Capstone Project

Applied Data Science Capstone by IBM/Coursera

Finding Optimal Locations for New Takeaway Coffee Shops

Table of Contents

Introduction: Business Problem

An established America coffee shop company, “Coffee on the go”, plans to open a number of takeaway shops (no-sitting) in London. Given their takeaway business model, the company needs to determine where the new outlets should be optimally located within the different boroughs in Central London, within Zone 1 (the area served by the underground system in London is defined by zones and zone 1 is the most central).

The main variable taken into account to pinpoint the ideal locations is to identify the areas in Central London with the highest number of potential customers “street traffic” (people walking in a given area). From a previous consumer survey, we know that customers of takeaway coffee shops usually purchase a coffee in the morning on their way to the workplace. Since the majority of workers in central London commute to their workplace by underground, takeaway coffee shops should be located ideally close or within minimum walking distance to underground stations.

The goal is to locate those shops in such a way that all the city underground stations are within minimal walking distance. Since there are lots of coffee shops in central London, we will try to detect locations that are not already crowded with existing outlets. We are also particularly interested in areas with no coffee shops in vicinity. We would also prefer locations close to underground stations that have a high “street traffic”, assuming that first two conditions are met.

We implement a K-Median model to get the optimal location of future shops. We will use this technique to generate a few most promising neighbourhood locations based on the above criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

We decided to use a grid of locations centred around underground stations to define our potential locations. Based on the definition of our problem, factors that will be appraised to take our decision are:

- Identification of underground stations locations in central London.
- Estimate the number of commuters exiting each station on average per day.
- Number of and distance to existing coffee shops in the stations' neighbourhood.

Data acquisition and data sources

The Following data sources will be needed to extract/generate the required information:

- The list of tube stations and their exact locations are obtained from Transport for London (TfL) web-site (https://tfl.gov.uk/info-for/open-data-users).
- The information about numbers of passenger exiting each London Underground station (or number of exits) are obtained as well from Tfl; exits are defined as number of passengers passing gates or ticket barriers going from the platforms to the street.
- The number of coffee shops and their location in every underground station proximity will be obtained using the Foursquare API.
- Coordinate of central London will be obtained using standard geocoding library functions.

Additional data insight

Stations geo-locations

It can be surprisingly hard to find a nicely structured dataset of stations and geo locations. Luckily some TfL libraries had some CSVs buried in it; otherwise the following web-sites provide and alternative source of structured data on stations geo location:

- https://www.doogal.co.uk/london_stations.php
- https://commons.wikimedia.org/wiki/London_Underground_geographic_maps/CSV

Passenger counts data

On the Transport for London (TfL) web-site https://tfl.gov.uk/info-for/open-data-users, under the Network statistics tab is possible to access passenger counts data. TfL collects information about passenger numbers entering and exiting London Underground stations, largely based on the Underground ticketing system gate data. Counts data is obtained during the autumn of each year and does not necessarily reflect whole-year annual demand. The data is adjusted to remove the effect of abnormal circumstances that may affect demand such as industrial action. We use data collected by TfL based on survey data up to 2017 and reconciled to Autumn 2017 counts. The data provides the number of exits for each underground station mapped by the survey; the data number of exits are reported by Time Period, namely: {Early, AM peak, Midday, PM Peak, Evening, Late, Total day}. Exits are defined as number of passengers passing gates or ticket barriers going from the platforms to the street. The exits number by time period for each station are provided alongside the unique station number, the Borough in which the station is located and the station name itself.