An IBM Data Science Professional Certificate

And

Personal

Capstone Project

Date: 16 September, 2020

Clustering 492 Indian Cities Based on Several Stats

Santanu Sikder

Clustering 492 Indian Cities Based on Several Stats

By ***Santanu Sikder***

*This is the **Report** section of the Capstone Project of the **IBM Data Science Professional Certificate Specialization**.*

It includes the complete information regarding the final data science project performed.

Contents

Clustering 492 Indian Cities Based on Several Stats

By ***Santanu Sikder***

1.) INTRODUCTION

a) Project Description

This project involves the analysis of 492 most prominent Indian cities for obtaining several inferences regarding the quality of living in developed and in underdeveloped cities.

Several stats available from the initial dataset have been used together with venue data to cluster the cities into two major clusters and hence the differences in the state of living was observed.

I, Santanu Sikder, have been the one and only person associated with this project and have performed all the data collection, transformation, analysis, visualization, etc.

My primary concern in this data science project has been population and how it gets affected from development and how does it affect it.

b) Importance of This Project

This project can draw some really very important observations about the Indian cities and their population, facilities, state of living, etc.

These observations can be used to improve public services as well as take significant decisions to uplift the underdeveloped cities.

Using the maps generated during this project, spotting out such types of cities becomes easier.

2.) DATA

The main statistical data about 492 Indian cities used in this project has been obtained from a CSV-format dataset which I came across during an internet search: *492-indian-cities-dataset.csv*

After some initial processing and cleaning, it was saved into another CSV file: *cleaned-492-indian-cities-dataset.csv*

Since this data was not enough for the purpose of this project, I used the Foursquare location data to fetch information regarding the types of venues within 20kms from the central coordinates of each city. The complete venue data has been saved into a JSON file: *venue-data-492-indian-cities-within-20km.json*

I extracted the various categories of venues found within the specified range from the above mentioned venue data and divided them amongst five major categories of venues. This major categories dataset along with a TOTAL column for the 492 cities has been saved into a CSV file: *major-categories-venue-data-492-indian-cities-within-20km.csv*

Finally, the statistical dataset and the major categories venue dataset for the 492 cities were joined together into a new dataset, which was the one used for analysis, visualization and building the ML model. This dataset was saved into the last CSV file of this project: *complete-major-categories-venue-data-492-indian-cities-within-20km.csv*

3.) METHODOLOGY

a) Tools Used

The platform used to carry out the complete data science process is Jupyter Lab (Dark themed, because I love it), so a big thanks to its developers.

The only programming language used is Python, as I am very familiar and comfortable with it.

The Python modules I used are:

i. pandas for data loading, processing, cleaning, saving, transforming and also visualization using its integration with matplotlib
ii. numpy for processing a histogram of the population
iii. matplotlib for adding titles, labels and ticks to the plot
iv. folium for generating maps and add markers for various cities
v. requests for using the Foursquare API
vi. json for converting JSON data into Python dictionaries and back into JSON string (for storing into file)
vii. sklearn for standard scaling the final dataset and building the ML model

MS Word 2013 is being used to generate this report and MS PowerPoint 2013 will be used to create the presentation.

b) Data Cleaning

The following procedure was followed for cleaning up the initial dataset (that containing the statistical data):

i. Removing the whitespace at the end of the name of each city

ii. Obtaining the female-to-male ratios (in terms of every thousand) of literacy, literacy rates and graduate populations into three new columns and removing the only male and only female columns (total 6 columns)

iii. Removing unnecessary columns corresponding to only male and only female populations under various age categories and the column corresponding to the district codes of the cities

iv. Converting the state names to title case

v. Separating out the latitudes and longitudes of the cities into different float data type columns from the string data type column of coordinates (location) and removing the location column

vi. Setting the column corresponding to the names of the cities as the index column of the dataframe

The following procedure was followed to clean the venue data:

i. After extracting the venue data from Foursquare into a JSON file, a dataframe was created and columns corresponding to different venue categories were added

