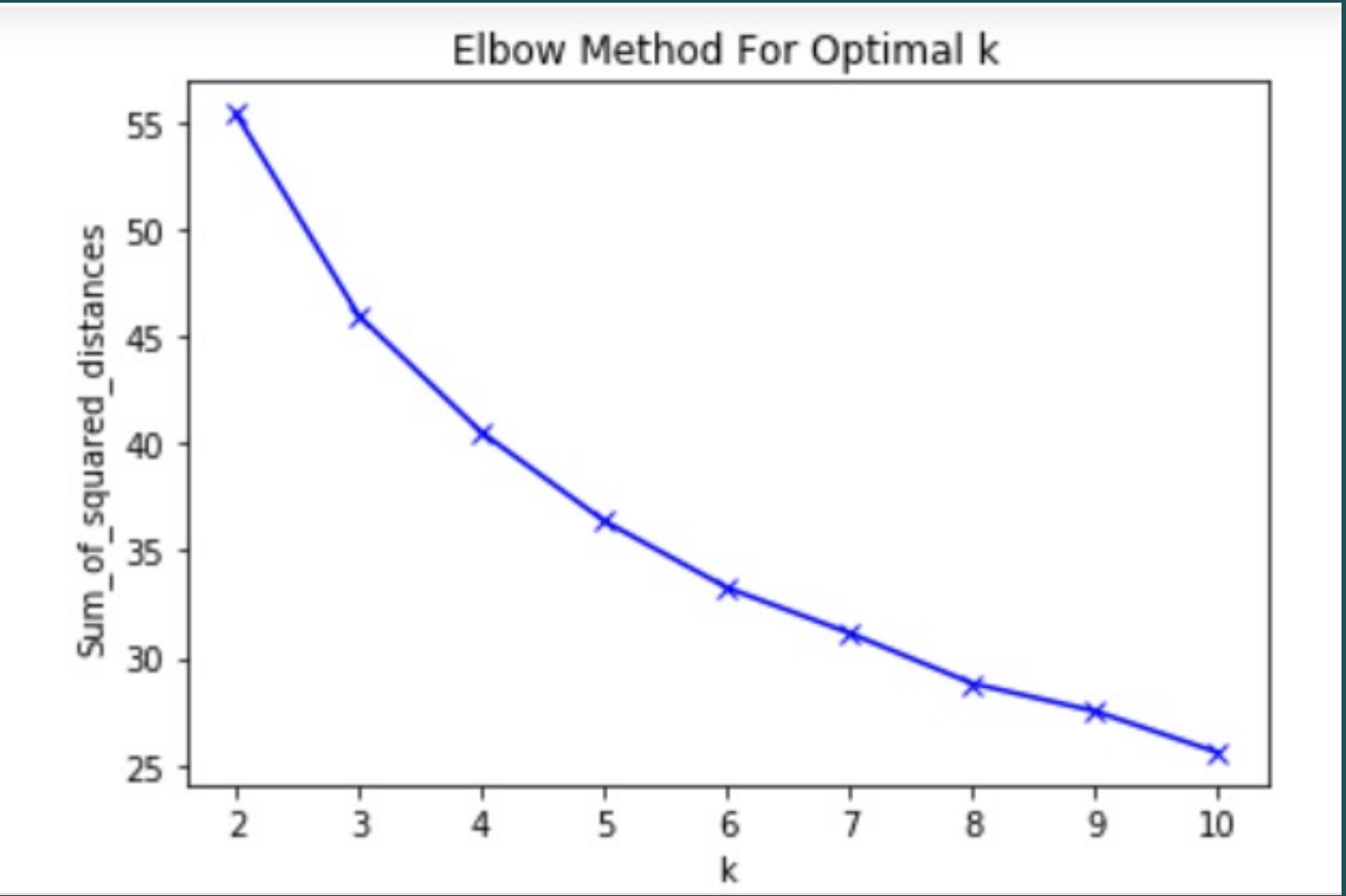
New Data

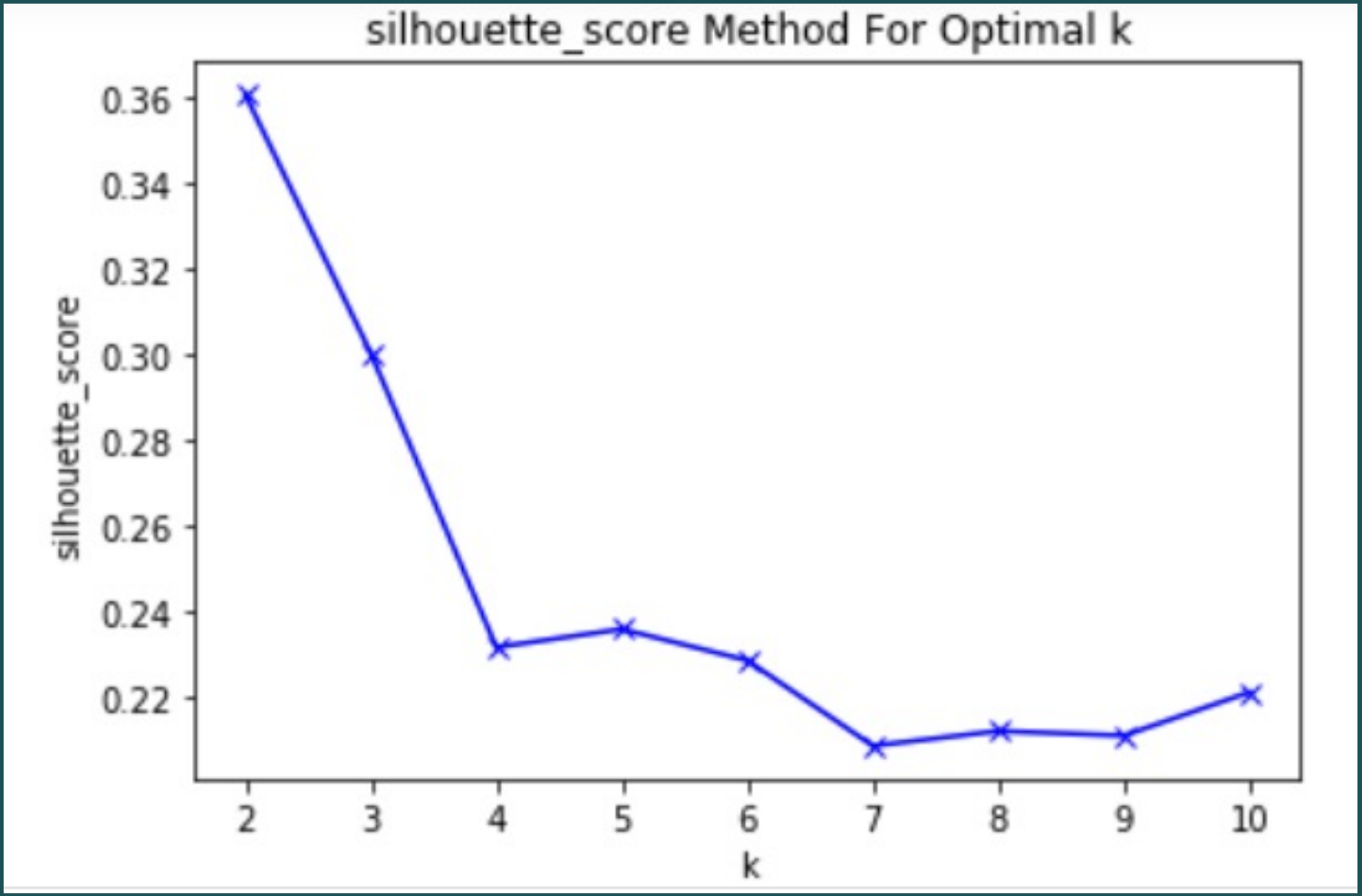
	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Sh Ser
0	0.083333	0.066667	0.0	0.101266	0.026316	0.090909	0.121951	0.090909	0.37
1	0.166667	0.066667	0.0	0.063291	0.026316	0.272727	0.048780	0.272727	0.05
2	0.250000	0.133333	0.4	0.215190	0.052632	0.272727	0.317073	0.363636	0.07
3	0.166667	0.333333	0.4	0.455696	0.236842	0.090909	0.146341	0.090909	0.16
4	0.000000	0.066667	0.0	0.025316	0.000000	0.090909	0.097561	0.090909	0.24

