Capstone Project - The Battle of Neighborhoods

Exploring neighborhoods in Pune, Maharashtra

By Nandhitha .V

Table of Contents

Capstone Project - The Battle of the Neighborhoods

I. Introduction

Pune also called Poona, is the second largest city in the Indian state of Maharashtra, after Mumbai. It is the ninth most populous city in the country with an estimated population of 6.4 million. Pune is ranked the number one city in India in the ease of living ranking index. But, for someone who is new to this state, it's puzzling to figure a neighborhood to settle.

I have recently got transferred to Pune for work. Which neighborhoods had got vegetarian restaurants? Which neighborhoods have shopping malls and theatres? Which neighborhoods are great for a coffee? Which neighborhoods are famous for its markets? Where are the ATMs? were the questions running on my mind as a new resident. So, in this capstone project, i am going to find a good neighborhood location in Pune to stay in by using data science methods and algorithms like clustering.

II. Data

The required data to do this project is as follows:

- Latitude and Longitude of neighborhood
- List of neighborhoods in Pune, Maharashtra
- Venue data(restaurants, ATM, theatres) of neighborhoods

Data sources

Following data sources will be needed to extract the required information:

- Web scraping data from Wikipedia to get information on Pune neighborhoods using Beautiful soup.

The link to the data is : 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune'

- Generating Longitude and Longitude coordinates of Pune as well the neighborhoods via geocoder, geopy package.

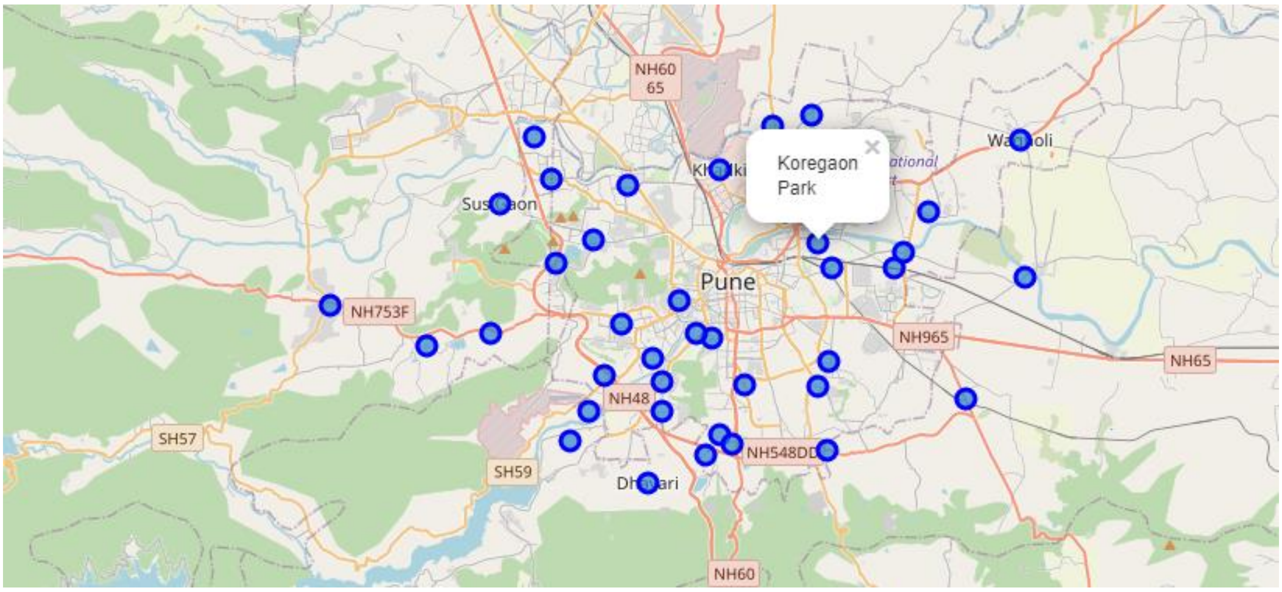
- Using Foursquare API to create a map of Pune and also to generate the venue data related to the neighborhood.