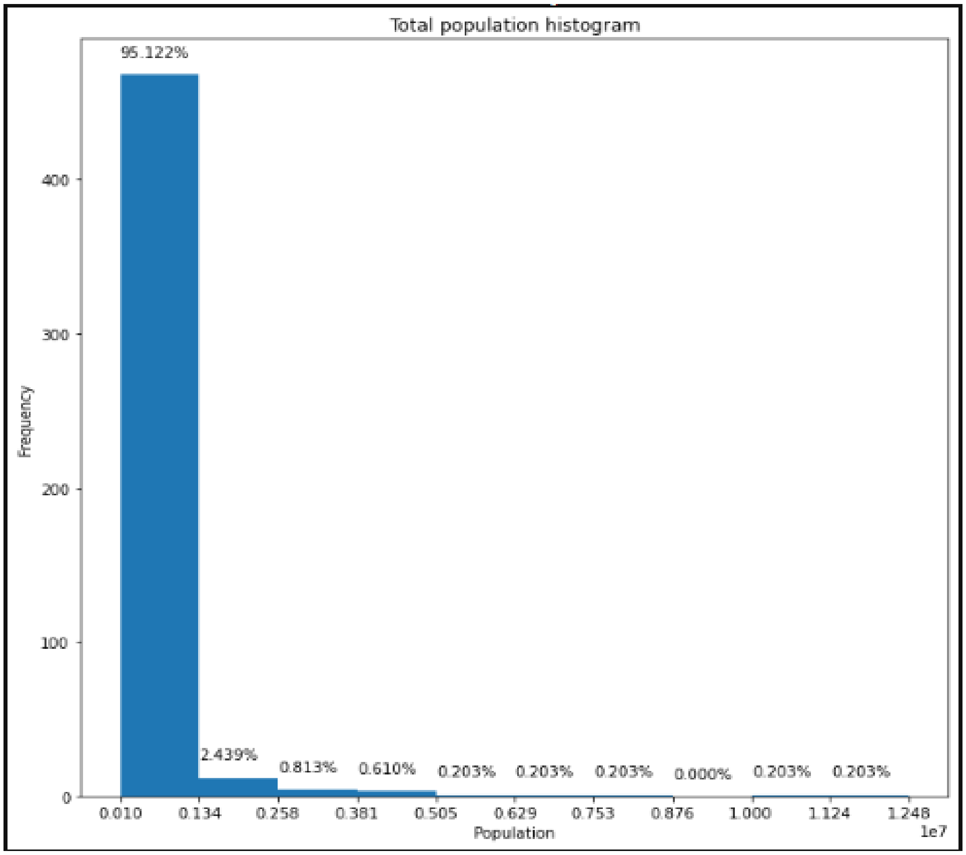
ii. After the above step was over, all those categories were grouped into 5 major categories by addition

iii. Finally a column for total venues per city was added

The statistical and venue datasets were joined to create the main one.

c) EDA

A very simple exploratory data analysis was done just to check the population distribution via a histogram:



This clearly tells that most of the cities (95 %+) have the lowest population amongst all the 492, i.e., between 0.1 million and 1.34 million or 1 lakh and 13.4 lakhs. This result will be helpful during post modelling analysis.

