IBM Data Science Capstone Project

Introduction

Background:

London is considered to be one of the world's most important global cities and has been called the world's most powerful, most desirable, most influential, most expensive, innovative, sustainable, most investment friendly and most popular for work city .London exerts a considerable impact upon the arts, commerce, education , entertainment, fashion, healthcare, media, professional services, research and development, tourism and transportation. So it seems fair to dive into the depths of this cultural pot boiler of a city. London is also one of the leading tourists destination and the city accounts for almost half of all inbound visitor spending in the U.K. The London Underground commonly referred to as the tube is the oldest and third longest metro system in the world. Over four million journeys are made every day on the tube network and approximately over a billion each year. London's most popular sport is undoubtedly football and it has five clubs in the English Premier League as of the 2019-2020 season: Arsenal, Crystal Palace, Chelsea, Tottenham Hotspur and the West Ham United. As the city is football crazy, so a sports bar culture is very popular among the Londoners. As the sports bar fuses good food with the adrenaline of sports, so it makes a good choice if anyone is interested in investing in the food and beverages industry.

Problem Statement:

To identify the areas of London where a sports bar can be opened .The area should be quite close to the action centre and well-connected too.

Target Audience:

Anyone interested in starting a food and beverage place while keeping in mind, the football culture of the city of London.

.

Data Section

London Underground, commonly referred to as the Tube, is the heart of the transportation system of the city of London. So, the location of the tube stations provides information about the footfall as well as the connectivity of the surrounding areas. For the clustering problem the below Wikipedia web page is scrapped and the corresponding data is extracted.

https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations

	Name	Latitude	Longitude	Platform / Entrance	Collected By	Collected On	Line	Step free
0	Acton Town	51.502500	-0.278126	Platform	User:Gagravarr	24/11/06	District, Piccadilly	NaN
1	Acton Central	51.50883531	-0.263033174	Entrance	User:Firefishy	08/05/2007	London Overground	NaN
2	Acton Central	51.50856013	-0.262879534	Platform	User:Firefishy	08/05/2007	London Overground	NaN
3	Aldgate	51.51394	-0.07537	Aldgate High Street entrance	User:Morwen	28/4/2007	Metropolitan	No
4	Aldgate East	51.51514	-0.07178	Entrance	User:Parsingphase	(2006)	District, Hammersmith & City	NaN

Moreover the Foursquare API to explore venue types surrounding each station is also used. Venues having a walk able distance of approximately 500 metres is explored using Foursquare . A screenshot of the foursquare url is as follows:

```
LIMIT = 100 # limit of number of venues returned by Foursquare API



radius = 500 # define radius

 # create URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}
&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    name_latitude,
    name_longitude,
    radius,
    LIMIT)
url # display URL
```

```
'https://api.foursquare.com/v2/venues/explore?&client_id=UWULHDGZWTE05NWFHPQ5GYMTNXCTOXVNFSSFFWEZNERKO0TN&
client_secret=YI3CYQVDOQJWWYK3QJKW5ZJZALS5M31XDVBTGONJISXIAHZZ&v=20180605&ll=51.55847,-0.10561&radius=500&
limit=100'
```

Methodology

Data Pre-processing:

The data scrapped from the Wikipedia page is cleaned and the following data frame is prepared.

	Name	Latitude	Longitude	Line
0	Acton Town	51.502500	-0.278126	District, Piccadilly
1	Acton Central	51.508835	-0.263033	London Overground
2	Acton Central	51.508560	-0.262880	London Overground
3	Aldgate	51.513940	-0.075370	Metropolitan
4	Aldgate East	51.515140	-0.071780	District, Hammersmith & City

Another web page is also scrapped to just get an idea about the usage of the tube stations. The following data-frame shows the number of yearly commuters.

```
df3=pd.read_html('https://en.wikipedia.org/wiki/List_of_London_railway_stations')
```

```
df4=df3[1]
```

