Exploring London Tube Neighborhood using web and Foursquare location data

Detailed Report

Table of Contents

Business Problem

Prospect of a Restaurant/Bar, close to Tube Stations in London, United Kingdom

Introduction

London Underground, better known as the Tube, has 11 lines covering 402km and serving 270 stations. The Tube handles up to 5 million passenger journeys a day. At peak times, there are more than 543 trains whizzing around the Capital

For this project, we want to look at the neighborhood surrounding the Tube stations and classify them based on the Restaurants and Bars closest to a station. By analyzing this data, we can classify stations and explore the opportunities to start up a new business. Given this scenario, we will go through the benefits and pitfalls of opening a restaurant/bar around the neighborhood

Target Audience

What type of clients or a group of people would be interested in this project?

- Business personnel who want to invest or open a restaurant/bar. This analysis will be a comprehensive guide to start or expand restaurants targeting the large pool of people availing the Tube in London
- Foodies and budget-conscious drinkers to find reasonable restaurants/bars on the way to home
- Budding Data Scientists, who want to explore more on the subject and discover new insight and solve various business problems

Data Preparation

Scraping London Underground Station data from Wikipedia

First used List of London Underground stations page from Wiki to scrap the table to create a dataframe

	Station	Photograph	Line(s)[*]	Local authority	Zone(s) [†]	Opened[4]	Main lineopened	Other name(s)[note 2]	Usage[5]
1	Acton Town	NaN	DistrictPiccadilly	Ealing	3	1 July 1879	NaN	Mill Hill Park: 1879–1910	5.99
2	Aldgate	NaN	Metropolitan[a]Circle	City of London	1	18 November 1876	NaN	NaN	8.47
3	Aldgate East	NaN	Hammersmith & City[d]District	Tower Hamlets	1	6 October 1884resited 31 October 1938	NaN	Commercial Road: Proposed before opening	13.17
4	Alperton	NaN	Piccadilly[h]	Brent	4	28 June 1903	NaN	Perivale-Alperton: 1903–10	2.82
5	Amersham	NaN	Metropolitan	Chiltern	9	1 September 1892	NaN	Amersham: 1892–1922Amersham & Chesham Bois: 19...	2.10

After little manipulation of the data, the dataframe is obtained as below

	Station	Lines	Local authority	Zones
1	Acton Town	DistrictPiccadilly	Ealing	3
2	Aldgate	MetropolitanCircle	City of London	1
3	Aldgate East	Hammersmith & CityDistrict	Tower Hamlets	1
4	Alperton	Piccadilly	Brent	4
5	Amersham	Metropolitan	Chiltern	9

Next used [Geo details of London Underground stations]() to get the Coordinates and created the below dataframe from the table

	Name	Latitude	Longitude	Line
1	Acton Town	51.502500	-0.278126	District, Piccadilly
2	Acton Central	51.50883531	-0.263033174	London Overground
3	Acton Central	51.50856013	-0.262879534	London Overground
4	Aldgate	51.51394	-0.07537	Metropolitan
5	Aldgate East	51.51514	-0.07178	District, Hammersmith & City

Joined the above 2 dataframes to obtain the dataframe as below

	Station	Lines	Local authority	Zones	Name	Latitude	Longitude	Line
0	Acton Town	DistrictPiccadilly	Ealing	3	Acton Town	51.502500	-0.278126	District, Piccadilly
1	Aldgate	MetropolitanCircle	City of London	1	Aldgate	51.51394	-0.07537	Metropolitan
2	Aldgate East	Hammersmith & CityDistrict	Tower Hamlets	1	Aldgate East	51.51514	-0.07178	District, Hammersmith & City
3	Alperton	Piccadilly	Brent	4	Alperton	51.54097	-0.30061	Piccadilly
4	Amersham	Metropolitan	Chiltern	9	Amersham	51.67435	-0.60732	Metropolitan

Dropped the common columns to obtain the final dataframe for analysis

	Station	Lines	Local authority	Zones	Latitude	Longitude
0	Acton Town	DistrictPiccadilly	Ealing	3	51.502500	-0.278126
1	Aldgate	MetropolitanCircle	City of London	1	51.51394	-0.07537
2	Aldgate East	Hammersmith & CityDistrict	Tower Hamlets	1	51.51514	-0.07178
3	Alperton	Piccadilly	Brent	4	51.54097	-0.30061
4	Amersham	Metropolitan	Chiltern	9	51.67435	-0.60732

