IBM – Coursera

Data Science Specialization

Capstone project - Final report

CLUSTERING OF DISTRICTS IN GUJARAT FOR REAL ESTATE INVESTMENTS BY ABC GROUP

Ligin Saju George – 2020

Table of Contents

I. INTRODUCTION

ABC Group is a multinational real estate Company which deals in building residential Societies in operating Countries. They are mainly focused in developing countries where the investments are cheaper compared to developed countries and the ROI is higher in the long run. They are looking for strategic investments in India where they can either buy properties or lands based on their analysis or construct residential societies in suitable cities. Their residential societies are fully luxurious and apt for upper middle class families or working professionals.

The Gujarat Government, as part of their flagship program which aims at attracting foriegn investments to their state, has invited ABC group to invest in Gujarat. The Government has offered to assure full support and help them establish as a real estate brand in Gujarat.

I. Problem Definition: ABC Group wants to cluster the districts of Gujarat based on the urban development and population growth so as to strategically decide on investing in Real estate.

The company wants to know which cities are highly developed with urban centres and growing population and which cities have the potential to grow in the long term. This is needed to decide on the type of real estate investment to be made in these cities.

II. DATA DESCRIPTION

The data needed for this analysis is as follows:

- The List of Districts of Gujarat with their population , Density of population and location attributes:

  This data is available on the wikipedia page titled, List of districts of Gujarat. (https://en.wikipedia.org/wiki/List_of_districts_of_Gujarat). This data can be extracted by the process of web scraping.

  The location attributes of each district can be extracted by the use of the python tool Geocoder.

- The popular venues in each district which represents the type of development the city has:

  This data can be extracted with the help of a foursquare API which helps to explore any location based on its location attributes.

III. METHODOLOGY

The aim is to find different clusters of districts which differ in terms of urbanisation and population growth. The assumption made here is that the number and variety of venues in a district is a metric to measure its level of Urbanisation. Because of this reason, a restaurant and an airport is viewed with the same weightage which can be a drawback of this study. The methodology used to cluster the dataset is k-means clustering. This is a process of segmenting different data points with respect to ‘k’ number of imaginary points set by the programmer. The clustering is done based on the distance between the data points and the imaginary points.

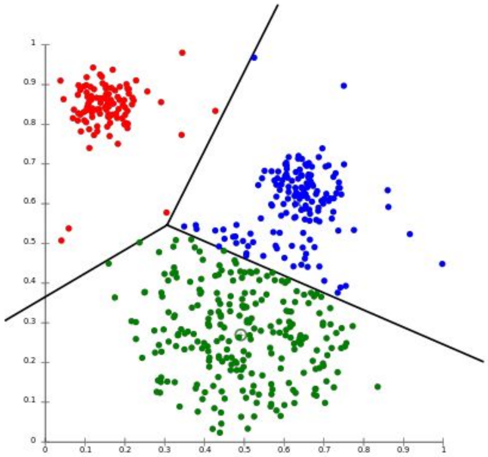
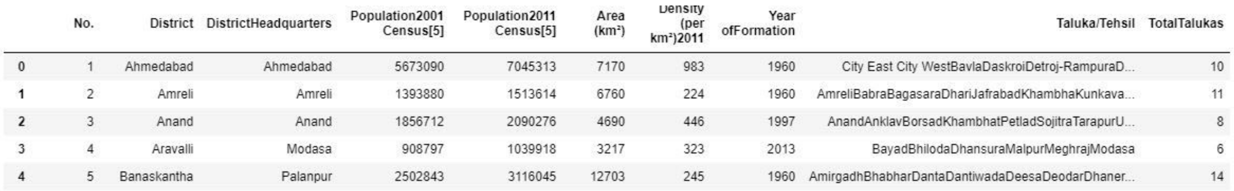


Figure 1: An Illustration of K-Means Clustering(K = 3)

