Explore India
Business Opportunities

Data Analysis Report

Coursera Capstone Final Project

Report by: Naren Neelamegam

Introduction

India is the second-most populous country, the seventh-largest country by area and the most populous democracy in the world. The trending population is around 1.35 billion. India has 29 states and 7 union territories with around 300+ popular cities. The goal of the analysis is to analyze the popular venues with the state capitals and popular cities in India along with the population data of each city/capitals to understand their influence over trending venues and find out the best possible business category to start a business in India.

This analysis answers the following for an entrepreneur who wants to start a business in India,

- What are the trending businesses across Indian state capitals?
- What are the trending businesses across Indian cities?
- Does population of a state capital or city influence the business trends?
- Should I start a business in a state capital or any cities in India?
- Which category of business should I consider?

Data Collection and Analysis

Indian State Capitals

Capital cities data are required for analyzing popular venues across state capitals. We will get this data from Wikipedia List of state and union territory capitals in India. We will only take State and Administrative Capital column from this data set.

Indian Cities and Population data

There are around 300+ popular cities in India which we can get from Wikipedia List of cities in India by population. Along with the cities list, we also get population data from this data set.

Note: This analysis is based on 2011 census population data. Though this might have increased approximately by 17% to 20% over the decade, we go with the currently available data set for this analysis. However, the whole analysis can be reused when the new dataset is available.

Location Data

To get the venue details of cities and capitals, we will require their latitude and longitude data. We will use python's geopy library to get these details. To save time, the location details are also captured on the first run to a csv file which you can download from GitHub repository.

Venue Data

To get the venue details of cities and capitals, we will use foursquare.com apis through developer account.

Data Cleaning and Structuring

There are two primary data sources we use for collecting Indian state capitals and cities details. Both of them are from Wikipedia pages, but with different column sets. The state capitals table has no population data and both of the data tables doesn't have geo location data. Though the Wikipedia data sets have a lot of details in different columns, we are only interested in City/Capital names, State names (required for data frame joins) and population data. We will dropout rest of the columns during data cleaning process. Also, we will have to remove subscript, superscript notations used in wiki pages for reference links. Python geopy library will be used to collect geo location of each city and the final dataset will have data like the below sample,

State	City	Latitude	Longitude	Population
Andhra Pradesh	**Amaravati**	**16.516910**	**80.500259**	**103000**
Bihar	**Patna**	**25.609324**	**85.123525**	**1684222**
Chandigarh	**Chandigarh**	**30.719402**	**76.764655**	**1028667**
Delhi	**New Delhi**	**28.614179**	**77.202266**	**249998**
Gujarat	**Gandhinagar**	**23.223288**	**72.649227**	**206167**

Since state capitals data and cities data are from different Wikipedia pages, there are high chances that we will be missing or having mismatched capital city names (for population data collection). We will use data frame merge method to compare both the data frames to find out missing capital cities and manually add missing details through a csv import. With the population category added to our dataset, we will use foursquare api to collect venue details of each Indian state capitals and cities and add them to our dataset for analysis. With the foursquare data added, our dataset will look like the below sample,

Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Mumbai	**18.938771**	**72.835335**	**Royal China**	**18.938715**	**72.832933**	**Chinese Restaurant**
Mumbai	**18.938771**	**72.835335**	**Town House Cafe**	**18.938550**	**72.833464**	**Bar**
Mumbai	**18.938771**	**72.835335**	**Sher-E-Punjab**	**18.937944**	**72.837853**	**Indian Restaurant**
Mumbai	**18.938771**	**72.835335**	**Britannia & Co.**	**18.934683**	**72.840183**	**Parsi Restaurant**
Mumbai	**18.938771**	**72.835335**	**Cafe Excelsior**	**18.937701**	**72.833566**	**Café**

Data Analysis

We will start with population analysis of Indian cities and state capitals by categorizing the population into High, Medium and Low populated areas. This will give us an insight of population across Indian state capitals and cities. Folium maps will come handy to plot the population stats on India map for visual analysis. With the venue data collected using Foursquare.com apis that are merged with both the datasets of Indian state capitals and cities, we will use KMeans clustering technique to cluster popular venues against each data points (locations) and find out differences and similarities of venues against population of the location.