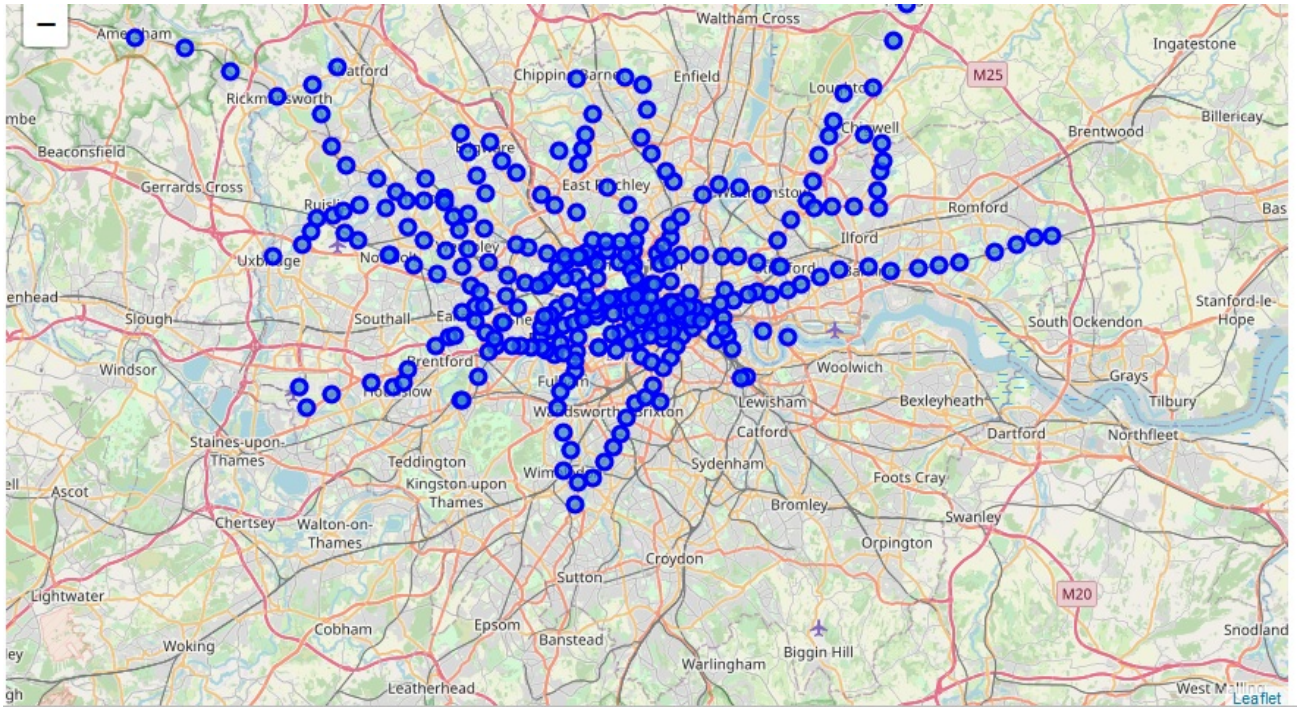
```
df4.rename(columns={'Station': 'StationName'}, inplace=True)
```

```
df4.head()
```

	Rank	StationName	Yearly passengers
0	1	Waterloo	98442742
1	2	Victoria	81356330
2	3	Liverpool Street	63004002
3	4	London Bridge	56442044
4	5	Euston	41911706

Data Visualization:

Using the coordinates of London and the Folium package, we can create a map of London with all its tube stations marked in blue.



Zooming in on the Stations:

Among the London Underground lines, into which the stations are divided, the Piccadilly line is one of the busiest. The Piccadilly line runs between Cockfosters in the suburban north and Action Town in the west, where it divides into two branches: one of these runs to Heathrow Airport and the other to Uxbridge in northwest London. So the line is quite diverse and helps commute more than two hundred and fifty million passengers yearly. The Piccadilly line serves many of London's key tourist attractions, including the British Museum (Russell Square), the numerous museums around South Kensington, Harrods (Knightsbridge), Hyde Park and Buckingham Palace(within walking distance of Green park station), Leicester Square(with its own station) and Covent Garden(also with its own station). The Piccadilly line runs beneath the famous Emirates stadium. So the Piccadilly line is widely used by tourists, football fans and professionals as well.

