IBM Capstone Report
Title: Locations for new coffee shops in London

Introduction / Business Problem

An entrepreneur wants to open a new coffee shop within London. There are already a lot of coffee shops so the entrepreneur wants to analyse which areas are underrepresented and have the lowest proportion of coffee shops.

People in London mainly travel using the tube, so the entrepreneur wants to analyse coffee shops based on their proximity to each tube and discover in which area the biggest opportunity lies.

The population density within a 500m radius of each tube station is unknown but there is data available about the number of people using each tube station. This third data source can be used once all of the other analysis is complete. When a shortlist of coffee shop locations has been calculated, the tube station usage data can be used to rank the potential locations in descending order of the number of people using each station.

The businessman wants to focus on Central London so we will use the zone variable to limit the train stations to zones 1 to 3.

Data

Geographical data about each of the tube stations in London and their latitude and longitude will be downloaded:
https://commons.wikimedia.org/wiki/London_Underground_geographic_maps/Tables

This consists of the values "station name", "latitude", "longitude", "zone"

id	latitude	longitude	name	display_name	zone	total_lines	rail
1	51.5028	-0.2801	Acton Town	Acton<br />Town	3	3	0
2	51.5143	-0.0755	Aldgate	NULL	1	2	0

There are 10 zones on the London Underground but we will focus on areas close to Central London, so the stations will be filtered to be between zones 1 and 3.

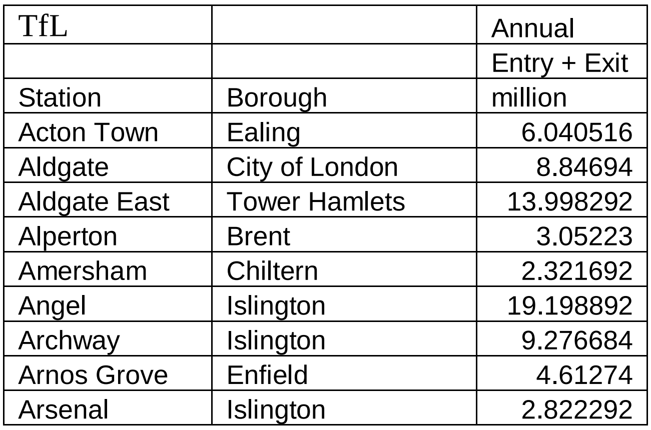
The Foursquare API will be used to analyse the proportion of coffee shops within a 500m radius of each tube station.

The venues/explore API will be used for the latitude and longitude of each tube station for venues within 500m. The venue category data will be used to calculate the saturation of coffee shops within a short distance to each tube.

This table can then be used to show the tubes which have the lowest density of coffee shops in proximity, then a k-means cluster analysis can be carried out to find similar stations based on the whole range of venue category types.

To analyse the potential number of customers in each shortlisted location a 3rd data source can be used – Transport for London exit figures for how many people have used the tube station: https://data.london.gov.uk/dataset/london-underground-performance-reports

The exit figures will be joined to the rest of the data to show which of the shortlisted store locations have the biggest potential customer base. There are some columns of data which we won't use, so the Pandas data frame will import a specific sheet from the online Excel file and drop the columns which aren't necessary.



TfL		Annual
		Entry + Exit
Station	Borough	million
Acton Town	Ealing	6.040516
Aldgate	City of London	8.84694
Aldgate East	Tower Hamlets	13.998292
Alperton	Brent	3.05223
Amersham	Chiltern	2.321692
Angel	Islington	19.198892
Archway	Islington	9.276684
Arnos Grove	Enfield	4.61274
Arsenal	Islington	2.822292

Methodology

A list of London stations with their location data was imported into a data frame from Wikimedia. The "zone" column was used to delete all stations outside of zone 3. We want to look for coffee shop locations close to Central London so only zones 1, 2 and 3 are desired.

The nominatim geolocation data was used to obtain the co-ordinates of London so that the map could be centred.

The longitude and latitude data for the tube stations in the data frame was used to plot points on the map representing each station.

The FourSquare API was used to explore venues within a 500 metre radius of each point. The points were the geolocation of each tube station. A function was used to apply the venue search to each tube station one by one.

This produced 10,748 results. The rows were converted to a one hot table to calculate the mean of each venue type to show how big a proportion coffee shops are, based on the other venues close to the tube station.

IBM Capstone Report
Title: Locations for new coffee shops in London

The coffee shop data was sorted in ascending order so that the tube stations with the fewest coffee shops appear at the top of the list.

To narrow down the search further, tube passenger data was imported into a dataframe and this was joined to the coffee shop data list. The stations without any coffee shops were then sorted in descending order based on the number of people using the station.

The top 10 opportunities were shown on a new map.

For further analysis, the station "Green Park" was used as an ideal candidate for choosing where to locate a new store. K-means clustering was used to look at the top 10 venue categories in proximity to each tube station. Once the K-means clustering had assigned a cluster number to Green Park, the data was filtered to produce all of the stations within the same cluster.

This resulted in a list of 90 tube stations that share common characteristics with Green Park. These were plotted on a final map.