III. Methodology

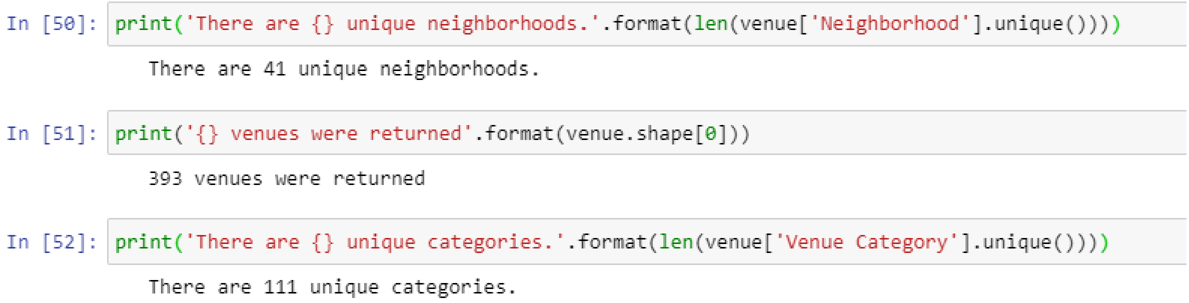
The project is done following the seven Data Science methodology phases. The methodology will include:

- Firstly, packages such as numpy, pandas, folium, k-means, geopy etc. which are required to do the project are imported.
- Generating the data required from Wikipedia 'https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune' by web scraping using Beautiful soup. Used functions like decompose for data cleaning which removes the unnecessary data from the website.
- The data collected includes the names of the neighborhoods in Pune and is converted into a data frame.
- The geocodes (latitude and longitude) of the neighborhoods are generated using the Nominatim from geopy.geocoders package.
- A data-driven map of Pune is generated using folium package by adding location markers to it.