So selecting the Piccadilly line for the Foursquare api calls would provide abundant information about the neighbourhoods along the Piccadilly line. Slicing the original data frame and creating a new data frame of the Piccadilly data. , we get the following :

Slicing the original dataframe and creating a new dataframe of the Piccadilly data.

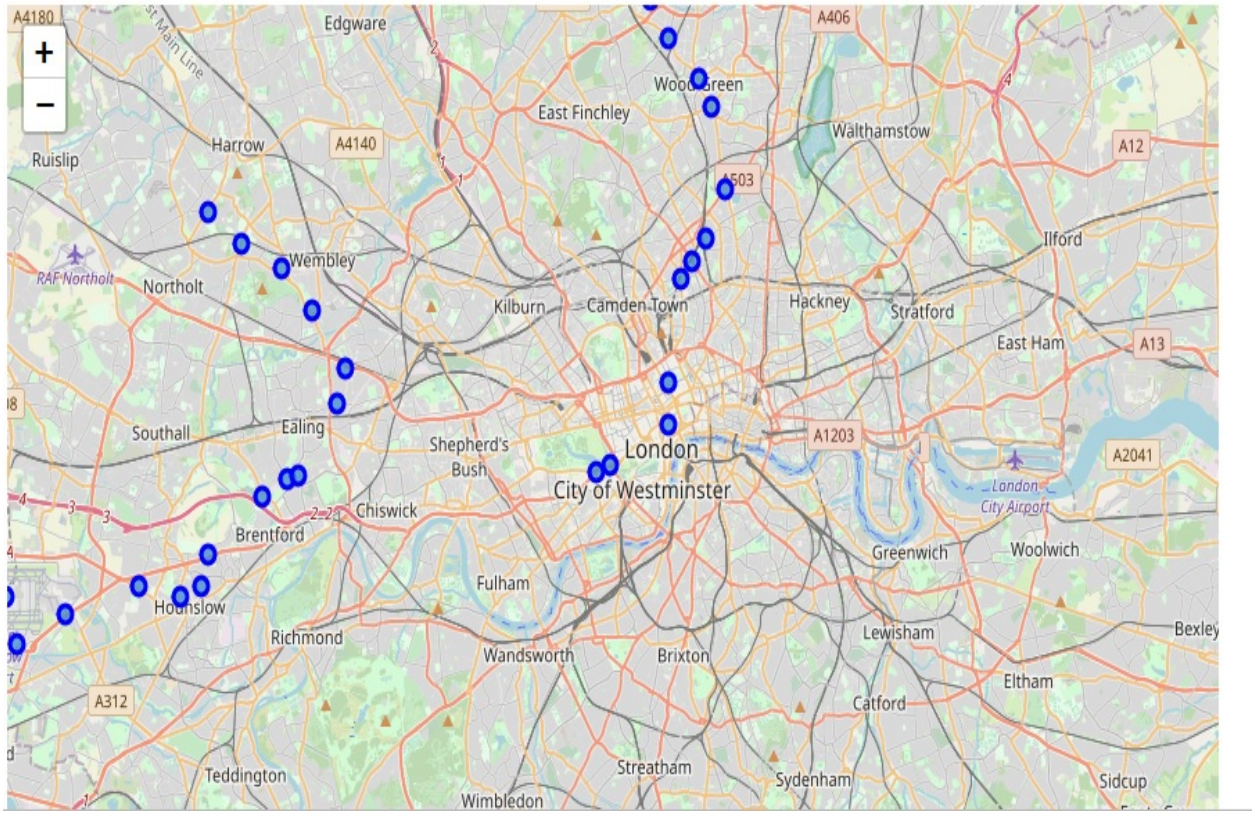
[27]:
```
Piccadilly_data = df2[df2['Line'] == 'Piccadilly'].reset_index(drop=True)
Piccadilly_data.head(15)
```

