EXPLORATION OF GEO SPATIAL DATA FOR NEW RESTAURANTS

Introduction: --

A foreign restaurant chain wants to start his restaurant chain in London. He wants to find the best place / places in London where he can start his business by opening the first few restaurants. The requirements of restaurant chain are listed below: --

1. The potential restaurant must be opened in London.
2. The area in which the restaurant must be opened should be a populated area of any chosen region in London. The population of the neighborhood should be more than 20 thousand people.
3. The restaurant should be opened in an area which already has at least 3 running restaurant in nearby area.
4. The Districts and the neighborhood names should be clearly articulated in the final report.
5. The latitude and longitudinal position where potential restaurant must be opened should be plotted as dots in a map of London so that the most central location can be detected by looking into the map.

Data collection and analysis: --

The data that is used as a source for finding the potential place for opening new restaurants are listed below: --

1) XML extracts from Wikipedia web pages containing postcode district for all regions of London along with its coverage area.

   The information extracted from these pages are: --

   a. London postcode districts in the form of alphanumeric post codes.
   b. Coverage places under each post codes.
   c. Post town (which is London always).
   d. Local Authority area.

   The wiki links for data collection is given below: --

   e. https://en.wikipedia.org/wiki/E_postcode_area
   f. https://en.wikipedia.org/wiki/EC_postcode_area
   g. https://en.wikipedia.org/wiki/N_postcode_area
   h. https://en.wikipedia.org/wiki/NW_postcode_area
   i. https://en.wikipedia.org/wiki/SE_postcode_area
   j. https://en.wikipedia.org/wiki/SW_postcode_area
   k. https://en.wikipedia.org/wiki/W_postcode_area
   l. https://en.wikipedia.org/wiki/WC_postcode_area

2) Co-ordinates from foursquare location data that will be used for finding the exact location in maps and for pointing the potential places where restaurants can be opened.

3) Census data of UK downloaded rom site https://www.nomisweb.co.uk. This data will be used for finding the population against each post code district of London.

4). An excel sheet containing the zip codes of London along with latitudes and longitudes.

Implementation Methodology: --

The implementation methodology used to generate the report showing the places where potential restaurants can be opened, are listed below: --

1. Extraction of data from different sources.
2. Cleaning the data extracted from various sources.
3. Transforming the data into a tabular report showing potential places
4. Plotting the places over the map of London.

Data Extraction: --

As a first step, data for London was retrieved in XML format from Wikipedia pages. There were 8 different web pages from where data was extracted. Each page contained the data for a region of London. The regions from where data was extracted were East London, East Central, North London, North West London, South West London, South East London, West and West Central London.

Data was also extracted from excel for UK postcode districts along with the latitudes and longitudes for every post code district. This excel was downloaded from https://www.freemaptools.com/download-uk-postcode-lat-lng.htm

To find the population against each post code district , data was downloaded from https://www.nomisweb.co.uk/census/2011/ks101ew

The co-ordinates for postcode districts were obtained from foursquare API. These co-ordinates were used to plot the places in London where potential restaurants can be opened.

The python libraries used for extraction of data was Request and Panadas.

The requests were made to Wikipedia pages to extract the xml's. Pandas reard_csv method was used to extract data from Excel files.

Examples are shown below: --

Get all the files from Wikipedia that contains the postcode and neighborhood information for london area for processing

```
sources_E=requests.get('https://en.wikipedia.org/wiki/E_postcode_area').text
sources_EC=requests.get('https://en.wikipedia.org/wiki/EC_postcode_area').text
sources_N=requests.get('https://en.wikipedia.org/wiki/N_postcode_area').text
sources_NW=requests.get('https://en.wikipedia.org/wiki/NW_postcode_area').text
sources_SE=requests.get('https://en.wikipedia.org/wiki/SE_postcode_area').text
sources_SW=requests.get('https://en.wikipedia.org/wiki/SW_postcode_area').text
sources_W=requests.get('https://en.wikipedia.org/wiki/W_postcode_area').text
sources_WC=requests.get('https://en.wikipedia.org/wiki/WC_postcode_area').text

print("data Requested")
```

```
data Requested
```