Map of Pune before Clustering

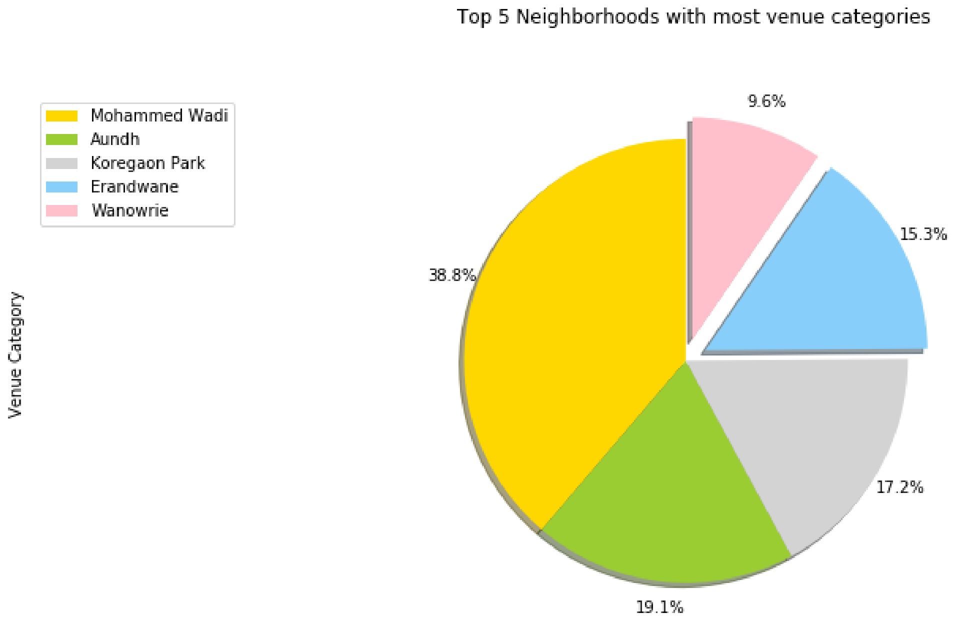


- The venues for each neighborhood is collected using the foursquare API by providing the credentials, latitude, longitude, version and URL.
- The number of unique neighborhoods, venues, venue categories in the data frame is printed.



```
In [50]: print('There are {} unique neighborhoods.'.format(len(venue['Neighborhood'].unique())))

    There are 41 unique neighborhoods.

In [51]: print('{} venues were returned'.format(venue.shape[0]))

    393 venues were returned

In [52]: print('There are {} unique categories.'.format(len(venue['Venue Category'].unique())))

    There are 111 unique categories.
```

- Top 5 Neighborhoods with most venue categories is depicted using a pie chart.



- A search query is written to find and explore the vegetarian restaurants in the neighborhood.

- The get_dummies method of pandas library is then used and the categories of venues is got.

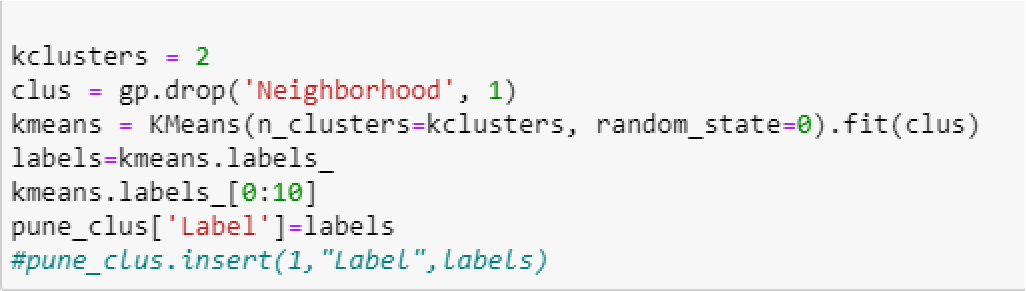
- Then group_by is used to Group rows of data frame by location and the mean of the frequency of occurrence of each category in every location is found.

- Top five categories of each neighborhood is printed with the frequency.

- A function called return_most_common_venues is used to return ten most common venues in each neighborhood for easy-understanding.

