COURSERA - IBM APPLIED DATA SCIENCE

Capstone Project – Week 5

INTRODUCTION

Pune district is the cultural capital of Maharashtra state in India. With it's roots firmly anchored in the ancient Maratha culture, Pune can boast of a highly intellectual, cosmopolitan population with a homogeneous composition of people originating from all *corners of India*

Pune has been ranked as the eighth most populous city in India, with an estimated population of about 7.2 million as of 2019. It has been ranked as one of the most livable cities in India several times.

Along with its extended city limits Pimpri Chinchwad and the three cantonment towns of Pune, Khadki and Dehu Road, Pune forms the urban core of the eponymous Pune Metropolitan Region (PMR). According to the 2011 census, the urban area has a combined population of 5.05 million while the population of the metropolitan region is estimated at 7.27 million. Situated 560 metres (1,837 feet) above sea level on the Deccan plateau on the right bank of the Mutha river, Pune is also the administrative headquarters of its namesake district.

More popular as the IT Capital of India, Pune is also known as the "Oxford of the East" due to the presence of several well-known educational institutions. The city has emerged as a *major educational hub in recent decades, with nearly half of the total international* students in the country studying in Pune. Research institutes of information technology, education, management and training attract students and professionals from India and overseas.

BUSINESS PROBLEM

Over the past decade, Pune city has seen a huge transformation in infrastructure due to the immense opportunities that it presents as one of the fastest developing cities of India. Hence, the platform for huge business opportunities and business friendly environment ends up attracting leading players into the market. This has thus resulted in a highly competitive market, and the costs of doing business is comparable to some of the highest in the country. This itself requires a thorough analysis of the business opportunities, and any investment must be strategically targeted so as to maximize returns,

Background

The analysis presented in the Capstone project maybe reviewed by any investor, business shop or venture capitalist to arrive at an informed decision on what business opportunity should be explored with an eye on location parameters

Business Problem

We expect that the following questions will have ready answers by the end of the analysis

- How is the entire city structured with regard to infrastructure, strategic locations, population?
- Which are the highly dense & well-developed areas that can be picked up for consideration
- Which are the areas in Pune that provide the highest business opportunities? And for what kind of business.

DATA

Data Sources

To make the analysis and observations more effective, it was decided to acquire data from multiple sources, following are the details:

From **Opendatasoft** API (https://public.opendatasoft.com/explore/?sort=modified&q=pune), Pune city data, along with neighbourhoods was retrieved. Following features were available to form the city map.

- **Country code:** Country information.
- **admin name 1:** State level details (Maharashtra, in this case).
- **admin name 2:** All cities included in Pune district.
- **admin name 3:** Defined neighbourhoods in Pune city.
- **place name:** The street, lane name or *peth* as locally called
- **postal code:** Zip code
- **Latitude:** The latitude value of the venue.
- **Longitude:** The longitude value of the venue.

Subsequently, **Foursquare's** explore API (https://api.foursquare.com/v2/venues) (which gives venues recommendations) was leveraged. With Pune as the center, and a radius of 500 m to scout around – each location/place was broken up into a series of venues and detailed information was procured to assist in identifying the tops spots within the city.

From Foursquare API (https://developers.zomato.com/api), following attributes for each venue were retrieved:
- **Name**: The name of the venue.
- **Category**: The category type as defined by the API.
- **Latitude**: The latitude value of the venue.
- **Longitude**: The longitude value of the venue.