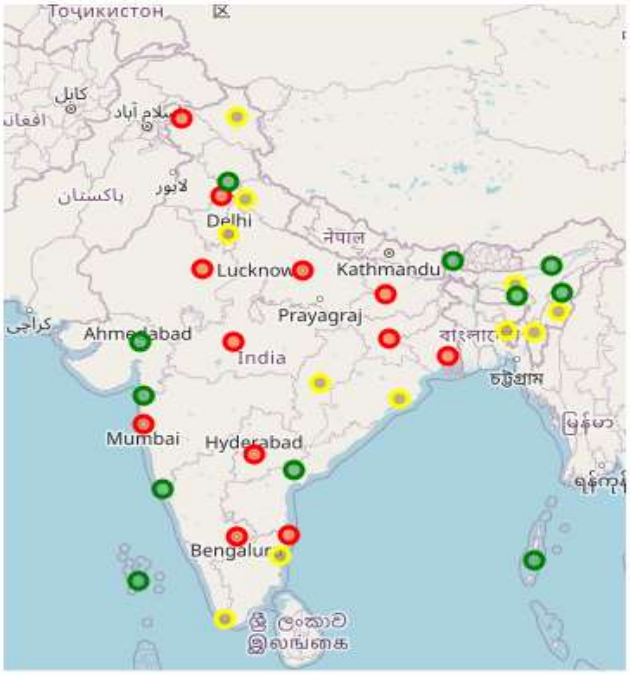
Finally, we will compare both the datasets of Indian state capitals and cities to identify which venue is more popular on both the datasets and are there any differences, similarities of venue categories between state capital cities and regular cities.

Exploratory Data Analysis

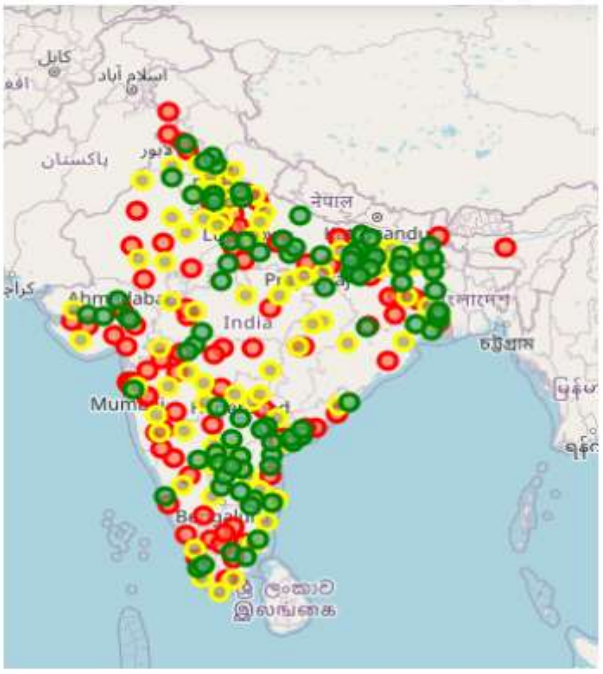
Population Data Analysis

As explained in the above section, we will start with population data analysis of Indian capital cities and other cities in India. First we will segment the population data into three categories as High, Medium and Low using data frame quanitle method with 34%, 33%, 33% partitions respectively. The result shows a common trend of population across Indian state capitals whereas a mixed share of population across other Indian cities. Color markings Red denotes high population, yellow for medium population, green for low population.