[27]:

	Name	Latitude	Longitude	Line
0	Alperton	51.540970	-0.300610	Piccadilly
1	Arnos Grove	51.616250	-0.133550	Piccadilly
2	Arsenal	51.558470	-0.105610	Piccadilly
3	Boston Manor	51.495371	-0.325573	Piccadilly
4	Bounds Green	51.607000	-0.124180	Piccadilly
5	Caledonian Road	51.548500	-0.118020	Piccadilly
6	Cockfosters	51.651170	-0.148110	Piccadilly
7	Covent Garden	51.513080	-0.124270	Piccadilly
8	Hatton Cross	51.466988	-0.423035	Piccadilly
9	Heathrow Terminals 1-2-3	51.471290	-0.452744	Piccadilly
10	Heathrow Terminal 4	51.459478	-0.446897	Piccadilly
11	Holloway Road	51.552900	-0.112780	Piccadilly

Using the folium package and the coordinates of the stations the Piccadilly line, we can create the following map:

```
map_Piccadilly
```



Zooming in further:

As we are interested in opening a sports bar, so we focus around the big ticket football clubs in London. Along the Piccadilly line, the area around the Emirates Stadium becomes an obvious choice. Choosing the area around the arsenal station within a radius of 500m (which would be a walk able distance) we get a result of 33 venues.

A sample of the json data and its corresponding data frame is as follows:

```
[34]: results = requests.get(url).json()
      results
```

```
[34]: {'meta': {'code': 200, 'requestId': '5e9fdeeabae9a2001b47e2f0'},
       'response': {'suggestedFilters': {'header': 'Tap to show:',
         'filters': [{'name': 'Open now', 'key': 'openNow'}]},
        'headerLocation': 'Islington',
        'headerFullLocation': 'Islington, London',
        'headerLocationGranularity': 'neighborhood',
        'totalResults': 32,
        'suggestedBounds': {'ne': {'lat': 51.5629700045,
          'lng': -0.09838547161098131},
         'sw': {'lat': 51.5539699955, 'lng': -0.11283452838901868}},
        'groups': [{'type': 'Recommended Places',
          'name': 'recommended',
          'items': [{'reasons': {'count': 0,
             'items': [{'summary': 'This spot is popular',
               'type': 'general',
               'reasonName': 'globalInteractionReason'}]},
            'venue': {'id': '4bd2d84541b9ef3b0cc4fee5',
             'name': 'The Arsenal Store',
             'location': {'address': '75 Drayton Park',
              'crossStreet': 'Highbury House',
              'lat': 51.55669516351197,
              'lng': -0.10609761727323813,
              'labeledLatLngs': [{'label': 'display',
                'lat': 51.55669516351197,
```

Cleaning the json and structuring it into a pandas dataframe.

[221]:
```
venues = results['response']['groups'][0]['items']

nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues =nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues.head()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/ipykernel_launcher.py:3: FutureWarning: pandas.io.json.json_normalize is deprecated, use pandas.json_normal
  This is separate from the ipykernel package so we can avoid doing imports until
```

[221]:

	name	categories	lat	lng
0	The Arsenal Store	Clothing Store	51.556695	-0.106098
1	DW Fitness First	Gym / Fitness Center	51.558208	-0.102262
2	Emirates Stadium	Soccer Stadium	51.555247	-0.108361
3	The North Bank	Soccer Stadium	51.555700	-0.108202
4	East Stand	Soccer Stadium	51.554861	-0.107310

[222]:
```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
33 venues were returned by Foursquare.
```

A sample of the top venues around the Arsenal station.