Clustering 492 Indian Cities Based on Several Stats

*By **Santanu Sikder***

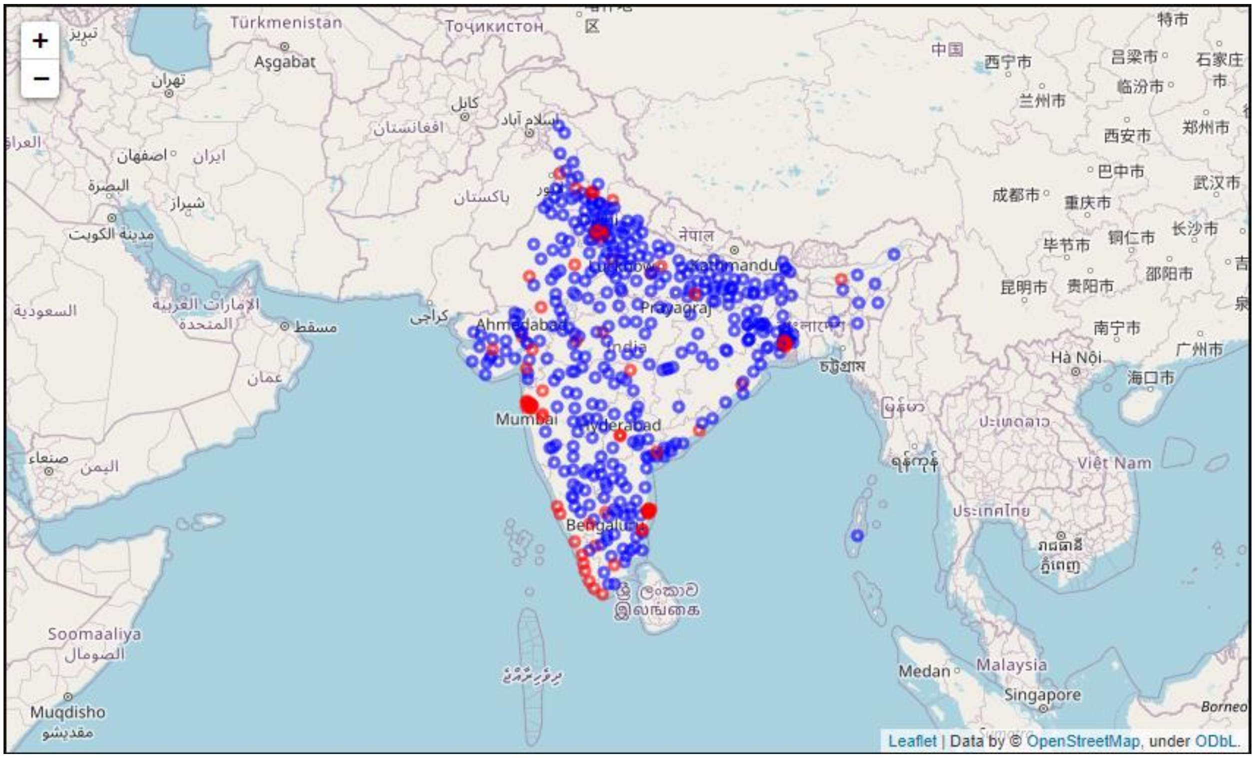
d) Machine Learning Operations

I have used clustering to segregate the cities into developed and underdeveloped cities. K-Means clustering was probably a very good choice for this as the data involved was also positional.

I used StandardScaler class for standard scaling before training.

I used Scikit Learn's KMeans class to run the KMeans clustering operation with 2 set as the number of clusters to create (n_clusters = 2). This operation was run 12 times (n_init = 12) to maximize the chances of getting the best clustering possible. I used the 'k-means++' (init = "k-means++"), which gave a very good output as the centroids were calculated smartly by the clustering engine.

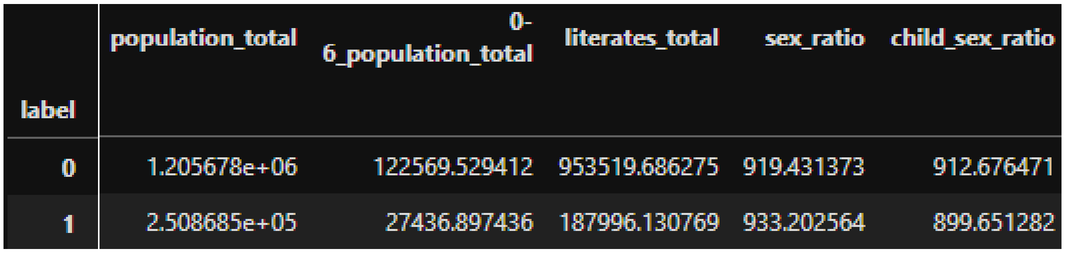
This is the map created after clustering all the location points (Map - 2):