Data preparation

The dataset came with a set of 791 rows and 13 features spanning out across all locations within Maharashtra (depicted below). An initial scan of missing values carried out, and the base data was cleaned up.
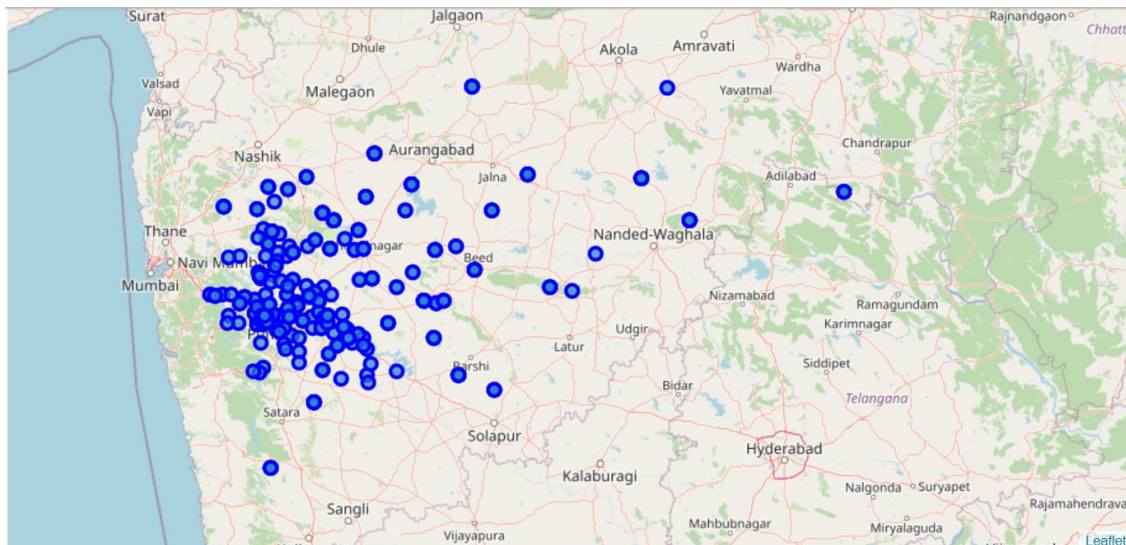
	country code	postal code	place name	admin name1	admin code1	admin name2	admin code2	admin name3	admin code3	latitude	longitude	accuracy	coordinates
0	IN	410403	Kaivalyadham	Maharashtra	16	Pune	521	Maval	NaN	19.4781	73.7845	1	19.4781,73.7845
1	IN	410405	Takave KH	Maharashtra	16	Pune	521	Maval	NaN	19.6594	73.8791	3	19.6594,73.8791
2	IN	410406	Shivali	Maharashtra	16	Pune	521	Maval	NaN	19.4781	73.7845	1	19.4781,73.7845
3	IN	410502	Kusur	Maharashtra	16	Pune	521	Junnar	NaN	18.5709	74.0691	4	18.5709,74.0691
4	IN	410502	Godre	Maharashtra	16	Pune	521	Junnar	NaN	19.3864	74.4703	3	19.3864,74.4703

The next strep involved eliminating the unwanted features and renaming columns to meaningful names, following is a depiction if the cleaned dataset

	Country	State	City	Neighbourhood	Place	Postalcode	latitude	longitude
0	IN	Maharashtra	Pune	Maval	Kaivalyadham	410403	19.4781	73.7845
1	IN	Maharashtra	Pune	Maval	Takave KH	410405	19.6594	73.8791
2	IN	Maharashtra	Pune	Maval	Shivali	410406	19.4781	73.7845
3	IN	Maharashtra	Pune	Junnar	Kusur	410502	18.5709	74.0691
4	IN	Maharashtra	Pune	Junnar	Godre	410502	19.3864	74.4703

METHODOLOGY & DATA EXPLORATION

Next step was to get a quick view of how the underlying dataset panned out in terms of actual locations. The folium library was used to review a quick plot (see below)

An immediate observation arising out of the map was that the locations are far spread out across the entire district – this is turn would render any observation very difficult to analyse given the huge expanse of land.

It was hence decided to filter out the dataset and include only places having neighbourhood within Pune & Pune City. Setting this filter immediately reduced the dataset size to 110 rows.

Next up, using the , Foursquare's explore API (https://api.foursquare.com/v2/venues) a list of at least 100 venues across a radius of 500 meters was searched for. The data was consolidated in a common merged dataframe.

```
(491, 9)
```

	PostalCode	Place	Neighborhood	PlaceLatitude	PlaceLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	411001	Pune New Bazar	Pune City	18.5372	73.8551	Shopper's Stop	18.534332	73.851900	Department Store
1	411002	Shukrawar Peth (Pune)	Pune City	18.5372	73.8551	Shopper's Stop	18.534332	73.851900	Department Store
2	411030	Narayan Peth	Pune City	18.5333	73.8667	Spree Lotus Valley	18.532172	73.863680	Hotel
3	411030	Narayan Peth	Pune City	18.5333	73.8667	AISSMS Dual Football Ground	18.530317	73.865011	Soccer Field
4	411030	Narayan Peth	Pune City	18.5333	73.8667	The Millers	18.531992	73.870609	Restaurant

