The Battle of Neighborhoods

A study of Realty and Cuisine variation across neighbourhoods in Pune!

ABSTRACT

A look at the relative affluence in the city as reflected by the property rates and choice of culinary dining options

Sunil Sidhanty

IBM DATA SCIENCE

Contents

Introduction/Business problem

I live in Pune now and have seen this city transform from a sleepy pensioner's town from my student days to a bustling metropolis. It is traditionally a centre of Maharashtrian culture and city of learning, called as slightly derivative "Oxford of the east". It has a thriving cosmopolitan centre, thanks to strong Automotive and Tier 1 network as well as a significant base for Indian IT service industry. The city has a strong German connection as well

It was my pleasure to relocate to Pune after a 25 year gap and the changes have been astonishing.

When I was new to the city recently, I struggled to find cuisine to suit my taste. I also had no idea of the geography and spent a lot of time to find a place to live. My intent in this exercise is two-fold:

1. For perspective house buyers (like me), I want to understand the trend and variety of neighbourhoods
2. For **Tourists and new international students** in Pune I hope to create a cluster of restaurants and cuisines to help them enjoy the city.

Description of the data

- *Data Source 1*: I will use foursquare data about Pune venues and then filter for restaurants

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Handewadi	18.485162	73.931680	Marrakesh	18.509430	73.928296	Middle Eastern Restaurant
2	Handewadi	18.485162	73.931680	Rolls Mania	18.509609	73.928547	Fast Food Restaurant
3	Handewadi	18.485162	73.931680	Kokkita	18.502177	73.937133	Indian Restaurant
4	Handewadi	18.485162	73.931680	City Cafe	18.509279	73.928375	Fast Food Restaurant
5	Handewadi	18.485162	73.931680	Vaibhavi Food Court	18.510085	73.928015	Indian Restaurant
...	...	...	...	...	...	...	...
1066	Senapati Bapat Road	18.525574	73.829548	Aware Maratha Khanawal	18.512948	73.845200	Indian Restaurant
1067	Senapati Bapat Road	18.525574	73.829548	Fish Curry Rice	18.510713	73.829623	Seafood Restaurant
1068	Senapati Bapat Road	18.525574	73.829548	Chinese Room Oriental	18.510482	73.834788	Chinese Restaurant

- *Data Source 2*: Also, I will use the find neighbourhoods of Pune and their property prices from https://www.moneycontrol.com/property-rates/pune/99acres-residential.html?ver=1. The coordinates of the neighbourhoods will be calculated via GeoPy service moneycontrol.com/property-rates/pune/99acres-residential.html?ver=1

magicbricks

Pune	Capital Values Rate/Sq ft (INR)	Updated on
Akurdi	5741	Mar 2018
Alandi	3900	Mar 2018
Ambe Gaon Budruk	4850	Mar 2018
Anand Nagar	6350	Mar 2018
Aundh	9160	Mar 2018
Balewadi	6907	Mar 2018
Baner	7323	Mar 2018
Bavdhan	6900	Mar 2018
Bhosari	5180	Mar 2018
Bhosle Nagar	13925	Mar 2018
Bhugaon	5200	Mar 2018
Bibwewadi	7650	Mar 2018
Boat Club Road	13900	Mar 2018
Camp	8800	Mar 2018
Chakan	3350	Mar 2018
Charholi	5400	Mar 2018
Chikhali	4700	Mar 2018
Dhankawadi	5050	Mar 2018
Dhanori	5360	Mar 2018
Dhayari	5130	Mar 2018
Dighi	4550	Mar 2018
Erandwane	12477	Mar 2018
Fatima Nagar	6705	Mar 2018
Hadapsar	5900	Mar 2018
Handewadi	4677	Mar 2018
Hinjewadi	5743	Mar 2018
Kalewadi	5565	Mar 2018
Kalyani Nagar	9400	Mar 2018
Karve Nagar	8650	Mar 2018
Katraj	5600	Mar 2018
Keshav Nagar	5570	Mar 2018
Kharadi	6550	Mar 2018
Kiwale	4900	Mar 2018
Kondhwa	5600	Mar 2018

Methodology

In this section, I will describe the data analysis and how I used the data to yield the results.