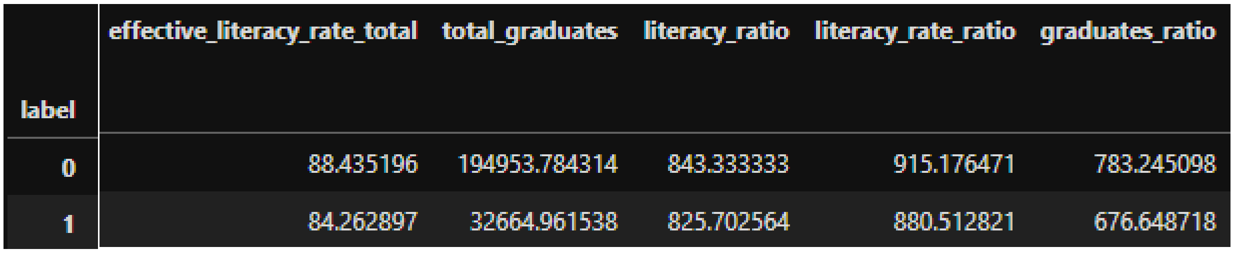
4.) RESULTS

Note: In the following content, Red and Blue markers respectively denote the WELL DEVELOPED (label 0) and the LESS DEVELOPED/UNDERDEVELOPED (label 1) cities as of the final run of the clustering algorithm. So any further run of the notebook might result in completely interchanged results, but the ultimate idea would remain more or less the same regarding the insights drawn about the two different categories of cities.

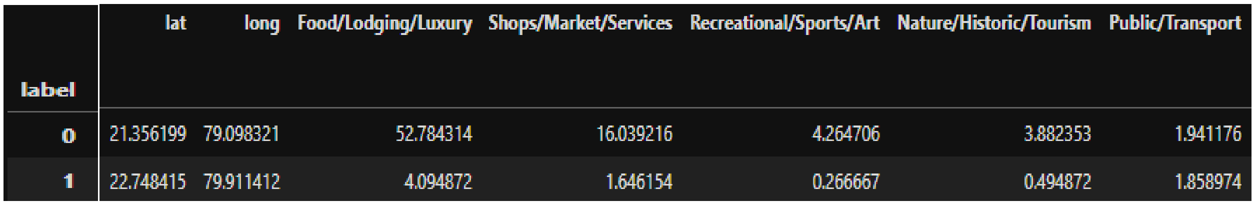
Apart from the population distribution, the following average-based observations were made:



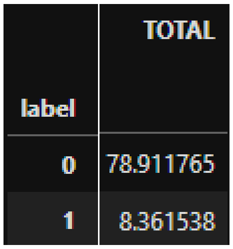
label	population_total	0-6_population_total	literates_total	sex_ratio	child_sex_ratio
0	1.205678e+06	122569.529412	953519.686275	919.431373	912.676471
1	2.508685e+05	27436.897436	187996.130769	933.202564	899.651282



label	effective_literacy_rate_total	total_graduates	literacy_ratio	literacy_rate_ratio	graduates_ratio
0	88.435196	194953.784314	843.333333	915.176471	783.245098
1	84.262897	32664.961538	825.702564	880.512821	676.648718



label	lat	long	Food/Lodging/Luxury	Shops/Market/Services	Recreational/Sports/Art	Nature/Historic/Tourism	Public/Transport
0	21.356199	79.098321	52.784314	16.039216	4.264706	3.882353	1.941176
1	22.748415	79.911412	4.094872	1.646154	0.266667	0.494872	1.858974

