TIONG GHEE TAN

Abstract

The project uses data modeling cluster techniques to identify unique venues within a neighborhood that can be subsequently used for accommodation or travel recommendations.

Applied Data Science CapStone Project

Battle of the Neighbourhood – London Edition

Contents

[Introduction 2](#_Toc31276878)

[Overview 2](#_Toc31276879)

[The Solution 2](#_Toc31276880)

[The Data 2](#_Toc31276881)

[Methodology 3](#_Toc31276882)

[Retrieval of Postal Code and Neighbourhood 3](#_Toc31276883)

[Retrieval of Latitude and Longitude Information 3](#_Toc31276884)

[Retrieval of Venue Information 4](#_Toc31276885)

[Data Transformation 5](#_Toc31276886)

[Model and Evaluation 5](#_Toc31276887)

[Results 7](#_Toc31276888)

[Deployment 9](#_Toc31276889)

[References 11](#_Toc31276890)

# Introduction

## Overview

I visited London recently and found that the accommodations are very expensive as compared to the other European countries. When I was choosing an Airbnb apartment, I had difficulty choosing which area to stay. With this project, I hope to be able to help other tourists, like myself, choose a neighbourhood to stay in based on selected preferences. For example, Chinese restaurants, theatres, museums, etc. As there are too many [neighborhoods in London](https://en.wikipedia.org/wiki/London_postal_district), I will just focus on Western Districts.  
  
The targeted customers for my project will be tourists who are new to London, tourists who are trying to identify a new neighbourhood to stay, or simply for tourists who are looking for a new neighbourhood to visit.  
  
Using the solution, the customers select a list of venues based on their preferences. Using the venues selected, the customers will be recommended neighbourhood(s) to visit.

## The Solution

The solution will use clustering algorithm, Kmeans, to form clusters based on venues. Subsequently, customers'' selections will be scored against this model to identify cluster that the belongs to. The neighbourhoods belonging to that cluster will be recommended to the customer.

## The Data

The following data sources are used for the project:

* [Wiki List of postal codes of Western London](https://en.wikipedia.org/wiki/W_postcode_area). From this Wiki page, I will be able to extract the list of Postal code and Neighborhood in Western London.
* [Google Map Platform](https://developers.google.com/maps/documentation/javascript/geocoding). Using the Google Map API, I will be able to retrieve Latitude and Longtitude Data based on Postal Code.
* [Foursquare](http://localhost:8888/notebooks/Desktop/projects/CourseaCapstone/www.foursquare.com). Finally, using the Lat and Long value, I will be able to retrieve popular venues within a neighborhood.

These data will be merged to form a Analytic dataset that will contain top 10 venues for each neighbourhood. This dataset will be the basis for running Kmeans.

# Methodology

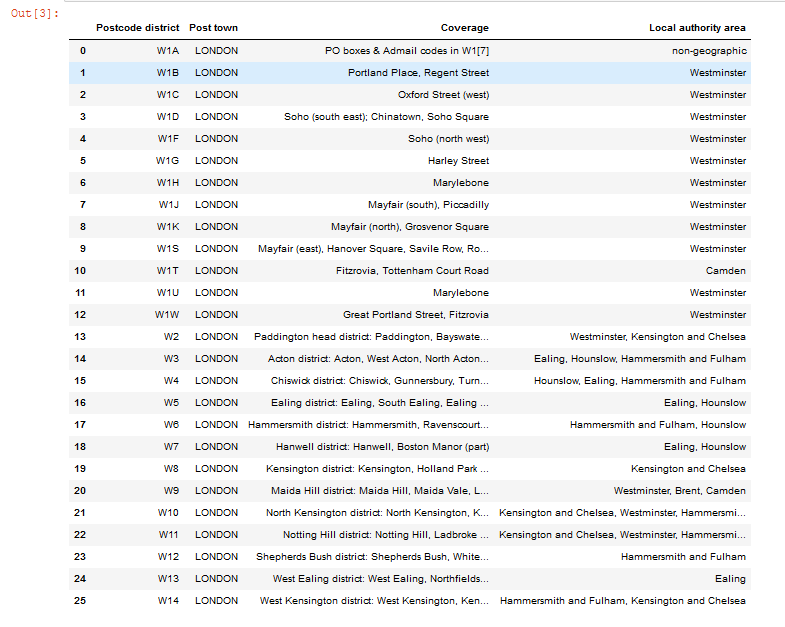
For this project, we need to associate neighbourhood data with venues. To achieve this, data from Wiki, Google Map, and Foursquare will be used. This section explains how the data is extracted and transformed.