Data Processing and Normalization

Clustering

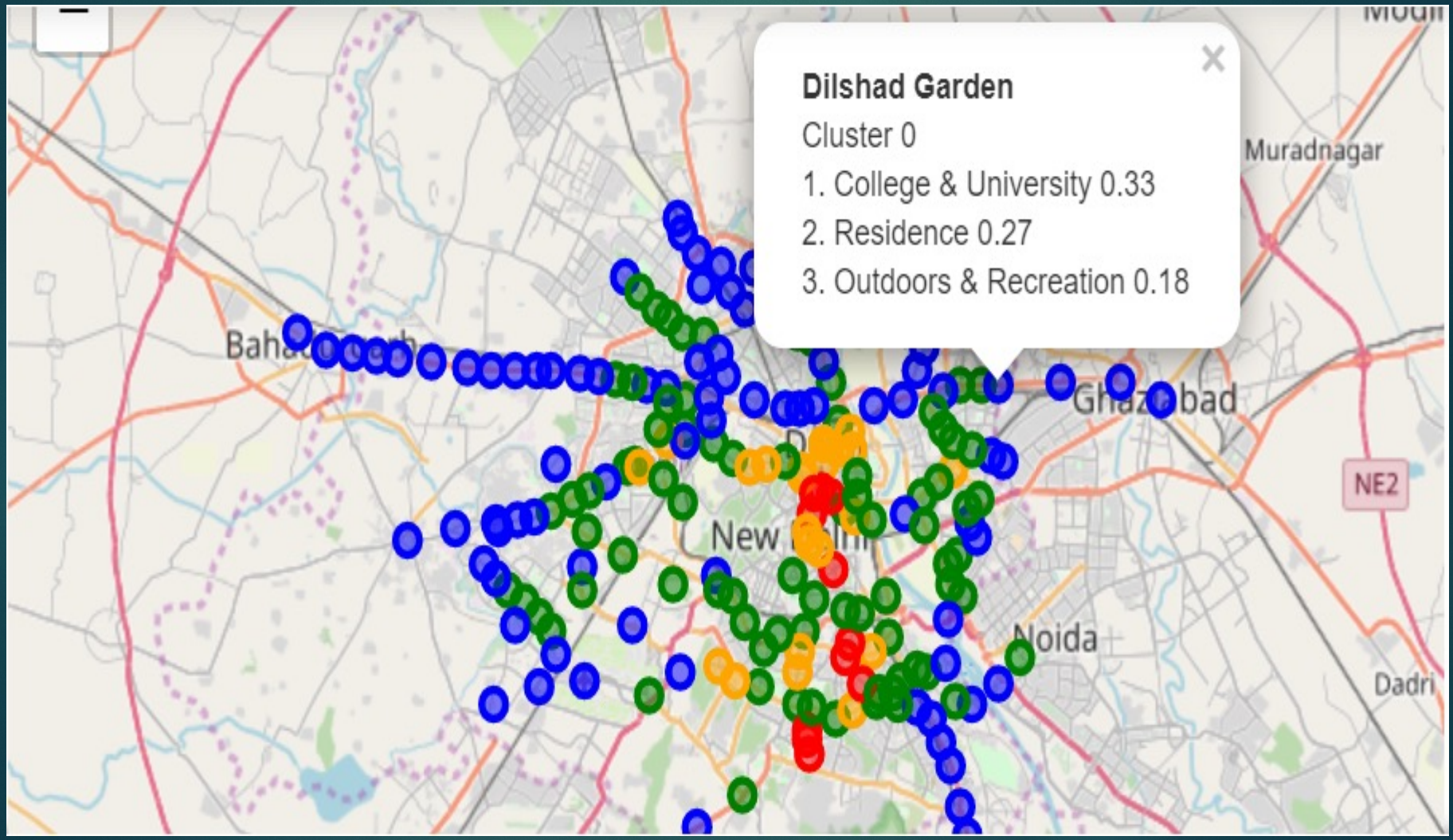
- So we have the data and using an unsupervised clustering Algorithm known as KMeans Clusteing we will Create clusters. To get started You may know that 'K' Means Numbers of clusters in which data is to be categorized. Now problem is we dont know how to evaluate the best K value so we will run the Kmeans algorithm on different values of K. Then we will use some measures or metrics to see which K value will be the most suitable.
- First we will use the 'ELBOW METHOD' to check the most appropriate value of K.
- Second we use the ‘Silhoutte Score Method’.


