Capstone Project

Battle of Neighborhood - Harrow (London)

Final Project: Neighborhood Recommendation to start up Indian Restaurant based on K-Means Algorithm

Version 1.0

Author: Unesh Alex, Rajendran Rayappan

# Introduction

When you consider the demography of London, you will see a rainbow — a community of multi-color multi-ethnic groups. From the South of the River Thames and across to the other side of it, the diversity of London is the strength every business from across the walks of life has taken opportunity of Growth.

In London Borough of Harrow, 63.8% of its population from the BME (Black and Minority Ethnic) communities, with the largest group being of Indian ethnicity, 26.4% (specifically those from Gujarat and South India). But something that really lack its essence is the availability of a high-end Indian Restaurant that caters beyond the food but comes with the ambience of the service. With the arrays of restaurants in Harrow, there are peculiarly lacking the finest restaurant that can answer the call within certain areas of interest when it comes to Indian cuisines and flavors.

To get a very good location for a restaurant that meet this need, the Harrow area, which has dense Indian population is explored through clustering and segmentation based on the Borough Coordinates & Indian Population in Boroughs and proximity to supplies.

This project will rely on public data from Wikipedia and Foursquare. Libraries — For convenience, all the libraries are presented at the beginning

# Data Section

## data SETS: WIKI Pages

**Wiki page – List of London boroughs**

The London Area consists of 32 Boroughs and the “City of London”. Our data – Boroughs, area, Latitude, longitude, etc., will be from the wiki page : <https://en.wikipedia.org/wiki/List_of_London_boroughs>

The focus of this project will be the neighborhoods are that are within London Borough of Harrow

**Wiki page – Indian community of London**

The percentage of Indian population in various London Boroughs are collected from the wiki page: https://en.wikipedia.org/wiki/Indian\_community\_of\_London

**Wiki page – List of areas of London**

The London area data is scrapped from the wiki - List\_of\_areas\_of\_London: https://en.wikipedia.org/wiki/List\_of\_areas\_of\_London

The BeautifulSoup package is used to scrap the needed data from Wikipedia.

## data sets: Foursquare API

The Foursquare API will be used to obtain the Harrow Area venues for the geographical location data. These will be used to explore the neighborhoods of Harrow accordingly.

The venues within the neighborhoods of Harrow like the area’s restaurants and proximity to amenities would be correlated. Also, accessibility and ease of supplies would be considered as it relates to venues.

## data sets: Geocoder Package

In obtaining the location data of the locations, the Geocoder package is used with the ArcGIS geocoder to obtain the latitude and longitude of the needed locations.

These will help to create a new data frame that will be used subsequently for the Harrow areas.

## Python Library Files

The Below Libraries are used:

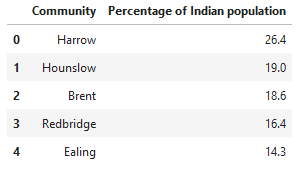
* Pandas - Library for Data Analysis
* NumPy – Library to handle data in a vectorized manner
* JSON – Library to handle JSON files
* Folium – Map rendering Library
* Matplotlib – Python Plotting Module
* Geopy – To retrieve Location Data
* Requests – Library to handle http requests
* Sklearn – Python machine learning Library

## Folium Library

Python visualization library is used to visualize the neighborhoods cluster distribution of Chicago city over an interactive leaflet map. Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python’s scientific libraries Pandas, NumPy and Scikit-learn.

## Data Preparation

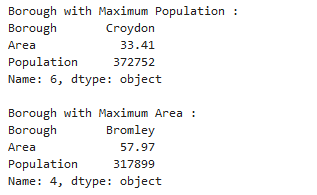
**London Indian Population:**



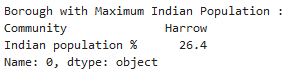
**London boroughs Data from Wiki Page**



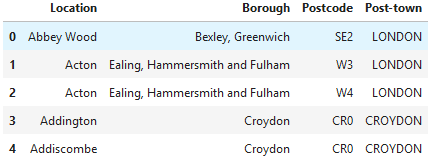
**London borough with Max Area & Population**



