**Clustering Cities in Gujarat State of India based on Demographic, Education and Location Data**

**1. Introduction/ Business Problem**

**Description & Discussion of the Background**

**Gujarat** is a state on the western coast of India with a coastline of 1,600 km (990 mi) and a population in excess of 60 million. It is the sixth largest Indian state by area and the ninth largest state by population. The economy of Gujarat is the fifth-largest state economy in India with ₹14.96 lakh crore (US$210 billion) in gross domestic product and a per capita GDP of ₹157,000 (US$2,200).

As you can see from the figures, Gujarat is a great state to live in. When we think of this from the perspective of an immigrant, we can ask the question：

1. Where is/are the best city/cities to live in Gujarat?

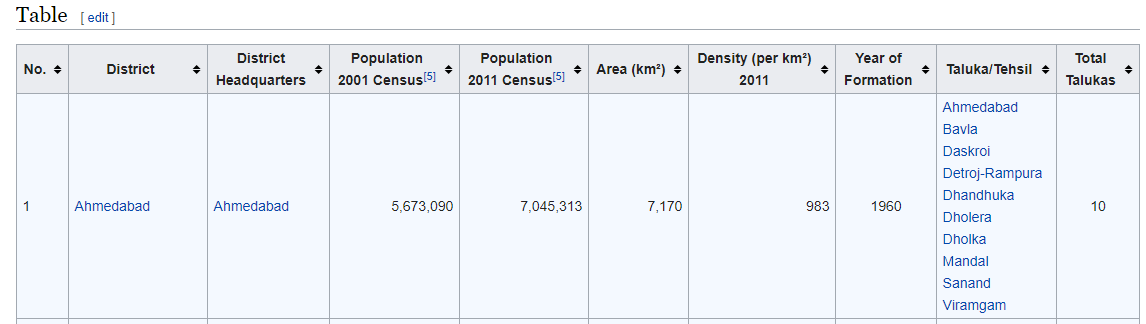
2. How can the cities be categorized according to demographic, educational and location data?

When we consider all these problems, we can cluster the cities according to the data described in the section below.

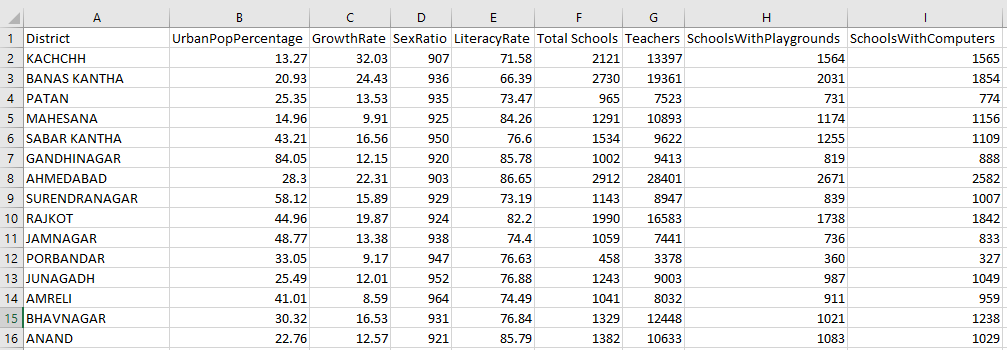
**2. Data Description**

To consider the problem, I have listed the data that will be used in the project as below:

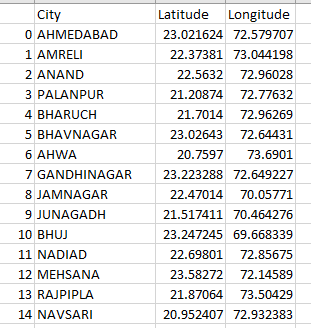
* List of Cities (District Headquarters) from District Table - https://en.wikipedia.org/wiki/List\_of\_districts\_of\_Gujarat



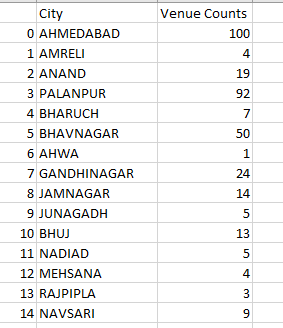
* Kaggle Dataset - 2015-16-District-wise Education Data India - https://www.kaggle.com/rajanand/education-in-india



* Geo Location Data - Geocoder Library/API



* Location Data - Total Venues per City - Foursquare API



**3. Methodology**

**Obtaining Data and Deciding What Features to Keep**

The data was obtained from various sources including Wikipedia, Kaggle, Geocoder Mapquest API and Foursquare API as mentioned above.