Load the excel file UK postcodes to read the data for ech postcode district along with ltaitude and longitude

```
df4=pd.read_csv("ukpostcodes.csv")
df4.head()
```

Read the file that has information for population in each PostCode District

```
df6=pd.read_csv("London_population.csv")
df6.columns=['PostCode District', 'Population','Percentage_population_by_postcode']
df6.head()
```

Data Cleansing: --

Once the data is loaded into panda's data frame the data was cleaned by implementing the following steps: --

1. Drop the unwanted columns from the data frame.
2. Insert column header in data frame for the pages requested from Wikipedia.
3. Drop the line break character in the data frame.
4. Filter the NaN records from the data frame.
5. Filter the records that contain only records related to London.

Some of the examples for data cleansing are shown below: --

Insert column header to the data frame

```
df3_E.columns=['PostCode District', 'PostTown','Neighborhood','Local Authority Area']
df3_E.head()
```

Remove the line Breaks from each column

```
df3_E.replace({  '\n' : ''}, regex=True, inplace=True)
df3_E.shape
```

Cleaning the data Frame

```
# droping the column names inside the data frame
df3 = df3[df3.Neighborhood != 'Coverage']
#Droping records labeled as NaN
df3=df3.dropna(how='any')
df3
jd=df3['PostCode District']
#jd
```

Data Transformation and report generation

This section discusses the process of joining the individual data frames and generating tabular report out of it. The process from the start is listed below: --

1. The data was retrieved from 8 Wikipedia pages. Each page contained the data of a region of London. The data from each frame was brought to individual data frame, cleaned and concatenated to a single data frame as shown below.

PostCode District	PostTown	Neighborhood	Local Authority Area
E98	LONDON	Non-geographic postcode district (News Interna...	Tower Hamlets
E77	LONDON	Non-geographic postcode district (NatWest, loc...	Tower Hamlets
E20	LONDON	Olympic Park district: Olympic Park, & parts o...	Newham, Waltham Forest, Hackney, Tower Hamlets
E18	LONDON	Woodford and South Woodford district: Woodford	Redbridge
E17	LONDON	Walthamstow district: Walthamstow and Leyton (...	Waltham Forest
E16	LONDON	Victoria Docks and North Woolwich district: Ca...	Newham
E15	LONDON	Stratford district: Stratford, West Ham (part)...	Newham, Waltham Forest

2. To retrieve the latitude and longitude of each Postcode District, data was loaded from excel file that had information of postcode, latitude and longitude information. This file was downloaded from free map tool website. The individual postcodes in the excel was rolled up to Postcode District. Then the mean of latitude and longitude was determined, and the data frame was created. An example is shown below

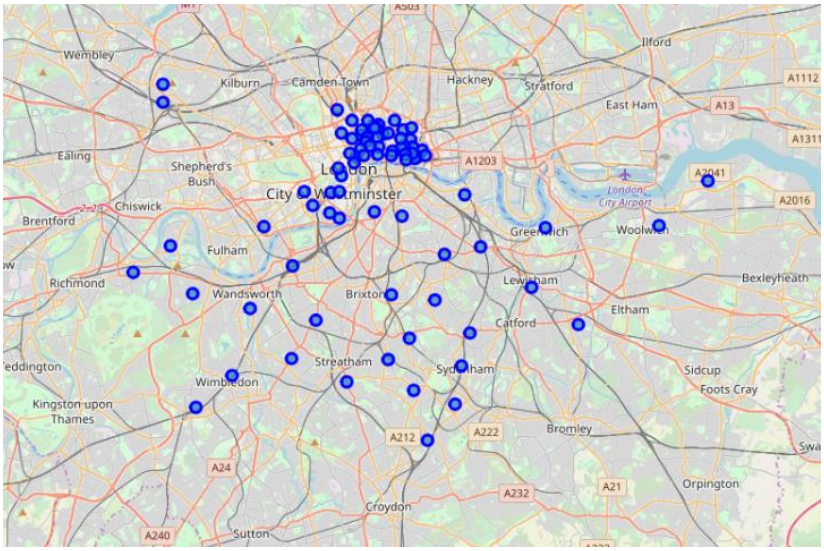
```
df4=df4.drop(['postcode','id'], axis=1)
df4.head()
```