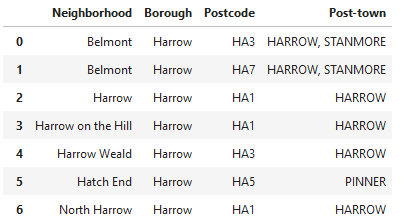
**London borough with Max Indian Population**



**London borough & Neighborhood Data exploration**

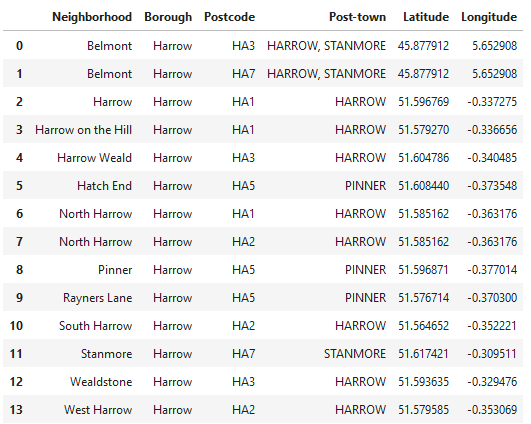


**Data Clean-Up**



**Geocoder - to obtain Coordinate for each area**

We are using geocoder to store the location data — latitude and longitude as follows. The obtained coordinates are then joined to create new data frame.



# Methodology

## Data Exploration

**Single Neighborhood**

 An initial exploration of a single Neighborhood within the London area was done to examine the Foursquare workability. The West Harrow is used for our Single Neighborhood exploration.



Let’s explore the top 100 venues that are within a 2000 meters radius of West Harrow. And then, let’s create the GET request URL, and then the URL is named.

Results are shown below:



The result is then cleaned up from json to a structured **pandas** data frame as shown below:



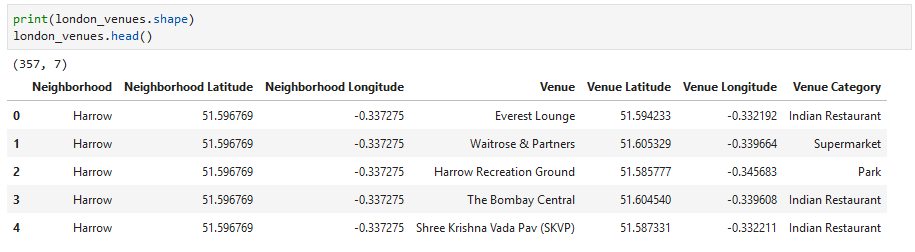
**Top 5 venues in West Harrow**



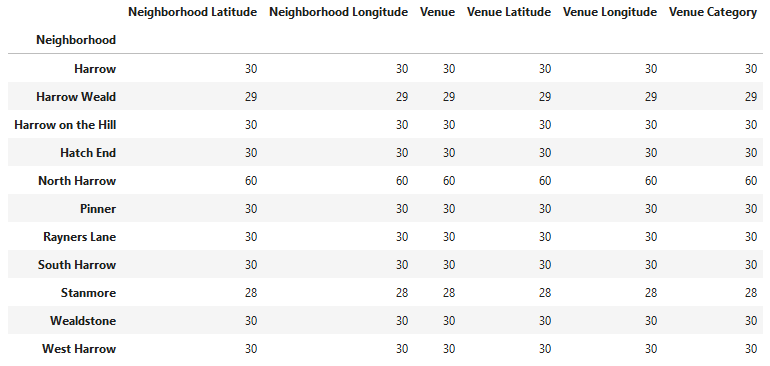
Interestingly, we have Indian restaurant as top venue in West Harrow.

**Multiple Neighborhood**

Now let’s explore (Multiple) Neighborhoods in the Harrow area. To do this, the function getNearbyVenues is used and it's created to repeat the same process for all neighborhoods.



