The Battle of the Neighbourhoods

A study of data pertaining to London venues

Shusanah Pillinger | Applied Data Science Capstone | 11 November 2019

CONTENTS

1 Background Introduction

I would like to go for a night out in London with a group of friends. We enjoy going to popular bars and pubs and tend to go to several venues on a night out. We live across London and will each travel to a neighbourhood by the underground train network - known as The Tube in London. The furthest we will travel is Zone 5 on the network. We will meet at a Tube station before proceeding to a bar or pub for drinks. After drinks we will probably head for food, so will head in a direction where restaurants are plentiful. We don't mind walking a bit further and can head home from a different Tube station if necessary. Which Tube station should we meet at for a successful night out?

2 Problem Definition

Problem 1: Which Tube stations are nearest the highest rated bars and pubs on Foursquare?

Problem 2: Of the Tube stations with at least 5 nearby bars or pubs, what are 10 most common venue categories nearby?

Problem 3: Which group of Tube stations offer the best choice for a meal after drinks?

Culminating in determining the answer to "Which Tube station should we meet at for a successful night out?"

3 Data

3.1 Data Requirements

This data science project required location information about the London Tube network which could be utilised with the Foursquare City Guide API to explore venues surrounding given locations.

Firstly, Tube station locations were required so that latitude and longitude coordinates and which Zone the stations are in could be determined. Information about London Underground stations can be found on Wikipedia [1] and Open Street Map [2]. Any duplicates and missing values needed to be investigated.

Venue data for a limited number of venues within 500m of Tube stations within Zones 1 to 5 was required so that specifically bar and pub data could be analysed further. This involved calls using the Foursquare API [3] to search nearby venues with the query “Bar”.

For each Tube station, a summary of the number of reviews and the average rating for pubs and bars in the surrounding area was required, so that those Tube stations requiring further exploration could be determined. This required data manipulation of the contents of a dataframe.

Then for the selected Tube stations with at least 5 bars or pubs, it was required to establish details of the 10 most common venue categories within 1000m of each Tube station, so that data could be grouped according to venue type. This venue data and categorisation was also collated from the Foursquare City Guide using their explore API.

Venue data pertaining to the nearby Tube station was clustered by the similarity of venue category nearby, so an area of London with plentiful restaurant choice could be determined.

3.2 Data Sources

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

[2] https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations

[3] https://developer.foursquare.com/

4 Methodology

4.1 Retrieval of London Tube Station Data

Details of all the Tube stations on the London Underground network were uploaded into a dataframe from a Wikipedia page. The data was found to include the station name, zone and lots of historical facts (which were discarded). This data was combined with latitude and longitude information from a source on Open Street Map.

4.2 Data Cleansing

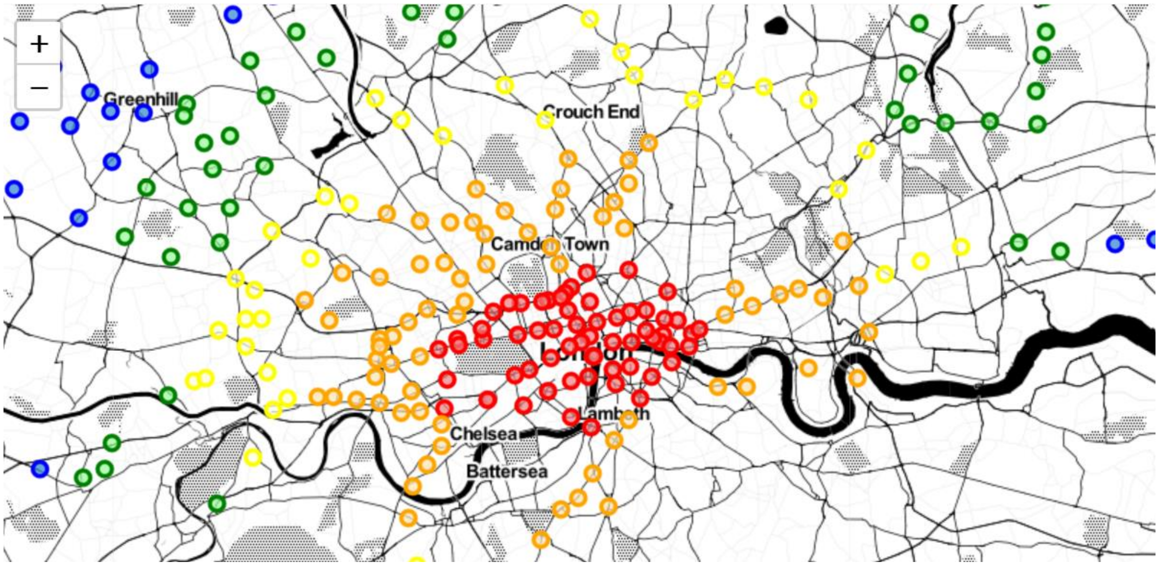
The dataset included Tube stations across the network. This was reduced to only stations in the inner 5 zones of the network. There were found to be discrepancies with the combined data: duplicate Tube stations, missing latitude and longitude values and some position data which included non-numeric values. The data was cleansed appropriately, and missing data manually included from pinpointing locations on Google Maps.