I. Data Cleaning:



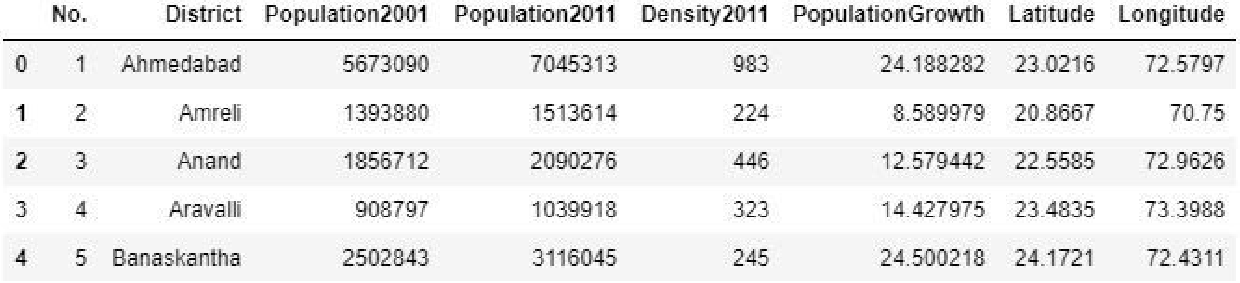
	No.	District	DistrictHeadquarters	Population2001 Census[5]	Population2011 Census[5]	Area (km²)	Density (per km²)2011	Year ofFormation	Taluka/Tehsil	TotalTalukas
0	1	Ahmedabad	Ahmedabad	5673090	7045313	7170	983	1960	City East City WestBavlaDaskroiDetroj-RampuraD...	10
1	2	Amreli	Amreli	1393880	1513614	6760	224	1960	AmreliBabraBagasaraDhariJafrabadKhambhaKunkava...	11
2	3	Anand	Anand	1856712	2090276	4690	446	1997	AnandAnklavBorsadKhambhatPetladSojitraTarapurU...	8
3	4	Aravalli	Modasa	908797	1039918	3217	323	2013	BayadBhilodaDhansuraMalpurMeghrajModasa	6
4	5	Banaskantha	Palanpur	2502843	3116045	12703	245	1960	AmirgadhBhabharDantaDantiwadaDeesaDeodarDhaner...	14

Figure 2: The head of initial table obtained from wikipedia

The initial table obtained from web scraping is shown above. We have the list of Districts (i.e 33) and important details like population in both 2001 and 2011, density per sq.km and area. The process of data cleaning is done to drop the unwanted columns. The percentage growth of population from 2001 to 2011 can be obtained from the growth equation and it is made as a new column.

$$Population\ Growth\ Percentage\ = \frac{Total\ Population\ in\ 2011 - Total\ Population\ in\ 2001}{Total\ Population\ in\ 2001} \times 100$$

The Latitude and Longitude of each district is required to find the venues in the locality. This is obtained by using the geocoder library. The final dataset which is obtained after cleaning is shown below.

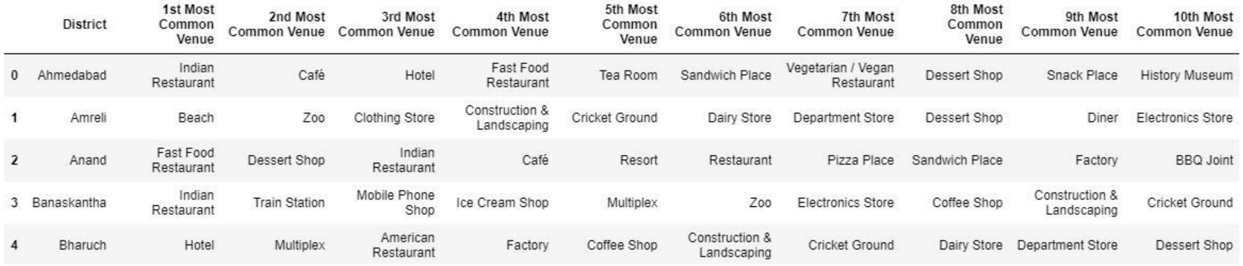


	No.	District	Population2001	Population2011	Density2011	PopulationGrowth	Latitude	Longitude
0	1	Ahmedabad	5673090	7045313	983	24.188282	23.0216	72.5797
1	2	Amreli	1393880	1513614	224	8.589979	20.8667	70.75
2	3	Anand	1856712	2090276	446	12.579442	22.5585	72.9626
3	4	Aravalli	908797	1039918	323	14.427975	23.4835	73.3988
4	5	Banaskantha	2502843	3116045	245	24.500218	24.1721	72.4311

Figure 3: The dataset after cleaning

II. Using Foursquare API to extract Venues:

The district venues are explored with the help of foursquare API. This is a commercial service which provides location details with the help of latitudes and longitudes data. The data extracted is used to map each district and their venues. The top ten venues of each district can help in understanding the type of district. For example, The top 10 venues in Ahmedabad are related to restaurants and food courts which indicates that the city is lively and populated. Similarly some districts have industries in their top 10 or beaches and related venues.