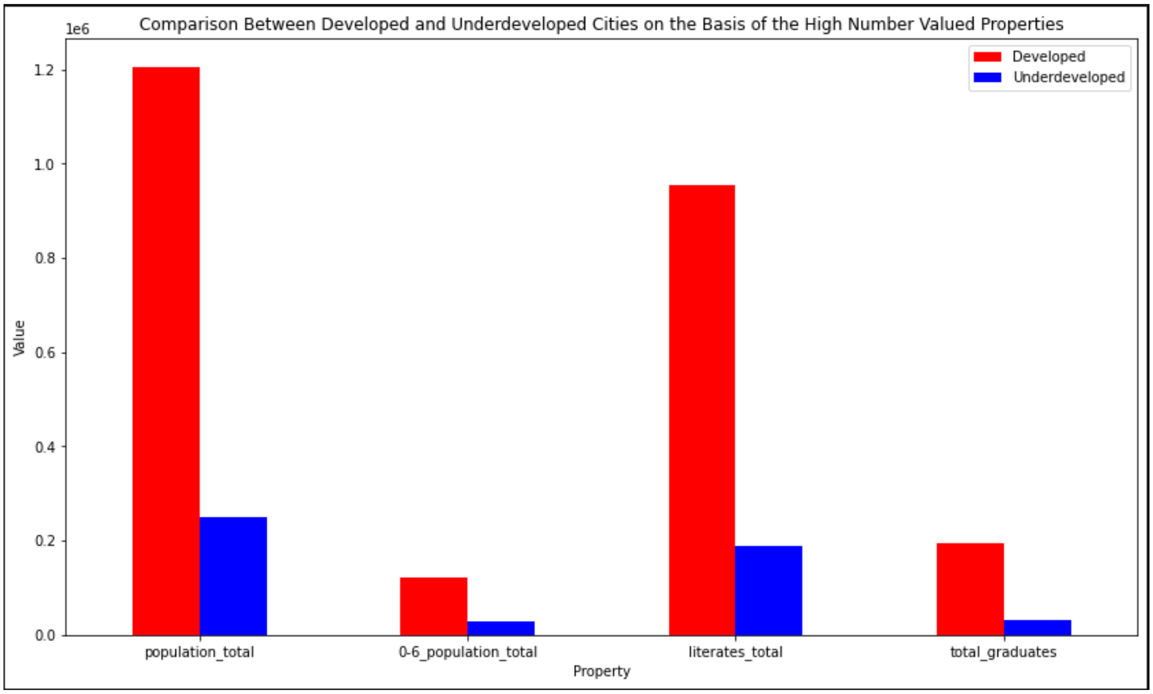


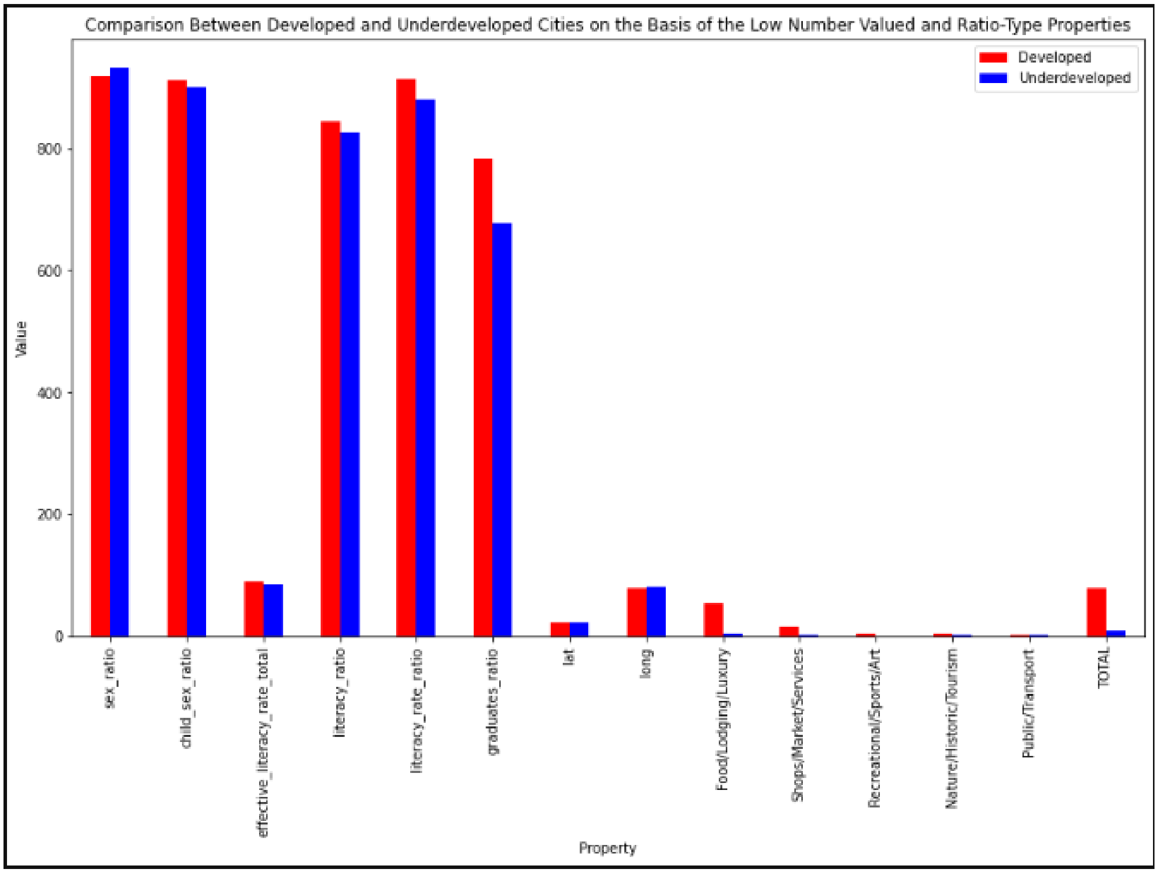
	TOTAL
label	
0	78.911765
1	8.361538

The summary to the above tabular data:

i. The Red cities have almost 5 times the average population as that of the Blue ones
ii. The average population of kids between 0 to 6 years of age is almost 5 times in case of the Red cities to that in the case of the Blue ones
iii. The Red cities have a very large population average of literates (more than 5 times as that of the Blue ones)
iv. The average sex ratios, child sex ratios, literacy rates, literacy rate ratios, literacy ratios, graduates ratios, geographical coordinates and public/transport facilities of both the classes of cities are almost similar
v. The average number of graduates in the Red cities is around 6 times that of the Blue cities
vi. In most of the classes of facilities, the Red cities are greatly ahead of the Blue ones. In fact, the Red cities even win the competition of the average all-round facilities provided

The above results have been visualized in the bar charts given below and they tell that the Red cities and the Blue cities in this case are respectively the well developed and the less developed cities. Map – 1 was modified to give Map – 2 (shown above) in this process.


Comparison Between Developed and Underdeveloped Cities on the Basis of the High Number Valued Properties
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Value
Developed
Underdeveloped
population_total
0-6_population_total
literates_total
total_graduates
Property


Comparison Between Developed and Underdeveloped Cities on the Basis of the Low Number Valued and Ratio-Type Properties
800
600
400
200
0
Value
Developed
Underdeveloped
sex_ratio
child_sex_ratio
effective_literacy_rate_total
literacy_ratio
literacy_rate_ratio
graduates_ratio