[37]:

	name	categories	lat	lng
0	The Arsenal Store	Clothing Store	51.556695	-0.106098
1	DW Fitness First	Gym / Fitness Center	51.558208	-0.102262
2	Emirates Stadium	Soccer Stadium	51.555247	-0.108361
3	The North Bank	Soccer Stadium	51.555700	-0.108202
4	East Stand	Soccer Stadium	51.554861	-0.107310
5	Arsenal Football Supporters Club	Sports Bar	51.559150	-0.103405
6	il guscio highbury	Italian Restaurant	51.558499	-0.098447
7	The Woodbine	Pub	51.559551	-0.098600
8	Café Beam	Café	51.559743	-0.098730
9	Gunners Pub	Sports Bar	51.558889	-0.098992
10	Gillespie Park Local Nature Reserve	Park	51.558873	-0.106168
11	Sobell Leisure Centre	Gym / Fitness Center	51.558375	-0.111615
12	Ashburton Triangle and Arsenal Museum	History Museum	51.556403	-0.107927
13	Wolkite Kitfo Restaurant & Cafe	Ethiopian Restaurant	51.556065	-0.111416
14	Bank of Friendship	Pub	51.558360	-0.098522
15	The Tollington	Pub	51.557764	-0.112685
16	DW Fitness First	Gym / Fitness Center	51.558228	-0.102019
17	Auld Triangle	Pub	51.562058	-0.105099
18	Bun & Bar	Burger Joint	51.558031	-0.098468
19	Domino's Pizza	Pizza Place	51.556213	-0.110868
20	Islington Ecology Centre	Park	51.558429	-0.106260

Exploring areas around Piccadilly line:

Exploring the venues along the Piccadilly line using Foursquare data, we get the following data frame.

A sample of the Piccadilly line dataframe:

```
[41]: print(Piccadilly_venues.shape)
      Piccadilly_venues.head()
```

```
(909, 7)
```

[41]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alperton	51.54097	-0.30061	The Gym	51.540819	-0.298715	Gym / Fitness Center
1	Alperton	51.54097	-0.30061	Sainsbury's	51.538431	-0.302540	Supermarket
2	Alperton	51.54097	-0.30061	Maru Bhajias	51.543873	-0.297200	Indian Restaurant
3	Alperton	51.54097	-0.30061	Subway	51.541707	-0.297996	Sandwich Place
4	Alperton	51.54097	-0.30061	East Pan Asian Restaurant	51.537700	-0.301996	Asian Restaurant

We get 189 unique venue categories along the Piccadilly line as follows:

[42]:
```
Piccadilly_venues.groupby('Neighborhood').count()
```

[42]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alperton	10	10	10	10	10	10
Arnos Grove	7	7	7	7	7	7
Arsenal	32	32	32	32	32	32
Boston Manor	7	7	7	7	7	7
Bounds Green	16	16	16	16	16	16
Caledonian Road	26	26	26	26	26	26
Cockfosters	24	24	24	24	24	24
Covent Garden	92	92	92	92	92	92
Hatton Cross	11	11	11	11	11	11
Heathrow Terminal 4	51	51	51	51	51	51
Heathrow Terminals 1-2-3	63	63	63	63	63	63
Holloway Road	37	37	37	37	37	37
Hounslow Central	45	45	45	45	45	45
Hounslow East	16	16	16	16	16	16
Hounslow West	13	13	13	13	13	13
Hyde Park Corner	100	100	100	100	100	100
Knightsbridge	61	61	61	61	61	61
Manor House	26	26	26	26	26	26

Analysing each neighbourhoods by hot encoding and choosing the top 5 venues we get the following sample:

[49]:
```
num_top_venues = 5

for hood in Piccadilly_grouped['Name']:
    print("----"+hood+"----")
    temp = Piccadilly_grouped[Piccadilly_grouped['Name'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----Alperton----
                 venue  freq
0          Supermarket   0.2
1    Indian Restaurant   0.2
2  Gym / Fitness Center   0.2
3          Bridal Shop   0.1
4     Asian Restaurant   0.1


----Arnos Grove----
               venue  freq
0           Bus Stop  0.29
1      Grocery Store  0.29
2  Chinese Restaurant  0.14
3           Beer Bar  0.14
4               Park  0.14


----Arsenal----
                 venue  freq
0                  Pub  0.12
1       Soccer Stadium  0.12
2                 Park  0.09
3  Gym / Fitness Center  0.09
4           Sports Bar  0.09
```

Analysing the neighbourhoods further, we get the following top 10 venues and convert it into a data frame as follows.

```
neighborhoods_venues_sorted.head()
```

[51]:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alperton	Gym / Fitness Center	Indian Restaurant	Supermarket	Fast Food Restaurant	Asian Restaurant	Bridal Shop	Sandwich Place	Exhibit	French Restaurant	Food Truck
1	Arnos Grove	Grocery Store	Bus Stop	Beer Bar	Park	Chinese Restaurant	Women's Store	French Restaurant	Food Truck	Food Court	Fish & Chips Shop
2	Arsenal	Soccer Stadium	Pub	Sports Bar	Gym / Fitness Center	Park	Café	Clothing Store	Fish & Chips Shop	Food Truck	Pizza Place
3	Boston Manor	Farm	Chinese Restaurant	Bus Stop	Canal Lock	Bar	Theater	Breakfast Spot	English Restaurant	Ethiopian Restaurant	Fruit & Vegetable Store
4	Bounds Green	Coffee Shop	Pub	Gourmet Shop	Grocery Store	Tennis Court	Café	Breakfast Spot	Campground	Convenience Store	Noodle House

Machine Learning tool:

The machine learning tool used in this project is K-means clustering. So far we have used the Foursquare API to explore the areas around the stations, get the most common venue categories in each neighbourhood and then use this feature to group the neighbourhoods into clusters.

The K-means clustering algorithm is used to complete the task.

Using elbow method, clustering is found to be optimum for k=5.

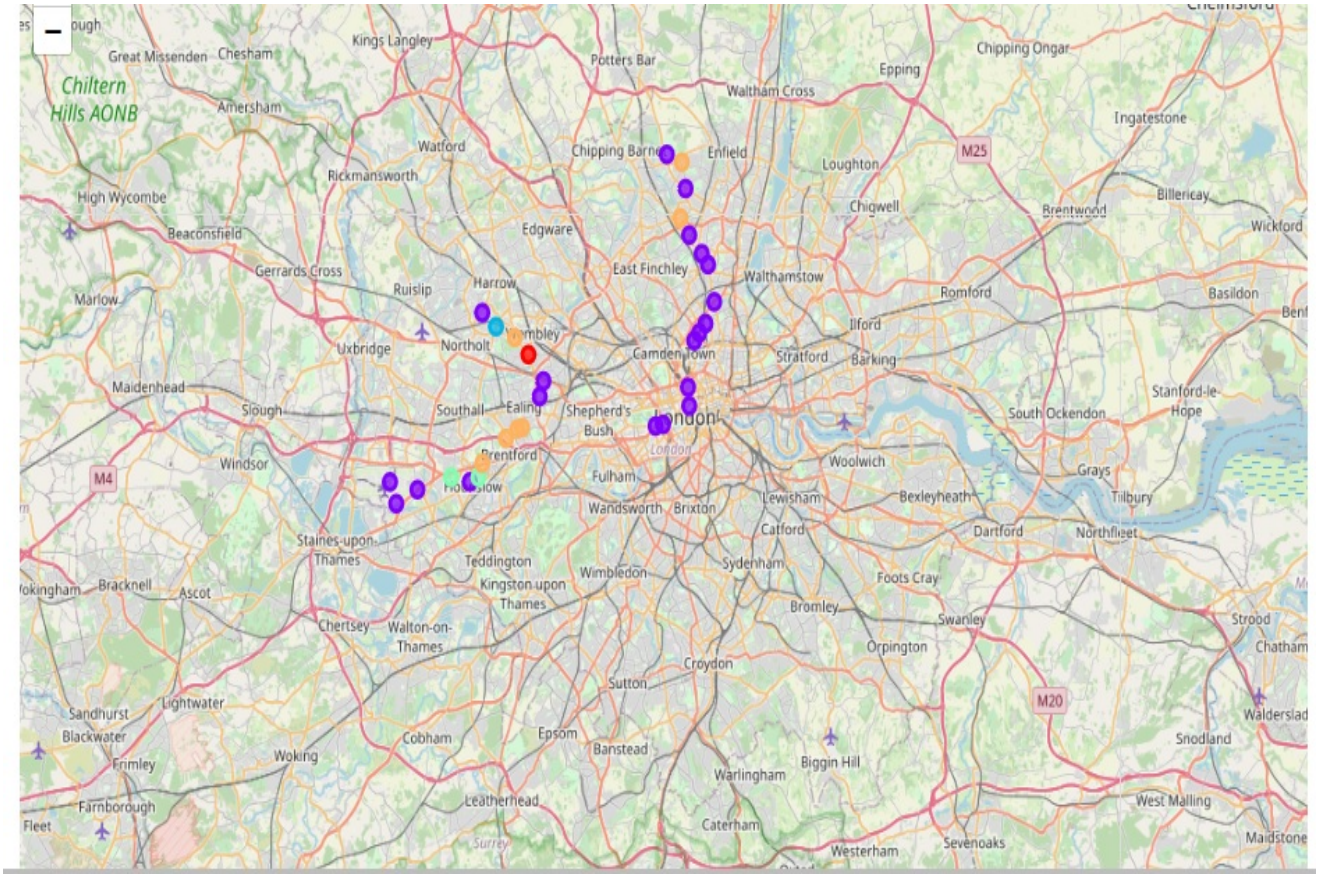
The following sample data frame is obtained using k-means.