	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Ahmedabad	Indian Restaurant	Café	Hotel	Fast Food Restaurant	Tea Room	Sandwich Place	Vegetarian / Vegan Restaurant	Dessert Shop	Snack Place	History Museum
1	Amreli	Beach	Zoo	Clothing Store	Construction & Landscaping	Cricket Ground	Dairy Store	Department Store	Dessert Shop	Diner	Electronics Store
2	Anand	Fast Food Restaurant	Dessert Shop	Indian Restaurant	Café	Resort	Restaurant	Pizza Place	Sandwich Place	Factory	BBQ Joint
3	Banaskantha	Indian Restaurant	Train Station	Mobile Phone Shop	Ice Cream Shop	Multiplex	Zoo	Electronics Store	Coffee Shop	Construction & Landscaping	Cricket Ground
4	Bharuch	Hotel	Multiplex	American Restaurant	Factory	Coffee Shop	Construction & Landscaping	Cricket Ground	Dairy Store	Department Store	Dessert Shop

Figure 4: The top 10 venues dataframe

III. Applying Clustering on the data:

The first clustering was done on the data which was created on the basis of the venues in each district. This is assumed as the metric to understand urban development. The number of clusters was set to 4.

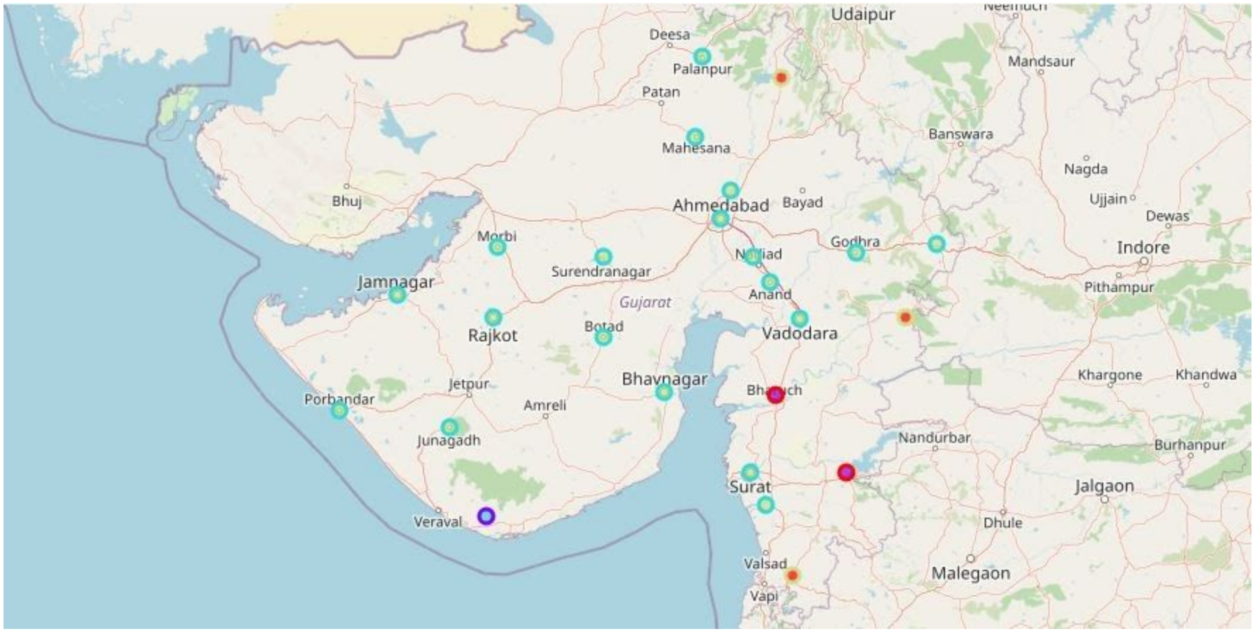


Figure 5: CLustering based on urban development(K = 4)

Now, the clustering is set to 3 because 4 is seen not to be segmenting the districts well and one cluster seems to be overcrowded with districts.