The number of venues returned for each neighborhood is then explored as follows:



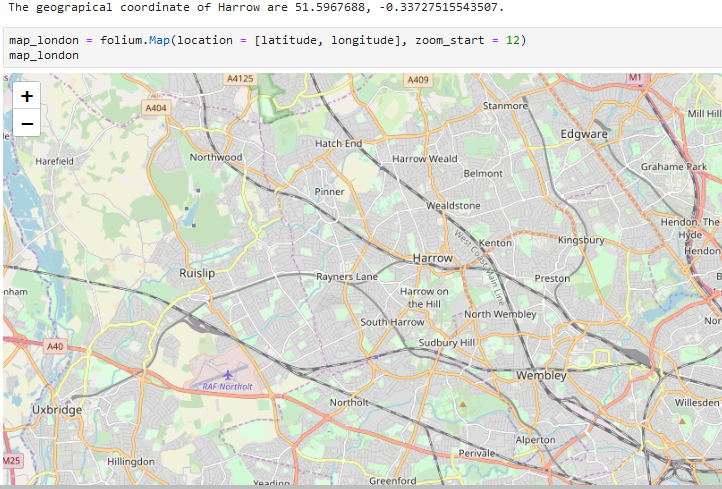
There are 48 unique categories.



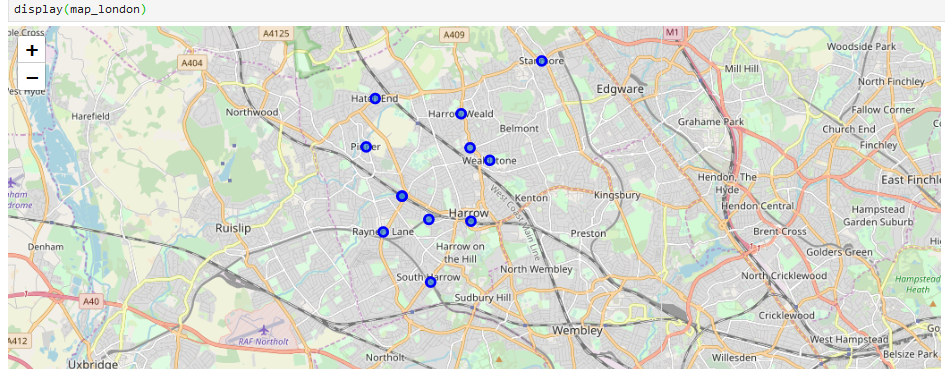
## Clustering

For this section, the neighborhoods in Harrow will be clustered based on the processed data obtained above.

Map Visualization — Using the geopy library, the latitude and longitude values of London is obtained.



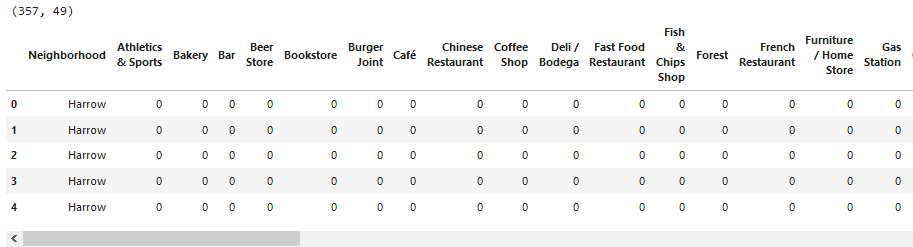
Harrow neighborhoods are then superimposed on top as shown below, still using the ‘folium’ library. Please note due to the location of the South East London, you might need to zoom to see the superimposed areas.



**Analyzing Each Neighborhood**

In this section, the objective is to check and explore the venues in each neighborhood.