[53]:

	Name	Latitude	Longitude	Line	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alperton	51.540970	-0.300610	Piccadilly	0	Gym / Fitness Center	Indian Restaurant	Supermarket	Fast Food Restaurant	Asian Restaurant	Bridal Shop	Sandwich Place	Exhibit	French Restaurant	Food Truck
1	Arnos Grove	51.616250	-0.133550	Piccadilly	4	Grocery Store	Bus Stop	Beer Bar	Park	Chinese Restaurant	Women's Store	French Restaurant	Food Truck	Food Court	Fish & Chips Shop
2	Arsenal	51.558470	-0.105610	Piccadilly	1	Soccer Stadium	Pub	Sports Bar	Gym / Fitness Center	Park	Café	Clothing Store	Fish & Chips Shop	Food Truck	Pizza Place
3	Boston Manor	51.495371	-0.325573	Piccadilly	4	Farm	Chinese Restaurant	Bus Stop	Canal Lock	Bar	Theater	Breakfast Spot	English Restaurant	Ethiopian Restaurant	Fruit & Vegetable Store
4	Bounds Green	51.607000	-0.124180	Piccadilly	1	Coffee Shop	Pub	Gourmet Shop	Grocery Store	Tennis Court	Café	Breakfast Spot	Campground	Convenience Store	Noodle House

Visualization of the clusters:

The folium package is used to create the following clustered map of Piccadilly line area, divided into 5 clusters.



Results:

Examining the individual clusters, we get the following data:

Cluster 1:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alperton	Gym / Fitness Center	Indian Restaurant	Supermarket	Fast Food Restaurant	Asian Restaurant	Bridal Shop	Sandwich Place	Exhibit	French Restaurant	Food Truck

Cluster 2:

A sample of the data frame is as follows.