Indian State Capitals - Population

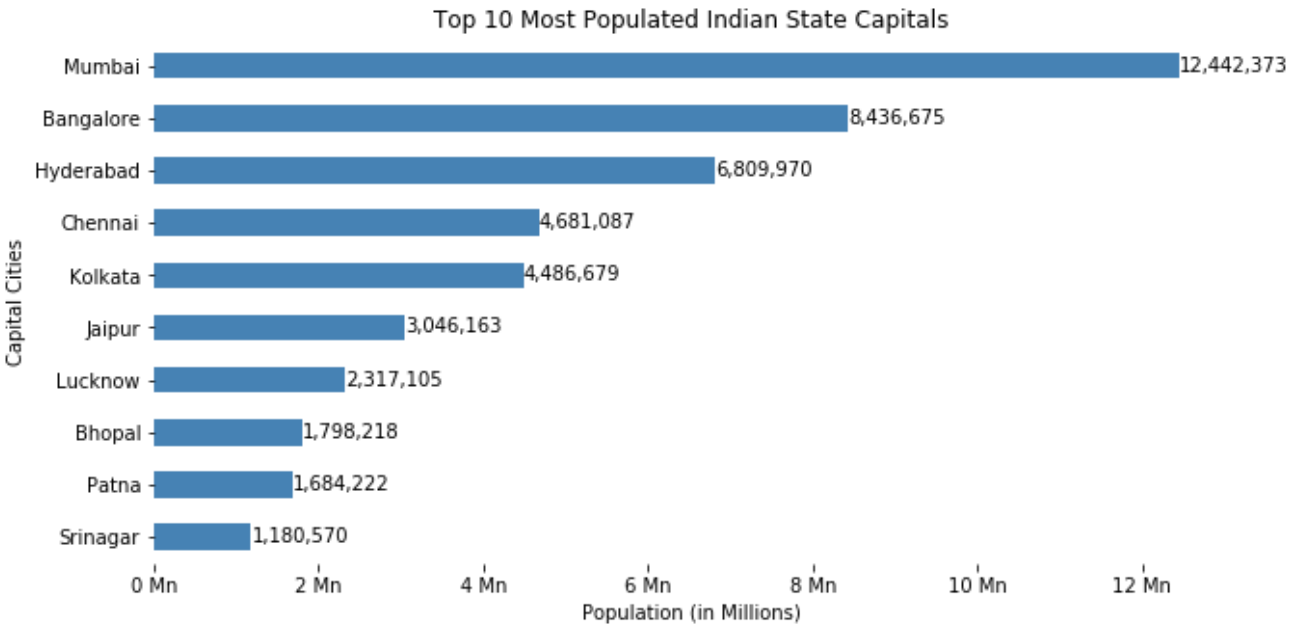


Indian Cities - Population

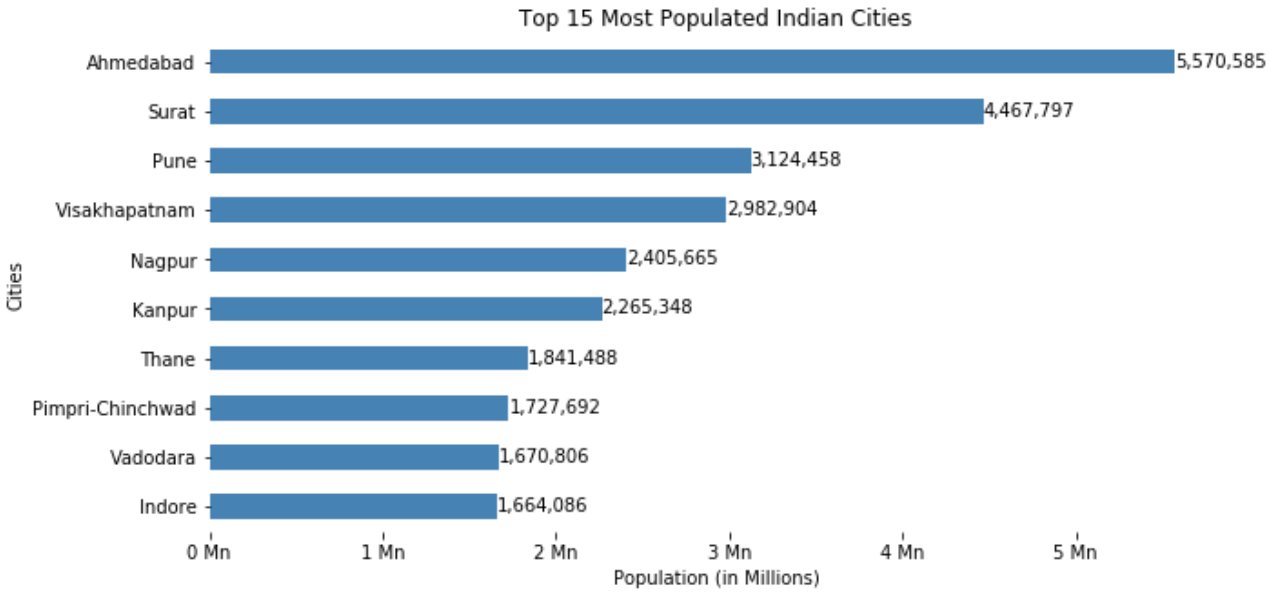


> "The result shows a common trend of population across Indian state capitals whereas a mixed share of population across other Indian cities"

Let’s have a look at the top 10 populated capital cities. India’s most popular City of Dreams - Mumbai tops with around 12 million people, followed by Silicon Valley of India - Bangalore, City of Pearls - Hyderabad and the Banking Capital of India- Chennai.