4.3 Data Storage

At various points during the analysis, output dataframes were saved to CSV files so that investigations of data could be carried out easily without having to rerun lots of notebook steps. This was particularly key with the Foursquare data, which could only be accessed by a limited number of API calls each day.

4.4 Mapping Tube Stations

A Folium Map was generated to show Tube stations across London, colour coded according to which zone the station was in. This highlighted that the dataset of Tube stations within zones 1-5 was still going to be too big to loop through within the Foursquare daily call quotas. It was established at this point that the dataset would be further reduced to just Tube stations within zone 1.



London Tube stations across the Undergound network.

Legend: Red – zone 1, Orange – zone 2, Yellow – zone 3, Green – zone 4, Blue – zone 5

4.5 Retrieval of Foursquare venue data

A developer account on the Foursquare developer portal was set up and developer credentials configured to access the underlying Foursquare venue data. The first Regular Endpoint calls retrieved a limited 5 results matching the Search query of "bar". There are a lot of bars in London and a limit was needed to restrict the number of results that would need to be looped through in a later stage (using the Premium Endpoint Explore feature, which had a lower daily quota for access).

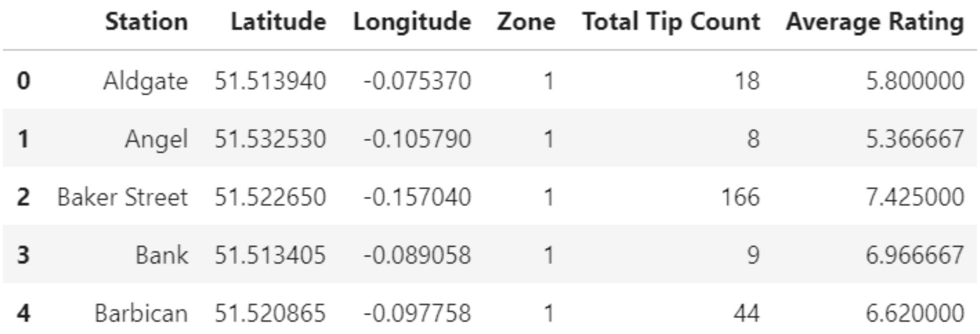
It was possible to retrieve details of bars within 500m of all 58 stations in zone 1.

4.6 Enhancement of Venue Data with Ratings and Tip Counts

Ratings and tip counts were queried using the get Venue Details endpoint of the Foursquare API for each venue. A function was written to add each bars' Foursquare rating and tip count to the details of the corresponding bar in a dataframe. Where no rating was available it was set to zero in this instance.