**One Hot Encoding**



**Check Indian Restaurants**



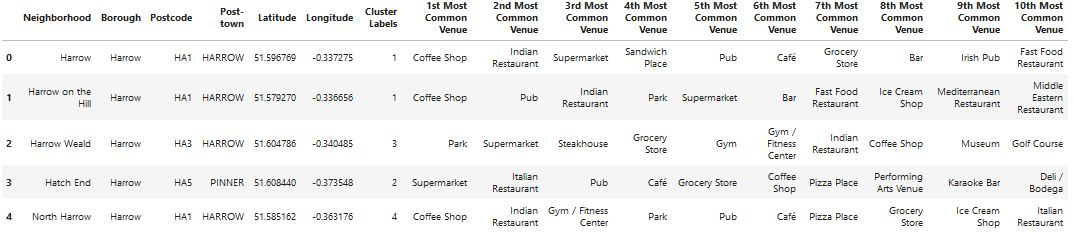
**Regrouping and Category Statistics**

We create a new panda data frame with 10 most common venues as shown below:

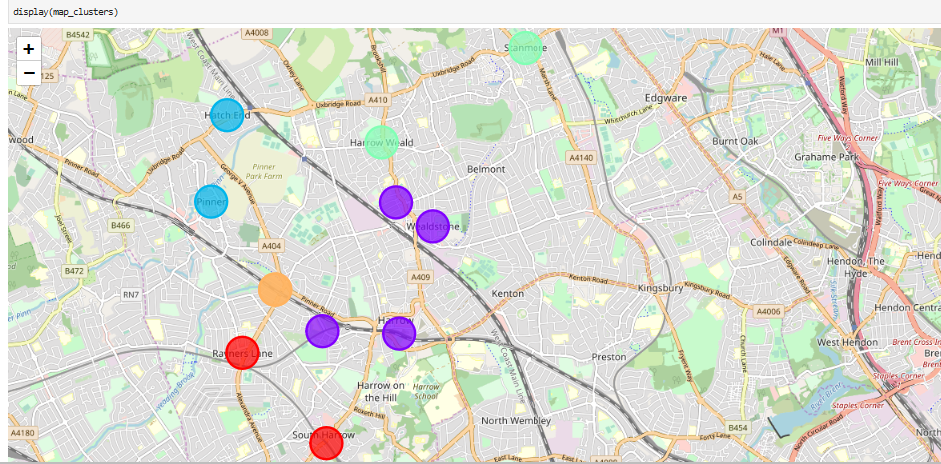


**Clustering of Neighborhoods**

We create the grouped clustering for the neighborhood as shown below:



**Visualizing the Resulting Clusters**



The individual clusters data frame is mentioned below:

# Cluster 1



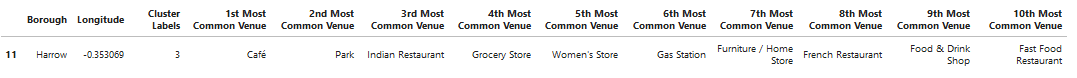
# Cluster 2



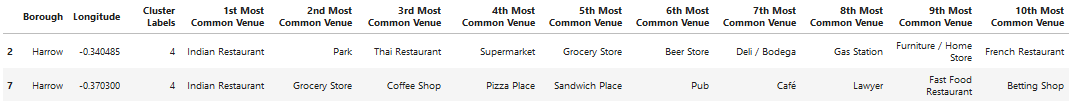
# Cluster 3



# Cluster 4



# Cluster 5



# Discussion and Conclusion

**The following are the highlights of Harrow Clusters:**

1. Coffee Shops, Indian restaurants & Italian restaurants are popular in the Harrow.
2. As for Indian restaurants, we can see the neighborhoods Rayner’s Lane, Harrow Weald and South Harrow which top in the Indians visited venues.
3. Also, to note, Harrow & Wealdstone are having second top Indians visited venues which are close to Rayner’s Lanes. But Harrow Weald is comparatively far from said 3 neighborhoods.

Considering all the analysis with the available Data - Below are the neighborhoods in the priority order to open Indian Restaurants:

1. Rayner’s Lane
2. Harrow Weald
3. South Harrow

Below are the few other factors in terms of data would have given better results with more insight into the best location.

1. Real Estate
2. Crime data
3. Income per capita
4. Traffic Data
5. More venues exploration with the Foursquare
6. Ratings and feedback of the current restaurants within the clusters

Thanks & Regards,

Unesh Alex, Rajendran Rayappan

[yuneshalex.r@gmail.com](mailto:yuneshalex.r@gmail.com)