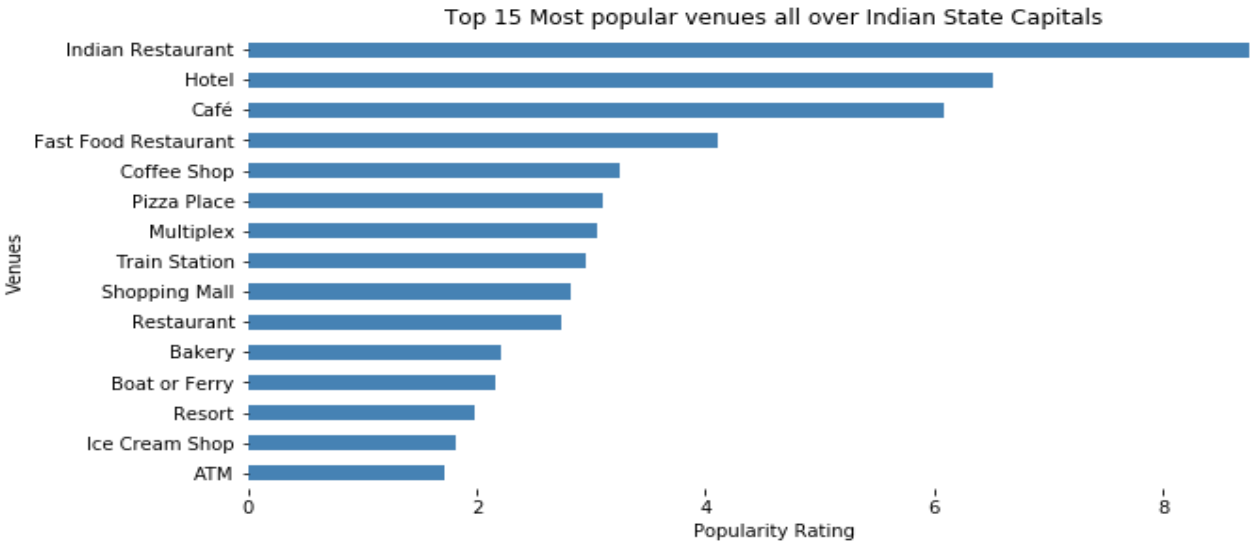


It’s not so surprising to see the above top list as those are the world popular cities that can attract more people. Let’s have a look at the other top 10 populated cities (excluding capitals) of India. The former capital of Gujarat Ahmedabad tops the list with around 5 million people, followed by diamond and textile hub Surat, Oxford of the East – Pune and City of Destiny Visakhapatnam.