After little more playing around with pandas, could get one well-arranged dataframe as below where duplicate records and removed and Latitude & Longitude values were standardized

	Station	Lines	Local authority	Zones	Latitude	Longitude
0	Acton Town	DistrictPiccadilly	Ealing	3	51.5025	-0.278126
1	Aldgate	MetropolitanCircle	City of London	1	51.51394	-0.07537
2	Aldgate East	Hammersmith & CityDistrict	Tower Hamlets	1	51.51514	-0.07178
3	Alperton	Piccadilly	Brent	4	51.54097	-0.30061
4	Amersham	Metropolitan	Chiltern	9	51.67435	-0.60732

Foursquare Location Data

Foursquare outlines high-level venue categories with sub-categories, used Foursquare API to explore food venues surrounding each station using **Food sub-category id (4d4b7105d754a06374d81259)**. Queried 100 popular spots in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance within each station.

Created a function which returns all the Food Venues by category for the neighborhood passed as an argument. The call returns a JSON file is converted into a dataframe

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Zones
0	Acton Town	51.5025	-0.278126	The Apple Tree Cakes	51.503174	-0.280474	Coffee Shop	3
1	Acton Town	51.5025	-0.278126	WP Fish X Chips	51.502580	-0.281240	Fish & Chips Shop	3
2	Acton Town	51.5025	-0.278126	London Transport Museum Depot	51.504175	-0.280622	Museum	3
3	Acton Town	51.5025	-0.278126	Gunnersbury Park Museum	51.499661	-0.286223	Museum	3
4	Acton Town	51.5025	-0.278126	Acton Centre	51.506608	-0.266878	Gym / Fitness Center	3

Since I am interested in Restaurants & Bars as popular spots, filtered the data frame with **Venue Category** containing the word **Restaurant** and **Bar**

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Zones
13	Acton Town	51.5025	-0.278126	casereccio	51.502436	-0.281674	Italian Restaurant	3
16	Acton Town	51.5025	-0.278126	The Corner Terrace	51.509839	-0.286459	Eastern European Restaurant	3
18	Acton Town	51.5025	-0.278126	Amigo's Peri Peri	51.508396	-0.274561	Fast Food Restaurant	3
28	Acton Town	51.5025	-0.278126	Persian Nights	51.508529	-0.282383	Middle Eastern Restaurant	3
29	Acton Town	51.5025	-0.278126	North China Restaurant	51.508251	-0.277435	Chinese Restaurant	3

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Zones
35	Acton Town	51.5025	-0.278126	The Chatsworth	51.508373	-0.276460	Cocktail Bar	3
53	Aldgate	51.51394	-0.07537	Discount Suit Company	51.516705	-0.075506	Cocktail Bar	1
96	Aldgate	51.51394	-0.07537	Kill The Cat	51.518658	-0.071516	Beer Bar	1
98	Aldgate	51.51394	-0.07537	BrewDog	51.509948	-0.080977	Beer Bar	1
105	Aldgate	51.51394	-0.07537	citizenM 7th Floor Skybar	51.510049	-0.076615	Hotel Bar	1

Methodology

Data Acquisition

- Wikipedia pages for Neighborhood
  https://en.wikipedia.org/wiki/List_of_London_Underground_stations

- Wikipedia pages for coordinates of the Neighborhood
  https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations

- Foursquare API to explore food venues surrounding each station using Food sub-category id (4d4b7105d754a06374d81259)

