A study in Pune City's Population and food-services data.

Coursera IBM Data Science Capstone

Ishan Daga

Introduction

Pune is a fast-growing and upcoming metropolitan city in Maharashtra, India. It is well known as the cultural capital of the state and the de-facto king of the region's Street food.

The sprawling and ever-growing population and city-scape are home to some of the finest joints in the Maharashtra and the number of places to choose from in every locality of the city increases year-on-year.

Owning a joint in the city is a sure shot form of a stable income, even if it were only a food-truck, the people of the city take to fresh ideas quite well and the street-food culture has only added to itself over the years.

This is a serious business and requires careful planning and consideration on the part of the investor/stakeholder to ensure maximum profits from the business.

Business Problem

If a business would like to open a restaurant/brewery/pub/fast food joint in the city, what would be the optimal place to open said service in order to receive maximum revenue/foot-fall?

Target Audience

Future & prospective Food service/restaurant owners and stake holders can get a clear picture of the city's habits and the population density in order to open a new place in a suitable locality.

Data

Data used

To solve this problem, the following data has been leveraged:

- Population Data by Wards (Neighborhoods)
- Ward Offices Map Vector Data (GeoJSON and Shape Vectors)
- Location-wise Venue data for each sub-locality in Pune
- Latitude and Longitude Location Data of Wards and Venues

Acquiring the data

The Population Data was pulled from the Pune Municipal Corporation's opendata webpage

http://opendata.punecorporation.org/Citizen/CitizenDatasets/Index

The Wards and their corresponding ward offices with the populations per ward are listed out in a PDF from the census of 2011 [most recent]. This PDF was then converted to a csv file via the free online tool: PDF to Tables
https://pdftables.com/blog/convert-pdf-to-csv

The resulting CSV file was then manually cleaned up and double-checked in Microsoft Excel, to ease the pre-processing in python. The final CSV file had only the Ward numbers, Ward Names, Ward Offices and Population Data.

The Location data for the Wards was pulled using the Google Maps Geocoding API sequentially and added to the dataframe.

The venue data was procured using the FourSquare API which is simpler to use and has a free option, both of which are not available with the more expansive Google Maps API. Foursquare however does boast more community contributions and is used by 100,000+ developers world-wide.

Methodology

After importing the data from the pre-processed csv file described in the previous section, the following dataframes are created:

- Ward Office, Population
- Ward Office, List of Wards, Latitude and Longitude Data

The Latitude and Longitude Data are pulled using the **Google Maps GeoCoding API**, iterating over every ward in the dataframe.

Data is manually added for a region of the city called “Camp” as it comes under military jurisdiction and was not part of the Corporation’s original dataset. The population was also manually entered after a simple google search.

The Data is pictured below:

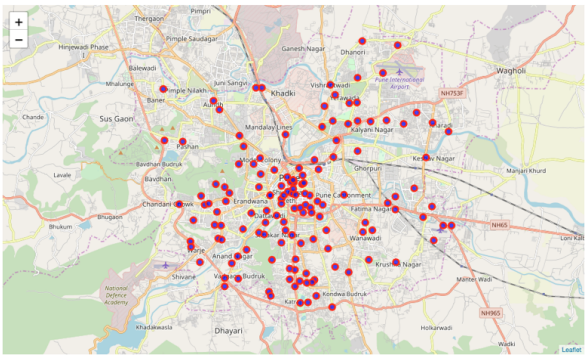
	Ward Office	Population
0	Aundh	181124
1	Bhavanipeth	177346
2	Bibvewadi	291446
3	Camp	171781
4	Dhankawadi	236648
5	Dholepatilroad	155413
6	Gholeroad	171678
7	Hadapsar	324751
8	Kasbavishrambaug	178484
9	Kothrud	209331
10	Sahakarnagar	205441
11	Sangamwadi	261957
12	Tilakroad	242290
13	Warje	233399
14	Yerwada/Nagarroad	239564