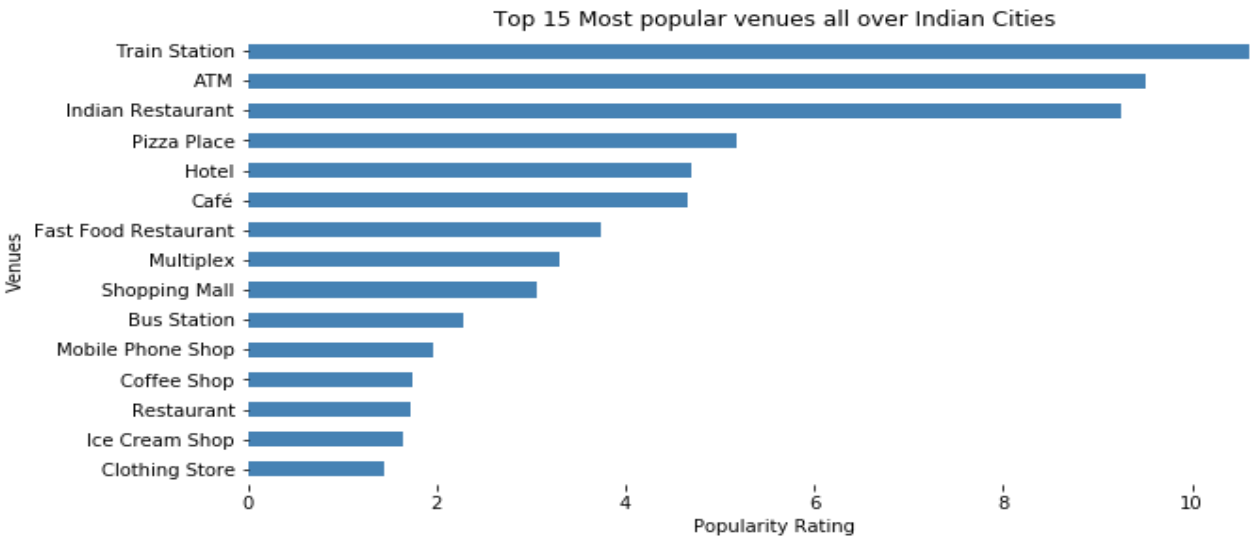


Venue Data Analysis

Our further analysis is going to be on how this population trend affects the popular venues in each of these cities. For venue analysis, we will have to collect venue details of each of these cities using foursquare.com explore api. The venue details collected for all the Indian state capitals are grouped together by the name of the city in the data frame to find the frequency of each of the venue category occurrence and the top 15 high occurrence of the venue category is plotted in following graph. The result show Indian Restaurant business is the leading player across Indian city capitals followed by Hotels and Cafes.



It's obvious to see the above top list in city capitals as it attracts more travelling people from other parts of states. Let's have a look at other cities of India (excluding capital cities),



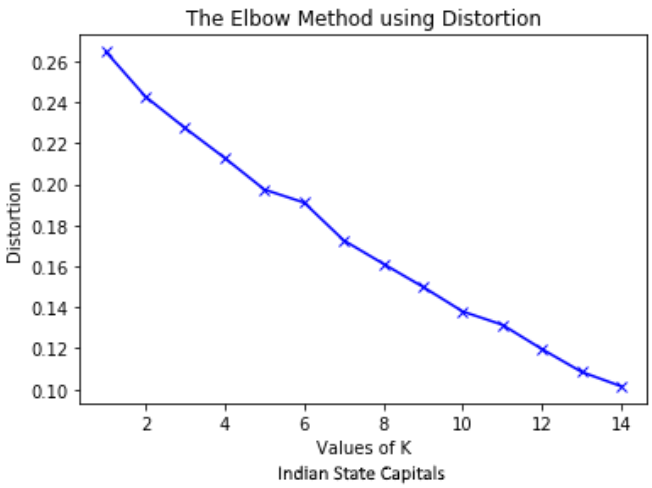
It differs a little from capital cities, however the first commercial entry in the above list starts with Indian Restaurant again, followed by Pizza Place, Hotel and Café. So, in common, other cities are not so different from capital cities from business perspective.

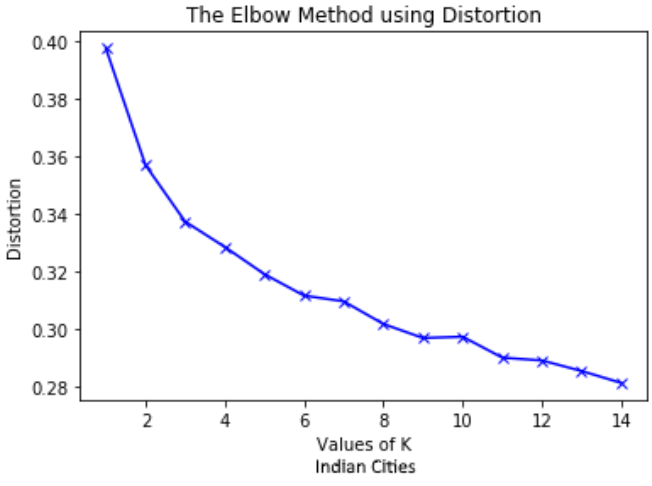
> "Indian Restaurant, Hotel and Café are most common venue categories on both Indian state capitals and other cities"

K-Means Clustering

There are many models for clustering, we will be using k-means clustering – one of the vastly used clustering model that especially useful if you need to quickly discover insights from unlabeled data. The data analysis reveals that "Indian Restaurant, Hotel and Café are the most common venue categories across Indian cities (including capital cities). In this section, we will analyze the population data influence over the venue data on both the data sets separately.