Exploratory data analysis and cleanup

- Filter data for Venue Category Restaurants & Bars

- Removed duplicate records and standardized the coordinates

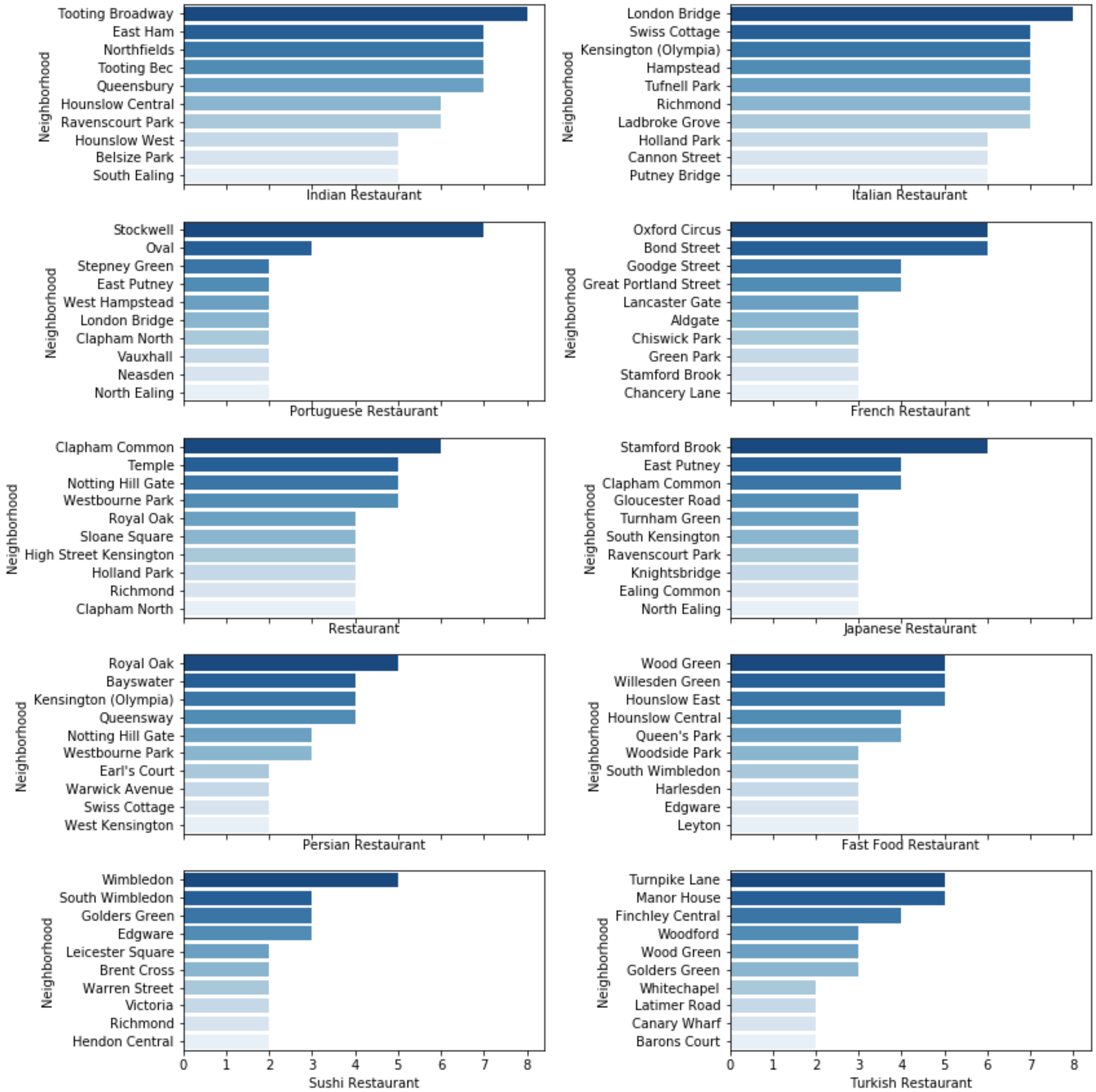
- Updated records with wrong coordinates

	Station	Lines	Local authority	Zones	Latitude	Longitude
0	Acton Town	DistrictPiccadilly	Ealing	3	51.5025	-0.278126
1	Aldgate	MetropolitanCircle	City of London	1	51.51394	-0.07537
2	Aldgate East	Hammersmith & CityDistrict	Tower Hamlets	1	51.51514	-0.07178
3	Alperton	Piccadilly	Brent	4	51.54097	-0.30061
4	Amersham	Metropolitan	Chiltern	9	51.67435	-0.60732