lat
long
Food/Lodging/Luxury
Shops/Market/Services
Recreational/Sports/Art
Nature/Historic/Tourism
Public/Transport
TOTAL
Property

5.) DISCUSSIONS

a) Observations

The following is all the observations made in summary:

i. A very high proportion of cities have an average population between 1 lakh (0.1 million) and 13.4 lakhs (1.34 million)
ii. A very high proportion of Blue cities can be seen on the map
iii. In the histogram, it is clear that the range 0.1 million to 1.34 million is the lowest one
iv. The Blue cities have less average number of literates and graduates, and also less average availability of facilities
v. Both the cities stand almost at the same spot in terms of the averages of the various kinds of ratios

b) Recommendations

The industries are advised to maintain their stake in the Red areas as they have an abundance of resources available and also plenty of human resource. Keeping in mind the population of youngsters, it is recommended to the cooperatives as well as the private service associations and businesses to set up appropriate and quality educational institutions.

It is also recommended to the public service providers to empower the Blue areas by improving their exposure to facilities and connectivity to the Red areas.

6.) CONCLUSION

This report concludes that amongst the prominent cities in India, there are a few very well developed ones. Although there is a higher population, higher literacy and better facilities in those well-developed cities, the female-to-male ratios in various aspects are more or less the same indicating the equal state of living of the females in both types of places.

Also, there is no specific state or region in India where a particular type of cities are heavily concentrated; all such cities are equally spread across India around center.