4.7 Summarising Tube Station Areas by Average Bar Rating and Total Tip Counts

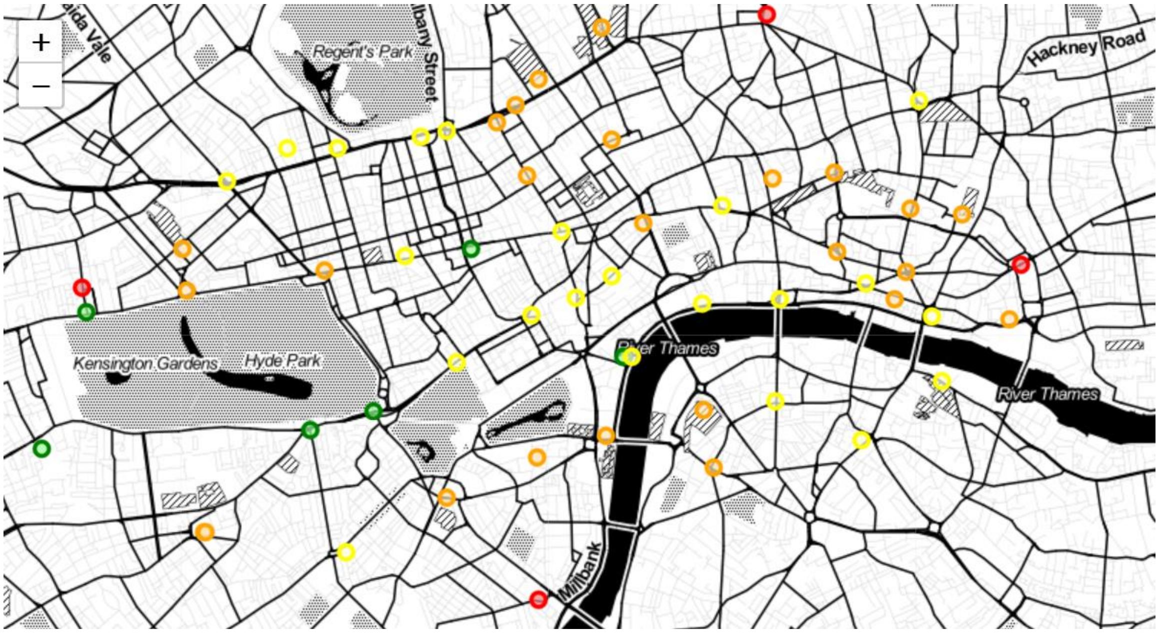
The venue data was manipulated and rolled up to a summary level by Tube station, providing results per Tube station. First 5 rows shown here:



	Station	Latitude	Longitude	Zone	Total Tip Count	Average Rating
0	Aldgate	51.513940	-0.075370	1	18	5.800000
1	Angel	51.532530	-0.105790	1	8	5.366667
2	Baker Street	51.522650	-0.157040	1	166	7.425000
3	Bank	51.513405	-0.089058	1	9	6.966667
4	Barbican	51.520865	-0.097758	1	44	6.620000

4.8 Mapping Average Rating of Bars Near Zone 1 Tube Stations

A Folium Map was generated to visualize the average rating of bars near each Tube station, within zone 1.



Average Rating of bars within 500m of London Tube stations across zone 1.

Legend: Red - rating ≤ 6, Orange – 6 < rating ≤ 7, Yellow – 7 < rating ≤ 8, Green – 8 < rating ≤ 9, Blue – rating > 9

4.9 Exploration of Venue Data around Top Rated Bar Areas

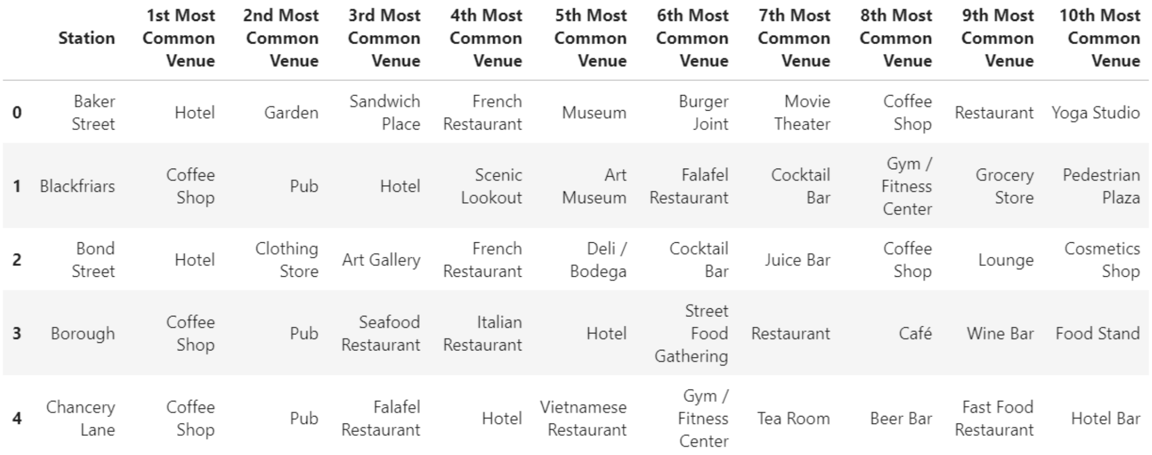
At this point it was decided to look at nearby venues for only Tube stations which had scored an average rating of 7 and above for their nearby bars. In the case of the map above, this corresponded to the yellow and green Tube stations. No Tube stations had nearby bars scoring 9 and over.

