Report for Capstone Project

Analysis of Cities and Population in India

Princi Jain

July 2020

Business Problem

Introduction :

- In today's era, everyone is on the move, we are constantly looking for opportunities for a successful career.
- Most of us are required to move from one city to other due to Job change.
- Even if some of us have business at a fixed place, we want to expand it to different places.

Target Audience :

- This Study will help you in determining the city of your choice where you can look for more job opportunities and yet find the similar background as your own city and can also choose state which is less populated as compared to others.
- Similarly, it can also help you identifying places where you can expand your business.

What we will do:

- This Analysis will include accessing some of the major cities in India and analyzing the population based on the states.
- It will also include diving those cities into similar groups based on some of different venues present in those cities.

How we will do:

- We will extract population of various cities of India and will determine the state with maximum population, we will also determine the State with maximum density.

- We will represent it in a graph on a color scale, describing the total as well as mean population of various states of India.
- We will also find the Different venues in cities and determines which cities have similar neighborhoods.

Data

Dataset 1 : Cities and State list along with Population

Sample Data:

Name of City	State	Population (2011)
Mumbai	Maharashtra	1,35,97,924
Delhi	Delhi	1,10,07,835
Bengaluru	Karnataka	84,25,970
Ahmedabad	Gujarat	72,08,200

Usage:

- We will group the data based on the states and will find the total and mean population of all states. We will plot graphs and Maps and will find interesting facts.

Source:

- https://www.downloadexcelfiles.com/wo_en/download-excel-file-list-cities-towns-india#.Xv23ZCgzY2w

Dataset 2: Latitude and Longitudes of the cities:

Sample Data:

Latitude	Longitude
10.2188344	92.5771329
15.9240905	80.1863809
27.6891712	96.4597226
26.4073841	93.2551303

Usage:

- We will merge both the data frames for further analysis.

Source:

- Google Geopy Package has been used for determining the latitudes and longitudes of different states and cities.

Dataset 3: Foursquare API data

Sample Data:

City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Mumbai	Maharashtra	CafÃ©	Ice Cream Shop	Hotel
Delhi	Delhi	Indian Restaurant	Hotel	Lounge
Bengaluru	Karnataka	Hotel	Ice Cream Shop	Park
Ahmedabad	Gujarat	CafÃ©	Dessert Shop	Indian Restaurant
Hyderabad	Telangana	Indian Restaurant	Bakery	Ice Cream Shop

Usage:

- We will find data of different venues, will convert it into data frame and draw out analysis.

Source:

- Foursquare API.

Dataset 4: Geo Json File of Different States in India:

Sample Data:

Usage:

- We will use this to create Choropleth Maps for representing population data.

Source:

- Google

Methodology

Exploratory data analysis:

1. Population Analysis:

- The first dataset consisted three columns viz Name of City, State and Population
- Import this into Data frame and perform grouping based on states.
- Then, Two Frames were created representing total and mean population of different states.
- These can be represented into bar charts to draw some interesting facts:

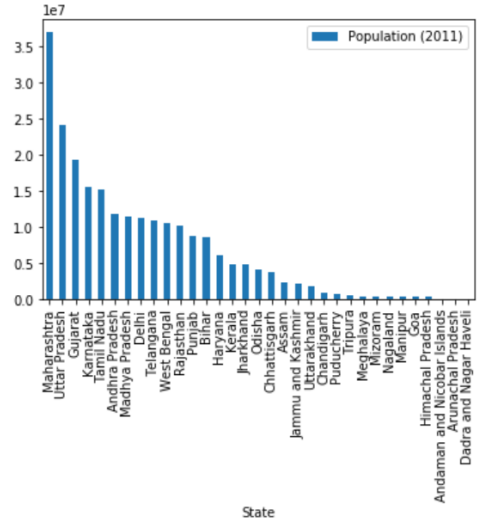

Fig: Total Population

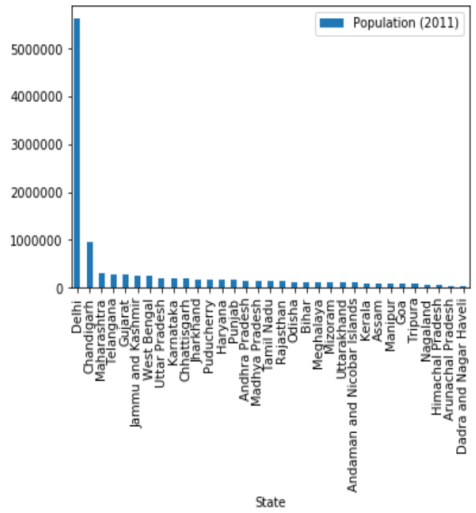

Fig: Mean Population

- We represent both graphs into Choropleth, representing population density in a region on a scale of color:

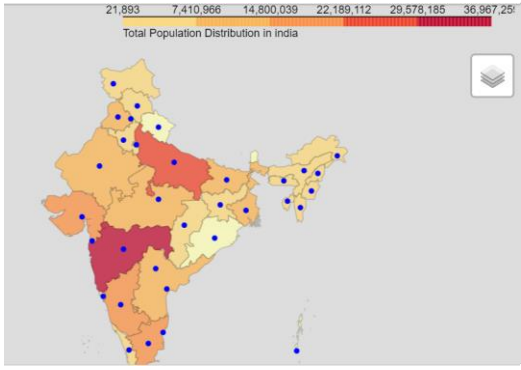

Fig: Total Population

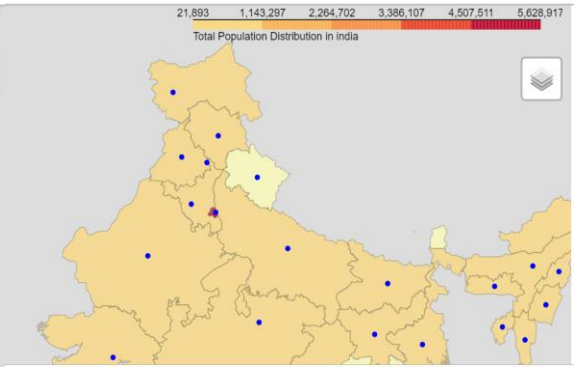

Fig: Mean Population

2. City Analysis:

- The initial data frame is then used to determine the Latitude and Longitude of different Cities using geopy package.
- These co-ordinates are saved into a csv file and imported as a new data frame:

	Name of City	State	Population (2011)	Latitude	Longitude
0	Mumbai	Maharashtra	13597924.0	18.938771	72.835335
1	Delhi	Delhi	11007835.0	28.651718	77.221939
2	Bengaluru	Karnataka	8425970.0	12.979120	77.591300
3	Ahmedabad	Gujarat	7208200.0	23.021624	72.579707
4	Hyderabad	Telangana	6809970.0	17.388786	78.461065

Fig: Combined Data frame

- Both data frames are combined into one and foursquare API is used to determine at most 50 venues for each city.
- Now, The venue details are merged with above data frame and following result is drawn:

	City	State	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mumbai	Maharashtra	18.938771	72.835335	Britannia & Co.	18.934683	72.840183	Parsi Restaurant
1	Mumbai	Maharashtra	18.938771	72.835335	Food for Thought	18.932031	72.831667	Café
2	Mumbai	Maharashtra	18.938771	72.835335	Starbucks	18.932190	72.833959	Coffee Shop
3	Mumbai	Maharashtra	18.938771	72.835335	Marine Drive	18.941221	72.823261	Scenic Lookout
4	Mumbai	Maharashtra	18.938771	72.835335	Royal China	18.938715	72.832933	Chinese Restaurant

Fig: City and Venue Data frame

- One hot encoding is performed for each unique Category per city to draw the following result:

	Zoo	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Andhra Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	A & C St
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig: One Hot Encoded Data frame

- Next, Data frame is grouped based on City and by taking the mean of the frequency of occurrence of each category:

  (Note : This Data frame is further used to apply ML Algorithm)

	City	Zoo	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Andhra Restaurant	Antique Shop	Arcade	Art Gallery
0	Achalpur	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.25
1	Achhnera	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00
2	Adalaj	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.022727	0.00
3	Adilabad	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00
4	Adityapur	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00

Fig: Mean Frequency Data frame

- Based on Frequency, This Data frame can be converted into the new Data frame representing top 10 most common places for each city :

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Achalpur	ATM	Mobile Phone Shop	Snack Place	Art Gallery	Women's Store	Fireworks Store	Factory	Falafel Restaurant	Farm	Farmers Market
1	Achhnera	ATM	Toll Booth	Indian Restaurant	Platform	Border Crossing	Mughlai Restaurant	Bed & Breakfast	Fireworks Store	Factory	Falafel Restaurant
2	Adalaj	Indian Restaurant	Café	Restaurant	Multiplex	Sandwich Place	Pizza Place	Coffee Shop	Fast Food Restaurant	Arcade	Chinese Restaurant
3	Adilabad	Pharmacy	Food	Fireworks Store	Fabric Shop	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field
4	Adityapur	Hotel	Café	Sandwich Place	Indian Restaurant	Ice Cream Shop	Bakery	Pizza Place	Market	Multiplex	Italian Restaurant