	latitude	longitude	PostCode District
1543	60.225522	-1.573860	ZE2
1544	60.220492	-1.605311	ZE2
1545	60.240982	-1.647017	ZE2
1546	60.225056	-1.557837	ZE2
1547	60.225420	-1.570269	ZE2

3. The data frames generated at point 1 and point 2 were merged to generate a single data frame containing postcode districts, latitude, longitude, neighborhood and local authority area. An example is shown below

	PostCode District	latitude	longitude	PostTown	Neighborhood	Local Authority Area
0	EC1A	51.521118	-0.105546	LONDON	St Bartholomew's Hospital	City of London, Islington
1	EC1M	51.521406	-0.102435	LONDON	Clerkenwell, Farringdon	Islington, Camden, City of London
2	EC1N	51.520027	-0.108930	LONDON	Hatton Garden	Camden, City of London
3	EC1P	51.524502	-0.112088	LONDON		non-geographic
4	EC1R	51.524967	-0.108433	LONDON	Finsbury, Finsbury Estate (west)	Islington, Camden

4. The data frame generated from point 3 was plotted into a map using Folium library. This was done to visualize the data points in the map.
5. An excel was downloaded from nomisweb. This excel contained information had information about Postcode district and population of London regions.



6. The Foursquare API calls were used retrieve the venues against each neighborhood and arrange them in a data frame. An example of the data frame is given below: --

The code below executes the above function for each neighborhood and creates a new dataframe called *London_venues*.

```
']: London_venues = getNearbyVenues(names=df5['Neighborhood'],
                                    latitudes=df5['latitude'],
                                    longitudes=df5['longitude']
                                   )
#
```

```
St Bartholomew's Hospital
Clerkenwell, Farringdon
Hatton Garden

Finsbury, Finsbury Estate (west)
Finsbury (east), Moorfields Eye Hospital
St Luke's, Bunhill Fields
Shoreditch
Broadgate, Liverpool Street
Old Broad Street, Tower 42

Bank of England
Guildhall
Barbican
St Mary Axe, Aldgate
Lloyd's of London, Fenchurch Street
Tower Hill, Tower of London

Monument, Billingsgate
Cornhill, Gracechurch Street, Lombard Street
Fetter Lane
St Paul's
Mansion House
```

7. Data Frame shown is point 6 was further refined to select the venues that are only tied to a venue category of restaurants.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
St Bartholomew's Hospital	51.521118	-0.105546	Iberica Farringdon	51.520833	-0.104727	Spanish Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Sushi Tetsu	51.523348	-0.104015	Sushi Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	The Modern Pantry	51.522878	-0.103649	Modern European Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Luca	51.522017	-0.101703	Italian Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Comptoir Gascon	51.519270	-0.103192	French Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	St. John Bar and Restaurant	51.520437	-0.101382	English Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Dukan 41	51.519211	-0.109040	Falafel Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Bleeding Heart Restaurant	51.519286	-0.106892	French Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Anglo	51.520536	-0.109351	Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	KIN	51.521416	-0.110005	Asian Restaurant
St Bartholomew's Hospital	51.521118	-0.105546	Polpo	51.520015	-0.102210	Italian Restaurant