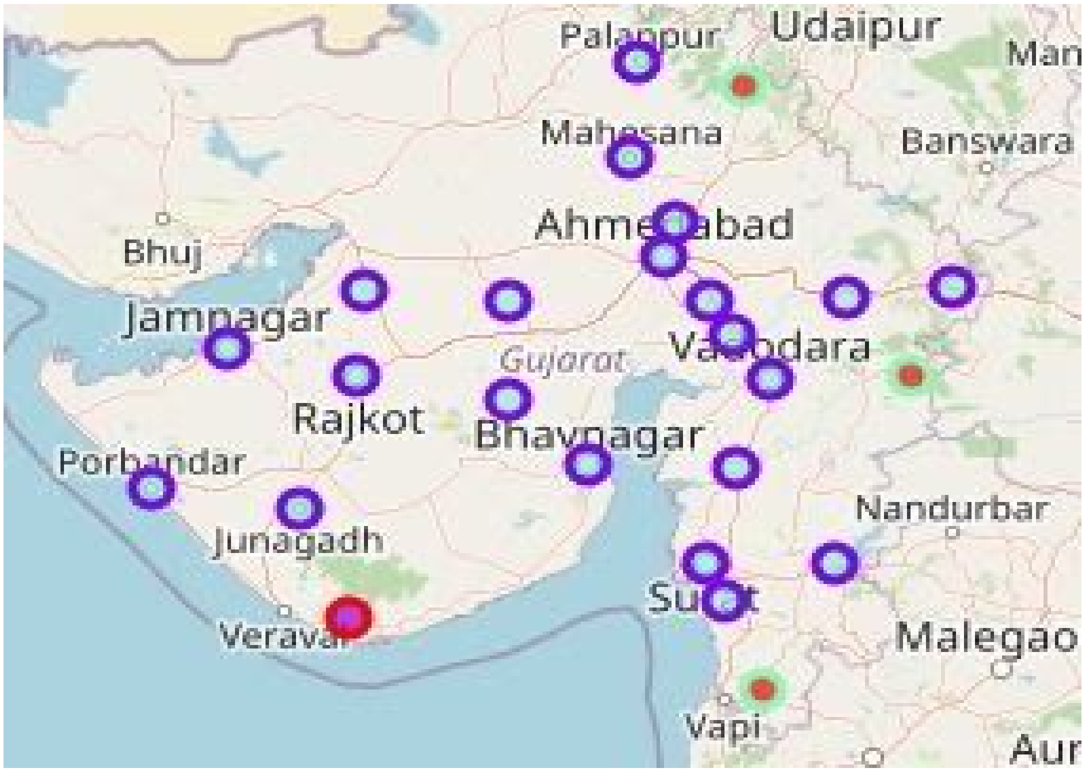


Figure 6: Clustering with respect to Urban Development(K = 3)

To incorporate the population growth, population growth percentage, population of 2011 and density of population is added to the venues dataset. It is then scaled and then clustered with k = 3.