28 of the zone 1 Tube stations had 5 bars nearby with an average of 7 and above.

The Foursquare Explore endpoint was utilized to retrieve details of up to 50 venues within 1000m of these 28 Tube stations. 1400 venues were inserted into a dataframe for analysis. The data had 197 unique venue categories.

4.10 Determination of Most Common Venue Categories

One hot encoding, grouping by station, taking the mean of the frequency of occurrence of each category and resorting revealed the top 10 most common categories of venues within 1000m of each Tube station. Top 5 results shown:



	Station	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	Baker Street	Hotel	Garden	Sandwich Place	French Restaurant	Museum	Burger Joint	Movie Theater	Coffee Shop	Restaurant	Yoga Studio
1	Blackfriars	Coffee Shop	Pub	Hotel	Scenic Lookout	Art Museum	Falafel Restaurant	Cocktail Bar	Gym / Fitness Center	Grocery Store	Pedestrian Plaza
2	Bond Street	Hotel	Clothing Store	Art Gallery	French Restaurant	Deli / Bodega	Cocktail Bar	Juice Bar	Coffee Shop	Lounge	Cosmetics Shop
3	Borough	Coffee Shop	Pub	Seafood Restaurant	Italian Restaurant	Hotel	Street Food Gathering	Restaurant	Café	Wine Bar	Food Stand
4	Chancery Lane	Coffee Shop	Pub	Falafel Restaurant	Hotel	Vietnamese Restaurant	Gym / Fitness Center	Tea Room	Beer Bar	Fast Food Restaurant	Hotel Bar

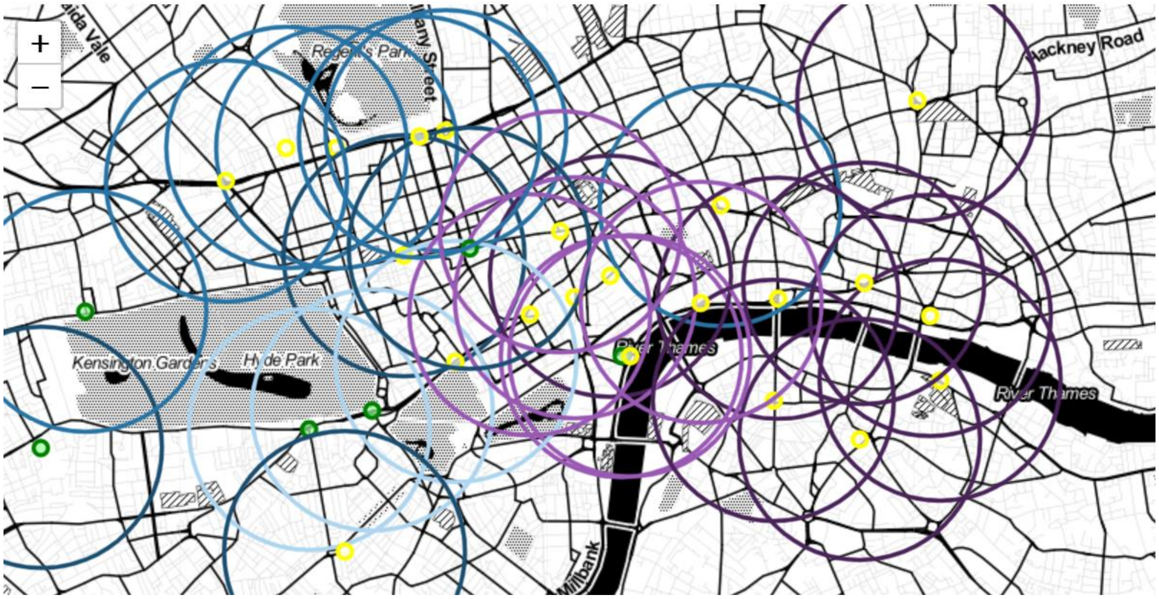
4.11 Clustering of Venue Category Data using k-Means Clustering

k-Means clustering was performed on the ranked categories for each Tube station using k=5. The five resulting clusters were examined to give them more appropriate labels based on a human interpretation of the groups. The clusters were named as follows:

Cluster	Name	Notes
1	Shopping & Restaurants	Appeared to have a range of venues featuring shops, different continental restaurants and hotels which may also have restaurants.
2	Food & Drink	Had an abundance of eating and drinking establishments.
3	Theatre	Predominantly the theatre and hotel district.
4	Business & Restaurants	Had a lot of coffee shops and fitness centres suggesting it may be providing services to businesses. There are also a range of restaurants.
5	Upmarket & Restaurants	Had more of an upmarket feel to it with boutique shopping and fancy restaurants.

4.12 Visualisation of Highly Rated Bars Alongside Clusters of Nearby Venue Categories

A Folium Map was generated to show the Tube stations in Zone 1 which had local bars with an average rating of 7 and above, alongside a circle showing the 1000m radius of each Tube station coloured according to the cluster determined by k-Means Clustering.



Tube stations across zone 1 with average rating of bars within 500m of 7 and above, overlaid with categorised cluster of venues within 1000m

Legend: Dark Blue - Shopping & Restaurants, Medium Blue – Food & Drink, Light Blue - Upmarket & Restaurants, Light Purple - Theatre, Dark Purple - Business & Restaurants

The blue circles are more associated with restaurant options than the purple areas.

5 Results and Discussion

5.1 Which Tube stations are nearest the highest rated bars and pubs on Foursquare?

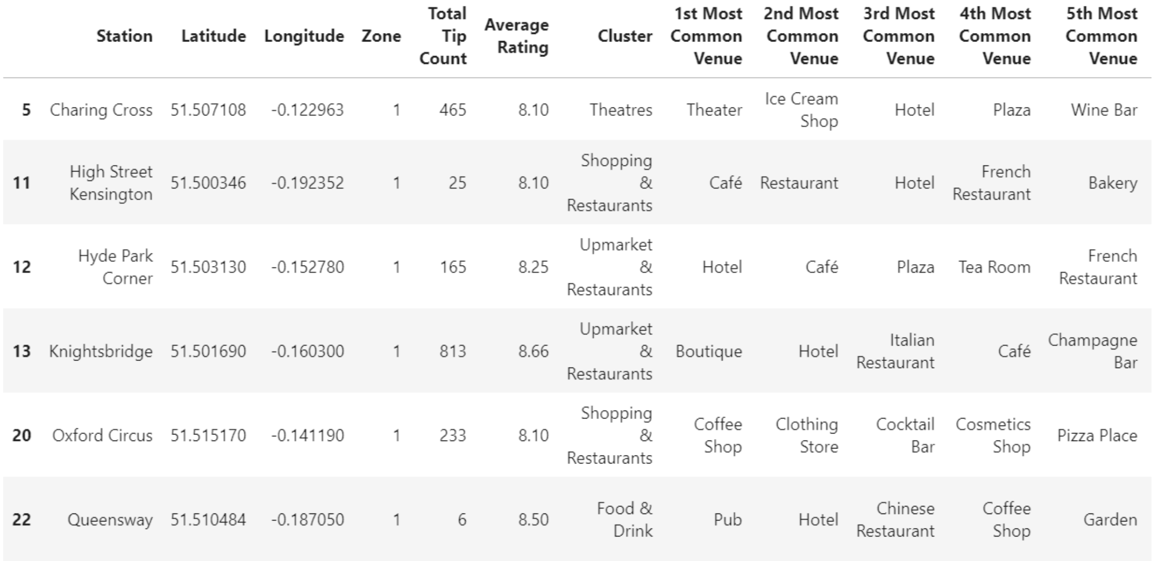
The top 6 rated Tube stations areas in Zone 1 of London, all having an average rating of over 8, are:

- Knightsbridge - 8.7
- Queensway - 8.5
- Hyde Park Corner - 8.2
- Charing Cross - 8.1

- Oxford Circus - 8.1
- High Street Kensington - 8.1

5.2 WHAT ARE 10 MOST COMMON VENUE CATEGORIES NEARBY?

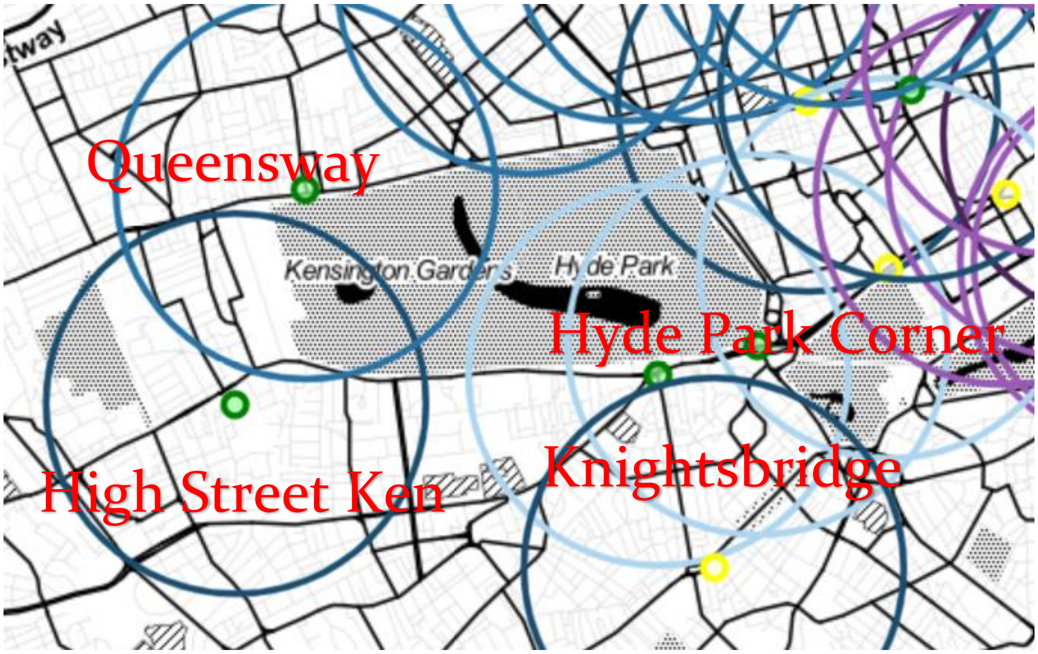
5 most common shown. See notebook for full details of top 10.



	Station	Latitude	Longitude	Zone	Total Tip Count	Average Rating	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Charing Cross	51.507108	-0.122963	1	465	8.10	Theatres	Theater	Ice Cream Shop	Hotel	Plaza	Wine Bar
11	High Street Kensington	51.500346	-0.192352	1	25	8.10	Shopping & Restaurants	Café	Restaurant	Hotel	French Restaurant	Bakery
12	Hyde Park Corner	51.503130	-0.152780	1	165	8.25	Upmarket & Restaurants	Hotel	Café	Plaza	Tea Room	French Restaurant
13	Knightsbridge	51.501690	-0.160300	1	813	8.66	Upmarket & Restaurants	Boutique	Hotel	Italian Restaurant	Café	Champagne Bar
20	Oxford Circus	51.515170	-0.141190	1	233	8.10	Shopping & Restaurants	Coffee Shop	Clothing Store	Cocktail Bar	Cosmetics Shop	Pizza Place
22	Queensway	51.510484	-0.187050	1	6	8.50	Food & Drink	Pub	Hotel	Chinese Restaurant	Coffee Shop	Garden

5.3 WHICH GROUP OF TUBE STATIONS OFFER THE BEST CHOICE FOR A MEAL AFTER DRINKS?

The overlay of clusters and the Tube stations which have bars which have an average rating of over 7, shows blue circles for restaurant related clusters, light purple for theatre related and dark purple for business related. The green coloured Tube stations, having bars rated over 8, around Hyde Park fall neatly into restaurant areas but the fact that they surround the park means that moving between areas to explore other areas will involve walking through sparsely populated areas.