## Retrieval of Postal Code and Neighbourhood

The Wiki page, [Wiki List of postal codes of Western London](https://en.wikipedia.org/wiki/W_postcode_area), provided Postal code and Neighbourhood information for Western London. Using the Python libraries BeautifulSoup and Tabulate, the information is extraction and stored into a dataframe. “W1A” is removed as it is used for PO boxes and Admail. The two columns that contains required are “Postcode district” and “Coverage”, they have been renamed to “PostalCode” and “Neighborhood” respectively.

The final output for this pre-processing is a table with twenty-five postal code.

*Scrapped Data from Wiki*



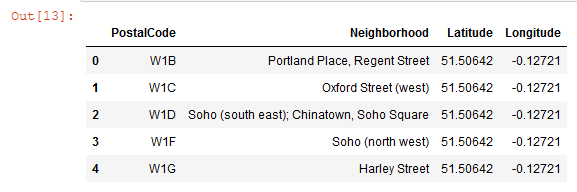
## Retrieval of Latitude and Longitude Information

The Latitude and Longitude information are provided by Google Map Platform. To retrieve the information, this API is used,

g = geocoder.arcgis('{}, United Kingom, London'.format(postal\_code))

The call is rather unstable so to mitigate any potential problem, the API is called repeatedly until the Latitude and Longitude values are returned.

*Latitude and Longitude retrieved based on Postal Code*

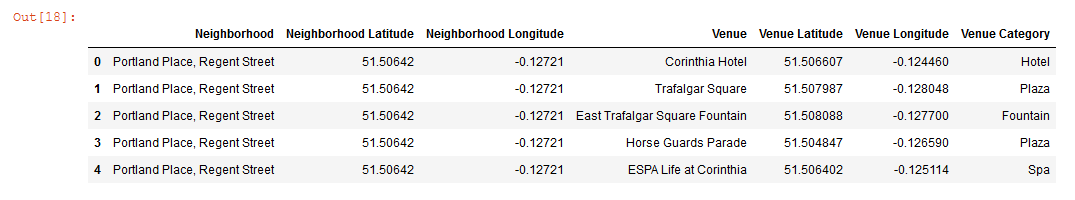


## Retrieval of Venue Information

Finally, the data required for this project is venue information provided by Foursquare API. To make use of the API, an account must be registered in [www.foursquare.com](http://www.foursquare.com).

For the purpose of this project, the radius of search has been set to 5oo meters and the venues returned is limited to 100.

*Venue Category from Foursquare API*



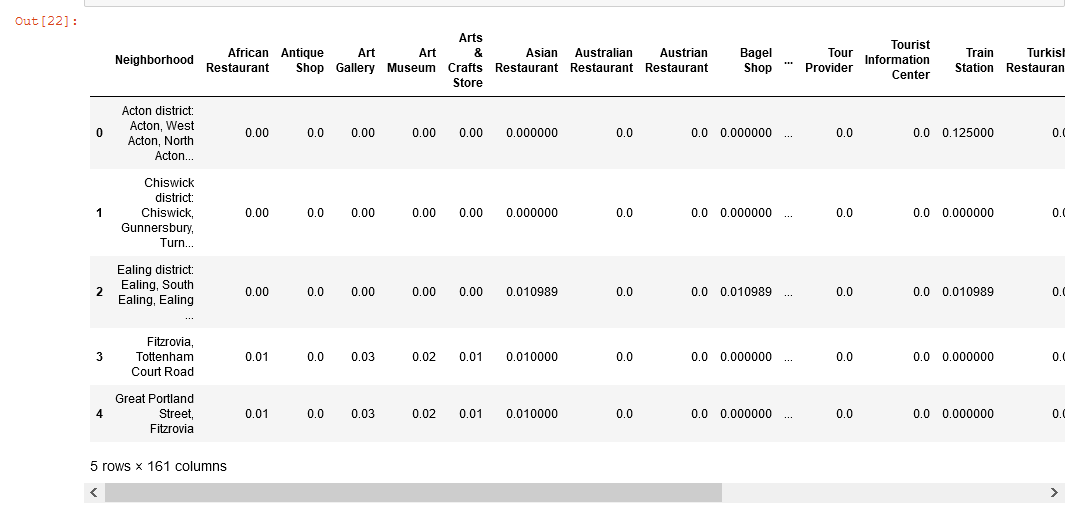
## Data Transformation

Finally, the data retrieved are combined into a format that is required for clustering.