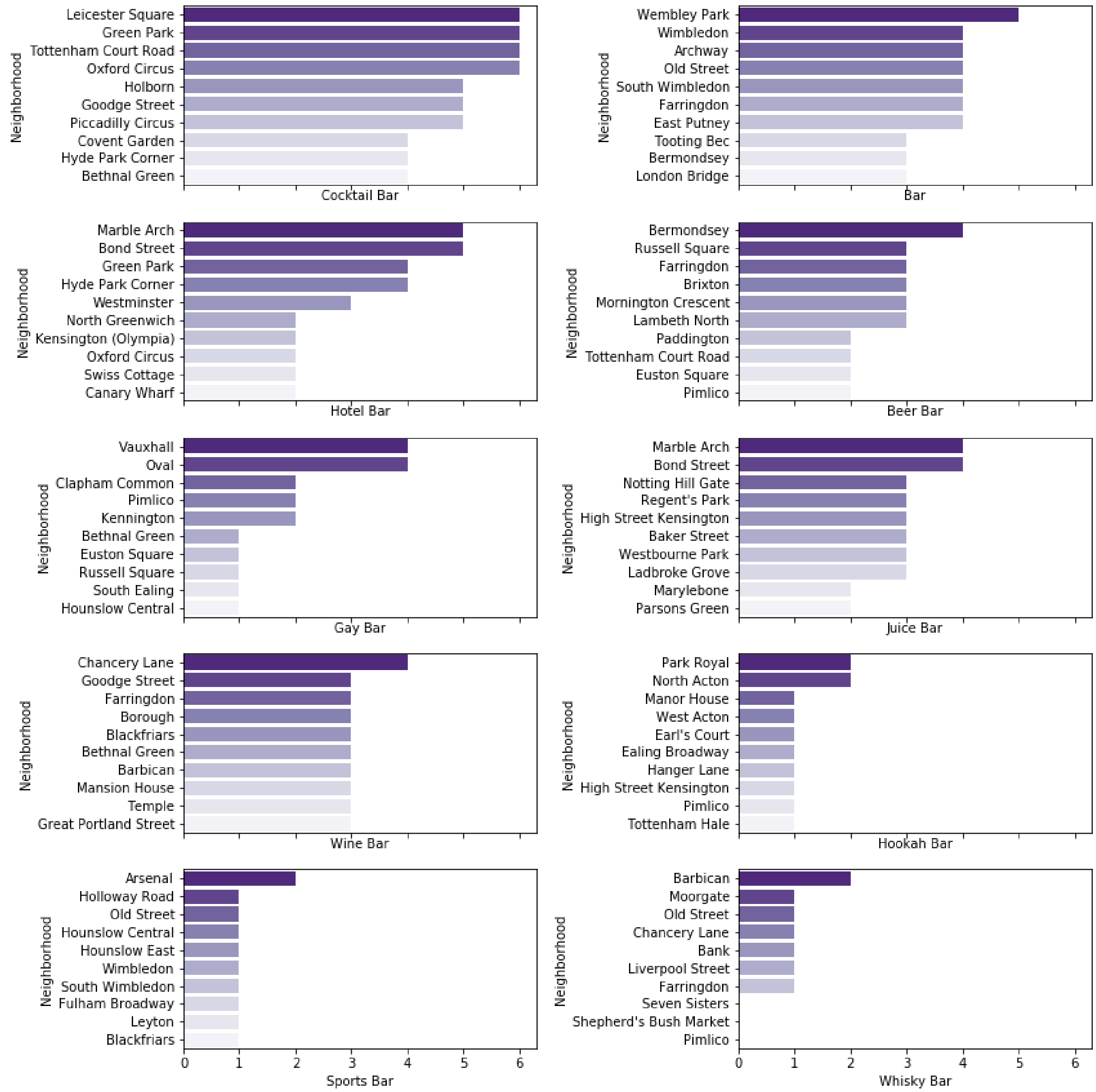
- Final dataset created by combining Wiki data with Foursquare Location data for 100 popular spots in a 1000m radius around each station

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Zones
0	Acton Town	51.5025	-0.278126	The Apple Tree Cakes	51.503174	-0.280474	Coffee Shop	3
1	Acton Town	51.5025	-0.278126	WP Fish X Chips	51.502580	-0.281240	Fish & Chips Shop	3
2	Acton Town	51.5025	-0.278126	London Transport Museum Depot	51.504175	-0.280622	Museum	3
3	Acton Town	51.5025	-0.278126	Gunnersbury Park Museum	51.499661	-0.286223	Museum	3
4	Acton Town	51.5025	-0.278126	Acton Centre	51.506608	-0.266878	Gym / Fitness Center	3

