**Introduction**

**The problem**

Can we classify the London underground stations to better understand how they are being used?

**Background**

London is a huge bustling city of 9 million residents. The day to day movement of this population is tied heavily with the famous London underground system and so by categorising each station we can learn more about the population and how it flows through the city. Once categorised we would know what each station is being primarily used for and from which other stations people will be commuting.

**Interest**

Knowing the answer to this problem could help a start up business choose which area to open in or perhaps a town planner to understand the flow of the London population when considering how to further extend the rail system.

**Data**

**Data source**

In order to search the vicinity of each station we must know it’s location preferably in the form of latitude and longitude. The following url provides the names, locations and tube lines for every station in a convenient table: <https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations>

**Data cleaning**

There 270 London underground stations, whereas this list contains 302 indicating duplicates which in this case are because of stations that can be accessed from different lines. Once duplicates are considered a couple of locations will need to be converted to a float since some give some additional information in string format.

**Methodology**

To classify a station we look at the venues within a 500m radius with Foursquare query. We use the very practical **categoryID** index to categorise each venue. The primary categories that every venue must come under are:

* Arts & Entertainment
* College & University
* Event
* Food
* Nightlife Spot
* Outdoors & Recreation
* Professional & Other Places
* Residence
* Shop & Service
* Travel & Transport

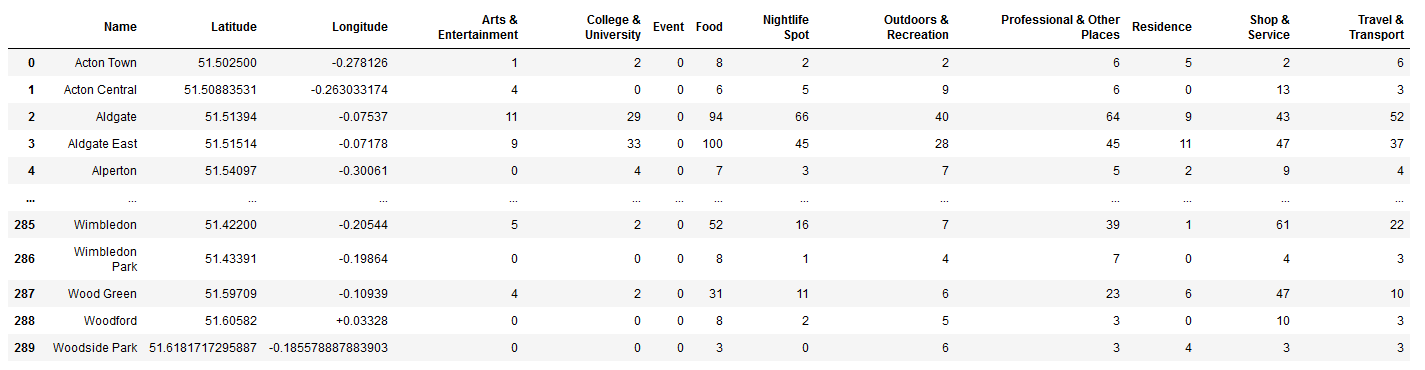
We count how many of each category each station has in its vicinity. Naturally a station in the very centre of London will have many food, arts, nightlife and event venues. Whereas a station on the outskirts will be in a residential area with far less surrounding venues altogether. These will be being used primarily for the daily commute of the local residents and won’t necessarily be a hub of activity. We could analyse the stations by location alone based on distance from the centre, but counting the venues will show stations that have a bustling commercial surrounding regardless of location.

Having scraped and cleaned the data we insert it into a pandas dataframe. Now we can use scitkit learn’s kmeans object to cluster the stations based on the counted venue types. We choose k=3 to determine the commercial, residential and the middleground business like stations that will attract a lot of commuters but not on the scale of the central stations.Once clustered we use the folium library to display our results on an adjustable map.