I began by scraping neighbourhood and property value from the money control website. For this, I used the pandas read function. I had to clean the resulting data frame in terms of unnecessary information or data that could not be handled in a data frame. The entire web scraped contents contained the data table that once cleaned gave me the following information

```
10]: pune1_data = df[0]
     pune1_data.head()
```

10]:

	Pune	Capital Values Rate/Sq ft (INR)	Updated on
0	Akurdi	5741	Mar 2018
1	Alandi	3900	Mar 2018
2	Ambe Gaon Budruk	4850	Mar 2018
3	Anand Nagar	6350	Mar 2018
4	Aundh	9160	Mar 2018

I. Data Transformation: convert absolute property values to slabs

Since we wanted to cluster the data eventually, I decided to convert the absolute property values into relative slabs like this:

Property Value range Rate in INR/Sq. ft	Category
3000 to 5000	Category 1
5000 to 7000	Category 2
7000 to 9000	Category 3
9000 to 11000	Category 4
Greater than 11000	Category 5

	Area	Rate	Category
1	Alandi	3900	1
2	Ambe Gaon Budruk	4850	1
14	Chakan	3350	1
16	Chikhali	4700	1
20	Dighi	4550	1
...	...	...	...
21	Erandwane	12477	5
34	Koregaon Park	11350	5
36	Law College Road	15000	5
55	Prabhat Road	15540	5

Then, I enabled geopy functions by installing the conda-forge geopy package. I used the nominatim function to add geospatial data to the data frame, that is the latitude and the longitude for all the neighbourhoods.

II. Data Transformation: Neighbourhoods that don't have valid coordinates

I discovered that geopy did not return valid results for some of the neighbourhoods. To fix this I identified these neighbourhoods and dropped them from the dataframe

```
This area code does not have Coordinates:  Ambe Gaon Budruk
This area code does not have Coordinates:  Charholi
This area code does not have Coordinates:  Dhankawadi
This area code does not have Coordinates:  Mohamadwadi
This area code does not have Coordinates:  Salunke Vihar
This area code does not have Coordinates:  Tingre Nagar
This area code does not have Coordinates:  Wanwadi
This area code does not have Coordinates:  Erandwane
```

Here was the resultant cleansed dataframe:

	Area	Rate	Category	Latitude	Longitude
1	Alandi	3900	1	18.677245	73.898113
14	Chakan	3350	1	46.540644	17.271490
16	Chikhali	4700	1	20.935774	79.729506
20	Dighi	4550	1	26.882192	78.149700
24	Handewadi	4677	1	18.485162	73.931680
...	...	...	...	...	...
12	Boat Club Road	13900	5	32.860278	-97.426164

A casual look at this data highlights another issue. The values returned by geopy are wildly off. In a city the neighbourhoods can't be this far-off

III. Data Transformation: Neighbourhoods that have erroneous data

I decided to create a bounding box around Pune and would only use data that falls within this box. The boundary conditions were identified by using google-maps