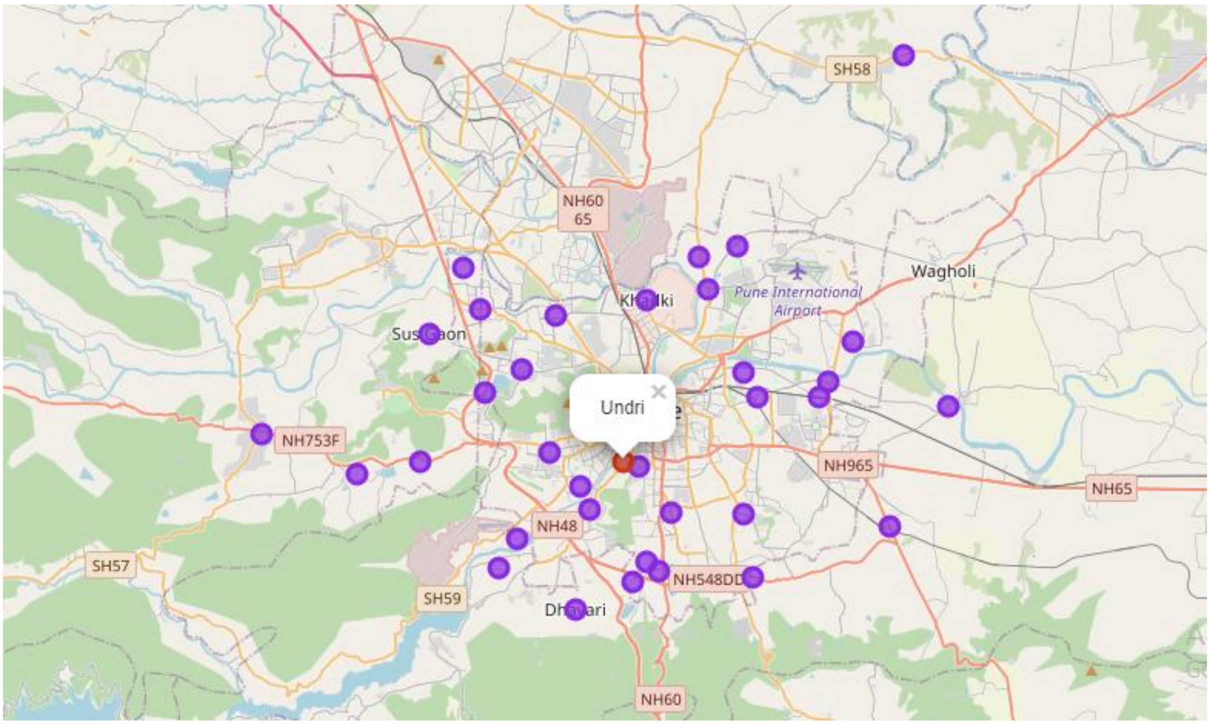
Clustering the Neighborhood

- K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups called clusters. So, here k-means clustering is performed on the most common places of the neighborhood with k=2.Labels are assigned for each cluster to differentiate.



```
kclusters = 2
clus = gp.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(clus)
labels=kmeans.labels_
kmeans.labels_[0:10]
pune_clus['Label']=labels
#pune_clus.insert(1,"Label",labels)
```

Map of Pune after Clustering



- Map of Pune before and after clustering is compared and visualized.

- Clusters are examined to derive cool insights from it.

IV. Results

The results of the modelling gives the below clusters.

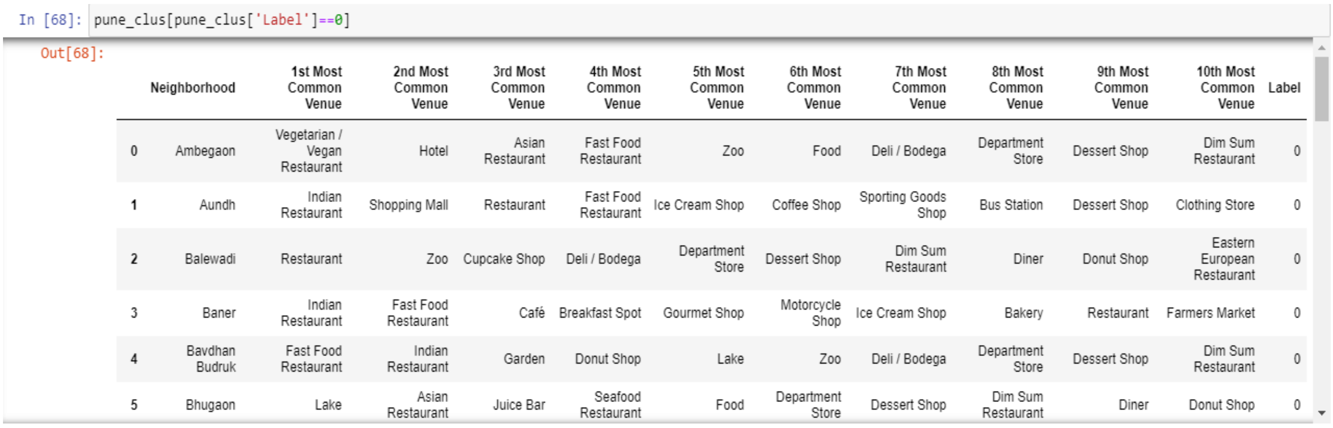
The k-means clustering was performed with k=2.The data was divided into two clusters with label-0 and label-1.The total number of neighborhoods were 41.

- Cluster 0 - All the neighborhoods in the cluster have diverse venues mostly restaurants, coffee shops, malls, department stores etc. as the first two common venues.

