Battle of Neighborhoods

A study of realty and culinary variation across neighborhoods in Pune!

IBM DATA SCIENCE: Capstone Project

Sunil Sidhanty

The Business Problem

- Pune is a bustling metropolis and was voted most liveable city in India. It is traditionally a centre of Maharashtrian culture and "Oxford of the east". Thanks to strong Colleges, Automotive and Tier 1 network as well as a significant base for Indian IT service industry, Pune has a significant influx of migrants, both Indian and International.
- When I was new to the city, 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.

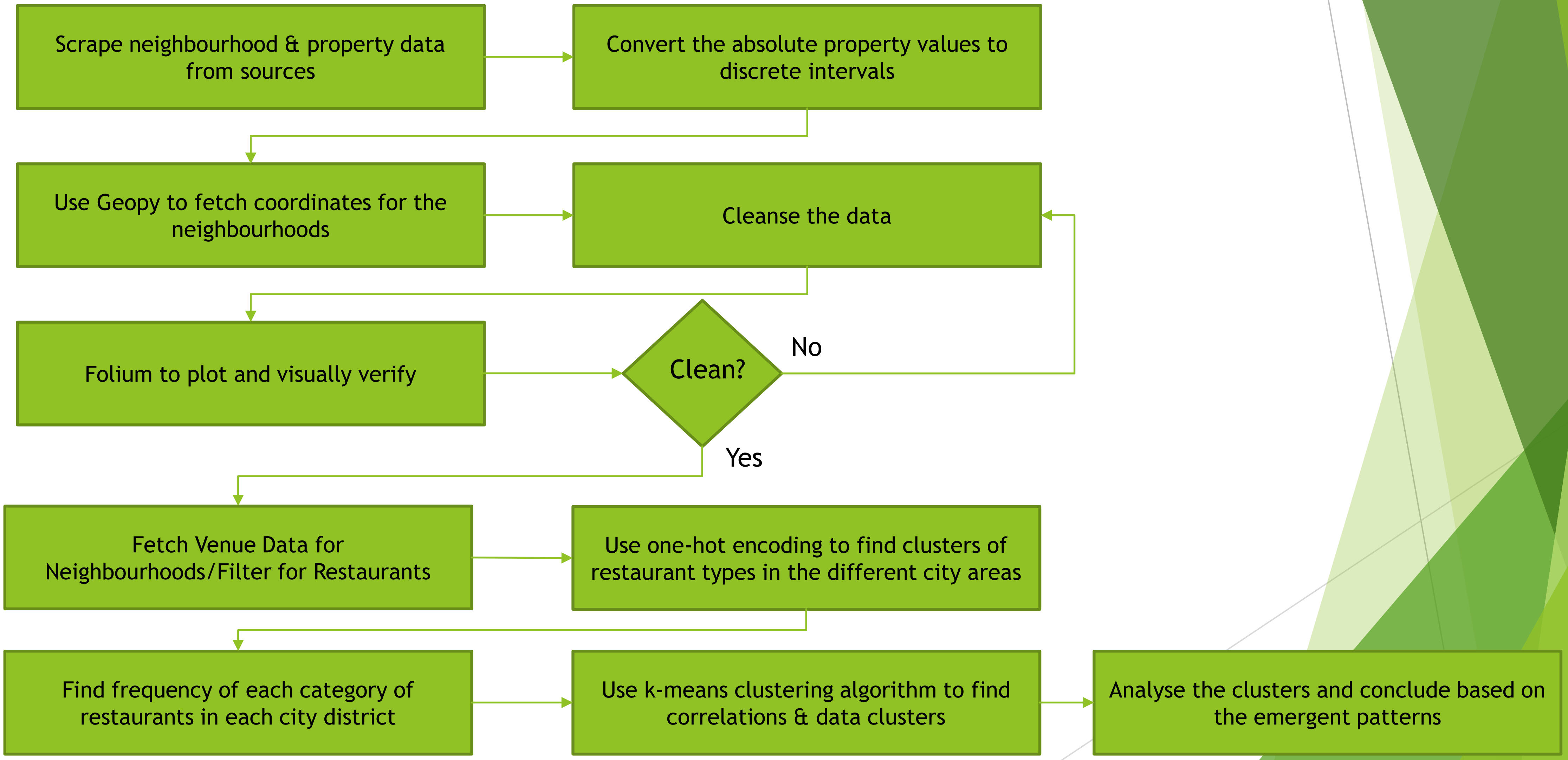
Property prices

Understand the trend and variety of neighbourhoods for **perspective house buyers**

Culinary Variation

Create a cluster of restaurants and cuisines for **Tourists and new international students**

My Methodology



Data Sources

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

I will use foursquare data about Pune venues and then filter for restaurants.

