Indian Cities And What Different People Could Get From Them

IBM Data Science Capstone Project

NAMAN RASTOGI

MAY 29, 2020

Introduction

Background

India is a developing country with a very diverse group of people. Along with the state in which it is in development, population is also a major factor of how different categories of people tend to take decisions concerning various situations from the decision of finding a good place/city to live pertaining to their needs, to the decision of how a particular city might affect a certain business or a company. In a developing country like India, these decisions become very crucial in transforming the lives of people and the overall growth of country as a whole.

Problems / Questions To Consider

Different categories of people like normal people who might want to transit to a different city based on how the other city is similar to another, which business could be more profitable in a particular city or how the city's current state might affect one's decision or business, view the city or a region with different perspectives for obvious reasons. So, it becomes important for them to have a clear idea of how a particular city is ideally suited for their needs. It becomes important to know as to how the overall cities of India are distributed in terms of development and demands pertaining to different groups of people and how they can fulfil their respective needs by knowing the current state of where these cities stand and potentials that these cities have for satisfying their needs. So, to reveal the insights for such people seeking answers, this project takes into account the venues of the city as a way of identifying patterns of how certain cities are similar to one another and what their distribution is in terms of various factors that have impact on people's decisions. This project in a way, also lays emphasis on how a particular city may be lacking in terms of a particular aspect, because this also plays an important role in getting important insights of their characteristics.

Audience

The primary audience of this investigational project might include, people who want to take decision to make a transit to another city, depending on how that another city is similar to his/her current city of residence, or how the another city may be better in terms of standard

of living and various opportunities it might have to offer, or businessmen who want to know how a particular city could be potentially an ideal centre for a certain business to be pursued there.

This project as a whole, based on how various cities are different and or similar to one another, provides recommendations to a supposedly bigger audience based on what their perspectives and needs are.