I

```
The dataframe has 14 Ward Offices and 144 Wards.
```

	Ward Name	Latitude	Longitude
0	Dhanori	18.596759	73.896851
1	Vidhyanagar Lohagaon	18.594668	73.917508
2	Tingre Pumping Station	18.576718	73.893968
3	Kalas Vishrantwadi	18.572605	73.878208
4	Nanasaheb Parulekar Vidhyalaya	18.567113	73.880916
5	Yerwada Prizen Press	18.562621	73.889375

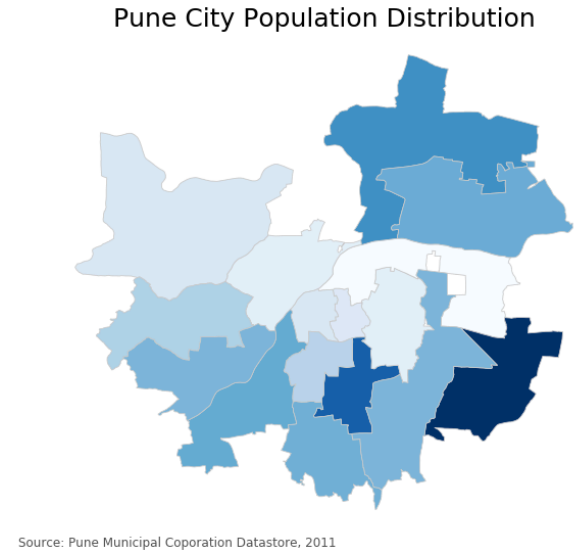
We can now visualize all the processed Wards on a Map using the **Folium Maps API**, using markers for each Ward.



Choropleth

This map only gives us the location co-ordinates for each of the wards and not much can be interpreted from this. We can instead use the **GeoPandas** library and a **GeoJSON** file of Pune to create a **choropleth** of the city's population by Major Ward Offices.

Before generating the Map, some pre-processing was required to correct spelling errors and allow the **GeoJSON** to match up with our Population Data Frame so that the map could be generated accurately.



This is a much better visual representation of the population distribution in Pune and can be cross referenced with other legends/maps to draw appropriate conclusions about the city's behavior.

Segmenting & clustering

Next, we will retrieve venue data by ward and segment and cluster the wards according to frequency of occurrence of the food/service industry in the city.

This data is pulled via the **FourSquare API** which is extensive, powerful, well organized and easy to use. We iteratively retrieve data for each Ward, storing the **category** of the **venue**, its **name** and **co-ordinates**. The resulting dataframe is pictured below:

	Ward	Ward Latitude	Ward Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dhanori	18.596759	73.896851	Reddy's	18.596037	73.895525	Indian Restaurant
1	Dhanori	18.596759	73.896851	Krushnai Dosa Point	18.596062	73.896681	Food Truck
2	Vidhyanagar Lohagaon	18.594668	73.917508	Four Points by Sheraton	18.590423	73.917115	Hotel
3	Tingre Pumping Station	18.576718	73.893968	Reddys Vadapav Center	18.574442	73.892433	Breakfast Spot
4	Tingre Pumping Station	18.576718	73.893968	Mad Momos	18.575221	73.895084	Café

Data Clean-up

There is a lot of Data here that we will not require to continue our analysis. The names of the locations hold no value, the Ward names have been repeated and there are several categories which are not of relevance to our current objective. All these values are dropped and using **One-Hot encoding**, we sort out the venues into a new dataframe where each ward has all venue categories with 1's if the venue is present and 0's if it is absent. A part of the Dataframe is shown below:

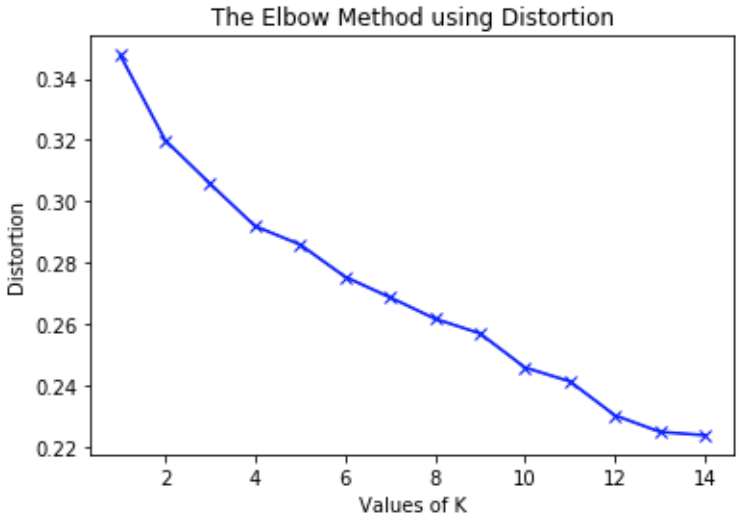
	Ward	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Bistro	Breakfast Spot	Brewery	Burger Joint	Burrito Place	Café	Chinese Restaurant	Chocolate Shop	Cocktail Bar	Coffee Shop	Crepe
0	Dhanori	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Dhanori	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Vidhyanagar Lohagaon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Tingre Pumping Station	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
4	Tingre Pumping Station	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	

Next, we check for frequency of each of the venue types in each ward and create a new Dataframe that is representative of this, to better visualize what we are dealing with.

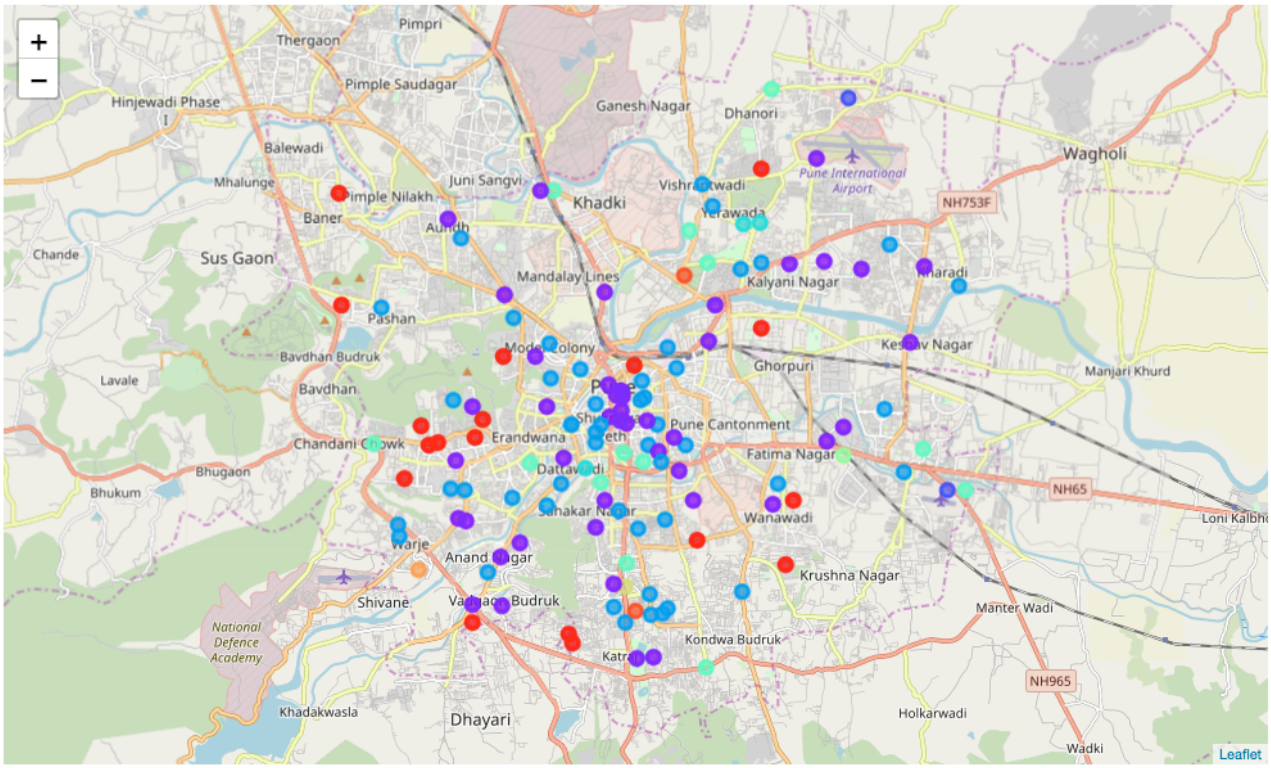
	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agakhan Pallace	Vegetarian / Vegan Restaurant	Bakery	Italian Restaurant	Café	Brewery	Fast Food Restaurant	Restaurant	Breakfast Spot	South Indian Restaurant	Thai Restaurant
1	Agam Mandir Katraj	Coffee Shop	Café	Vegetarian / Vegan Restaurant	Food Court	Food	Fast Food Restaurant	Dumpling Restaurant	Donut Shop	Diner	Dim Sum Restaurant
2	Anandnagar -Higne Khurd	Gourmet Shop	Fast Food Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Cocktail Bar	Food	Dumpling Restaurant	Donut Shop	Diner	Dim Sum Restaurant
3	Aundh I.T.I	Indian Restaurant	Mexican Restaurant	Fast Food Restaurant	Snack Place	Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Donut Shop	Diner	Dim Sum Restaurant
4	AundhGaon	Indian Restaurant	Fast Food Restaurant	Restaurant	Food Court	Korean Restaurant	Chocolate Shop	Snack Place	Mexican Restaurant	Diner	Dim Sum Restaurant