1

	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

I will then use geopy Nominatim to get latitude & longitude coordinates

2

	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

Foursquare API will be used to retrieve venues for the neighbourhood

3

Data Transformation & Cleansing

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

	a	Rate	Category
	di	3900	1
	k	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

Since we wanted to cluster the data eventually, I decided to convert the absolute property values into relative categories

1

2

```
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
```

Some Neighbourhoods that don't have valid coordinates. We filter these out

	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

Some neighbourhoods return values but these are erroneous. We filter these out by creating a bounding box of valid coordinates

3

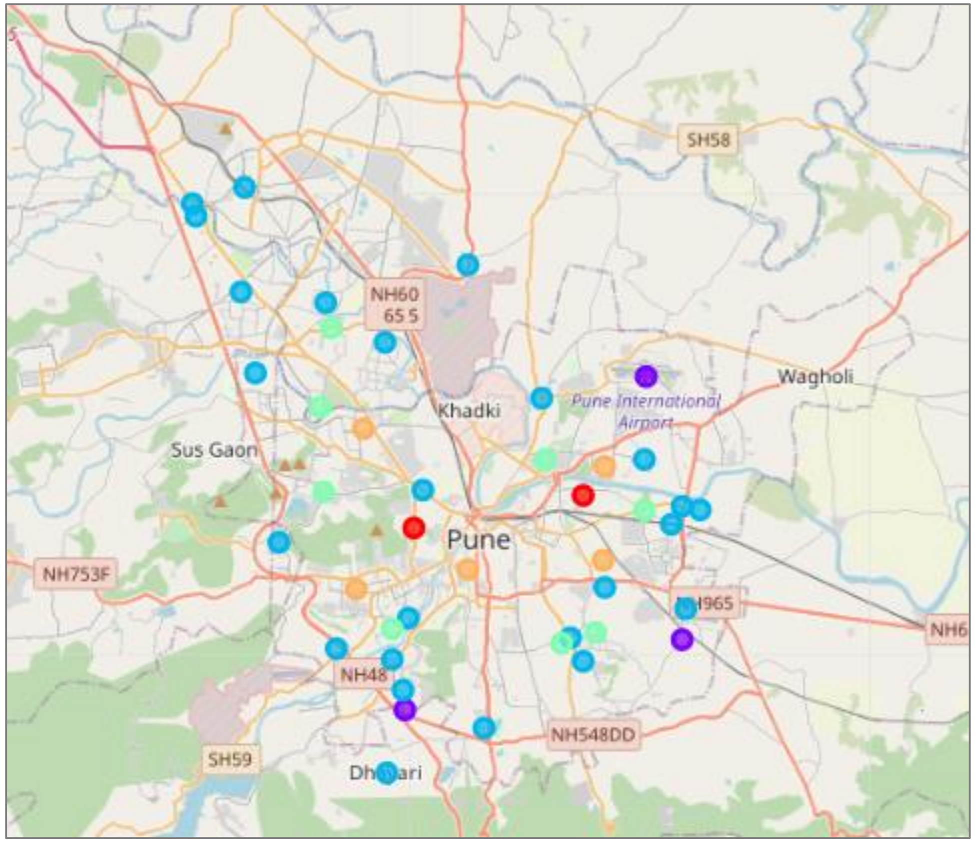
4

```
# Now we see a fresh error as the shapes of two dataframes are different.
# The suburb "Dhayari" does not have venue data. Lets drop it


pune_merged.drop(pune_merged[pune_merged['Area'] == "Dhayari"].index, inplace = True)
pune_merged.shape
```

One neighbourhood does not have any foursquare venues. We will filter this out

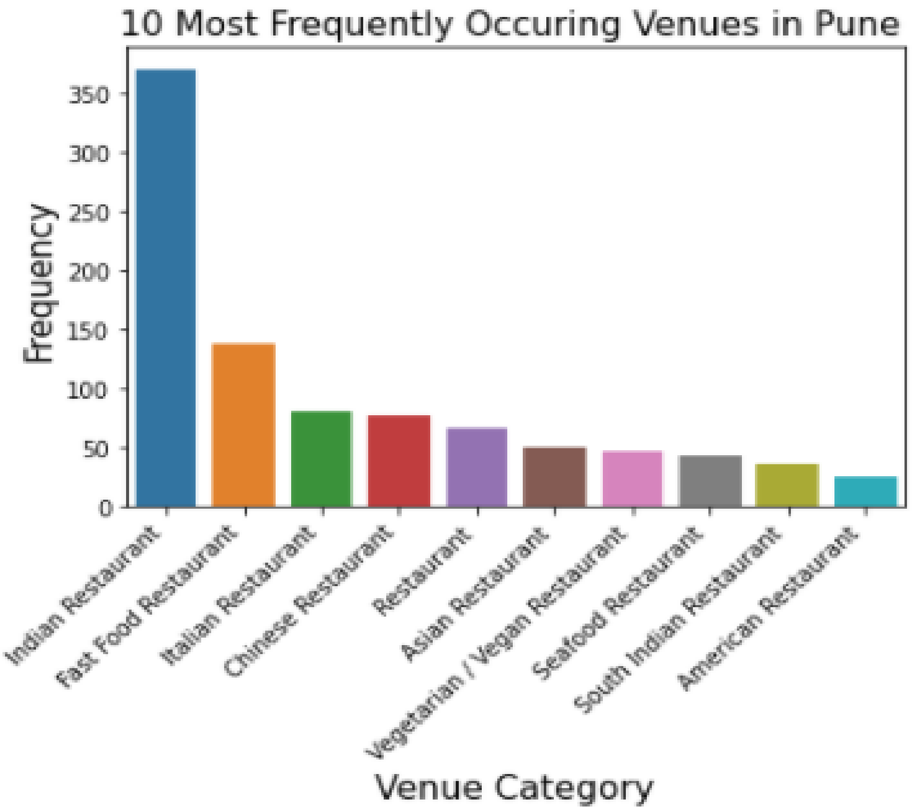
Geo-spatial distribution of property prices