Results

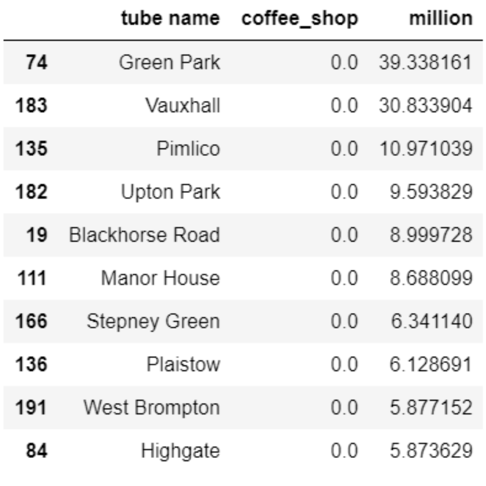
The station usage data shows that there are a very large number of potential opportunities in London. In 2017 alone there were nearly 3 billion entrances and exits of London tube stations.

There were 207 stations in zones 1 to 3 and the map shows that these are spread out across many different communities.

Based on the venue data from Foursquare, there are many stations which don't have a coffee shop in close proximity. Green Park was the most lucrative station without a coffee shop because it is used around 39 million times per year.

The top 10 candidate locations are shown on the map as being in very different locations, so further analysis of the local community and business costs would be warranted.

The top 10 store locations with passenger numbers:



	tube name	coffee_shop	million
74	Green Park	0.0	39.338161
183	Vauxhall	0.0	30.833904
135	Pimlico	0.0	10.971039
182	Upton Park	0.0	9.593829
19	Blackhorse Road	0.0	8.999728
111	Manor House	0.0	8.688099
166	Stepney Green	0.0	6.341140
136	Plaistow	0.0	6.128691
191	West Brompton	0.0	5.877152
84	Highgate	0.0	5.873629

IBM Capstone Report
Title: Locations for new coffee shops in London

Using Green Park as a candidate location, k-means clustering was used to create 5 different clusters based on the local venue data. All of the tube stations which were allocated the same cluster as Green Park were then plotted on the map because if they share the same cluster, there may well be characteristics of those other locations which also make them an ideal candidate for a new coffee shop. 90 potential locations were then plotted on the final map.

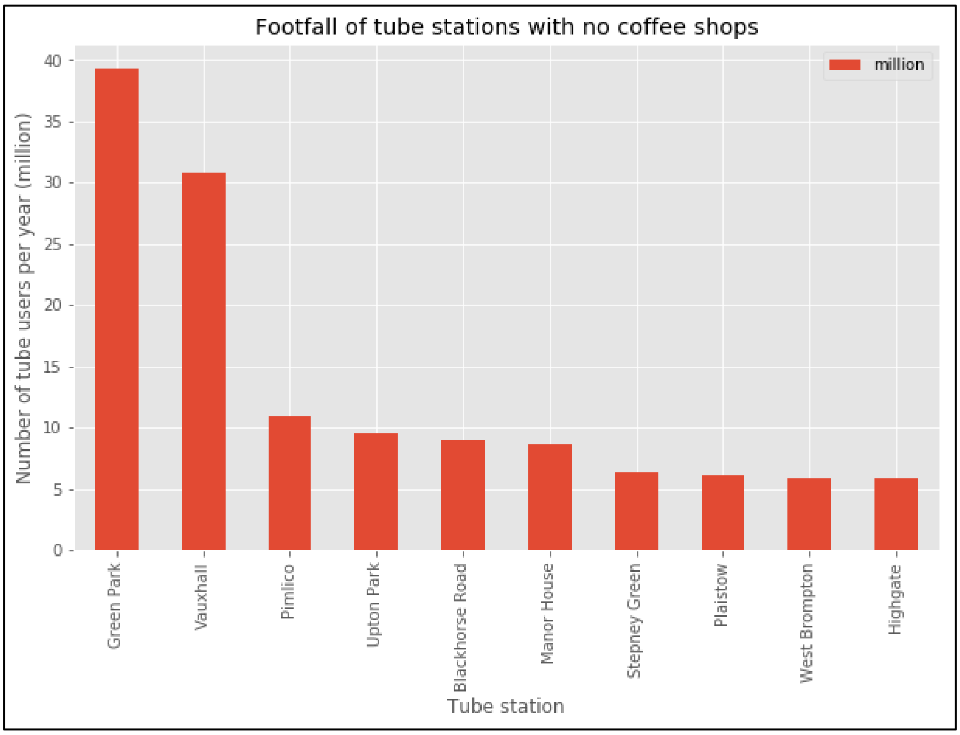
Discussion

Whilst the free access to the Foursquare data was useful to contribute to the analysis, the venue data doesn't appear to be very accurate and has lots of missing venues. A very large number of tube stations were determined to not have any coffee shops in close proximity, but a search on Google Maps shows that some coffee shops do exist. Green Park was identified in the data as being the most lucrative location for a new store, yet Google Maps shows that there is a Starbucks Coffee store right next to the station.

The tube location and passenger number data sources are of good quality, so a new study should be carried out using a different venues data source before a clear recommendation can be made.

Conclusion

The top 10 identified coffee shop locations have combined annual passenger numbers of 133 million per year. However, the Foursquare data about coffee shop venues in close proximity to each tube appears to be very inaccurate so the local areas should be analysed using a separate tool such as Google Maps to determine whether the area can support an additional coffee shop.


Footfall of tube stations with no coffee shops
million
Number of tube users per year (million)
Green Park
Vauxhall
Pimlico
Upton Park
Blackhorse Road
Manor House
Stepney Green
Plaistow
West Brompton
Highgate
Tube station