k-means Clustering

Here we will cluster the wards by occurrence of each venue type, using a **k-means clustering** algorithm. To optimize this algorithm, the **elbow method** was used to find the optimal number of clusters for the least distortion. The graph for the optimization run is below:



Clearly, after **k=10**, the graph becomes linear and has a small delta(slope), indicating exponentially decreasing results. Thus 10 clusters were selected as the optimal number. After running the clustering algorithm, the clusters are then visualized on a map using the **Folium API:**



Results

The results from the **k-means clustering** can be interpreted as follows:

- Cluster 1,2,3: Colored in Purple [1], Blue [2] and Red [3], in descending order: High presence/density of Food-related venues.
- Remaining clusters: Moderate to low presence of the industry.

This, over-layed with the **Choropleth** Map can give us a clear-cut picture of the optimal place to open a new café/bar, where to station a food truck and the best areas for restaurants.

Discussion

Comparing the two final maps of the Population Choropleth and k-means clustered venues, it is clear that the Hadapsar constituency has the most place available for growth, based on location popularity and population data of the region.

Second in line would be the Pune Airport Ward, under the Viman Nagar Ward Office, whose population is 2nd highest, with a purple and a blue [high presence/popularity] of food-venues.

These areas are optimal and may provide a decent foot-fall to any who wish to open a food-service based company/venue. The airport is also a high traffic area for non-localites which makes it even more desirable for high profile venues. This can be confirmed with real-world data as there are multiple 4,5-star hotels and restaurants in both areas.

If the business wishes to be a trail-blazer and set a new trend/ experiment with an area with relatively less exposure, but a decent population to support the business, the Aundh-Baner constituency shows promise with 2-3 high profile locations in a relatively large area, which would be appropriate as competition in the locations would be significantly lower, compared to the center of the city with the highest concentration of food-related venues and a moderate population.

Limitations and Future Possibilities for Research

This project considers only population data and frequency of venues to draw its conclusions. It is meant to be a placeholder/starting point for future work that could involve real estate prices [both to rent and buy], ratings of the venues in different localities, spending data from the venues, travel data for the city, population growth rates for each area and the median per capita income of each wards' households. Using this data, if made available, would improve the suggestions and give more concrete evidence on the location to set up a food-related business.

Conclusion

Thus, by following the Data Science Methodology as described in the course, we have successfully identified a relevant problem to which there is a large target audience, found data and processed it and using visualizations, we have drawn conclusions and answered the question set forth in the beginning.

We have manually scraped data from a pdf, used multiple different APIs to collect location and venue data on the city. Cross referenced this data with the population data. Used k-means clustering to segment different localities based on venue occurrence and created visually appealing maps for the customer's benefit and to support our conclusions.

There is scope for more data to be added to provide further evidence to support the claims made, and as and when available, can be added to the project and the evidence updated regularly and iteratively as is appropriate according to the Data Science Methodology.

The findings put forth in the project are sure to help the stakeholders make informed decisions on where to open their respective businesses.