

Capstone Project - The Battle of the Neighborhoods

Clustering of neighborhoods in Mumbai

Introduction: Business Problem

In this project I will try to segment the neighborhoods of Mumbai using K-means clustering.

Since there is no proper data on the neighborhoods in Mumbai, I took the list from Wikipedia and scraped it with Beautiful soup and converted it into a table, populating the location attributes separately. For the missing values, I populated a separate csv file which I used in the program using the Dropbox API. Then the Foursquare API was used to get details on common venues in the city.

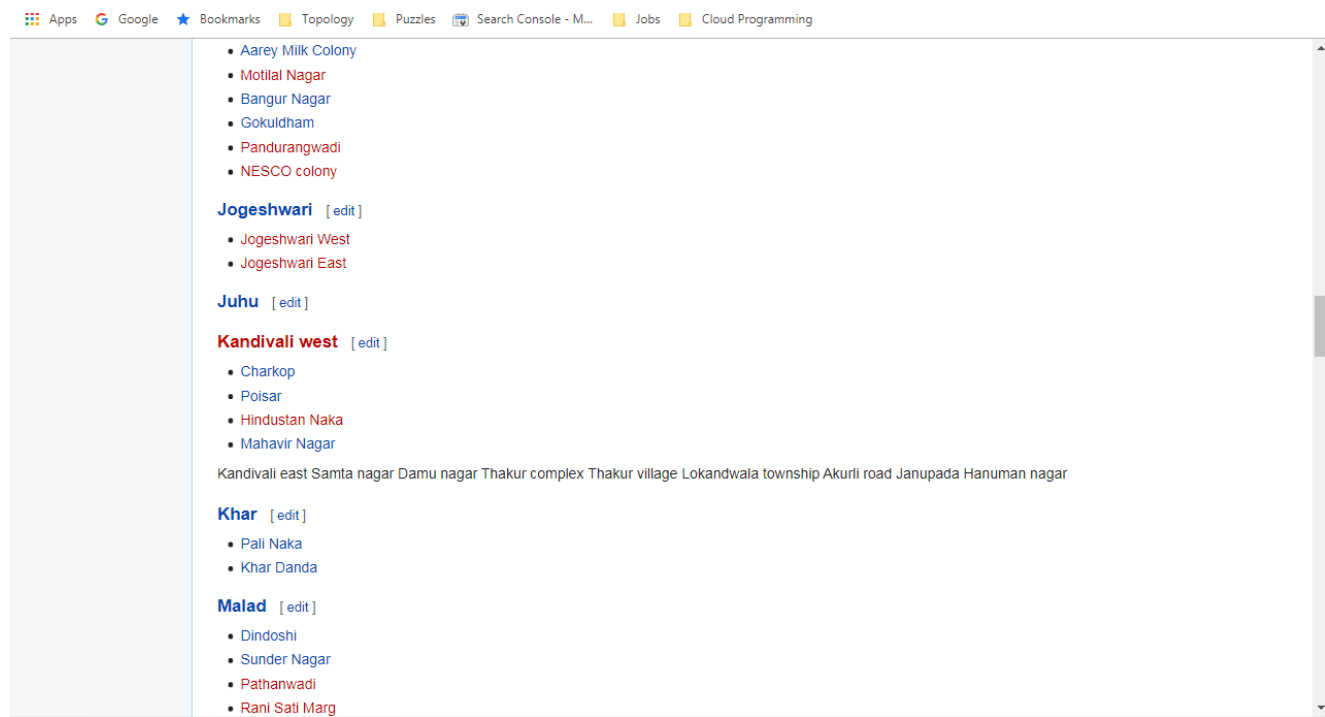
Data

Data sources used: Wikipedia page for list of neighborhoods, type of venues from Foursquare API and location co-ordinates from Geocoder

Methodology

Data Preparation

The list on Wikipedia looked like this. The data had clear categories apart from a couple of issues and there were 200 odd items on the list. The first task was to convert the list to a table.