For further analysis venue data and population data are combined together and to work out K-Mean clustering on our datasets, we will need to find optimal K value for our model. We will go with Elbow method using Distortion for both capital cities and other cities.

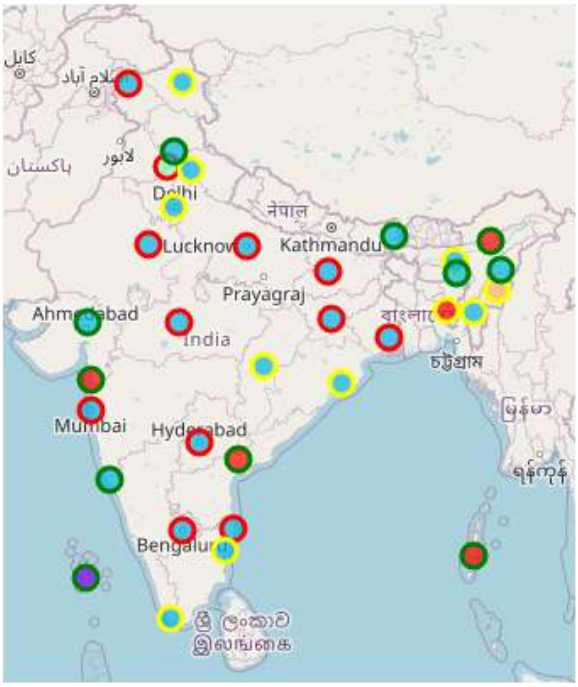




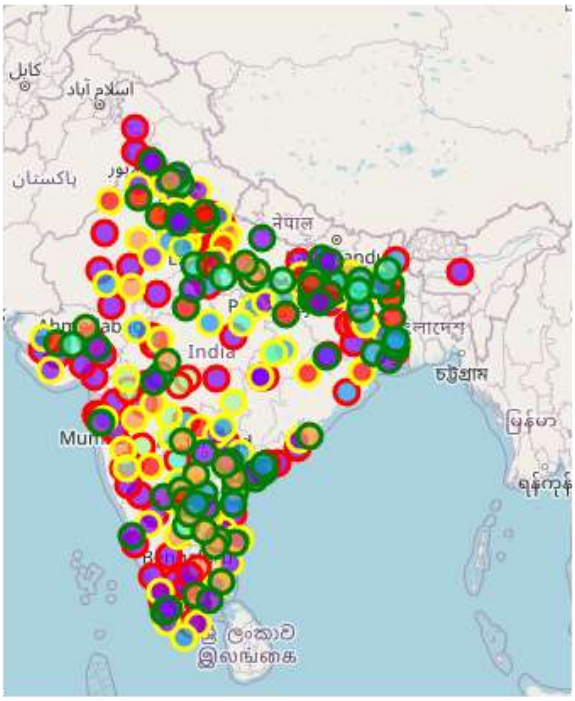
The Elbow method shows an optimal K value 5 for capital cities dataset and K value 6 for other Indian cities dataset. The following maps shows the data points after applying K-Means model clustering on

venues data which shows venue categories are mostly similar across Indian capital cities and also the most populated capital cities. However, venue cluster on Indian cities show a mixed venue trend.