This is the format of the data required for clustering:

1. Venue Categories should be in a single row based Neighborhood instead of multiple rows. In total, there are 160 venue categories.
2. Mean of frequency of venue is calculated for each venue categories.
3. Finally, the Neighborhood column is removed.

*Clustering Data*

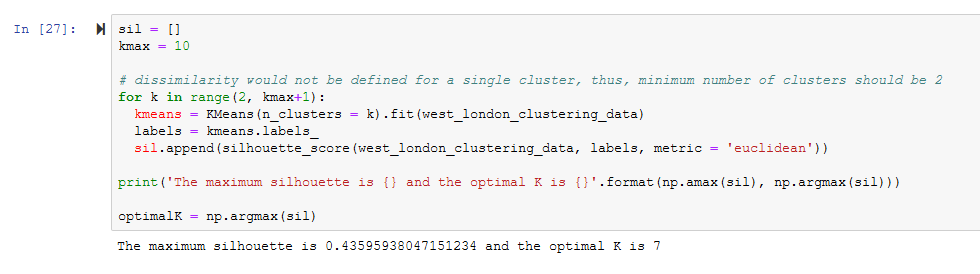


# Model and Evaluation

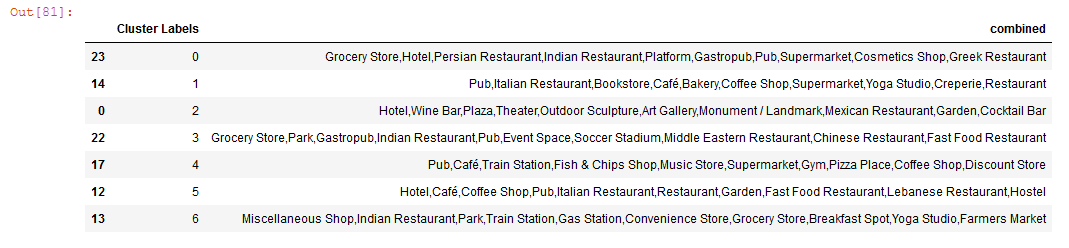
Kmeans clustering algorithm is used for this project. The parameter required for this algorithm is the number of clusters. To identify the optimal cluster, the model is executed several times with different values of n and the silhouette is used to evaluate the model.

The optimal cluster is n is 7 and the corresponding silhouette is 0.436.

*Identifying the optimal number of cluster*



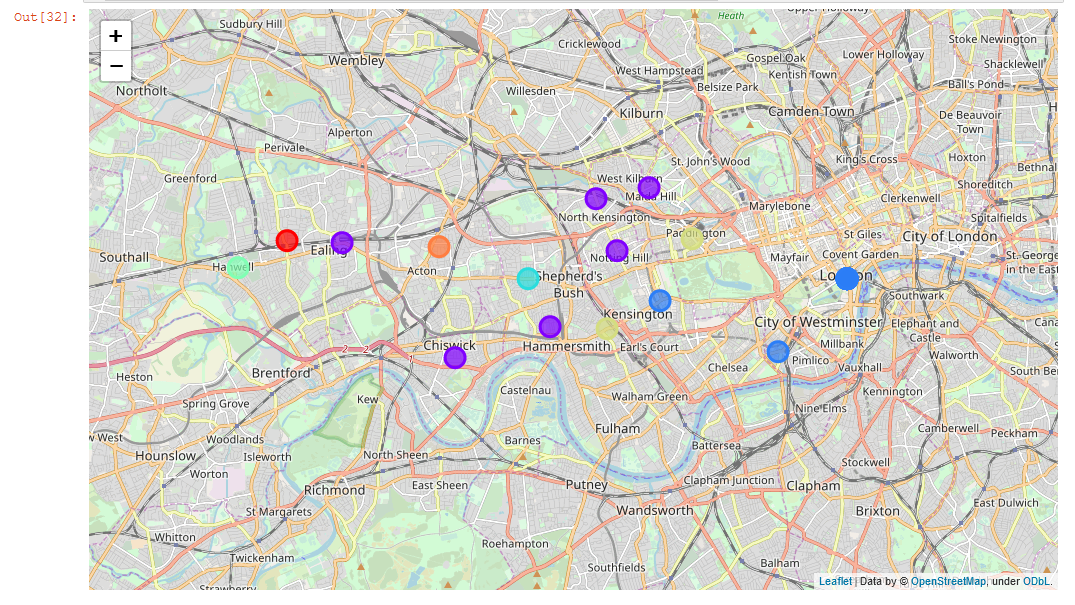
The cluster and its associated venues are as follows:



# Results

As presented previously, the optimal K is 7.

To visual the location of the clusters, I use folium map.



From the map, there are 5 neighbourhoods in cluster 1, the largest cluster. The neighbourhoods have the following venues: Pub, Italian Restaurant, Bookstore, Café, Bakery, Coffee Shop, Supermarket, Yoga Studio, Creperie, Restaurant.

The second largest cluster is 2, with 3 neighbourhoods. The neighbourhoods have the following venues: Hotel, Wine Bar, Plaza, Theatre, Outdoor Sculpture, Art Gallery, Monument / Landmark, Mexican Restaurant, Garden, Cocktail Bar.

Finally, the third largest cluster is 5, with 2 neighbourhoods. The neighbourhoods have the following venues: Hotel, Café, Coffee Shop, Pub, Italian Restaurant, Restaurant, Garden, Fast Food Restaurant, Lebanese Restaurant, Hostel.

The rest of the clusters are as follows:

|  |  |
| --- | --- |
| Cluster | Venues |
| 0 | Grocery Store, Hotel, Persian Restaurant, Indian Restaurant, Platform, Gastropub, Pub, Supermarket, Cosmetics Shop, Greek Restaurant |
| 3 | Grocery Store, Park, Gastropub, Indian Restaurant, Pub, Event Space, Soccer Stadium, Middle Eastern Restaurant, Chinese Restaurant, Fast Food Restaurant |
| 4 | Pub, Café, Train Station, Fish & Chips Shop, Music Store, Supermarket, Gym, Pizza Place, Coffee Shop, Discount Store |
| 6 | Miscellaneous Shop, Indian Restaurant, Park, Train Station, Gas Station, Convenience Store, Grocery Store, Breakfast Spot, Yoga Studio, Farmers Market |

My analysis:

Cluster 1, 4, 6 are where most of the locals lives or visits. There are no hotel or tourist attraction but there many local amenities like grocery Store, pub, restaurants, etc.

Cluster 0, 2, 3, 5 are tourist area as there are hotel, museum, theatres, stadium, and restaurants.

I tried to use Word Cloud to gain further insights into each cluster.

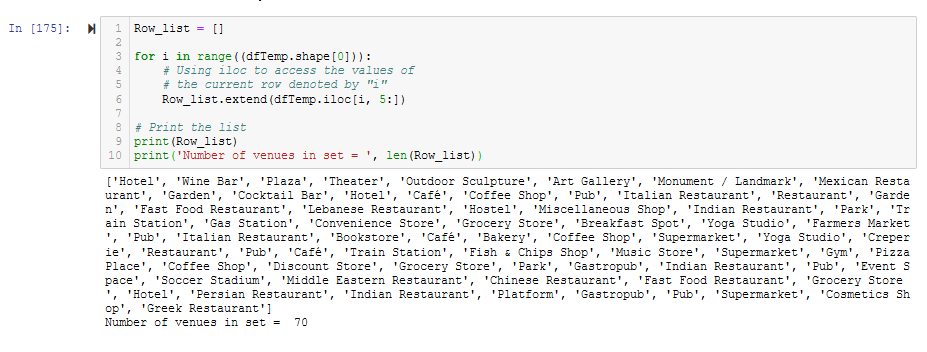
|  |  |
| --- | --- |
| In Cluster 0, the keywords are Hotel, Store/Shop and Restaurant  In Cluster 1, the keywords are Bakery, Café, Bookstore, Pub and Restaurants.  In Cluster 2, the keywords are Bar, Theatre, Gallery, Monument and Hotel.  In Cluster 3, the keywords are Park, Stadium and Restaurant.  In Cluster 4, the keywords are Café, Shop/Store, Pub and Train Station.  In Cluster 5, the keywords are Restaurants, Hostel, and Garden.  In Cluster 6, the keywords are Store, Park and Station. | A screenshot of a cell phone  Description automatically generated |

# Deployment

Here’s an example of how the Kmeans model can be used.

