## Introduction / Business Problem

A businessman wants to open a new coffee shop within London. There are already a lot of coffee shops so the businessman 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 businessman 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"

Eg: Acton Town, 51.5028, -0.2801, 3

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 – TfL exit figures for how many people have used the tube station:
https://data.london.gov.uk/dataset/london-underground-performance-reports