Fig: Most Common Places Data frame

Machine Learning Applied:

1. K-means Clustering using multiple values of K:

- Now, the k-means clustering is applied on Mean frequency data frame for k=1 to 12 elbow method is used to identify best fit for k:

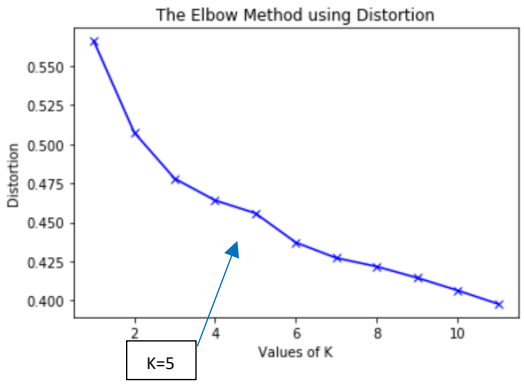

Fig: Distortions

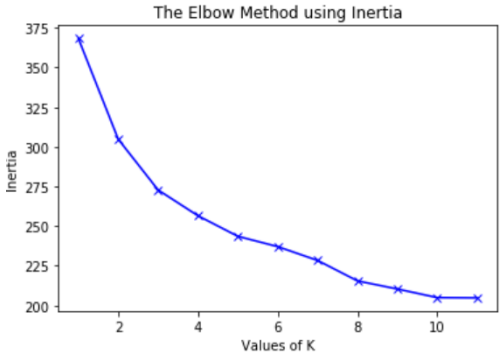

Fig: Inertia

- From Above Figures, It is concluded that k=5 will be best fit and now K-means is performed again to draw the output.

2. K-means Clustering using K=5:

- Data is divided into 5 said clusters and following results are found:

	City	State	Population	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th M Comn Venue
292	Achalpur	Maharashtra	112293.0	21.241445	77.425757	0.0	ATM	Mobile Phone Shop	Snack Place	Art Gallery	Women's Store	Fireworks Store	Factory	Falafe Restau
1093	Achhnera	Uttar Pradesh	22781.0	27.176058	77.758213	0.0	ATM	Toll Booth	Indian Restaurant	Platform	Border Crossing	Mughlai Restaurant	Bed & Breakfast	Firewo Store
1202	Adalaj	Gujarat	11957.0	23.164692	72.582105	4.0	Indian Restaurant	Café	Restaurant	Multiplex	Sandwich Place	Pizza Place	Coffee Shop	Fast F Restau
277	Adilabad	Telangana	117167.0	19.500000	78.500000	0.0	Pharmacy	Food	Fireworks Store	Fabric Shop	Factory	Falafel Restaurant	Farm	Farme Marke
191	Adityapur	Jharkhand	174355.0	22.782355	86.159003	0.0	Hotel	Café	Sandwich Place	Indian Restaurant	Ice Cream Shop	Bakery	Pizza Place	Marke

Fig: Cities with Cluster Labels

Results

Analysis of Outcome:

- Each cluster is analyzed based on Frequency:

	1st Most Common Venue				2nd Most Common Venue				3rd Most Common Venue			
	count	unique	top	freq	count	unique	top	freq	count	unique	top	freq
Cluster Labels												
0	547	118	Indian Restaurant	56	547	127	Indian Restaurant	44	547	117	Women's Store	38
1	138	25	Train Station	87	138	31	Train Station	43	138	45	Women's Store	34
2	167	11	ATM	151	167	41	Women's Store	58	167	30	Fish & Chips Shop	70
3	36	6	Mobile Phone Shop	29	36	13	Women's Store	12	36	7	Women's Store	17
4	141	13	Indian Restaurant	128	141	57	ATM	18	141	57	Train Station	16

Fig: Cluster Names with Frequency

- Each city is assigned with a Cluster Name:

	City	State	Population	Latitude	Longitude	Cluster Labels	Cluster Names	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1019	Piriyapatna	Karnataka	14924.0	12.337668	76.102711	1	Food Hubs	Indian Restaurant	Café	Women's Store	Fireworks Store	Fabric Shop	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
1020	Adalaj	Gujarat	11957.0	23.164692	72.582105	1	Food Hubs	Indian Restaurant	Café	Pizza Place	Multiplex	Restaurant	Sandwich Place	Coffee Shop	Fast Food Restaurant	Speakeasy	Cricket Ground
1021	Nandgaon	Maharashtra	11517.0	20.333058	74.670309	3	Stations	Train Station	Platform	Women's Store	Fireworks Store	Fabric Shop	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
1022	Barh	Bihar	10803.0	25.450163	85.702765	1	Food Hubs	ATM	Market	Train Station	Hotel	Women's Store	Fireworks Store	Factory	Falafel Restaurant	Farm	Farmers Market
1023	Chhapra	Gujarat	10147.0	20.925219	72.930057	1	Food Hubs	Bakery	Pizza Place	Indian Restaurant	Platform	Coffee Shop	Restaurant	Electronics Store	Multiplex	Flea Market	Playground
1024	Panamattom	Kerala	10032.0	9.598157	76.746078	1	Food Hubs	Indian Restaurant	Bakery	Bus Station	Waterfall	Concert Hall	Breakfast Spot	Art Gallery	Bar	Fireworks Store	Farm
1025	Bageshwar	Uttarakhand	9079.0	30.008672	79.930297	1	Food Hubs	Market	Women's Store	French Restaurant	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Fireworks Store
1026	Adyar	Karnataka	7034.0	12.868623	74.929288	1	Food Hubs	Lounge	Hotel	Seafood Restaurant	Ice Cream Shop	Indian Restaurant	Gym	Fast Food Restaurant	Snack Place	Multiplex	Café

Fig: Cities with Assigned Cluster Labels

Visual Results:

- Maps are drawn and Each city has been represented as a marker:

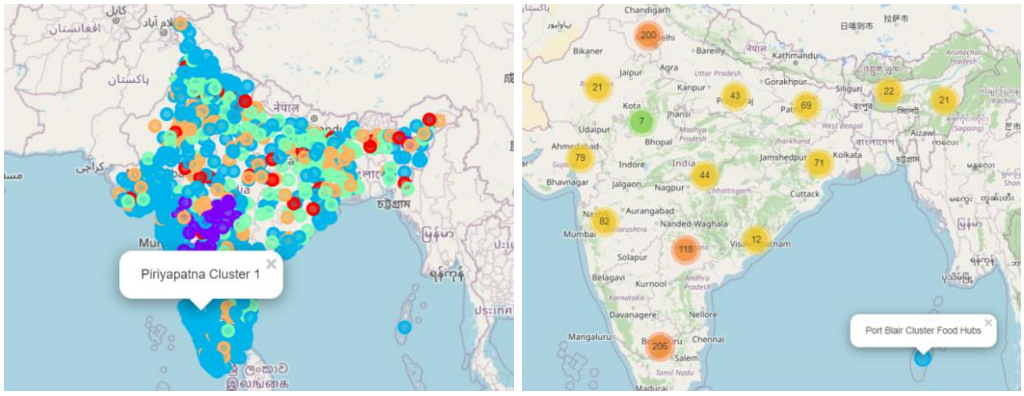

Fig: Maps with Different Clusters names and Labels

- Every Cluster is represented by a separate color code.

Discussions

- India is a big count with second highest population across world. The total population densities of the states in total can vary. Because of the complexity, very different approaches can be used in clustering and classification studies.

- I used the K-means algorithm as part of this clustering study. Where I tested the data set using Elbow method and set the optimum value of K to 5. However, only 33 state coordinates are being used, for more detailed and accurate guidance. the data set can be expanded, and the details of the neighborhood can also be included.

- Following results are drawn from the above analysis, cities are mainly clustered into groups which are:

	1st Most Common Venue				2nd Most Common Venue				3rd Most Common Venue			
	count	unique	top	freq	count	unique	top	freq	count	unique	top	freq
Cluster Names												
ATM Hubs	167	11	ATM	151	167	41	Women's Store	58	167	30	Fish & Chips Shop	70
Food Hubs	547	118	Indian Restaurant	56	547	127	Indian Restaurant	44	547	117	Women's Store	38
Indian Cuisine Restaurants	141	13	Indian Restaurant	128	141	57	ATM	18	141	57	Train Station	16
Mobile Stores and Women utility	36	6	Mobile Phone Shop	29	36	13	Women's Store	12	36	7	Women's Store	17
Stations	138	25	Train Station	87	138	31	Train Station	43	138	45	Women's Store	34

Fig: Cities counts per Cluster

Conclusion:

- People are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

- This study will help them in establishing business at new place or will also help them in choosing a city where they might be interested in working.