We have 7 clusters with 10 venues each. All the venues are combined to form 70 venue list. Using set function, duplicates are removed, and the final list contains 46 venues.

*Combined list of venues from 7 clusters*

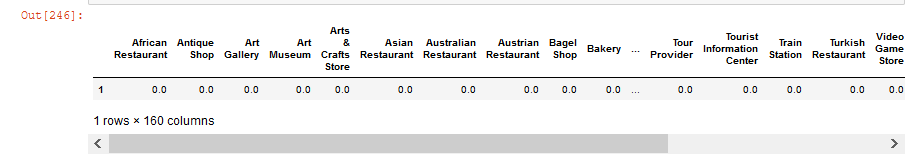


*List without Duplicates*

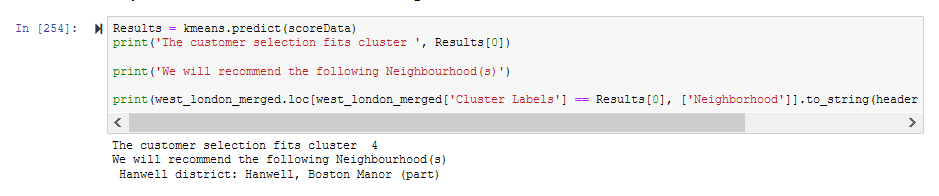


Assume that we will have an interface to let a customer choose 10 venues from a list of 46, this is how we will make use of his/her selections.

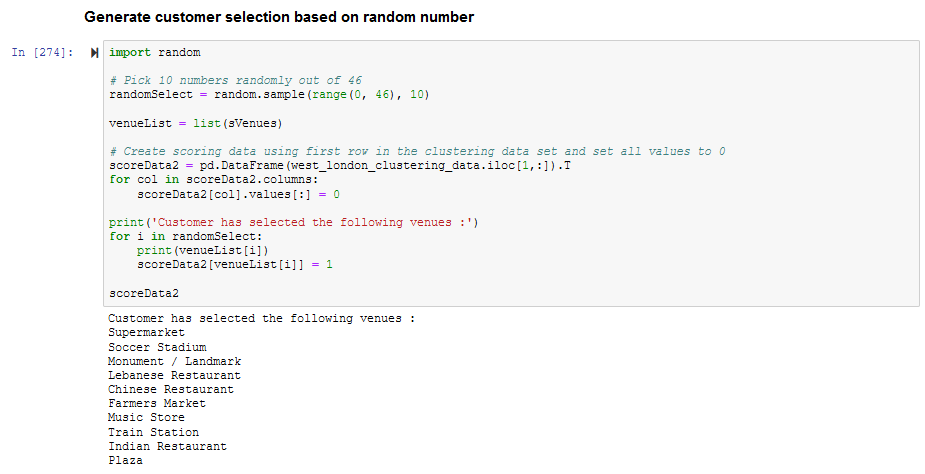
Update the customer select to a dataframe of 160 venues. Customer selection will be set to 1.



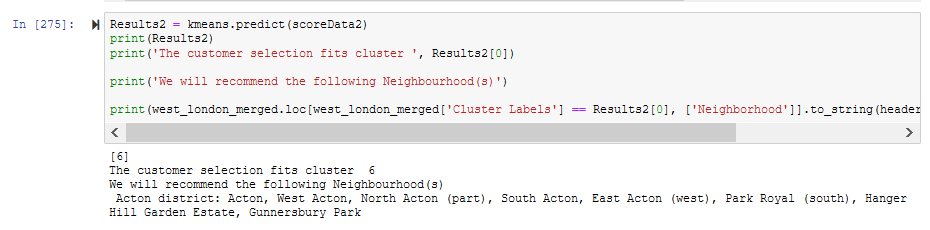
Score the dataframe using the Kmeans model, process the results.



Here is another where selection is made randomly using a generator.



The result is:



# References

Coursea notes and assignment notebooks are referred to when working on this project.