|  |  |
| --- | --- |
| A picture of a winding road and trees  India Growth Analysis through clustering | Abstract  What Clustering Based on Foursquare Neighbourhood Data Can Tell Us About India Growth Story?  Pranal Dongare  IBM Data Science Capstone Project |

Table of Contents

[Introduction 2](#_Toc534550223)

[Data 2](#_Toc534550224)

[Step 1 : Getting the co-ordinate details of Indian Cities 2](#_Toc534550225)

[Step 2 : Ranked List of Top 110 Indian Cities Based of Government of India’s Livability Index 2](#_Toc534550226)

[Step 3 : Data Cleanup and Manipulation 3](#_Toc534550227)

[Step 4 : Retrieval of Foursquare Neighbourhood Data 3](#_Toc534550228)

[Methodology 3](#_Toc534550229)

[Step 1 : Data Cleanup 3](#_Toc534550230)

[Step 2 : Foursqaure API call to get the venue details for the selected cities 4](#_Toc534550231)

[Step 3 : K-Means Cluster 4](#_Toc534550232)

[Discussion: 5](#_Toc534550233)

[Conclusion: 5](#_Toc534550234)

[References 6](#_Toc534550235)

## Introduction

India (Bharat, in local terms) is one of the fastest growing major, emerging economy in the world. It is pegged by many analyst to grow at more rapid pace (In Percentage Terms) than any other major economy in near future. This along with beneficial demographic dividend makes India great option for investment.

However, given it’s large size and diversity, it will be interesting to see how much different cities in India are similar or dissimilar to each other. This will also be very interesting analysis as we might get few insights about India’s growth story along the way.

Hence, in this project, I will pursue to segment major Indian cities on the basis of their Foursquare neighbourhood data and then try to identify any trends or insights from the segmentation. We might not get any particular insights after the exercise but I am sure journey is worth taking.

So, I invite you to start this journey with me! Let’s go!

## Data

Data is new oil! Of course! Let me explain the data which I will be using in this project.

### Step 1 : Getting the co-ordinate details of Indian Cities

First of all, the complete list of major Indian Cities along with details of their latitude and longitude information is created in excel file. I have referred below Github repository to create the information.

<https://gist.github.com/gsivaprabu/5336570>

The sample data from the excel sheet “Indian Cities Coordinates” is something like below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **City** | **Latitude** | **Longitude** | **State** |
| 1 | Port Blair | 11.67 N | 92.76 E | Andaman and Nicobar Islands |
| 2 | Adilabad | 19.68 N | 78.53 E | Andhra Pradesh |
| 3 | Adoni | 15.63 N | 77.28 E | Andhra Pradesh |
| 4 | Alwal | 17.50 N | 78.54 E | Andhra Pradesh |
| 5 | Anakapalle | 17.69 N | 83.00 E | Andhra Pradesh |

The link to entire dataset is [this](Indian%20Cities%20Coordinates.xlsx).

### Step 2 : Ranked List of Top 110 Indian Cities Based of Government of India’s Livability Index

Next step was to get the list of top 110 cities in India as per the Government of India releases “City Livability Index 2018”. The list was found at below location and was converted into Excel file.

<https://indianexpress.com/article/india/full-list-of-indias-ease-of-living-index-see-where-your-city-is-ranked-5306185/>

The sample data from the excel sheet “Indian Cities Livability Index” is below:

|  |  |
| --- | --- |
| **No** | **City** |
| 1 | Pune |
| 2 | Navi Mumbai |
| 3 | Greater Mumbai |
| 4 | Tirupati |
| 5 | Chandigarh |

The entire data can be found [here](Indian%20Cities%20Livability%20Index.xlsx).

A side note : Pune, my home city, has been ranked first on the list! ☺

### Step 3 : Data Cleanup and Manipulation

Now, comes the exciting part. We will start with these two datasets as base and get to the Python and Pandas for further data manipulation.

As a cleanup, we will cleanup table containing latitude and longitude details and then left join two tables on “Indian Cities Livability Index”.

This will be our input table to get the “Foursquare” neighbourhood data for the cities.

### Step 4 : Retrieval of Foursquare Neighbourhood Data

Once we have “Foursquare” data, we will cluster the cities based on the closeness to each other. For example, cities having identical weightage for features like “Pubs” or “Airport Lounge” would be clustered together.

Once we have this data ready. We can then proceed with the K-Means clustering and analysis with the help of Folium chart.

## Methodology

Step 1 : Data Cleanup

As mentioned above, “Left Join” on two tables was done to ensure that only those cities which are present in both the dataframes are used for further analysis.

Below is the code snippet for the same:

*merged\_indian\_cities\_df = pd.merge(indian\_cities\_initial\_df,indian\_livable\_cities\_df,on='City')*

After above step out of 108 original cities 80 were left for usage for further analysis.

During this step, we also ensured that we didn’t have any NaN data present. Code snippet and output for the same is below:

*Code : print (merged\_indian\_cities\_df.count())*

Output : *City 80*

*Latitude 80*

*Longitude 80*

*dtype: int64*

### Step 2 : Foursqaure API call to get the venue details for the selected cities

During this step we called the Foursquare API used in previous assignment. This was used to retrieve the information on the venues and top venues for each of the cities in the dataframe. Out of 80 cities, we received data for 65 cities and only these cities were kept in the dataframe for further processing.