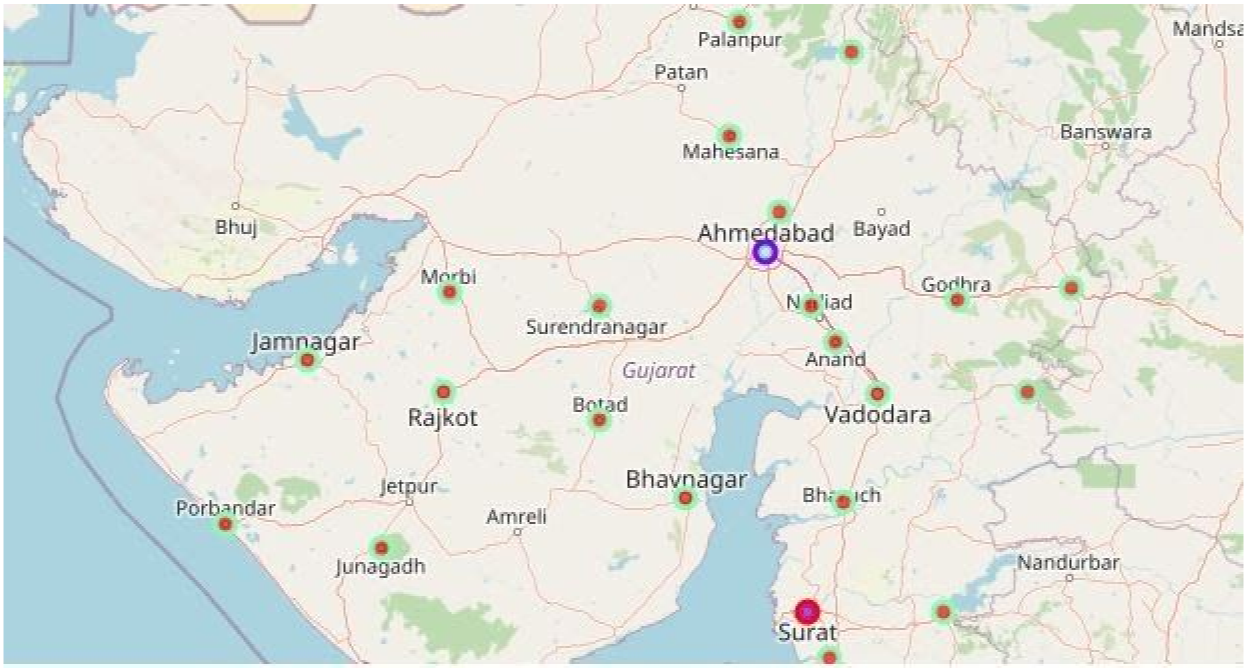


Figure 7: Clustering with respect to Urban development and Population growth(K = 3)

To see the clusters based on population alone , the data related to population is scaled and clustered.

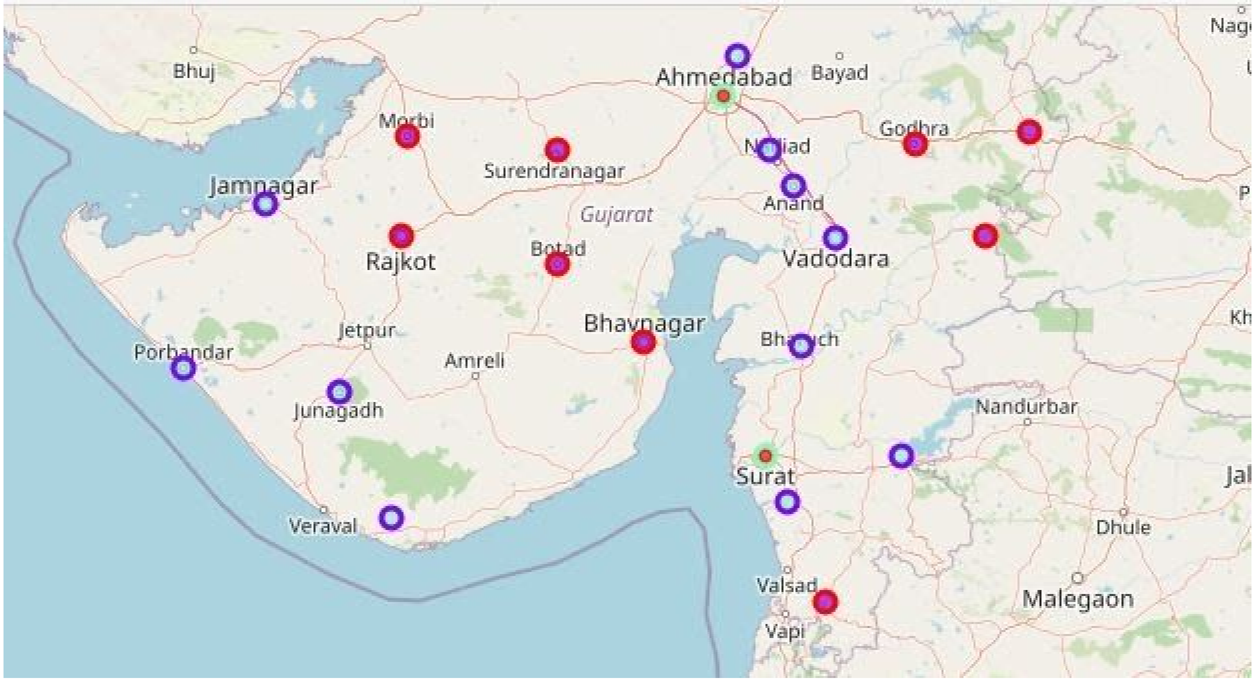


Figure 8: Clustering based on Population Growth (K = 3)

Based on these clustering, a matrix was created with the results. In order to overcome the accumulation of districts in the lower cluster of population and urban development, a sub-clustering was done which differentiated the districts into two in the lower cluster.

IV. RESULTS

The results obtained was converted to a table with rows featuring each district and columns featuring each clustering process. These are:

- Population Growth Alone.
- Urban Development Alone
- Population Growth and Urban Development Together.
- Sub Cluster of Population and Urban Development together for the Low cluster.

