**Coursera Capstone**

## Capstone Project - The Battle of Neighborhoods

**Data Description**

**By : Tushar Nirmal**

# Data Source -

Following sources of data will be used for getting the required information for our capstone project : -

1. **The list of neighborhoods of Pune, India from Wikipedia url :**

**“https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Pune"**

1. **The coordinates ( latitude , longitude ) of the neighborhoods of Pune from Open street Map APIs from url –**

**“http://nominatim.openstreetmap.org”**

1. **From foursquare.com , we need to determine the proximity of various amenities as per requirement .Mainly we will need the following data :**

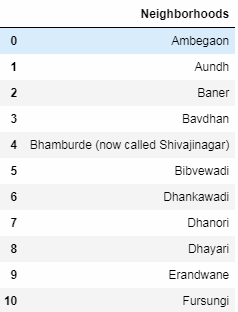
* **Details of all the venues of Pune neighborhoods within 1000 meters of the coordinates**
* **Details of office venues of Pune neighborhoods within 1000 meters of the coordinates .**
* **Details of high school venues of Pune neighborhoods within 1000 meters of the coordinates .**
* **Details of university venues of Pune neighborhoods within 1000 meters of the coordinates .**

**Data Description :**

How we will use the above data in order to provide solution to our business problem ?

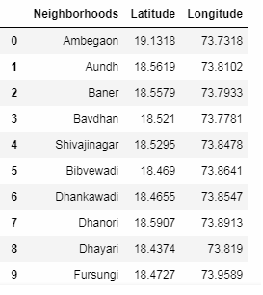
* Step – 1

Firstly , we need to have the name of all the neighborhoods of Pune in order to find the best one for opening a restaurant . Same can be obtained from a wikipedia url which contains all the names. These names are fetched using beautifulsoup package and is stored in a dataframe . After successfully fetching the neighborhoods name from Wikipedia , the dataframe will look as :



* Step – 2

The dataframe obtained above contains names of neighborhoods of Pune . To get the latitudes and longitudes of these neighborhoods , we will use the Open street Map APIs . Once we get all the coordinates , we will add these coordinates to our dataframe . There are few neighborhoods for which we are not able to fetch coordinates . We will manually add these coordinates by searching on google . Once that is complete , dataframe will look as :



* Step – 3

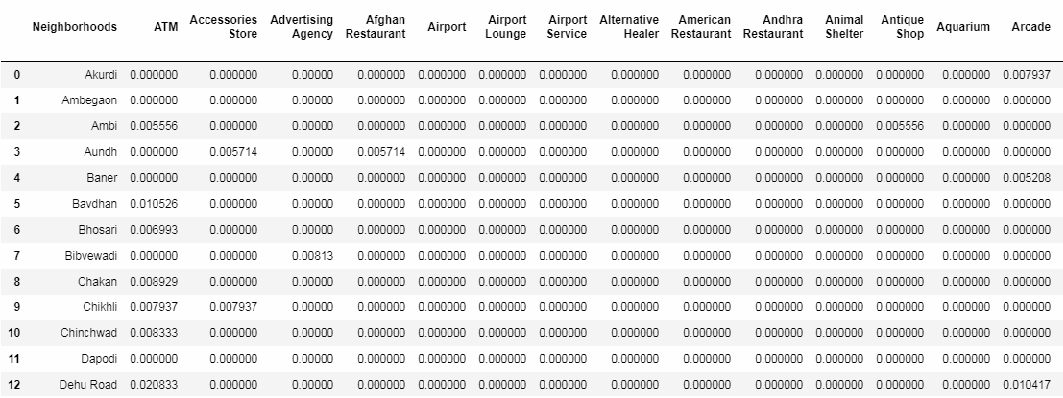
Using Foursquare.com , we will find proximity of various amenities from each neighbourhood of Pune . The resulting dataframe will contain following columns – neighborhood name , neighborhood latitude and longitude , venue name ,venue latitude and longitude and venue category .