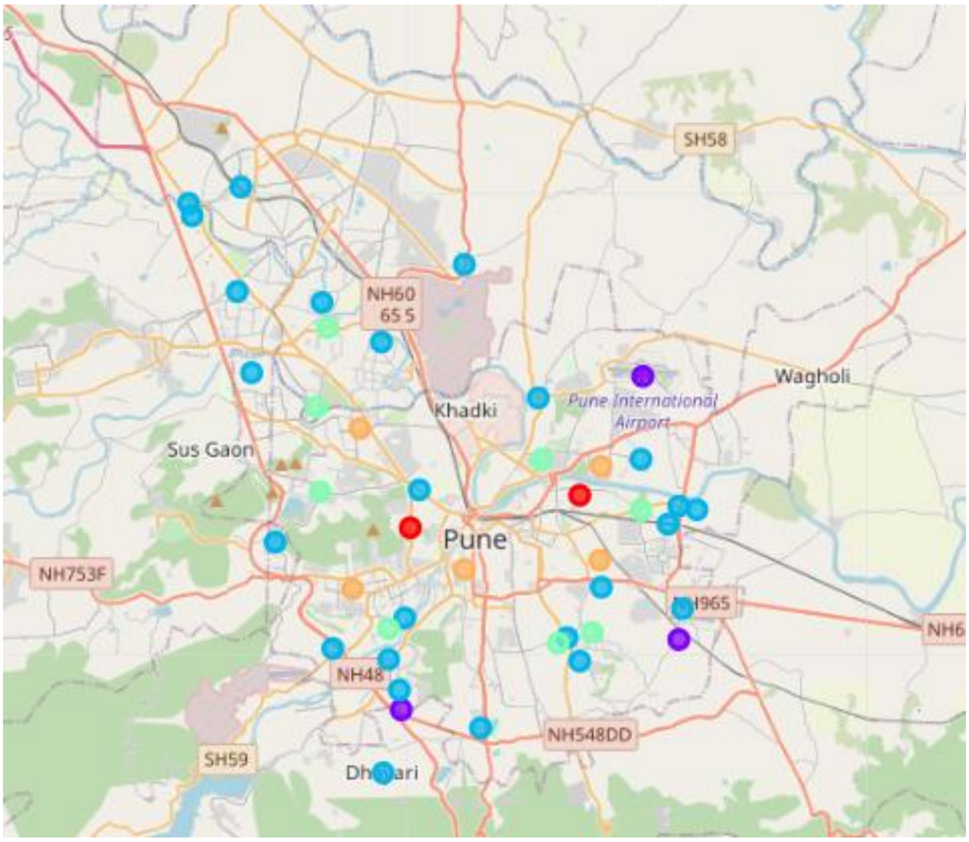
The code and results look like this:

```
pune_df = pune_cln.loc[(pune_cln['Latitude'] >=18.4) &
          (pune_cln['Latitude'] <= 18.65) & (pune_cln['Longitude'] >=73.74) &
          (pune_cln['Longitude'] <= 73.98)]
pune_df
```

	Area	Rate	Category	Latitude	Longitude
24	Handewadi	4677	1	18.485162	73.931680
37	Lohegaon	4600	1	18.580330	73.918386
46	Narhe	4747	1	18.460143	73.826010
0	Akurdi	5741	2	18.648642	73.764708
3	Anand Nagar	6350	2	18.478490	73.821326
5	Balewadi	6907	2	18.582027	73.768983
7	Bavdhan	6900	2	18.520954	73.778087
8	Bhosari	5180	2	18.621009	73.850130
19	Dhayari	5130	2	18.437398	73.819043

Now this data looks visually correct and passes the smell test!

To verify this, I used the folium package and my data frame, I then created a map of Pune city. To make it more interesting, I also decided to add colour coding based on category



Already we see a lovely heatmap-ish clustering! The areas on the outskirts (Purple) are Category 1 (Cheapest to live in) and the ones in the centre (RED) are the most expensive Category 5.
The near concentric circles show a clear gradation: The cost of your property increases as you go towards the city centre

---

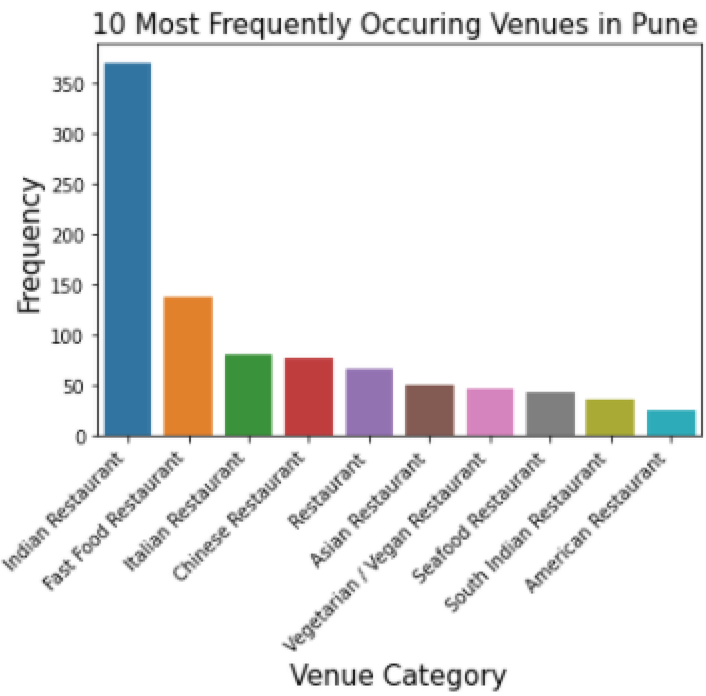
I then started working on the Foursquare venue data. I first did a dry-run for the “Akurdi", to see if the venues retrieved from foursquare passed the smell test.