8. The data frame was aggregated to find the number of restaurants against each neighborhood.

Neighborhood	latitude	longitude	Restaurant_Count
Anerley district: Anerley, Crystal Palace (par...	51.412299	-0.059364	1
Balham district: Balham, Clapham South, Wandsw...	51.446305	-0.149193	11
Bank of England	51.516353	-0.091683	32
Barbican	51.519522	-0.093892	12
Barnes district: Barnes	51.476309	-0.243405	3
Battersea head district: Battersea, Clapham South	51.467968	-0.164021	9
Belgravia, Chelsea (part), area between Sloane...	51.492500	-0.150732	26
Belgravia, north of Eaton Square, Knightsbridg...	51.498073	-0.156462	19
Blackfriars	51.512531	-0.099990	34
Bloomsbury, British Museum, Southampton Row	51.519518	-0.125812	18

9. The data was then retrieved from excel sheet that contained the census data for London Region. This data was merged with the data frame generated at point 8 to show the population of each Neighborhood.

PostCode District	latitude	longitude	Population	Percentage_population_by_postcode	Neighborhood	Restaurant_Count
EC1A	51.521118	-0.105546	894	2.69	St Bartholomew's Hospital	25
EC1M	51.521406	-0.102435	2473	7.45	Clerkenwell, Farringdon	21
EC1N	51.520027	-0.108930	2650	7.98	Hatton Garden	28
EC1R	51.524967	-0.108433	4885	14.71	Finsbury, Finsbury Estate (west)	24
EC1V	51.526695	-0.098121	12765	38.44	Finsbury (east), Moorfields Eye Hospital	14

10. Now the following rule is applied according to the requirement of the restaurant chain to generate a tabular report containing areas where restaurants can be opened: --
    a. Filter out all records where population of a postcode district is less than 20000.
    b. Filter out the places who contribute less than 5% of total population of a region for London.
    c. Finally select only those places where restaurant count is more than 3.0

11. **The final report is shown below**

PostCode District	Neighborhood	latitude	longitude	Restaurant_Count	Population	Percentage_population_by_postcode
SW19	Wimbledon district: Wimbledon, Colliers Wood, ...	51.423924	-0.203416	15	77676	8.88
SE18	Woolwich district: Woolwich, Royal Arsenal, Pl...	51.484328	0.072814	3	77384	7.83
SW11	Battersea head district: Battersea, Clapham South	51.467968	-0.164021	9	71717	8.20
SE15	Peckham district: Peckham, Nunhead, South Berm...	51.472701	-0.065687	8	64359	6.51
SW17	Tooting district: Tooting, Mitcham (part)	51.430737	-0.164644	9	64215	7.34
SW18	Wandsworth district: Wandsworth Town, Southfie...	51.451171	-0.191439	6	58992	6.74
NW11	Golders Green district: Golders Green, Temple ...	51.578417	-0.197487	3	32684	5.93

Plotting the report: --

12. The report is plotted into map and shown to find the central place where the restaurants can be opened on first place.

Results

The results that was evident is listed below: --

1. There were 7 potential places where the restaurant chain can start their business in London. These places were: --
    a. Wimbledon district
    b. Woolwich district
    c. Battersea head district
    d. Peckham district
    e. Tooting district
    f. Wandsworth district
    g. Golders Green district
2. Out of the 7 places four places were from South West London neighborhood, two from South East and one from North West London.
3. Restaurant count was very high in Wimbledon district. It was 15 in number and has the highest populated neighborhood in London.

Recommendation and Observation

Based on the results that are obtained, it is recommended that along with a districts population density and restaurant counts the average income of the people living in the neighborhood can also be considered. This is a data that is difficult to obtain, and a masked data of income can lead to erroneous results.

The population density when taken on whole (both rural and urban) was not giving a meaningful result, it was observed that only when Rural population was accounted for the report the results made more sense.

Conclusion

This analysis can be used by any restaurant chain that wants to open their business in foreign country. The data it should use, is of existing restaurant counts and the population density. There were 7 prospective locations determined in London to open a restaurant.

References

1. Wikipedia
2. Foursquare Api
3. Census data UK
4. Freemaptools