There was an attempt to obtain Geocoding data related to the cities using Google Maps API but because of API usage restrictions it didn't prove fruitful.

Over 650 irrelevant features were discarded from the Kaggle Dataset which contained about 680 features. Samples containing data from states other than Gujarat were discarded.

Only the venue counts in the 15 km radius of the 33 cities were obtained using the Foursquare API. The Categorical counts for those venues were available like the counts of cafes and restaurants or counts of libraries and movie theatres. But the categorical venue counts were not considered because for example, not all cities contained a museum venue.

**Combining Data from Various Sources and Data Wrangling**

It was a difficult task combining data from four different sources. The data from Wikipedia was in HTML format. The Kaggle dataset was in csv format. The data obtained from the two API's was in JSON format.

Four separate Pandas DataFrames were created from the data obtained from these four different sources and then they were combined into a single DataFrame.

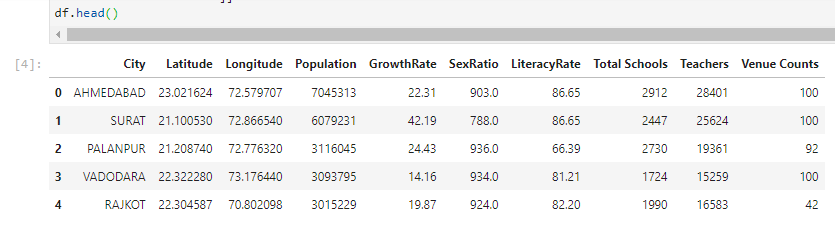
Some columns in the separate DataFrames were of the wrong datatypes so they had to be converted to numerical datatypes after some wrangling like removing commas from numbers. Some samples containing NaN values were removed.

All four Pandas DataFrames were created with one feature shared by all and that was 'City' which contained the values - City names. The dataframes were combined based on the city names.

Some samples(cities) had to be discarded because one of the four DataFrames didn't contain data for that city. After this data for 25 city headquarters for the Gujarat State remained.

After this, the combined DataFrame was sorted according to Population in descending order and the 5 least populated cities were discarded.

Below is the screenshot for the combined DataFrame.



**KMeans clustering - Clustering the 20 cities of Gujarat State**

The KMeans clustering algorithm was used to cluster the 20 cities of the Gujarat state. The number of clusters was chosen to be 4 and the number of iterations was chosen to be 20.

The data was normalized using the MinMaxScaler from the sklearn library in Python and then fed to the KMeans cluster object to be fitted.

Following is the partial resulting DataFrame with Cluster Labels included:

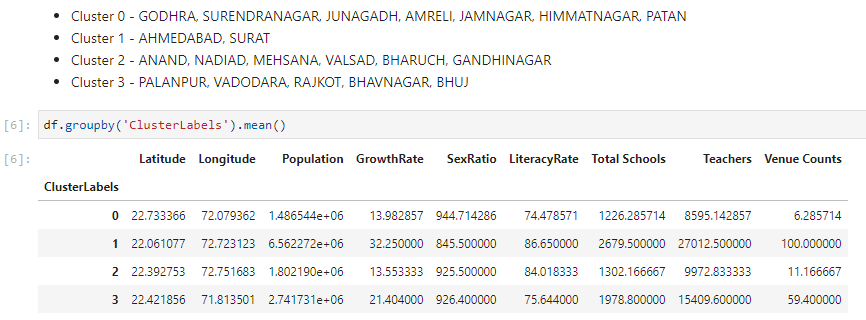


**4. Results**

**Clustered Cities**

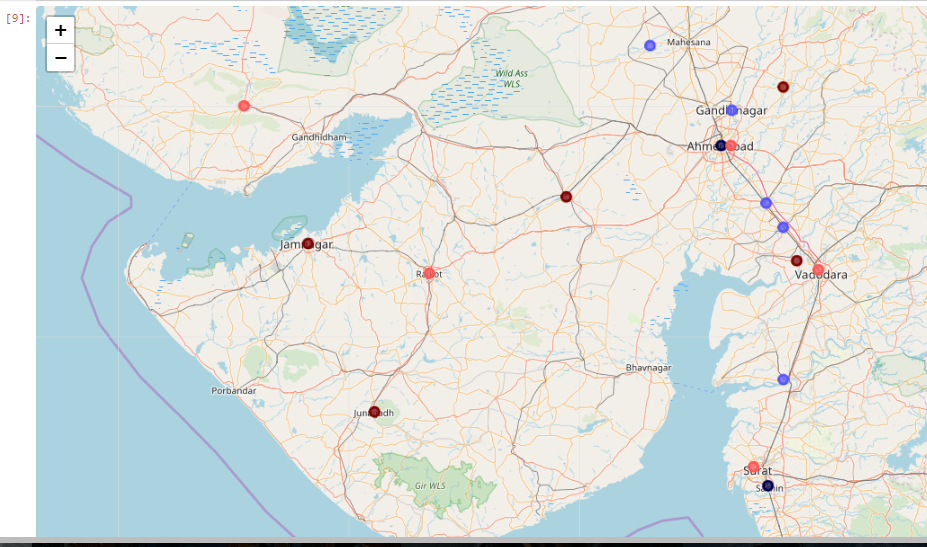
* Cluster 0 - GODHRA, SURENDRANAGAR, JUNAGADH, AMRELI, JAMNAGAR, HIMMATNAGAR, PATAN
* Cluster 1 - AHMEDABAD, SURAT
* Cluster 2 - ANAND, NADIAD, MEHSANA, VALSAD, BHARUCH, GANDHINAGAR
* Cluster 3 - PALANPUR, VADODARA, RAJKOT, BHAVNAGAR, BHUJ

**Cluster Analysis Based on Feature Mean Values Per Cluster**

****

* (Tier 0 Cities) - Cluster Label 0 - Low Population, High Sex Ratio, Low Growth Rate, Low Literacy Rate, Low Schools and Teachers, Low Venue Counts
* (Tier 1 Cities) - Cluster Label 2 - Low Population, High Sex Ratio, Low Growth Rate, High Literacy Rate, Low Schools and Teachers, Low Venue Counts
* (Tier 2 Cities) - Cluster Label 3 - Medium Population, High Sex Ratio, Medium Growth Rate, Low Literacy Rate, Medium Schools and Teachers, Medium Venue Counts
* (Tier 3 Cities) - Cluster Label 1 - High Population, Low Sex Ratio, High Growth Rate, High Literacy Rate, High Schools and Teachers, High Venue Counts

**Visualizing the Clustered Cities using Folium**

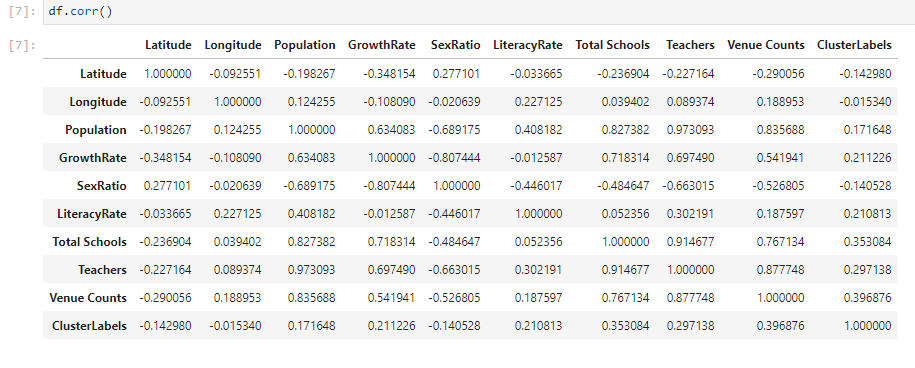
****

**5. Discussion**

**Some Observations Based on Feature Correlation Data**

.

.



1. No Feature is directly correlated to the Cluster Label
2. High Population equals Higher number of schools, teachers and venue counts
3. There is a negative correlation between High Population and SexRatio of Male:Female
4. Venue Counts has the largest correlation value for the Cluster Label

**6. Conclusion**

In this project, I grouped cities of the Gujarat State of India based on various kinds of data including Demographic Data, Education Data and Location Data. I used the K-Means algorithm as part of this clustering study.

The Latitude, Longitude Co-ordinates for the cities obtained from using the Geocoder library in Python were not accurate for some cities. This led to the Foursquare API providing inaccurate venue counts for some cities.

The Foursquare API also set a limit of 100 venues per city which further decreased the accuracy of the data used.

The venue counts were just one feature among the seven features used to cluster the data and might not have had as much of an influence on the final cluster labels as required for this project.

The city clusters met most of my personal perceptions as highly developed cities were placed in one cluster whereas the not-so-developed cities were grouped together in a single cluster.

**Ahmedabad and Surat are the best cities of Gujarat to immigrate to.**

**7. Future Scope**

More relevant features can be used to cluster the cities and more features that include economic data and immigration data may be included.