	name	categories	lat	lng
0	Shawarma King	Middle Eastern Restaurant	18.649124	73.765480
1	SKS Katthi Kabab Rolls	Asian Restaurant	18.648234	73.765805
2	Cafe Aroma	Café	18.650155	73.766566
3	Café Crème	Café	18.648282	73.765939
4	Hotel Annapurna	Indian Restaurant	18.648775	73.764911

A quick google search proved that the data is reasonable!

Then, retrieved the foursquare data for all venues in the neighbourhoods. The result was a list of 2914 venues all over Pune city. I filtered for restaurants and ended up with 1070 restaurants.

I plotted a bar chart with the frequency of the 10 most frequently occurring restaurants in the whole city, using seaborn/matplotlib packages.




```
<Figure size 1296x504 with 0 Axes>
```

We can conclude that while Pune is a cosmopolitan city the predominant cuisine is Indian

While Pune is a cosmopolitan city, its Indian roots are indeed very strong, and an overwhelming portion of the cuisine is Indian or Indianized!

To find clusters of restaurant types in the different city areas, I first transformed the data frame with the restaurant venues, by one-hot encoding (0/1), as seen in the picture below.

	Neighborhood	American Restaurant	Andhra Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	...	North Indian Restaurant	Parsi Restaurant	Punjabi Restaurant	Restaurant	Re
0	Handewadi	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	Handewadi	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

Next, I used grouping to show the frequency of each category of restaurants in each city district.

	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
0	Akurdi	Indian Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Dumpling Restaurant
1	Anand Nagar	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Seafood Restaurant	Restaurant	Chinese Restaurant	French Restaurant	Falafel Restaurant	Indian Chinese Restaurant
2	Aundh	Indian	Restaurant	South Indian	Fast Food	Chinese	Vegetarian / Vegan	Middle Eastern	Asian Restaurant	English	Italian Restaurant

Now, with all this data, I ran a k-means clustering algorithm from the scikit-learn package which is an unsupervised machine learning algorithm. I tried a few different values to see the clustering and ended up with k to be 5.

Results

And here already comes the result:

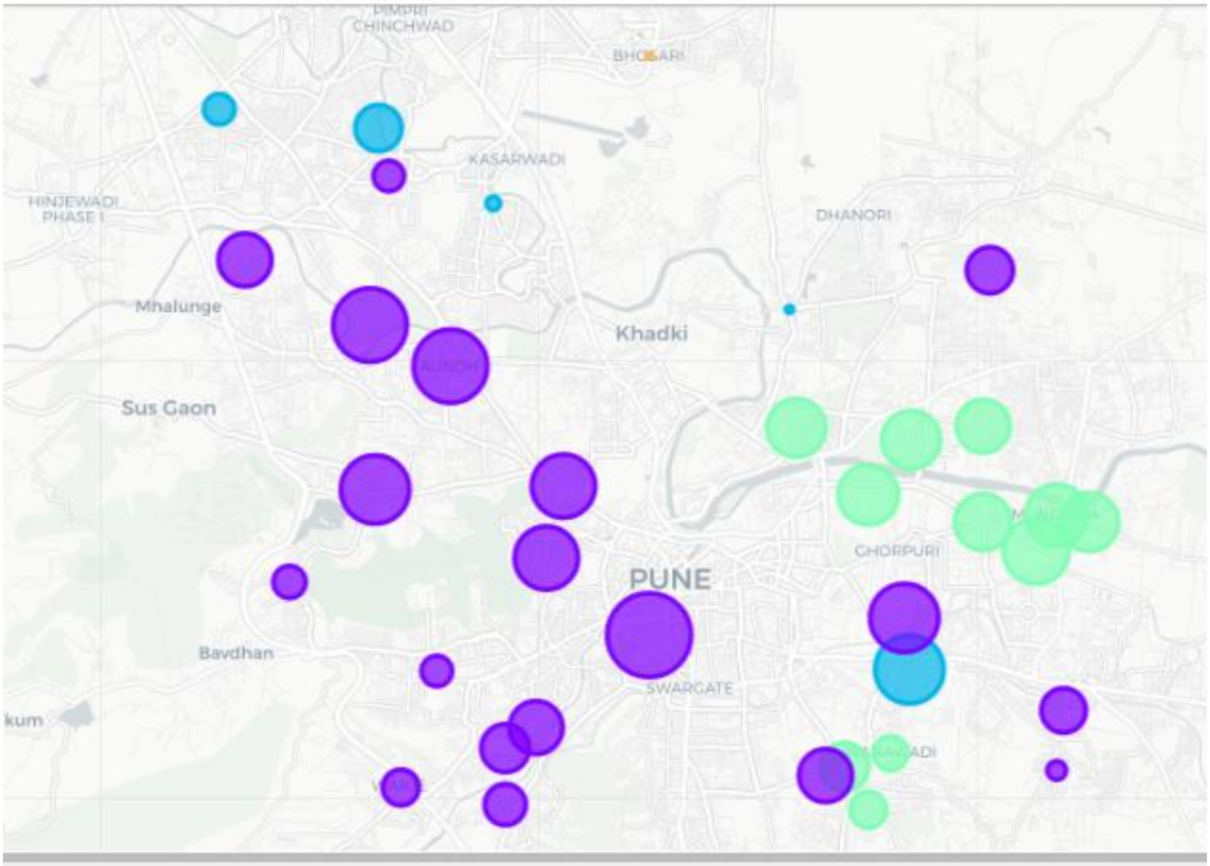
	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	Akurdi	Indian Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Dumpling Restaurant
1	1	Anand Nagar	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Seafood Restaurant	Restaurant	Chinese Restaurant	French Restaurant	Falafel Restaurant	Indian Chinese Restaurant
2	1	Aundh	Indian Restaurant	Restaurant	South Indian Restaurant	Fast Food Restaurant	Chinese Restaurant	Vegetarian / Vegan Restaurant	Middle Eastern Restaurant	Asian Restaurant	English Restaurant	Italian Restaurant
3	1	Balewadi	Indian Restaurant	Fast Food Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant	American Restaurant	Punjabi Restaurant	Seafood Restaurant	Asian Restaurant	French Restaurant
4	1	Bavdhan	Indian Restaurant	Asian Restaurant	Seafood Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant	Mediterranean Restaurant	Falafel Restaurant	Indian Chinese Restaurant	Greek Restaurant
5	4	Bhosari	Indian Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Italian Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Fast Food Restaurant	Falafel Restaurant	English Restaurant

I then decided to merge this with the realty prices.

	Neighborhood	Rate	Category	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 Most Common Venue	9th Most Common Venue	10th Most Common Venue
24	Handewadi	4677	1	18.485162	73.931680	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Chinese Restaurant	Restaurant	Greek Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Falafel Restaurant	Indian Chinese Restaurant
37	Lohegaon	4600	1	18.580330	73.918386	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Chinese Restaurant	Dumpling Restaurant	American Restaurant	Tex-Mex Restaurant
46	Narhe	4747	1	18.460143	73.826010	2	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Italian Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Falafel Restaurant	English Restaurant
0	Akurdi	5741	2	18.648642	73.764708	1	Indian Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Dumpling Restaurant
3	Anand Nagar	6350	2	18.478490	73.821326	1	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Seafood Restaurant	Restaurant	Chinese Restaurant	French Restaurant	Falafel Restaurant	Indian Chinese Restaurant

What we see in the table are the city districts and their most common venues, and they now have been assigned five different cluster labels.

A visual representation using folium was the logical next step!



As remarked earlier, the cuisine data is homogenous. But does throw up a few interesting observations!

Cluster 0 – Katraj and the Vegetarian cluster

	Neighborhood	Category	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 Most Common Venue	9th Most Common Venue	10th Most Common Venue
29	Katraj	2	0	Vegetarian / Vegan Restaurant	Restaurant	English Restaurant	Indian Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Fast Food Restaurant	Falafel Restaurant	Eastern European Restaurant

This is a bit lonesome and off to a corner of the city. I suspect this is an outlier.