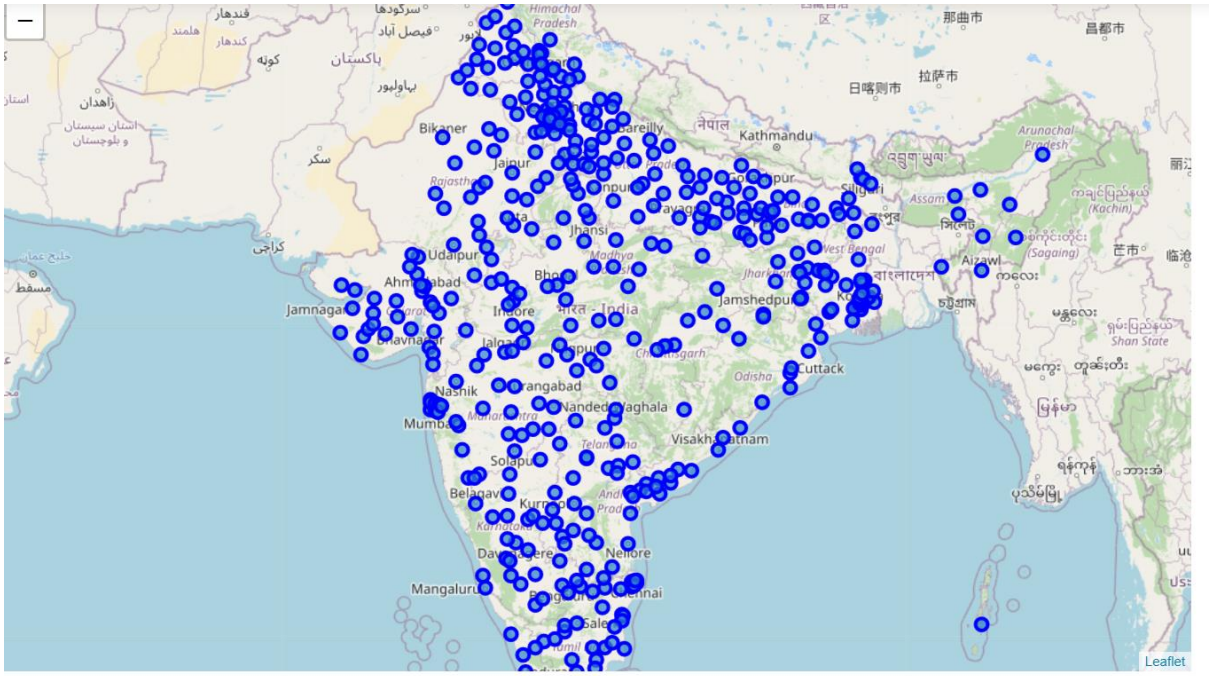
Data

The dataset required for this project and analysis has been gathered from Kaggle and it is curated by merging the census 2011 of Indian Cities with Population more than 1 Lac and City wise number of Graduates from the Census 2011, to create a visualization of where the future cities of India stands today. It lists the top 500 Indian cities.

Methodology

Preparation / Pre-processing

In this stage of my project, I gathered the dataset from the mentioned source and queried it for the names of the cities and their geospatial coordinates (which was provided in that dataset). Then I extracted the coordinates of India using the geopy library of python, and this process is termed as geocoding. Then using all such information, I mapped out all the cities of my dataset.



Then, using the Foursquare API, I queried for the top venues (according to the user ratings) for each city of my dataframe. Then after extracting the category of a venue, I moved onto to clean the json to restructure my dataframe. So, in this way I got a dataframe which contained venue category for each city that the original dataset contained. Foursquare API returned in total of 1163 venues with

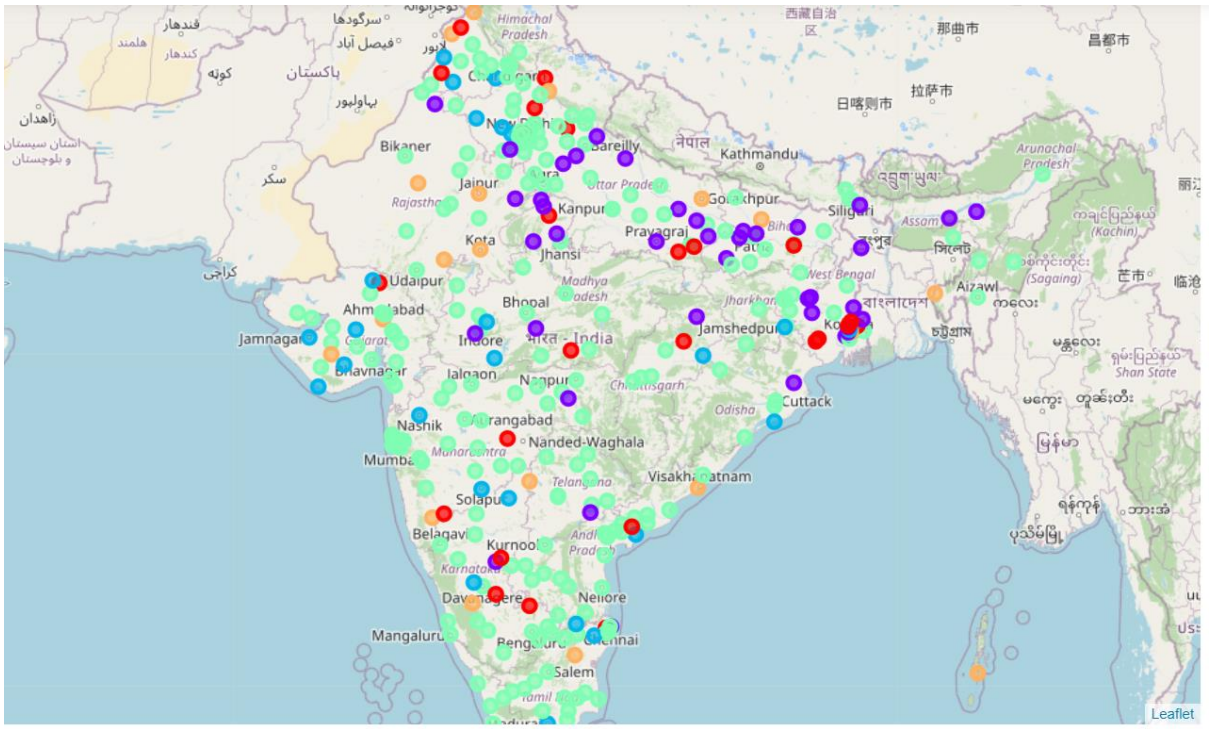
185 unique venue categories. For the purpose of analysis, I used top 5 venues for each city as a defining characteristic to get a measure of similarity or difference between the cities.