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

Foursquare APIs give us a good view of the "Restaurant" views across Pune



Foursquare APIs returned **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.

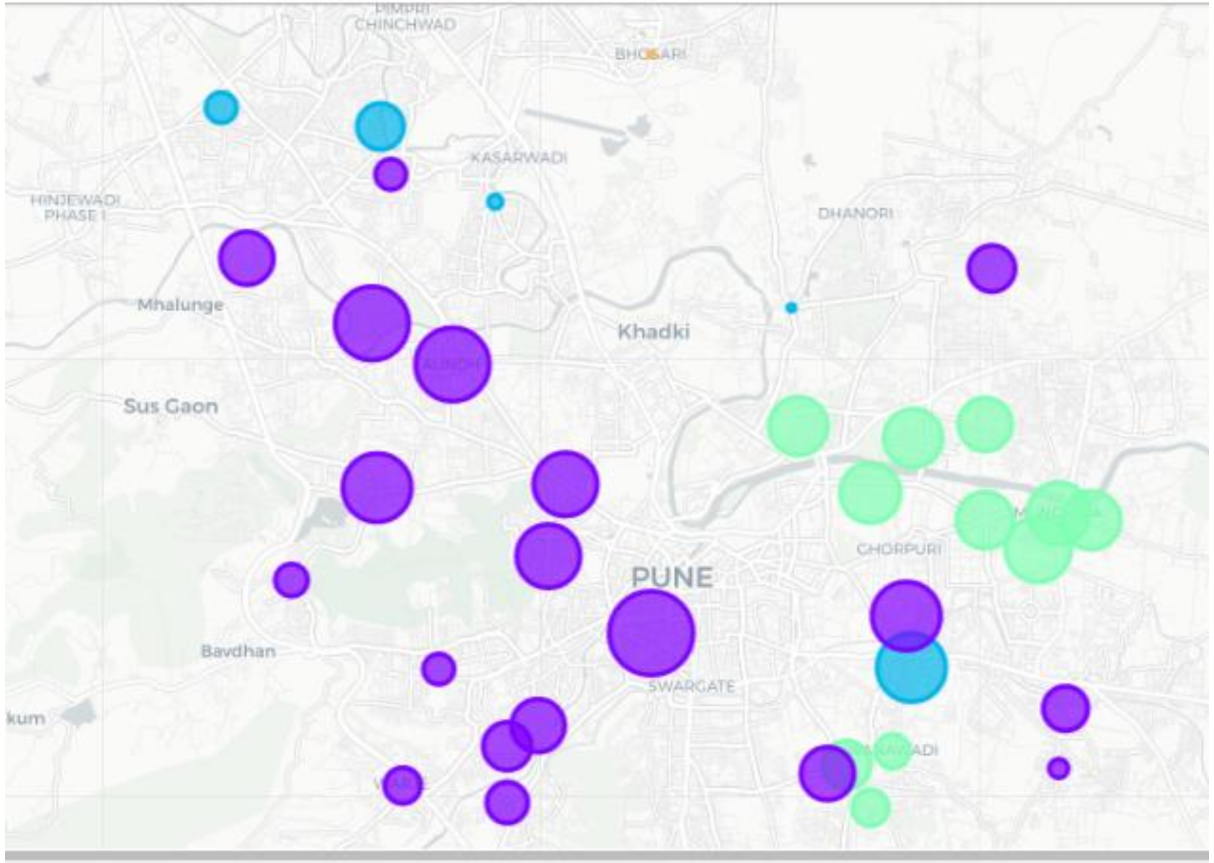
While Pune is a cosmopolitan city, its Indian roots are indeed very strong, and an overwhelming portion of the cuisine is Indian or Indianized!

Merging the frequency of venues with Realty property rates

	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

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.

Plotting the k-means clusters of like clusters



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

Discussions

- 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
- Foursquare API data also seems a bit dated. While I did a google match to verify a few random data points to verify the data, but I suspect we can get better datasets commercially
- I also plan to explore other data sets published here http://opendata.punecorporation.org/Citizen/CitizenDatasets/Index
- My original intent was to create a choropleth map, but did not find accurate data shapes

Conclusions

Cluster 0 - Katraj and the Vegetarian cluster	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
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.
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
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.

- 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!
- 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