- Cluster 1 - The neighborhood in this cluster have other venues like furniture store, zoo etc.

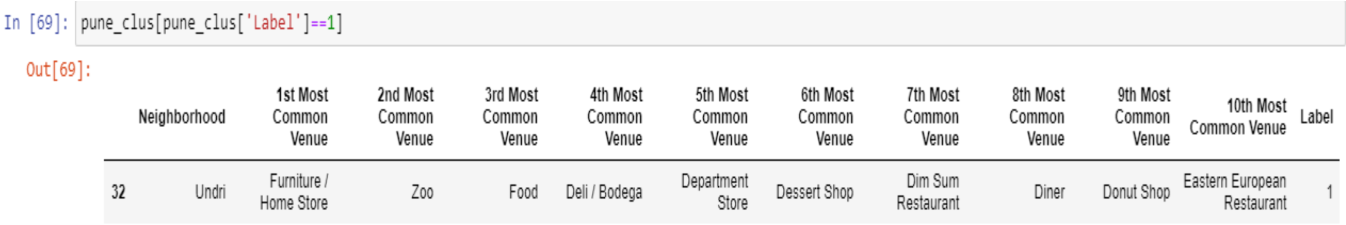
On examining the clusters, it is clear that 40 neighborhoods are similar and thus it falls under the same cluster with the label-0. The only neighborhoods which is different from the other are ' Undri ' and thus it is falls into another cluster with label-1.

Examining each cluster



```
In [68]: pune_clus[pune_clus['Label']==0]
Out[68]:
```

	Neighborhood	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	Label
0	Ambegaon	Vegetarian / Vegan Restaurant	Hotel	Asian Restaurant	Fast Food Restaurant	Zoo	Food	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	0
1	Aundh	Indian Restaurant	Shopping Mall	Restaurant	Fast Food Restaurant	Ice Cream Shop	Coffee Shop	Sporting Goods Shop	Bus Station	Dessert Shop	Clothing Store	0
2	Balewadi	Restaurant	Zoo	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Donut Shop	Eastern European Restaurant	0
3	Baner	Indian Restaurant	Fast Food Restaurant	Café	Breakfast Spot	Gourmet Shop	Motorcycle Shop	Ice Cream Shop	Bakery	Restaurant	Farmers Market	0
4	Bavdhan Budruk	Fast Food Restaurant	Indian Restaurant	Garden	Donut Shop	Lake	Zoo	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	0
5	Bhugaon	Lake	Asian Restaurant	Juice Bar	Seafood Restaurant	Food	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Donut Shop	0



```
In [69]: pune_clus[pune_clus['Label']==1]
Out[69]:
```

	Neighborhood	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	Label
32	Undri	Furniture / Home Store	Zoo	Food	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Donut Shop	Eastern European Restaurant	1

V. Discussion

In this project, Analysis of Pune neighborhoods recommendations based on venue categories like restaurants, mall, coffee shops etc. has been presented. This will be a great recommendation for visitors like me who are new to the place to find out nearby venues of interest and in deciding a place to stay in.

- The generated results shows the 2 clusters associated with the neighborhoods. It is evident that Cluster 0 is the most representative of the Pune city. Just looking at this cluster, it shows there are 80% of restaurants, grocery stores, coffee shops, electronic stores which will make life easier if we stay in this neighborhood.

- On the other hand, as we observe Cluster 1, the percentage is less than 40% .There are not many restaurants in these area. But this place is has got other spots like Furniture store, Zoo etc.

VI. Conclusion

Using the combination of data from the URL and the foursquare API, we were able to collect, clean, analyse, discover and examine the venues of the neighborhoods in Pune to find the best neighborhood to live in.

Since all the neighborhoods were similar with respect to the most common venues, and also likely to live , I would find a good place to stay in any of it other than the neighborhood 'Undri' which didn't seem to be a residential and accessible place.