Capstone Project - The Battle of Neighborhoods

Week 2

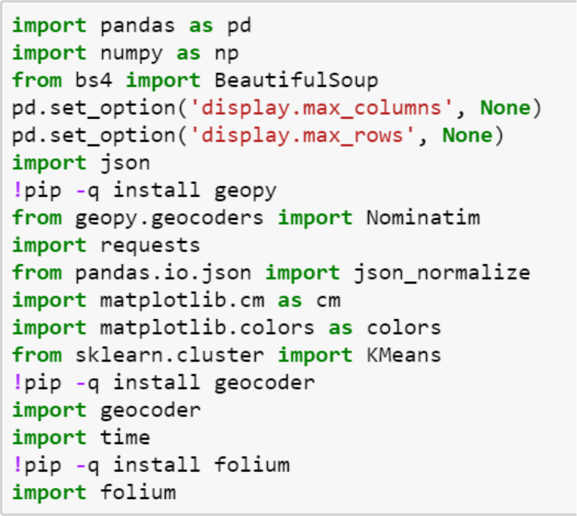
1. Introduction

1.1.Defining the Problem & its Background

London is one of the most ethnically diverse cities in the world. In this populated and diverse city, there are many restaurants. With a variety of cuisines and dining spots, people have an abundance of options and dishes to choose from. No matter what you are craving or where you are from, chances are you can find what you are looking for in this city! For example, Asian, Afghan, Indian, African and American. As a result, opening a new restaurant in London can be a challenge. This project aims at finding a good place to open a new African restaurant with organic and healthy menu. Hence, target groups are mainly Africans as well as people looking for healthy food.

1.2.Required Tools

To import the required tools and libraries, the following codes were used:



```
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
import json
!pip -q install geopy
from geopy.geocoders import Nominatim
import requests
from pandas.io.json import json_normalize
import matplotlib.cm as cm
import matplotlib.colors as colors
from sklearn.cluster import KMeans
!pip -q install geocoder
import geocoder
import time
!pip -q install folium
import folium
```

Fig. 1: Imported tools

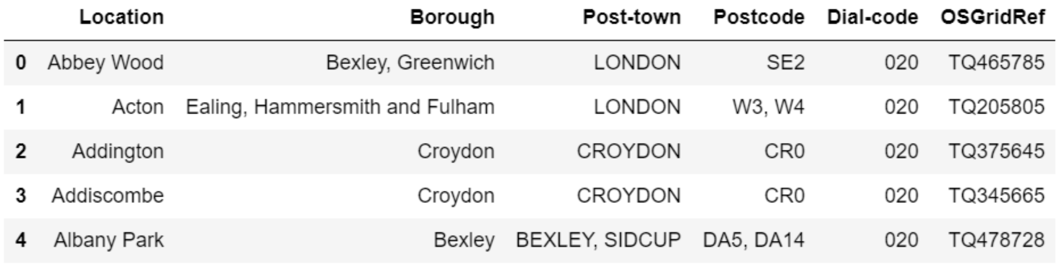
2. Data Collection & Preparation

The City of London is made up of several Planning Districts or neighbourhoods. Hence, we need to gather the required data from different sources. In this project, all the information needed will be extracted from the collected data from open and public resources including Wikipedia and Foursquare. To explore the venues in the neighbourhoods of London, we need geographical location data. We utilize Foursquare API for this purpose. Other information like postal code will be obtained using the data from Wikipedia.

2.1.Scraping the Wikipedia Page: List of Areas of London

As the first data source, the Wikipedia page on the list of areas of London[1] was scraped to obtain the required information. The following table shows the gathered information for the first five rows:

Table 1: Obtained information from Wikipedia Page: List of Areas of London



	Location	Borough	Post-town	Postcode	Dial-code	OSGridRef
0	Abbey Wood	Bexley, Greenwich	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

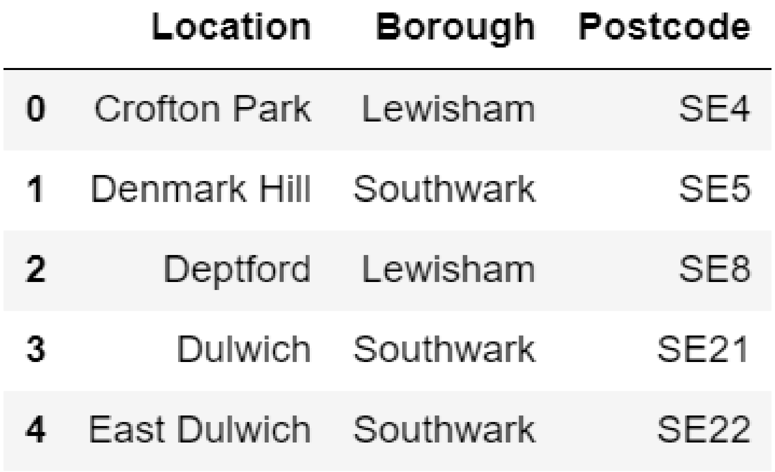
Since we are using free services of Foursquare API, there is a limitation regarding the number of possible calls. Hence, instead of whole city, just South East London was considered.

[1] https://en.wikipedia.org/wiki/List_of_areas_of_London

2.2.Scraping the Wikipedia Page: Demography of London

Table 2 gives the obtained data from the Wikipedia page for the demography of London[2].

Table 2: Obtained information from Wikipedia Page: Demography of London



	Location	Borough	Postcode
0	Crofton Park	Lewisham	SE4
1	Denmark Hill	Southwark	SE5
2	Deptford	Lewisham	SE8
3	Dulwich	Southwark	SE21
4	East Dulwich	Southwark	SE22

2.3.Location Data: Geocoder & arcgis_geocoder

This section aims at obtaining the latitude and longitude of the locations of interest using geocoder and arcgis_geocoder.