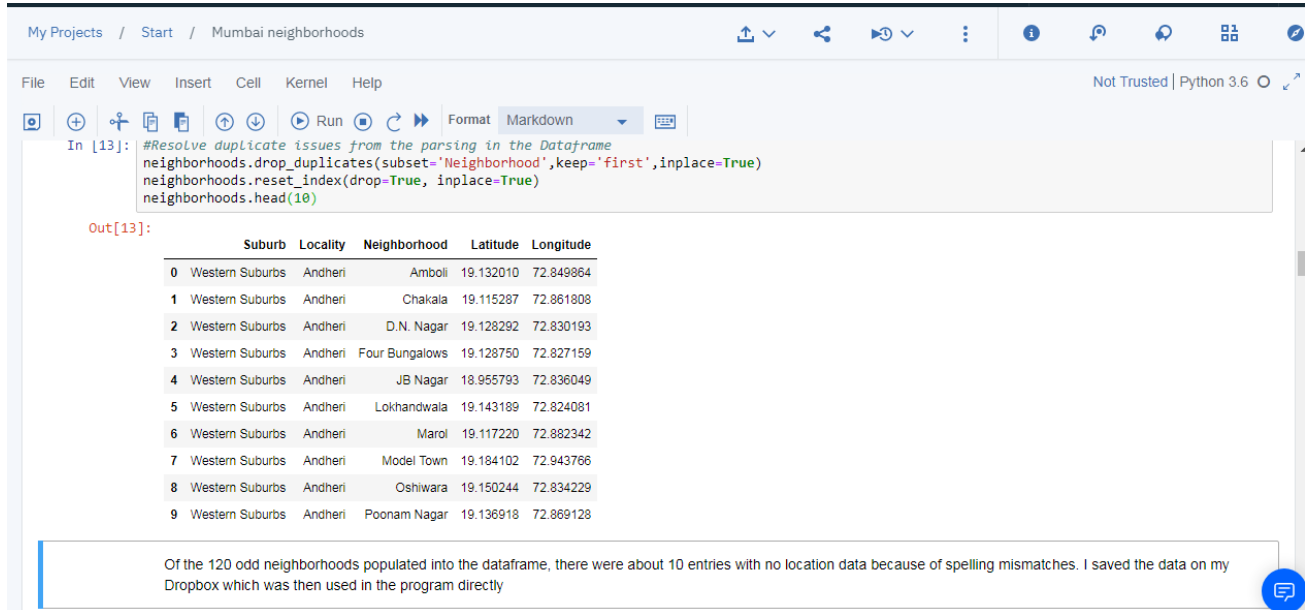


A nested for loop with break statements worked to scrape the page and ultimately produce a data-frame as below.

There were a few missing locations with the output from the Geocoder. I populated the missing values and the indexes on a csv file and stored it in my Dropbox account. Then I created a Dropbox instance which directly read the file from my account and updated the data-frame.

There was an issue with the Geocoder output as every time the function was called, different sets of neighborhoods were missing. So the file was saved to pickle as location data doesn't change.

The final output was a table with three levels of neighborhood attribute.



The screenshot shows a Jupyter Notebook interface with the following content:

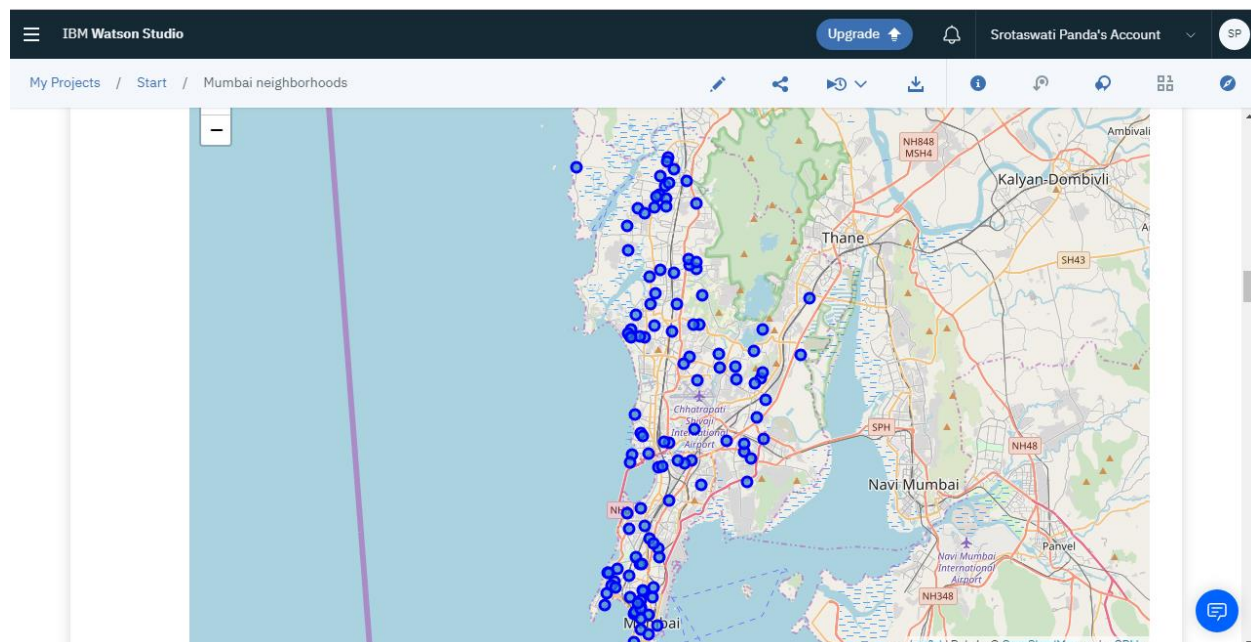
```
In [13]: #Resolve duplicate issues from the parsing in the Dataframe
neighborhoods.drop_duplicates(subset='Neighborhood',keep='first',inplace=True)
neighborhoods.reset_index(drop=True, inplace=True)
neighborhoods.head(10)
```

Out[13]:

	Suburb	Locality	Neighborhood	Latitude	Longitude
0	Western Suburbs	Andheri	Amboli	19.132010	72.849864
1	Western Suburbs	Andheri	Chakala	19.115287	72.861808
2	Western Suburbs	Andheri	D.N. Nagar	19.128292	72.830193
3	Western Suburbs	Andheri	Four Bungalows	19.128750	72.827159
4	Western Suburbs	Andheri	JB Nagar	18.955793	72.836049
5	Western Suburbs	Andheri	Lokhandwala	19.143189	72.824081
6	Western Suburbs	Andheri	Marol	19.117220	72.882342
7	Western Suburbs	Andheri	Model Town	19.184102	72.943766
8	Western Suburbs	Andheri	Oshiwara	19.150244	72.834229
9	Western Suburbs	Andheri	Poonam Nagar	19.136918	72.869128

Of the 120 odd neighborhoods populated into the dataframe, there were about 10 entries with no location data because of spelling mismatches. I saved the data on my Dropbox which was then used in the program directly