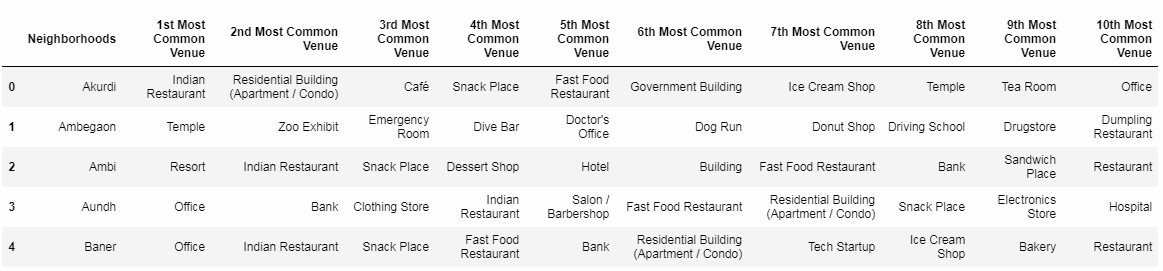
* Step – 4

We want to know the which category is the most common in each neighborhood . To know the top most category for a neighbourhood , we will do it in two sub steps :

* + - Firstly , we will group by the dataframe by neighbourhood and convert each category in the column “Venue Category” from categorical to numerical (0 if category is not present for that neighborhood and 1 if present ) and then taking the mean for all the columns for each neighborhood . As a result , the resulting dataframe indicates the frequency of that venue’s category for that neighborhood . The greater the value , the greater the number of entries for that category . The dataframe will look like –

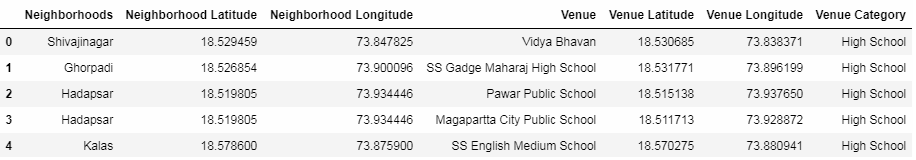


* + - Secondly , for each neighborhood , we will sort the row in the descending order in order to get the most common venues (we have written a function which returns the values of the most common venues for each neighborhood). Therefore the resulting dataframe will look like –



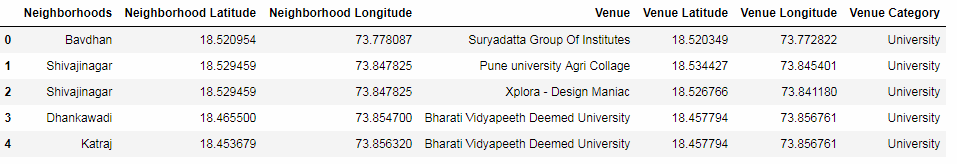
* Step – 5

Using Foursquare.com , we will find proximity of all the high schools from all the neighborhood of Pune . The resulting dataframe will have the coordinates of high schools along with coordinates of neighborhood –



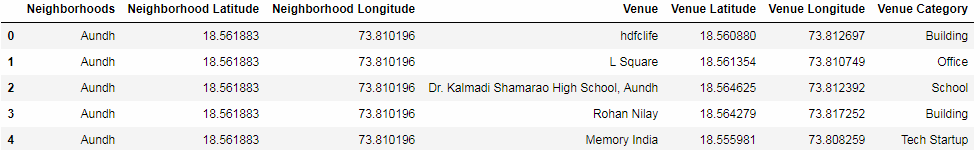
* Step-6

Using Foursquare.com , we will find proximity of all the university from all the neighborhood of Pune . The resulting dataframe will have the coordinates of universities along with coordinates of neighborhood –



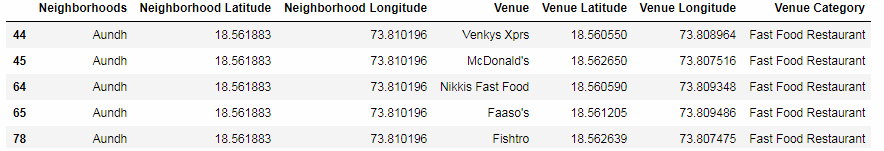
* Step – 7

Using Foursquare.com , we will find proximity of all the offices from all the neighborhood of Pune . The resulting dataframe will have the coordinates of offices along with coordinates of neighborhood –



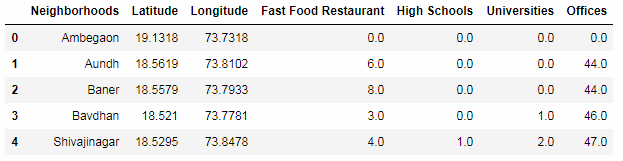
* Step -8

From the dataframe obtained in step-3 ,we fetch all the rows which has venue category as fast food restaurant . Thus the resulting dataframe will look as below –



* Step -9

We write a function to add columns (the value of each element is the count of the category for that neighborhood ) of above dataframes into the original dataframe (dataframe which contains coordinates and names of neighborhoods . )



* Step-10

When we multiply these rows with their weights and sum each row , we get a score for each neighbourhood. The neighbourhood with the maximum score is our desired neighborhood for opening the Fast Food Restaurant .