Cluster 1 – Indian-Asian-Fast food Cluster

This is the largest grouping in Pune. While there is a smattering of international cuisines most of the food is Indian based. Interestingly, irrespective of affluence this combination spread of cuisine seems to be predominant

	Neighborhood	Category	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 Most Common Venue	9th Most Common Venue	10th Most Common Venue
24	Handewadi	1	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Chinese Restaurant	Restaurant	Greek Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Falafel Restaurant	Indian Chinese Restaurant
37	Lohegaon	1	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Chinese Restaurant	Dumpling Restaurant	American Restaurant	Tex-Mex Restaurant
0	Akurdi	2	1	Indian Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Dumpling Restaurant
3	Anand Nagar	2	1	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan	Asian Restaurant	Seafood Restaurant	Restaurant	Chinese Restaurant	French Restaurant	Falafel Restaurant	Indian Chinese

Cluster 2 - the Indian-North-Indian-Fast food Cluster

What differentiates this cluster from Cluster 1 is that we see some North Indian cuisine appear here. Also, from an economic perspective this cluster is very homogeneous and made up entirely of category 2.

	Neighborhood	Category	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 Most Common Venue	9th Most Common Venue	10th Most Common Venue
46	Narhe	Category1	2	Indian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Italian Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Falafel Restaurant	English Restaurant
22	Fatima Nagar	Category2	2	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant	Italian Restaurant	Asian Restaurant	Falafel Restaurant	Indian Chinese Restaurant	Middle Eastern Restaurant	American Restaurant	Parsi Restaurant
50	Pimple Gurav	Category2	2	Fast Food Restaurant	North Indian Restaurant	Indian Restaurant	Chinese Restaurant	Restaurant	English Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Falafel Restaurant
57	Rahatani	Category2	2	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant	American Restaurant	North Indian Restaurant	Andhra Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Punjabi Restaurant	Indian Chinese Restaurant
71	Vadgaon Budruk	Category2	2	Indian Restaurant	Fast Food Restaurant	Seafood Restaurant	Restaurant	Vegetarian / Vegan Restaurant	English Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	Falafel Restaurant
74	Vishrantwadi	Category2	2	Indian Restaurant	Chinese Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Falafel Restaurant	Italian Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant	English Restaurant
77	Wakad	Category2	2	Indian Restaurant	Fast Food Restaurant	North Indian Restaurant	Asian Restaurant	Chinese Restaurant	Multicuisine Indian Restaurant	English Restaurant	Indian Chinese Restaurant	Greek Restaurant	French Restaurant

Cluster 3 - the Indian-Italian Cluster

Cluster 3 sees a strong presence of international cuisine (Italian etc) in terms of food but is widespread in terms of property prices. Geographically it is clustered towards the North East of the city and is reasonably contiguous

	Neighborhood	Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
23	Hadapsar	2	3	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
30	Keshav Nagar	2	3	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
33	Kondhwa	2	3	Indian Restaurant	Asian Restaurant	Fast Food Restaurant
38	Lullanagar	2	3	Indian Restaurant	Fast Food Restaurant	Asian Restaurant
45	Mundhwa	2	3	Indian Restaurant	Italian Restaurant	Chinese Restaurant
75	Wadgaon Sheri	2	3	Indian Restaurant	Italian Restaurant	Restaurant
39	Magarpatta	3	3	Indian Restaurant	Italian Restaurant	Chinese Restaurant

Cluster 4 - Bhosari Cluster

I suspect this is an outlier as well. Bhosari seems to be inclined towards Thai food and is a small area of the city.

Conclusion

We achieved the goals presented at the outset

1. The property prices are inherently segregated based on distance from city centre. There is a weak co-relation to the types of cuisines and affluence of the areas, which is great news!
2. The food is predominantly Indian and derivative Indian (South Indian, Punjabi etc.). But interestingly there are distinct secondary clusters between other cuisines like Italian. As permitted by the first conclusion

Discussion

This is my first data science project and I have dabbled in programming after many decades. I have also leapfrogged from earlier more fundamental languages (I am looking at you C!) to new shiny Python, and the change is mindboggling. This assignment needed to be "assembled" rather than built. I decided to work with an Indian dataset and suspect that western data-sources like geopy are not very accurate when it comes to tier 2 Indian cities! This resulted in lot of effort to cleanse the data and that helped me learn the mechanics of the language better!