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
2	Arsenal	Soccer Stadium	Pub	Sports Bar	Gym / Fitness Center	Park	Café	Clothing Store	Fish & Chips Shop
4	Bounds Green	Coffee Shop	Pub	Gourmet Shop	Grocery Store	Tennis Court	Café	Breakfast Spot	Campground
5	Caledonian Road	Café	Rental Car Location	Park	Theater	Coffee Shop	Chinese Restaurant	Restaurant	Pub
6	Cockfosters	Italian Restaurant	Café	Turkish Restaurant	Greek Restaurant	Platform	Bakery	Coffee Shop	Fish & Chips Shop
7	Covent Garden	Theater	Coffee Shop	Sushi Restaurant	Cocktail Bar	Hotel	Burger Joint	Music Store	Dessert Shop
8	Hatton Cross	Hotel	Convenience Store	Pub	Rental Car Location	Bus Station	Fast Food Restaurant	Trail	Gym / Fitness Center
9	Heathrow Terminals 1-2-3	Airport Lounge	Coffee Shop	Department Store	Clothing Store	Restaurant	Pharmacy	Grocery Store	Pub
10	Heathrow Terminal 4	Airport Lounge	Coffee Shop	Hotel	Cosmetics Shop	Airport Service	Italian Restaurant	Bookstore	Airport Terminal
11	Holloway Road	Café	Soccer Stadium	Pub	Coffee Shop	Ethiopian Restaurant	Gym / Fitness Center	Women's Store	Malay Restaurant

Cluster 3:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
27	Sudbury Hill	Hotel	Train Station	Metro Station	Grocery Store	Women's Store	Exhibit	French Restaurant	Food Truck	Food Court	Fish & Chips Shop

Cluster 4:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Common Venue
13	Hounslow East	Indian Restaurant	Bus Stop	Asian Restaurant	Hostel	Grocery Store	Metro Station	Dessert Shop	Convenience Store	Coffee Shop	Supermar
14	Hounslow West	Indian Restaurant	Hotel	Bus Stop	Platform	Pharmacy	Asian Restaurant	Park	Grocery Store	Event Space	Food Cou

Cluster 5:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
1	Arnos Grove	Grocery Store	Bus Stop	Beer Bar	Park	Chinese Restaurant	Women's Store	French Restaurant	Food Truck	Food Court
3	Boston Manor	Farm	Chinese Restaurant	Bus Stop	Canal Lock	Bar	Theater	Breakfast Spot	English Restaurant	Ethiopian Restaurant
19	Northfields	Park	Café	Bus Stop	Italian Restaurant	Brewery	Kebab Restaurant	Grocery Store	French Restaurant	Diner
20	Oakwood	Greek Restaurant	Grocery Store	Indian Restaurant	Paper / Office Supplies Store	Bus Stop	Café	Golf Course	Metro Station	Chinese Restaurant
21	Osterley	Bus Stop	Convenience Store	Thai Restaurant	Hockey Field	Metro Station	Coffee Shop	Chinese Restaurant	Sandwich Place	Bus Station
24	South Ealing	Pizza Place	Coffee Shop	Indian Restaurant	Pub	Gay Bar	Breakfast Spot	Mediterranean Restaurant	Bus Stop	Grocery Store
28	Sudbury Town	Indian Restaurant	Pub	Fast Food Restaurant	Bus Stop	Middle Eastern Restaurant	Train Station	Park	Metro Station	Women's Store

Discussion:

Studying the above clusters we can infer that the area around the Piccadilly line stations have a flurry of food and beverages options.

As our problem was to find an F&B option which would target the football crazy fans of London, so the area around the famed Emirates Stadium in arsenal station would be an apt solution. We already know from that the area around the underground stations are very well connected, so to open a sports bar around the arsenal station provides solution to our problem.

The following image gives the best F&B option:

	Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
2	Arsenal	Soccer Stadium	Pub	Sports Bar	Gym / Fitness Center	Park	Café	Clothing Store	Fish & Chips Shop

Conclusion:

The data analysis can also be extended to explore areas around other famous football clubs of London i.e. Chelsea, Tottenham Hotspurs, West Ham United and many others .