Venues Clustered - Indian State Capitals

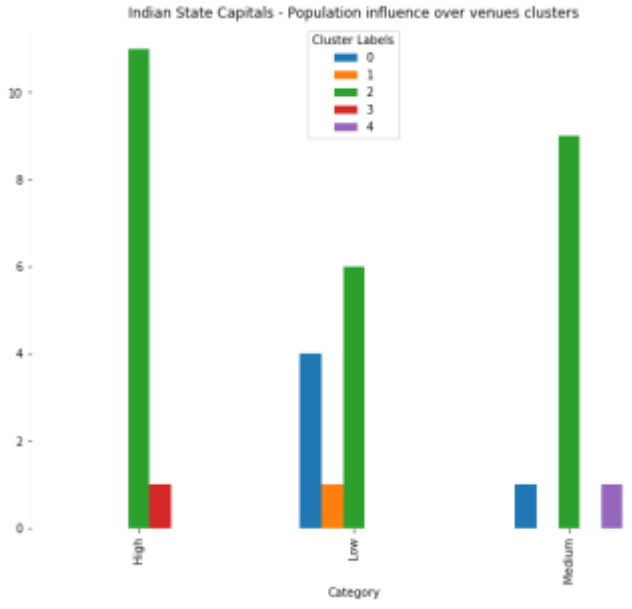


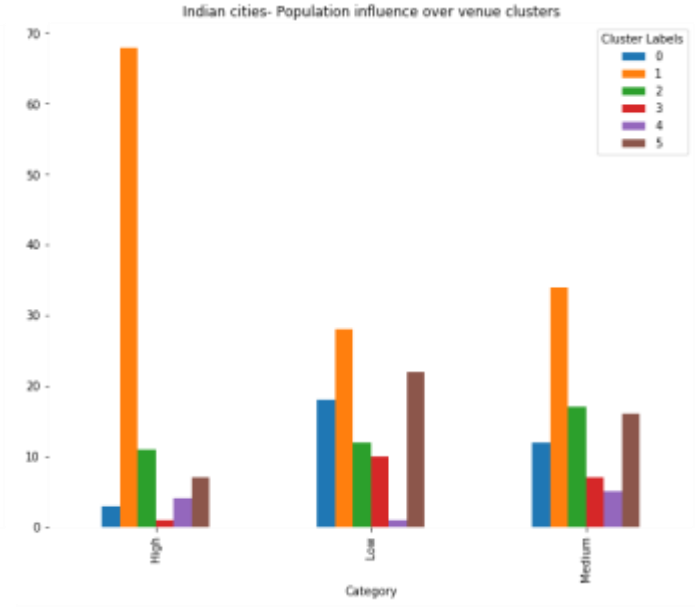
Venues Clustered - Indian Cities



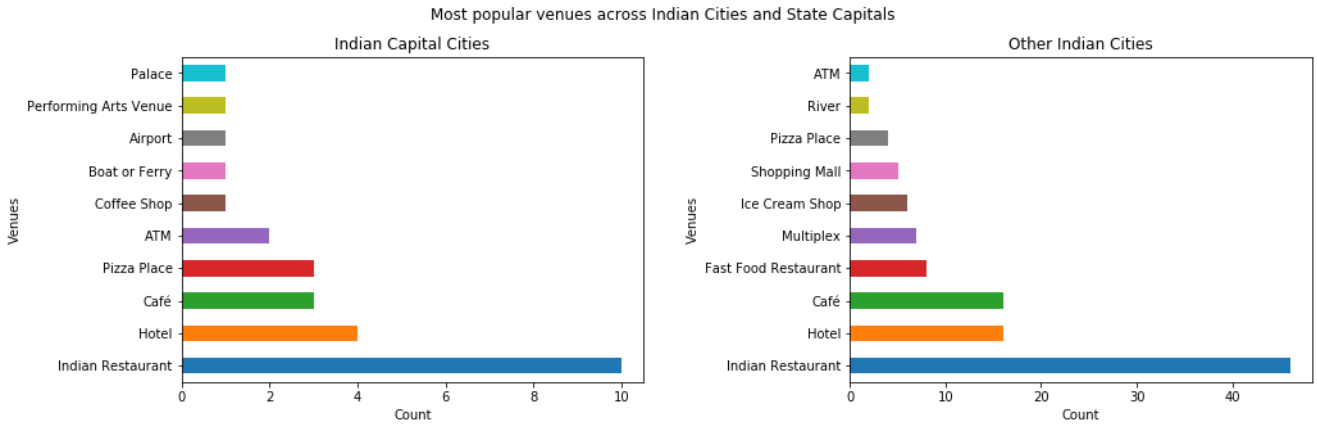
**Borders of the markers indicate population category of the city*

Too many data points on the map of Indian cities above may seems to be confusing to analyze the venue popularity. So, let's plot a simple bar graph to look into more details.