The code snippent is below:

***Code*** *: city\_venues = getNearbyVenues(names=merged\_indian\_cities\_df['City'],*

*latitudes=merged\_indian\_cities\_df['Latitude'],*

*longitudes=merged\_indian\_cities\_df['Longitude']*

*)*

***Output :*** *City City Latitude City Longitude \*

*0 Hyderabad 17.4 78.48*

*1 Hyderabad 17.4 78.48*

*2 Hyderabad 17.4 78.48*

*3 Hyderabad 17.4 78.48*

*4 Hyderabad 17.4 78.48*

*Venue Venue Latitude Venue Longitude \*

*0 Cafe Bahar 17.399595 78.478566*

*1 Hotel Megacity 17.401133 78.475990*

*2 Rahul tiffins & fast food centre 17.403185 78.480070*

*3 Bajrang Pan Shop 17.402279 78.477165*

*4 Chutneys 17.404134 78.481566*

***Venue Category (This is column of interest)***

***0 Indian Restaurant***

***1 Hotel Bar***

***2 Breakfast Spot***

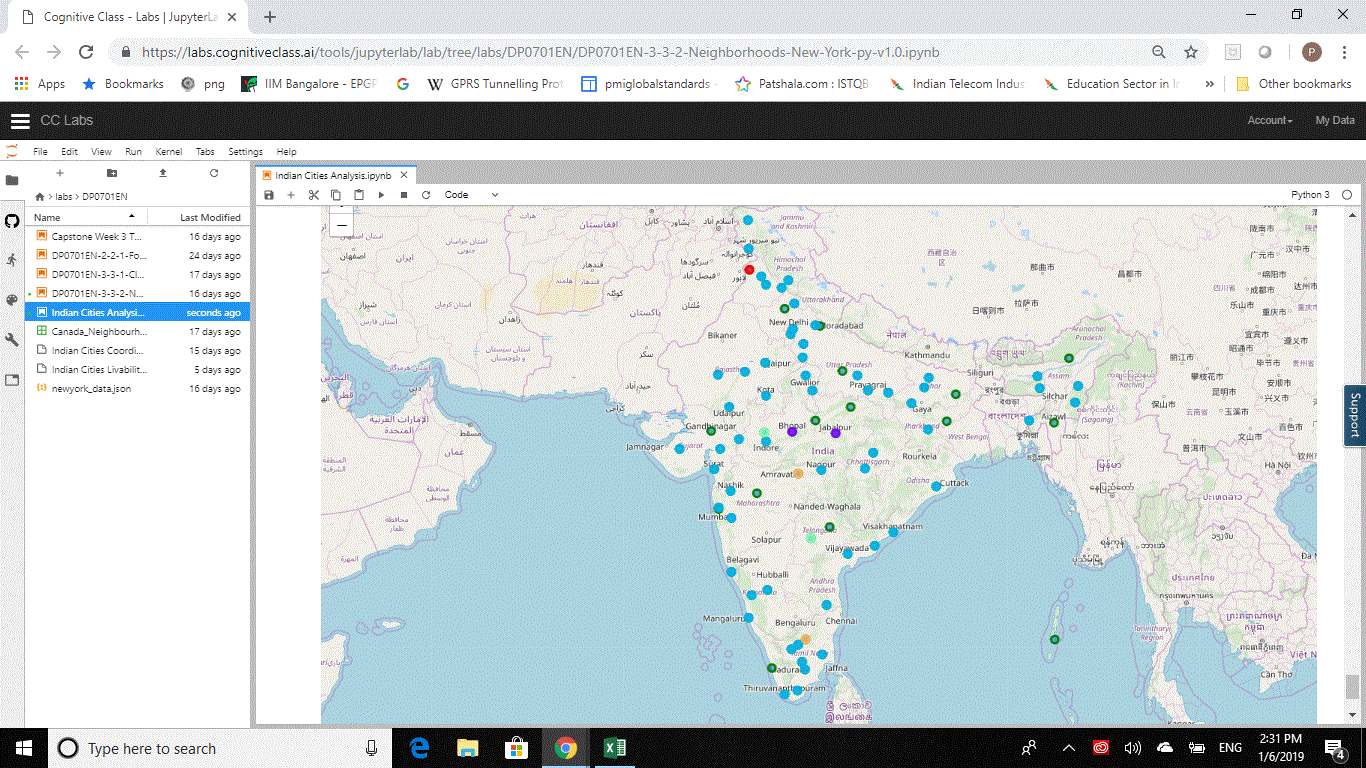
***3 Smoke Shop***

***4 Vegetarian / Vegan Restaurant***

### Step 3 : K-Means Cluster

We used K-Means clustering machine learning algorithm to identify the similarity and dissimilarity amongst the 65 Indian cites.

The output of the K-Menas clustering is below:



## Discussion:

From the above exercise, we can observe that most of the Indian cities lie in the same cluster (Marked by Blue). The spread is also not skewed and is fairly distributed across the Indian territory.

## Conclusion:

In conclusion, we can say that at least as per data captured by “Foursquare”, Indian growth is homogeneous across the country and we see no major differences in any particular geographical areas of country. Ideally we would expect cities on West and South of the country being clustered together (Indicating more developed category) than do cities to the East and North.

However, we should be aware that this analysis is based **only** on the data collected from “Foursquare” venue details, reviews and category of the venue. We should expect significant change in the analysis once we start adding other datapoints like, Number of Industries, IT Companies, Wages, Electrification, Crime data and further.

One another important conclusion which can be drawn from the analysis is that given equal chance and infrastructure to grow, citizens across India show inclination towards similar sort of venues. Hence regardless of diversity of culture and language, at the aspirational level, there is homogeneity in the wide and large country like India. This can help policy makers better informed decisions.

## References

1. Latitude and Longitude Data : <https://gist.github.com/gsivaprabu/5336570>
2. Indian Cities Livability Index : <https://indianexpress.com/article/india/full-list-of-indias-ease-of-living-index-see-where-your-city-is-ranked-5306185/>