- Plotted a Bar chart with the number of Top 10 Restaurants by Station



- Plotted a Bar chart with the number of Top 10 Bars by Station



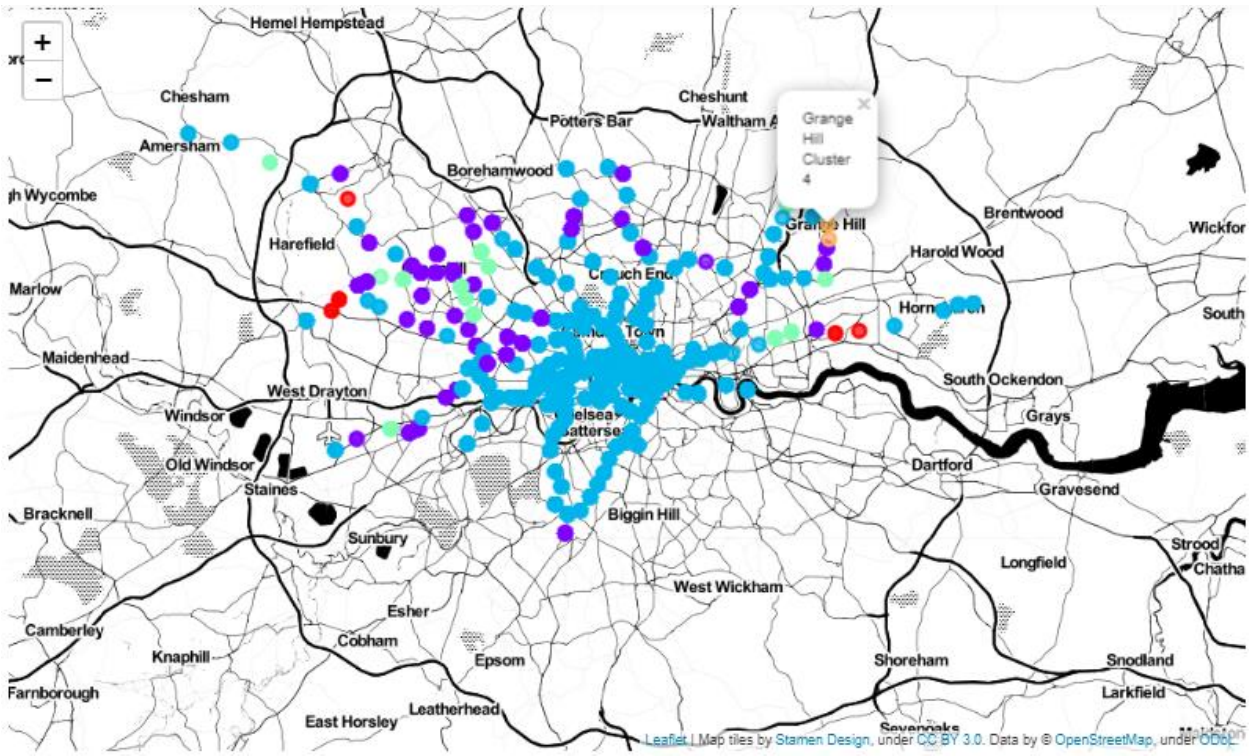
Modelling

One-hot encoding is done on the Venue Category for better fitment to ML algorithm and prediction. The Venue data is then grouped by the Station (Neighborhood) and the mean of the values are calculated. To help in finding similar Station (Neighborhood) by Restaurants and Bars, used K-means unsupervised clustering algorithm with a cluster size of 5

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Zones	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5 C
13	Acton Town	51.5025	-0.278126	casereccio	51.502436	-0.281674	Italian Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Re
16	Acton Town	51.5025	-0.278126	The Corner Terrace	51.509839	-0.286459	Eastern European Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Re
18	Acton Town	51.5025	-0.278126	Amigo's Peri Peri	51.508396	-0.274561	Fast Food Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Re
28	Acton Town	51.5025	-0.278126	Persian Nights	51.508529	-0.282383	Middle Eastern Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Re
29	Acton Town	51.5025	-0.278126	North China Restaurant	51.508251	-0.277435	Chinese Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Re