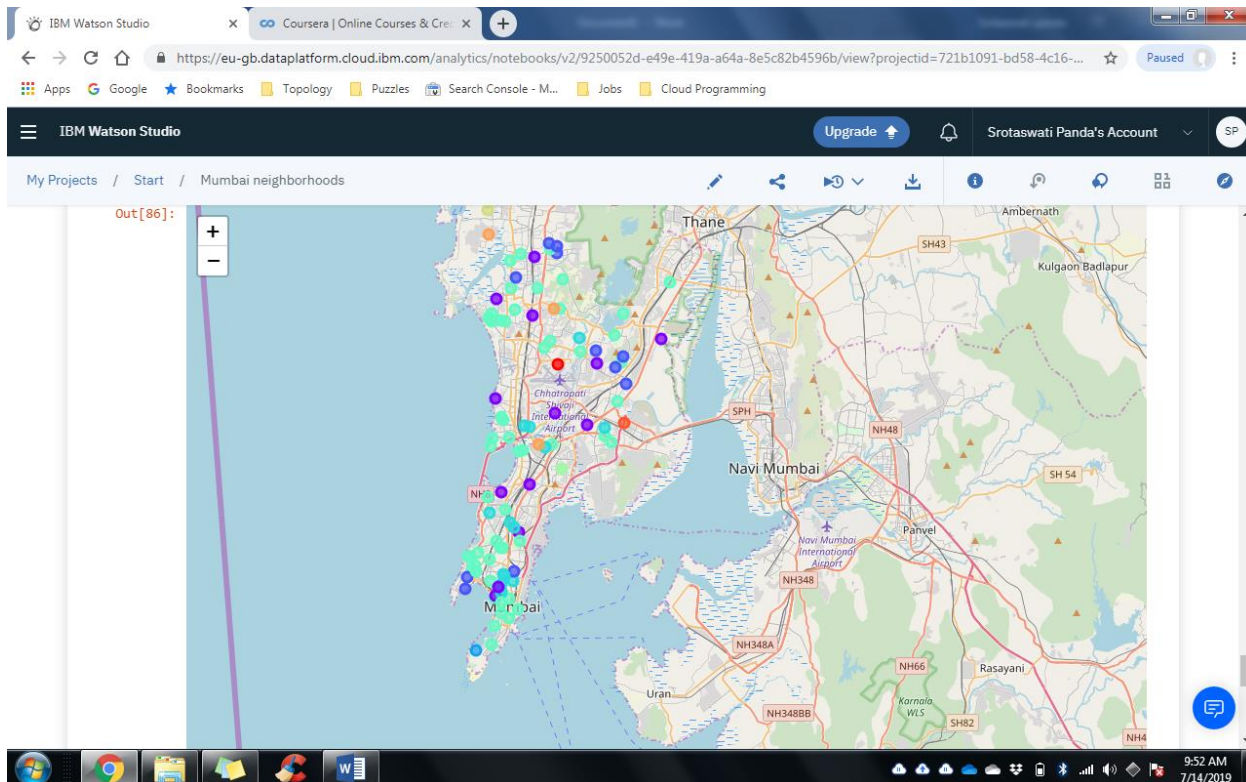
Using the tabular data, a map showing the neighborhoods in Mumbai was then straightforward



K-Means clustering

The foursquare API was then used to get the data on venues in the city. There were about 180 different categories of various venues. There were a couple of issues with this step as some neighborhoods had no data and some venue category names were ambiguous. The foursquare and main table were cleaned to remove any clashes.

Using one-hot encoding, dummy variables of the venue categories were created, and 5 clusters were selected. 3 clusters simply divided the data into whether an Indian restaurant was in the vicinity or not and an outlier cluster with the airport.



Results and Discussion

Red clusters or Cluster 0 is an outlier neighborhood which has an airport

Purple clusters or Cluster 1, Blue clusters or Cluster 4 are neighborhoods which have more Indian, fast-food, falafel restaurants and farmers markets among the most common venues. These neighborhoods are mostly in West Mumbai and probably cover neighborhoods with high disposable incomes.

Blue clusters or Cluster 2 neighborhoods have fields, gyms, more parks, pubs. These neighborhoods cover mostly Central Mumbai. The proximity of these venues suggests that accessibility to residential neighborhoods in East and West Mumbai may have been a prime reason for these venue locations.

Green clusters or Cluster 1 neighborhoods are closer to beaches, parks, Italian, Chinese, Vegan restaurants