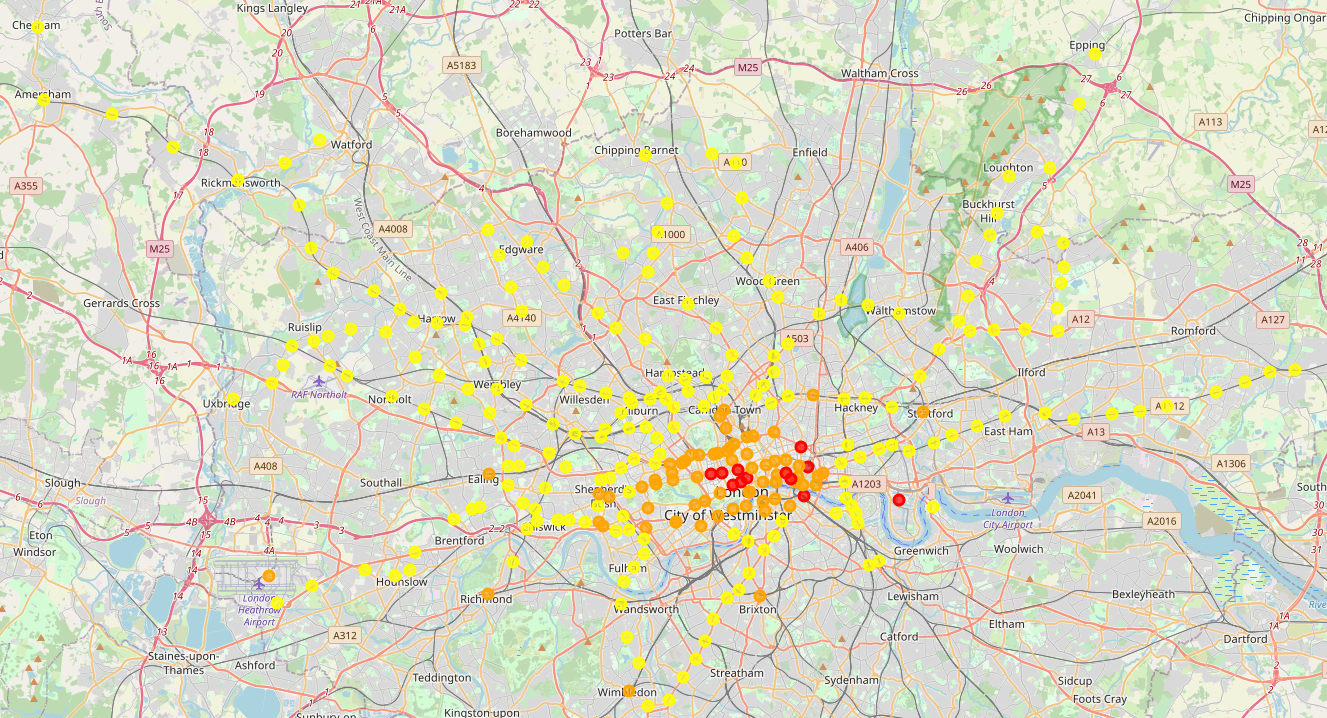
**Results**

Once the final dataframe was cleaned and ready for clustering it was already clear that the volume of local venues was clearly going to be a larger factor than the types since there was such a great polarisation in the venue count of each stations

.



Reviewing the map we mostly see what we expected, the central stations being the hubs of commercial activity and the outer stations being primarily residential stations.



Once surprising fact is the ratio of the types of stations. There are 212, 66 and 12 stations of each respective category. Meaning just 12 stations will have to deal with the majority of the incoming population.

Another interesting note is that there are notable outliers away from the central part of the city that are surrounded by a lot of venues. To the west, Heathrow airport is one of them and understandably so. But Canary wharf station is a lone red dot to the east not near any other stations but seems to be a magnet for local excitement. It must be attracting many many commuters and tourists without help from nearby stations and so would suggest the need for new stations in that area. Also as a startup business owner I would be hesitant to start near this station as it seems full of competition and yet not so easily accessible to the public.