2.4.Location Data: Foursquare API

The geographical location data for the South East London Area venues was obtained utilizing the Foursquare API[3].

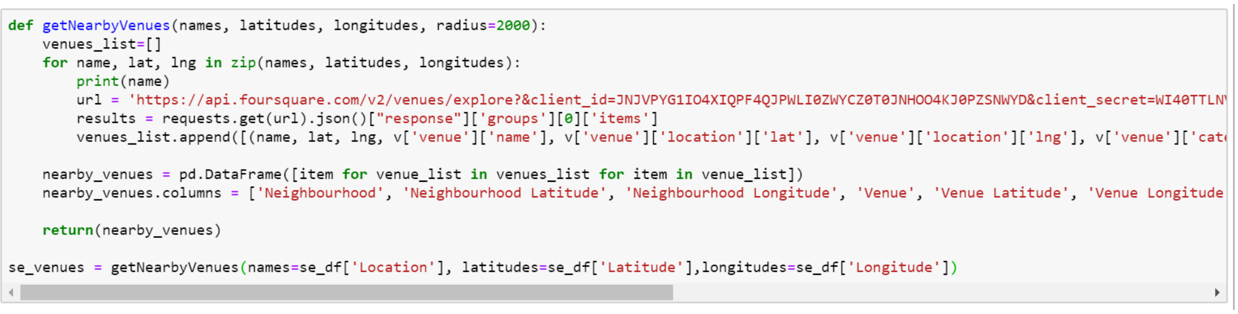
[2] https://en.wikipedia.org/wiki/Demography_of_London

[3] https://api.foursquare.com/v2/venues/explore?&client_id=JNJVPYG1IO4XIQPF4QJPWLI0ZWYCZ0T0JNHOO4KJ0PZSNWYD&client_secret=WI40TTLNV14IBYFMUNQPT5XGAOS1I4TDS5BU25QUKZ0H5CF3&v=20180605&ll=51.46196000000003,-0.007539999999949032&radius=2000&limit=100

3. Methodology

3.1.Data Processing

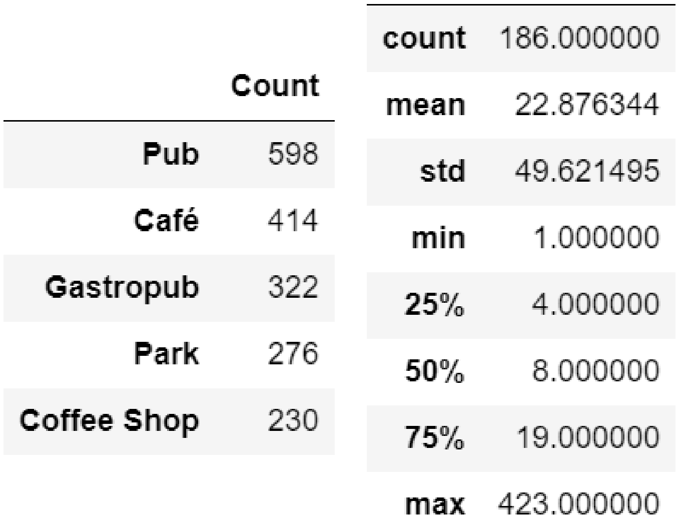
To explore the Neighborhoods in the South East London area, `getNearbyVenues` is utilized. Then, the `getNearbyVenues` is used on each neighbourhood. The following figures shows the codes for this purpose.



```
def getNearbyVenues(names, latitudes, longitudes, radius=2000):
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id=JNJVPYG1IO4XIQPF4QJPWLI0ZWYCZ0T0JNHOO4KJ0PZSNWYD&client_secret=WI40TTLN
        results = requests.get(url).json()["response"]['groups'][0]['items']
        venues_list.append([(name, lat, lng, v['venue']['name'], v['venue']['location']['lat'], v['venue']['location']['lng'], v['venue']['cat

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighbourhood', 'Neighbourhood Latitude', 'Neighbourhood Longitude', 'Venue', 'Venue Latitude', 'Venue Longitude

    return(nearby_venues)

se_venues = getNearbyVenues(names=se_df['Location'], latitudes=se_df['Latitude'],longitudes=se_df['Longitude'])
```

Fig. 2: Codes for exploring the neighborhoods

The next step was determination of the number of venues returned for each neighbourhood as well as checking the number of unique categories that can be returned for the venues. Fig. 3 depicts the obtained results.



	Count
Pub	598
Café	414
Gastropub	322
Park	276
Coffee Shop	230

count	186.000000
mean	22.876344
std	49.621495
min	1.000000
25%	4.000000
50%	8.000000
75%	19.000000
max	423.000000

Fig. 3: Obtained Number of venues returned for each neighbourhood

3.2.Clustering

This section employs the obtained data from previous section for the clustering problem. In this project, the k-Means algorithm was used for clustering. The optimum number of clusters was found employing the Elbow method and Sihouette Coefficient.

4. Results

Based on the obtained results,Pubs, Cafe and Coffee Shops are popular in the South East London. In case of restaurants, the most popular type is Italian. Further to the above, although Lewisham area has the most African population, restaurants in the top 5 venues are rare in the top 10 venues.

5. Conclusion & Recommendation

It can be concluded that the determined Clusters 2 and 3 are the best viable areas to open a brand new African restaurant. the proximity of these areas to other ameneties as well as their accessibility to the stations are paramount.

To achieve better results, it is recommended that more information like crime data for each area and traffic access collected and employed.