Visualization

Visualizing the 7 clusters in a London map using Folium library



Result

Took a sneak peak of the Food venues around London Tube Station neighborhood with data exploration concentrated on the Restaurants & Bars. Used data from web resources like Wikipedia along with python libraries and Foursquare API, to set up a realistic data-analysis scenario

Cluster 0

```
london_merged.loc[london_merged['Cluster Labels'] == 0, london_merged.columns[[1] + list(range(5, london_merged.shape[1]))]]
```

	Neighborhood Latitude	Venue Longitude	Venue Category	Zones	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	9t Cc
1207	51.54029	0.113216	Chinese Restaurant	5	0	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Res
6261	51.55364	-0.449054	Chinese Restaurant	6	0	Fast Food Restaurant	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Res
6264	51.55364	-0.448794	Fast Food Restaurant	6	0	Fast Food Restaurant	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Res
6813	51.56147	-0.445925	Chinese Restaurant	6	0	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fa Res
6814	51.56147	-0.446161	Indian Restaurant	6	0	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fa Res
6817	51.56147	-0.439185	Indian Restaurant	6	0	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fa Res
6818	51.56147	-0.439186	Chinese Restaurant	6	0	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fa Res
9117	51.62970512	-0.434700	Chinese Restaurant	6 & 7	0	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Res
13774	51.53824	0.092085	Chinese Restaurant	4	0	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Res
13776	51.53824	0.113216	Chinese Restaurant	4	0	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Res

Cluster 1

```
london_merged.loc[london_merged['Cluster Labels'] == 1, london_merged.columns[[1] + list(range(5, london_merged.shape[1]))]]
```

	Neighborhood Latitude	Venue Longitude	Venue Category	Zones	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	8 C
254	51.54097	-0.297200	Indian Restaurant	4	1	Asian Restaurant	Indian Restaurant	Fast Food Restaurant	Hookah Bar	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	Re
256	51.54097	-0.301996	Asian Restaurant	4	1	Asian Restaurant	Indian Restaurant	Fast Food Restaurant	Hookah Bar	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	Re
259	51.54097	-0.295594	Hookah Bar	4	1	Asian Restaurant	Indian Restaurant	Fast Food Restaurant	Hookah Bar	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	Re
266	51.54097	-0.296093	Fast Food Restaurant	4	1	Asian Restaurant	Indian Restaurant	Fast Food Restaurant	Hookah Bar	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	Re
470	51.61625	-0.132136	Beer Bar	4	1	Fast Food Restaurant	Indian Restaurant	Chinese Restaurant	Beer Bar	Vietnamese Restaurant	Greek Restaurant	English Restaurant	E Re
485	51.61625	-0.128742	Chinese Restaurant	4	1	Fast Food Restaurant	Indian Restaurant	Chinese Restaurant	Beer Bar	Vietnamese Restaurant	Greek Restaurant	English Restaurant	E Re
486	51.61625	-0.143378	Fast Food	4	1	Fast Food	Indian	Chinese	Beer Bar	Vietnamese	Greek	English	E

Cluster 2

```
london_merged.loc[london_merged['Cluster Labels'] == 2, london_merged.columns[[1] + list(range(5, london_merged.shape[1]))]]
```

	Neighborhood Latitude	Venue Longitude	Venue Category	Zones	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 Mo Comm Ven
13	51.5025	-0.281674	Italian Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Chinese Restaurant	Eastern European Restaurant	Italian Restaurant	Keb Restaur
16	51.5025	-0.286459	Eastern European Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Chinese Restaurant	Eastern European Restaurant	Italian Restaurant	Keb Restaur
18	51.5025	-0.274561	Fast Food Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Chinese Restaurant	Eastern European Restaurant	Italian Restaurant	Keb Restaur
28	51.5025	-0.282383	Middle Eastern Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Chinese Restaurant	Eastern European Restaurant	Italian Restaurant	Keb Restaur
29	51.5025	-0.277435	Chinese Restaurant	3	2	Japanese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Cocktail Bar	Chinese Restaurant	Eastern European Restaurant	Italian Restaurant	Keb Restaur