The resultant dataset was observed to be a combination of 491 venues & locations.
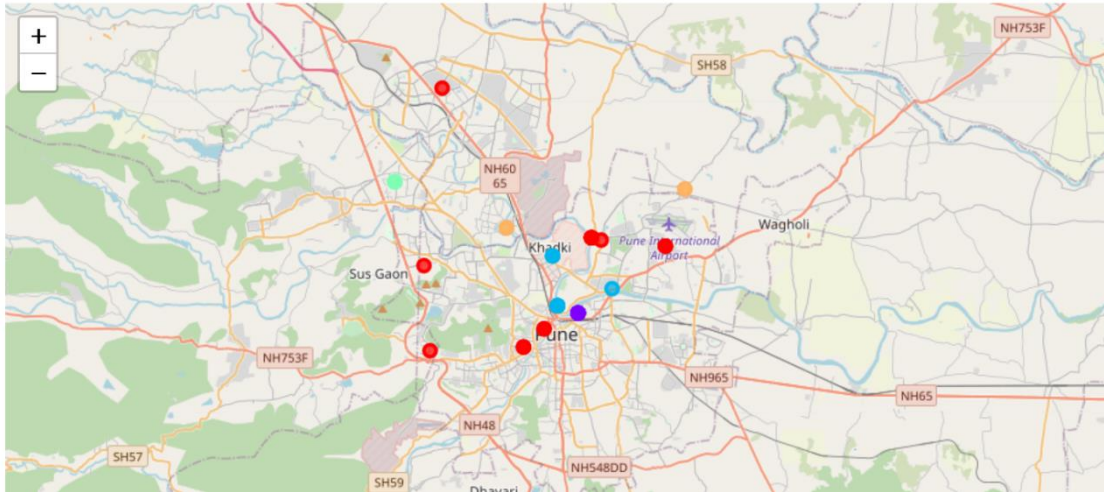
Categorization

To ease out the analysis process, it was decided to move each category to columns and combine it with the location data.

	PostalCode	Place	Neighborhoods	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
0	411001	C D A (O)	Pune City	Intersection	IT Services	Snack Place	Motorcycle Shop	Department Store	Electronics Store	Hotel Bar	Hotel	Gym / Fitness Center
1	411001	Dr.B.A. Chowk	Pune City	Intersection	IT Services	Snack Place	Motorcycle Shop	Department Store	Electronics Store	Hotel Bar	Hotel	Gym / Fitness Center
2	411001	Ghorpuri Bazar	Pune City	Intersection	IT Services	Snack Place	Motorcycle Shop	Department Store	Electronics Store	Hotel Bar	Hotel	Gym / Fitness Center
3	411001	N.W. College	Pune City	Intersection	IT Services	Snack Place	Motorcycle Shop	Department Store	Electronics Store	Hotel Bar	Hotel	Gym / Fitness Center
4	411001	Pune Cantt East	Pune City	Intersection	IT Services	Snack Place	Motorcycle Shop	Department Store	Electronics Store	Hotel Bar	Hotel	Gym / Fitness Center

Clustering

From an analytical standpoint – to arrive at a conclusion so as to point out locations where most of the venues were pronounced, it was decided to do a k-means clustering to report out the top 5 clusters.

Following is the map developed based on the clustered datapoints. Thus, there are 5 clusters in all and all 5 clusters are located well within the city limits.

The final step is now to classify each cluster and review each datapoint with regards to the venues/category of businesses in each other vis-à-vis their significance.

RESULTS & DISCUSSION

Based on location & clustering by venue/category – we have clustered our 110 locations into 5 different clusters. Each cluster is found to have it's own significance based on the businesses corresponding popularity.

Cluster 1

This cluster includes mostly Indian Restaurants, Coffee Shoppes, Fast Food joints and Ice-cream shops. Locations are mainly along the Central city areas (7 points) and some exterior locations (2 points)

Cluster 2

This cluster includes mostly Departmental Stores, Cinema Theaters, Hotel Bars and Gymnasiums. Locations are mainly along the Central city areas, Northern side of the city and some exterior locations (20 points)

Cluster 3

This cluster is fairly well-pronounced in the Central & Western sides of the city (20 locations) well-interspersed with Restaurants, Hotels, Outlet Malls, Soccer fields and Gymnasiums.

Cluster 4

This cluster includes mostly Indian Restaurants, Snack places & Shopping Malls (13 points)

Cluster 5

A smaller-sized cluster, this cluster includes mostly Mobile Phone shops, Vegetarioan Restaurants, Electronics shops – all placed along exterior locations.

CONCLUSION

A sizeable chunk of the business in Pune city is controlled by the Food & Beverages industry. Nearly 60% of all the businesses deal with cuisine of various kinds

An obvious and foregone conclusion towards Food & Beverage as a successful industry based on underlying datapoints and subsequent analysis is most certain.