Analysing The Cities

To analyse the data, I used a popular unsupervised machine learning algorithm called k-means clustering to partition observations into a specified number of clusters in order to discover underlying patterns. Specifically, I used the top 5 venue categories for each city (based on occurrences in the dataset) as each city's vector profile for finding similarities with other cities. The first step was to calculate the average frequency for each venue category across each city. Using a Pandas dataframe I converted each venue category into a boolean (yes/no) column using the Onehot encoding method, verifying that new dataframe's column count equalled the number of unique venue categories I identified during data preparation. Next I grouped rows by city mean of frequency for each category, and used that to find the five most common venues for each city. Then I moved on to apply K-means clustering algorithm. After trying out different k values (where k= number of clusters), I found the clusters to be most meaningful and interesting with around k=6. The output of the K-means algorithm is an array of cluster assignments for each row in the dataframe. With that I then stitched the cluster labels back into the dataframe and combined city location data in order to print out and visualize the results.

Results And Discussion

I used the folium library of python to give colors to the clusters and to map them out on the map of India. The result of generating the plot of the clusters was this:



Cluster 1: red circle

Cluster 2: purple circle

Cluster 3: blue circle

Cluster 4 : navy blue-green(most of them are of this color)

Cluster 5: light green circle

Cluster 6: Orange circle

The very first thing to notice about our analysis is that, although we have taken into account top 493 cities based on population, but the total number of unique venues that the foursquare API returned is much less when this number is viewed in terms of the number of cities that we took into account. From our clustering algorithm that we ran on our dataset, we could observe that, the cities which are falling in the cluster 4, are much more promising when it comes to the varieties that the people can get there if there are positively looking to make a transit to another city(this could also be validated from their population levels as they occur in top positions in terms of overall population). Also, the number of venues which are returned for these cities are more when compared to other ones (We could infer this from our dataframe), thus, also showing their potential to attract more people of diverse groups. These are the cities which are also the corporate hubs of India and attracts more job seekers. From the type and variety of venues, it could rightly be said that these cities would be preferred choices for people who may want a good standard of living and relatively more varieties. As far as other clusters are considered, broadly speaking, these are not very much promising for those people needing to transit to get a high standard living but at the same time potential centres for starting various business and opening more venues characterized by greater varities, possibly because they reasonably lack in terms of such factors. Cluster 5 is mainly about hotels whereas the last cluster cities are mainly concerned with pharmacy, suited for pharmacists and their relevant works and people with businesses concerned with medical/pharmacy. Cluster 2 cities are characterized by particularly restaurants. Cluster 3 cities have mainly ATMS in common showing more number of average daily digital transactions. Cluster 1 is mainly centred around more train stations. The other way of looking at this and also looking from their population levels, they are kind more in a situation where more demand for development needs to be fulfilled. And for a developing country like India, these observations seem fairly reasonable. India is a developing country and through the medium of this project, I tried to show how the overall cities of India are distributed in terms of development and demands pertaining to different groups of people and how they can fulfil their respective needs by knowing the current state of where these cities stand and potentials that the cities have for satisfying their needs. This project has a large scope of improvement because much more data could be fed into our algorithm to give a larger scale of analysis and visualization.

Conclusion

Although, I was able to fairly and reasonably address the problems and questions posed in the beginning of the report up to a certain degree of correctness, there are still areas where this project could be further worked upon and increase its scope in efficiently predicting the patterns with even more insights.

We saw that the number of venues for many cities were very less (much less compared to other developed countries), a big indicator for this could also be that many cities are in a very high demand of development (because they lack in such respect). With more data and analysis with some more important features, this could become more efficient in terms of revealing more insights of the cities.