Cluster 3

```
london_merged.loc[london_merged['Cluster Labels'] == 3, london_merged.columns[[1] + list(range(5, london_merged.shape[1]))]]
```

	Neighborhood Latitude	Venue Longitude	Venue Category	Zones	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
2162	51.62611	0.046300	Indian Restaurant	5	3	Italian Restaurant	Indian Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
2165	51.62611	0.043466	Italian Restaurant	5	3	Italian Restaurant	Indian Restaurant	Vietnamese Restaurant	German Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
3115	51.65429872	-0.519777	Indian Restaurant	7	3	Indian Restaurant	Vietnamese Restaurant	Greek Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
3118	51.65429872	-0.522199	Indian Restaurant	7	3	Indian Restaurant	Vietnamese Restaurant	Greek Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
3938	51.53931	0.050107	Indian Restaurant	3 & 4	3	Indian Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Greek Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
3939	51.53931	0.053797	Fast Food Restaurant	3 & 4	3	Indian Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Greek Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
3940	51.53931	0.050299	Indian	3 & 4	3	Indian	Fast Food	Vietnamese	Greek	Empanada	English	Ethiopian	Falafel

Cluster 4

```
london_merged.loc[london_merged['Cluster Labels'] == 4, london_merged.columns[[1] + list(range(5, london_merged.shape[1]))]]
```

	Neighborhood Latitude	Venue Longitude	Venue Category	Zones	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 I Com V
5309	51.61326	0.089149	Greek Restaurant	4	4	Greek Restaurant	Vietnamese Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	Fr Resta
5620	51.60288	0.089149	Greek Restaurant	4	4	Greek Restaurant	Vietnamese Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	Fr Resta

Here is how we can categorize the Clusters

- **Cluster 0 (Red):** Chinese and Vietnamese Restaurants are predominant with no Bars
- **Cluster 1 (Violet):** Fast Food Restaurants are most common on this cluster
- **Cluster 2 (Blue):** All Restaurants & Bars are concentrated in this cluster. French, Japanese Restaurants & Bars top the list

- **Cluster 3 (Light Green):** Most of the Indian Restaurants with few Bars are clustered here
- **Cluster 4 (Orange):** Only couple of Restaurants in this cluster

Discussion

- Restaurants are more common than the Bars in each of the Station Neighborhood
- Neighborhood of London Bridge and Tooting Broadway has the highest number of Restaurants
- Liver Pool Street, Wimbledon, Tootenham Court Road Tube Stations neighborhood is dominated by Bars
- Roding Valley, Grange Hill, Buckhurst Hill to name a few are the least used Tube stations of London as a result has the least number of restaurants as per the analysis
- High Street Kensington is one of the Stations where we have the most number of Restaurants & Bars. This station is served by the Circle (yellow) line and the District (green) line, both of which are very easy to use and well-connected to major attractions in the city, justifies the reason

Recommendations

- Avoid neighborhoods in **Cluster 2**, already high concentration of Restaurants and Bars, will have intense competition
- Stations under **Cluster 4** are least used as a result has only couple of Restaurants so better to avoid
- Open new Bars in the neighborhood of **Cluster 0** with little to no competition
- **Cluster 1 & 3** are also an option for new Restaurants or Bars with moderate competition if have unique selling propositions to stand out for competition. But avoid Fast Food Restaurants and Indian Restaurants respectively

Conclusion

We have a small glimpse of how real life data science projects look like using some of the frequently used python libraries to scrap web data, explored the neighborhood of London Tube Stations using Foursquare APIs, created visualizations using Folium and various python libraries to solve a business problem. My analysis was concentrated mainly on the possibilities of opening a Restaurant or Bar targeting the huge pool of daily commuters, around 5 million per day. Surprisingly, some of the results obtained match with my experience while in London.

All of the above analyses is dependent on the adequacy and accuracy of Foursquare data. Foursquare data is limited but can provide insights into a city's development. This data could be combined with other sources to provide more accurate results. A more comprehensive analysis and future work would need to incorporate data from other external databases.

Some drawbacks of this analysis are

- clustering is completely based on the most common venues obtained from Foursquare data
- results could potentially vary if we use some other clustering techniques like DBSCAN