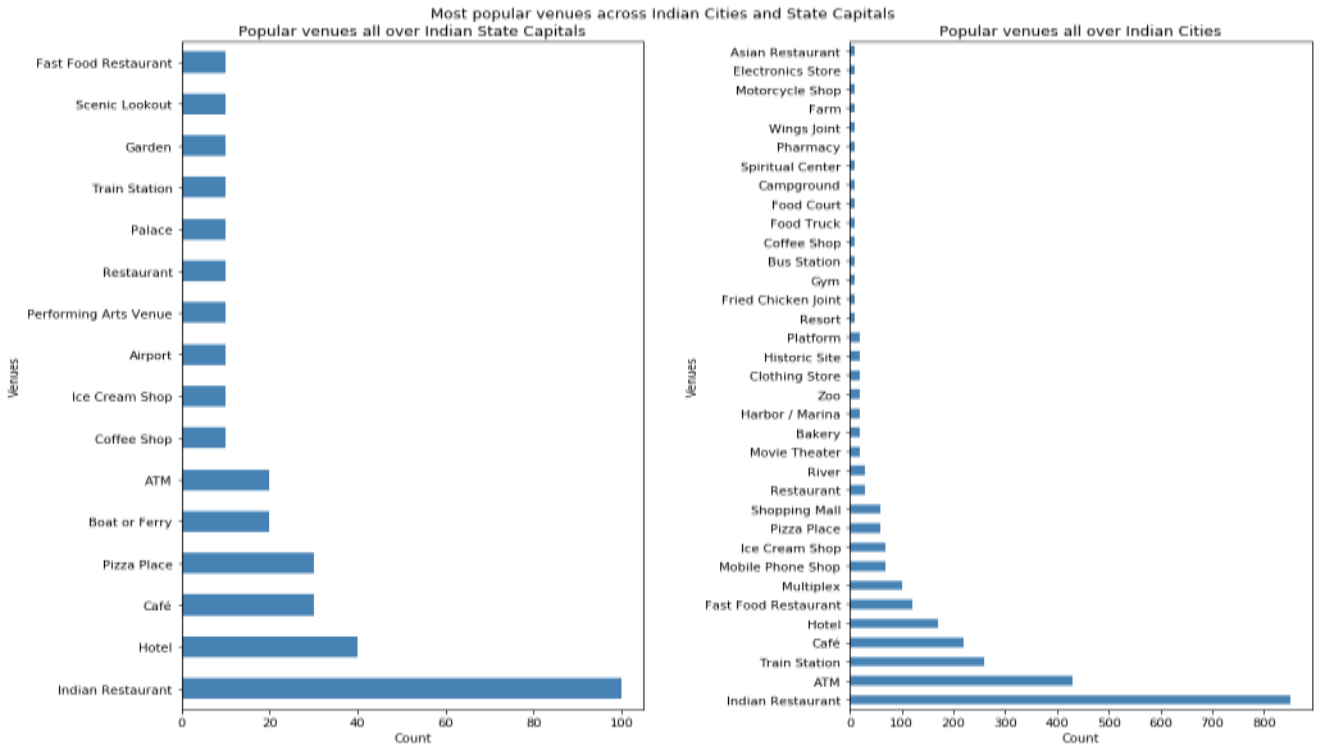
The bar graph above reveals a surprising fact that, regardless of population, cluster 2 (green bar) is common across Indian state capitals and cluster 1 (orange bar) is common across Indian cities. Let's have a look the 1st most common venues which are grouped in those clusters.



As we have seen earlier "Indian Restaurant, Hotels and Cafes" keep their top position in popular venues categories for Indian state capitals and Indian cities show, they are not so different from capital cities venue popularity.

> "Indian Restaurant, Hotel and Café are most common venue categories on both Indian state capitals and other cities regardless of population of the city"

Let's put all the top 10 popular venues together to find the frequency of each of the venue category to find out which business stands top. The below figure is the summary of all top 10 venues across cities on corresponding dataset based on the frequency of the venue occurrences. The chart reveals that "Indian Restaurant, Hotel and Café' keeps the top position in Indian state capitals and Indian cities, though looks different, considering commercial venues, it's no different from capital cities.



Conclusion

Though we have a hand full of businesses running in cities across India, the analysis shows business under “Indian Restaurant, Hotel and Café’ categories running successfully regardless of population, status of the cities. The best option to start a business could possibly be one of these top 3 categories. However, rest of the venues are not to be left out, comparing the venues across both data sets, it clearly indicates that there are more travelling people across cities (note Train Station have a position in both the datasets), hence the restaurants, hotels, cafe are leading on its position.

So facts identified for Business startups in India are,

1. Business under “Indian Restaurant, Hotel, Café” category, be it capital city or other cities
2. Population doesn’t have much impact on these top venues
3. Business close to Train stations, Airports
4. Business which deals with travelling people