Elbow Method For Optimal k
Sum_of_squared_distances
55
50
45
40
35
30
25
2
3
4
5
6
7
8
9
10
k


silhouette_score Method For Optimal k
silhouette_score
0.36
0.34
0.32
0.30
0.28
0.26
0.24
0.22
2
3
4
5
6
7
8
9
10
k

Clusters Visualization and Analysis

- Using Number of clusters as 4 we have visualize the map.



- **Results:-**
- **Cluster 0 (Blue)**
- Data Analysis Shows that **Blue cluster/ Cluster 0** have high value in almost all the fields of College & University, Professional & Other Places, Outdoors & Recreation, Shops & Service and also in other categories for some stations.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence
0	0.083333	0.066667	0.0	0.101266	0.026316	0.090909	0.121951	0.090909
1	0.166667	0.066667	0.0	0.063291	0.026316	0.272727	0.048780	0.272727
4	0.000000	0.066667	0.0	0.025316	0.000000	0.090909	0.097561	0.090909
6	0.000000	0.000000	0.0	0.025316	0.000000	0.000000	0.097561	0.000000
9	0.000000	0.000000	0.0	0.025316	0.000000	0.181818	0.073171	0.090909

Cluster 1 (Blue)

- Data Analysis Shows that **Green cluster/ Cluster 1** have high value in almost all the fields of Residence, Professional & Other Places, Outdoors & Recreation, Arts & Entertainment , Shops & Services and College & University than other categories.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence
2	0.250000	0.133333	0.4	0.215190	0.052632	0.272727	0.317073	0.363636
5	0.500000	0.200000	0.0	0.075949	0.078947	0.363636	0.170732	0.181818
11	0.333333	0.400000	0.2	0.265823	0.210526	0.363636	0.439024	0.272727
16	0.166667	0.400000	0.0	0.215190	0.026316	0.636364	0.146341	0.363636
19	0.166667	0.333333	0.0	0.075949	0.052632	0.181818	0.146341	0.272727

CLUSTER 2 (YELLOW)

- Data Analysis Shows that Yellow cluster/ Cluster 2 have high value in almost all the fields of Event, Residence, Outdoors & Recreation, Arts & Entertainment , Travel & Transport than other categories.
- Analysis shows us that this cluster is geographically centered at almost central part of Delhi and including stations like Kashmere Gate, Welcome which have interchanges of Travel and Transport facility.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence
3	0.166667	0.333333	0.4	0.455696	0.236842	0.090909	0.146341	0.090909
8	0.583333	0.133333	0.4	0.379747	0.263158	0.363636	0.439024	0.454545
10	0.250000	0.200000	0.6	0.354430	0.105263	0.454545	0.829268	0.363636
15	0.333333	0.400000	0.6	0.113924	0.236842	0.545455	0.560976	0.272727
17	0.250000	0.400000	0.6	0.430380	0.342105	0.454545	0.414634	0.363636

CLUSTER 3 (RED)

- Data Analysis Shows that Red cluster/ Cluster 3 have high value in almost all the fields of Event, Nightlife, Outdoors & Recreation, Professional & Other Places, College & University than other categories.
- Analysis shows us that this cluster consists of industrial and professional places like Faridabad, some Noida sectors.

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence
7	0.666667	0.666667	0.2	0.607595	0.526316	0.363636	0.317073	0.545455
12	0.416667	0.266667	0.2	0.582278	0.421053	0.181818	1.000000	0.181818
14	0.416667	0.266667	1.0	0.911392	1.000000	0.636364	0.658537	0.181818
18	0.166667	0.466667	0.0	0.367089	0.236842	0.363636	0.878049	0.090909
31	0.583333	0.266667	0.2	0.848101	0.421053	0.727273	0.268293	0.363636

Discussions

- Note that Foursquare data isn't very precise and accurate. The Foursquare counts for the venues that we can explore via Foursquare data. This doesn’t take into account other attributes that are essential and impotant to take while making clusters like Venue's size, services it provide, small shops or 'Rehdis' that are usually present almost everywhere. Moreover we do not take into account the Food while analysis as this is the attribute almost equal everywhere.
- Also using Geocoder has some limitations in how accurate the geocoding of an address will be returned. There is also ambiguity with some places due to similar names or changes in names.

Conclusion

- Foursquare data is limited but can provide insights into a city's development. This data could be combined with other sources to provide more accurate results.