The Clusters are termed as High(H), Medium(M) and Low(L). The sub clustering is divided as Low-High and Low-Low.

	Population Growth	Urban Development	Population and Urban Development	Sub- Cluster of Population and Urban Development
Ahmedabad	H	H	H	N.A
Amreli	L	L	L	LL
Anand	L	H	L	LL
Banaskantha	M	H	L	LL
Bharuch	L	H	L	LL
Bhavnagar	M	H	L	LL
Botad	M	H	L	LL
Chotta Udaipur	M	M	L	LL
Dahod	M	H	L	LL
Gandhinagar	L	H	L	LL
Jamnagar	L	H	L	LL
Junagadh	L	H	L	LL
Kheda	L	H	L	LL
Mehsana	L	H	L	LL

Morbi	M	H	L	LL
Navsari	L	H	L	LL
Panchmahal	M	H	L	LL
Porbandhar	L	H	L	LL
Rajkot	M	H	L	LL
Sabarkantha	L	H	L	LL
Surat	H	H	M	N.A
Surendranagar	M	H	L	LL
Tapi	L	H	L	LL
Vadodara	L	H	L	LH
Valsad	M	M	L	LL

Table 1: The result matrix with clustering on each feature

V. Discussion

1. Cluster A (HHH, HHM) :

   Districts: Ahmedabad, Surat.

   This is the top most cluster with high population growth and well established urban development. It can be assumed that these cities have high migration into it and generate jobs by business and Industries. These cities can also have high competition from other players and the investment needed would be high. But, this cluster promises returns as these centres need more spaces to live.

2. Cluster B(MHL-LL):

   Districts: Banaskantha, Bhavnagar, Botad, Dahod, Morbi, Panchmahal, Rajkot, Surendranagr.

This cluster has medium population growth and high urban development. This cluster has the potential to grow as the population will rise in the future. Deep analysis will show that the density in some of these cities are higher. These cities are well suited for investments seeing its growth potential. These investments can be in the buying of land for building societies at a later stage.

3. Cluster C(LHL-LH):

   Districts: Vadodara

   This cluster has well urbanised cities with low population.

   This adds up to cluster this city in an overall low, but in the top subcluster. In this case, Vadodara even though has low population growth rate, the density of population is at par with top clusters. This may be an indication that there is less availability of space and less migration to this city. In this case, buying of properties can be a good option here.

4. Cluster D(LHL-LL):

   Districts: Anand, Bharuch, Gandhinagar, Jamnagar, Junagadh, Kheda, Mehsana, Navsari, Porbandhar, Tapi.

   This Cluster has low population growth and decent urbanisation which leads to an overall low growth. These cities can be considered for long term investments seeing the potential. For example, Gandhinagar in this cluster is the capital city, it has high population density but lower population growth which results in a lower cluster. This can be treated as an outlier.

VI. LIMITATIONS

1. The urban centers were treated with the same weightage. This resulted in treating an airport and a coffee shape as the same. This can cause the cluttering or overcrowding of districts in one cluster because if weights were specified, the venues with higher weightage can move to a different cluster. An ambiguity in creating weights is about how to give the weights. This can lead to a bias.
2. The population data used was based on the 2011 national census. This is clearly a drawback as the picture can be a lot different today which is nine years later.
3. The results give an impression that there are a lot of missing features which should have been added to make a clear clustering. Getting the number of Companies in a district could have given insight into the amount of job creation, data on the present number of residential societies could have opened a possibility to include the competition into the picture. These can be incorporated in a further study.
4. There seems to be errors in the foursquare data obtained. For example, Ahmedabad has an airport which was not captured in the data.

VII. CONCLUSION

The clusters obtained opens the possibilities of investments in the state of Gujarat. ABC company now has a better view of the state which should be deepened with further analysis. The Cluster A has cities which are fast growing. The competition and the cost of investment is going to be comparatively high but the returns can be handy considering the amount of migrations happening and the increase in jobs and comfort of living. These features should be added in a further study. Availability of water, the climatic conditions etc should also be considered in a further study. Cluster B is a soil for strategic investments seeing the growth potential Cluster B and Cluster C can yield better ROI in the long run seeing the comparatively lesser cost of investments. Finally the Cluster D should be considered for purchase of land which should be cheaper now and further developing them into residential parks in the long run.

VIII. REFERENCE

- https://en.wikipedia.org/wiki/List_of_districts_of_Gujarat