5.4 Limitations

The restrictions on calls to the Foursquare API meant that the data was focused on just Zone 1 and only examining 5 bars per Tube station area. Without the restrictions it would be possible to look at the whole city and all of the bars in each area. This will produce different results.

The clusters determined by the k-means algorithm are still very broad and do not necessarily produce clusters of Tube stations that are geographically near each other. For example, Sloane Square is separated from Oxford Circus and nearby Stations by parks, despite being a similar type of area in terms of categorising venues.

Where Tube stations are very close together, there may be an element of double counting the same venues, giving an impression that there may be more choice than actually true.

Eliminating further exploration of Tube station areas where the bars have an average rating of less than 7, might miss some good nearby areas for food which might co-incidentally be not that far away from a different Tube station with a higher rating.

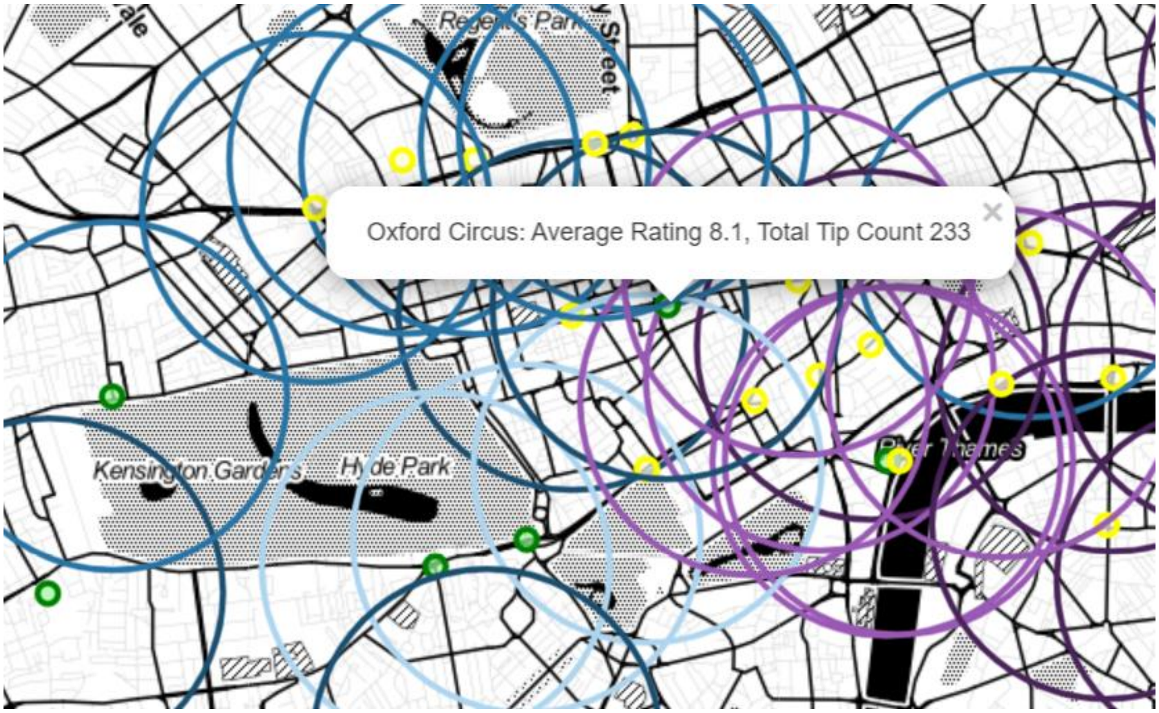
The average rating of the bars doesn't take into account the number of ratings (just that there are one or more ratings). With more time this could be refined.

6 Conclusions

Which Tube station should we meet at for a successful night out?

On balance, Oxford Circus has highly rated bars (with average rating of over 8), is in a restaurant area and is surrounded by lots of other close by Tube stations (with bars with an average rating of over 7) which themselves are in restaurant areas. It borders the

Theatre district, which is not devoid of restaurants either, and may offer a good atmosphere for the evening. Based on this preliminary analysis, our group would meet at Oxford Circus to start our evening out.


Regent's Park
Street
Oxford Circus: Average Rating 8.1, Total Tip Count 233
×
Kensington Gardens
Hyde Park
River Thames