Capstone Project

Applied Data Science Capstone by IBM/Coursera

Finding Optimal Locations for New Coffee Shops

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 new 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 fewer coffee shops in the 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 an optimal 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.

Methodology

The methodology used in this study consist in formulating and solving an optimization model using Decision Optimization in Watson Studio. The objective of the optimization model is to minimize the total distance from tube stations to the new coffee shops so that a commuter always gets to our new coffee shop easily. To achieve that we use a decision optimization approach. The goal is to locate those shops in such a way that 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 fewer coffee shops in the vicinity. We would also prefer locations close to underground stations that have a high "street traffic".

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. The methodology is set up in five simple steps as outlined below:

- Step 1: Import the docplex package (This package is preinstalled on Watson Studio). Note that the more global package docplex contains another sub package docplex.cp that is dedicated to Constraint Programming, another branch of optimization.
- Step2: Model the data: the data for this problem is quite simple; it is composed of the list of selected tube stations and their geographical locations. The criteria to select the tube stations will be discussed in the analysis section.
- Step 3: Prepare the data: We need to collect the list of tube stations locations and keep their names, latitudes, and longitudes. Also, we need to define how to compute the earth distance between 2 points To easily compute distance between 2 points, we use the Python package geopy.
- Step 4: Define number of shops to open: this is equivalent to create a constant that indicates how many coffee shops we would like to open, in our case we fixed this to 5 new shops.
- Step 5 Create the DOcplex model. The model contains all the business constraints and defines the objective. Namely, the objective is to minimize the total distance from tube stations to coffee shops so that a customer always gets to the new coffee shop easily.

In the following sections we will look in more details at the data and we will discuss the results more closely.

Analysis

The list of tube stations and their exact locations are obtained from Transport for London (TfL) web-site. The data downloaded as csv file contain the relevant information needed to identify and locate the tube station in central London, namely Station Name and Number, Latitude, Longitude, and Zone (a snapshot is reported in Table 1). The data collected include a total of 302 tube stations, for some stations some information is missing, because of lack of precise locations or station number, WEdecided to use only data on station with complete data available (no station from Zone 1 were missed).

Table 1: Tube Stations geo-locations

	ID	Latitude	Longitude	Station Name	Station Number	Zone	total_lines	rail
0	1	51.5028	-0.2801	Acton Town	500.0	3.0	2	0
1	2	51.5143	-0.0755	Aldgate	502.0	1.0	2	0
2	3	51.5154	-0.0726	Aldgate East	503.0	1.0	2	0
4	5	51.5407	-0.2997	Alperton	505.0	4.0	1	0
5	7	51.5322	-0.1058	Angel	507.0	1.0	1	0
6	8	51.5653	-0.1353	Archway	508.0	2.5	1	0
7	9	51.6164	-0.1331	Arnos Grove	509.0	4.0	1	0
8	10	51.5586	-0.1059	Arsenal	510.0	2.0	1	0
9	11	51.5226	-0.1571	Baker Street	511.0	1.0	5	0
10	12	51.4431	-0.1525	Balham	512.0	3.0	1	1
11	13	51.5133	-0.0886	Bank / Monument	513.0	1.0	4	0
12	14	51.5204	-0.0979	Barbican	501.0	1.0	3	0
13	15	51.5396	0.0810	Barking	514.0	4.0	2	1
14	16	51.5856	0.0887	Barkingside	515.0	5.0	1	0
15	17	51.4905	-0.2139	Barons Court	516.0	2.0	2	0
16	18	51.5121	-0.1879	Bayswater	517.0	1.0	2	0
19	21	51.5403	0.1270	Becontree	518.0	5.0	1	0
20	22	51.5504	-0.1642	Belsize Park	519.0	2.0	1	0
21	24	51.5270	-0.0549	Bethnal Green	520.0	2.0	1	0
22	25	51.5120	-0.1031	Blackfriars	521.0	1.0	2	0

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 (as described in Table 2).

Table 2: Number of exits for each underground station within a given borough

	Borough	Station Number	Station Name	Time Period	Number exiting
0	Barking	514.0	Barking	Early	1139
1	Barking	514.0	Barking	AM peak	5990
2	Barking	514.0	Barking	Midday	6226
3	Barking	514.0	Barking	PM Peak	11325
4	Barking	514.0	Barking	Evening	4925
5	Barking	514.0	Barking	Late	1903
6	Barking	514.0	Barking	Total day	31507
7	Barking	518.0	Becontree	Early	89
8	Barking	518.0	Becontree	AM peak	461
9	Barking	518.0	Becontree	Midday	1253
10	Barking	518.0	Becontree	PM Peak	2565
11	Barking	518.0	Becontree	Evening	1148
12	Barking	518.0	Becontree	Late	517
13	Barking	518.0	Becontree	Total day	6035
14	Barking	555.0	Dagenham East	Early	65
15	Barking	555.0	Dagenham East	AM peak	603
16	Barking	555.0	Dagenham East	Midday	900
17	Barking	555.0	Dagenham East	PM Peak	1946
18	Barking	555.0	Dagenham East	Evening	998
19	Barking	555.0	Dagenham East	Late	337

We aggregated the total number of exits of each station within a Borough; for a total of 32 Boroughs; it is evident that the boroughs with the highest aggregate exits' numbers are located within Zone 1 (see Table 3 for the number of exits per day by Borough). Westminster is the Borough with the highest number of exists per day with a total of 1029767 person crossing the gates to the streets of Westminster; followed by Camden, City of London, Lambeth, Southwark, etc.

Table 3: number of exits per day by Borough

	Borough	Time Period	Number exiting
31	Westminster	Total day	1029767
4	Camden	Total day	513541
5	City of London	Total day	425289
20	Lambeth	Total day	306749
27	Southwark	Total day	259328
28	Tower Hamlets	Total day	251313
23	Newham	Total day	248361
19	Kensington & Chelsea	Total day	233535
18	Islington	Total day	225838
11	Hammersmith & Fulham	Total day	183301
2	Brent	Total day	134398
12	Haringey	Total day	127585
6	Ealing	Total day	106211
1	Barnet	Total day	103118
30	Wandsworth	Total day	94630
29	Waltham Forest	Total day	88308
16	Hillingdon	Total day	67104
25	Redbridge	Total day	61971
13	Harrow	Total day	58368
22	Merton	Total day	58047
0	Barking	Total day	56715
10	Hackney	Total day	54772
9	Greenwich	Total day	43110
17	Hounslow	Total day	41350
26	Richmond	Total day	25093
8	Essex	Total day	24044
7	Enfield	Total day	23341

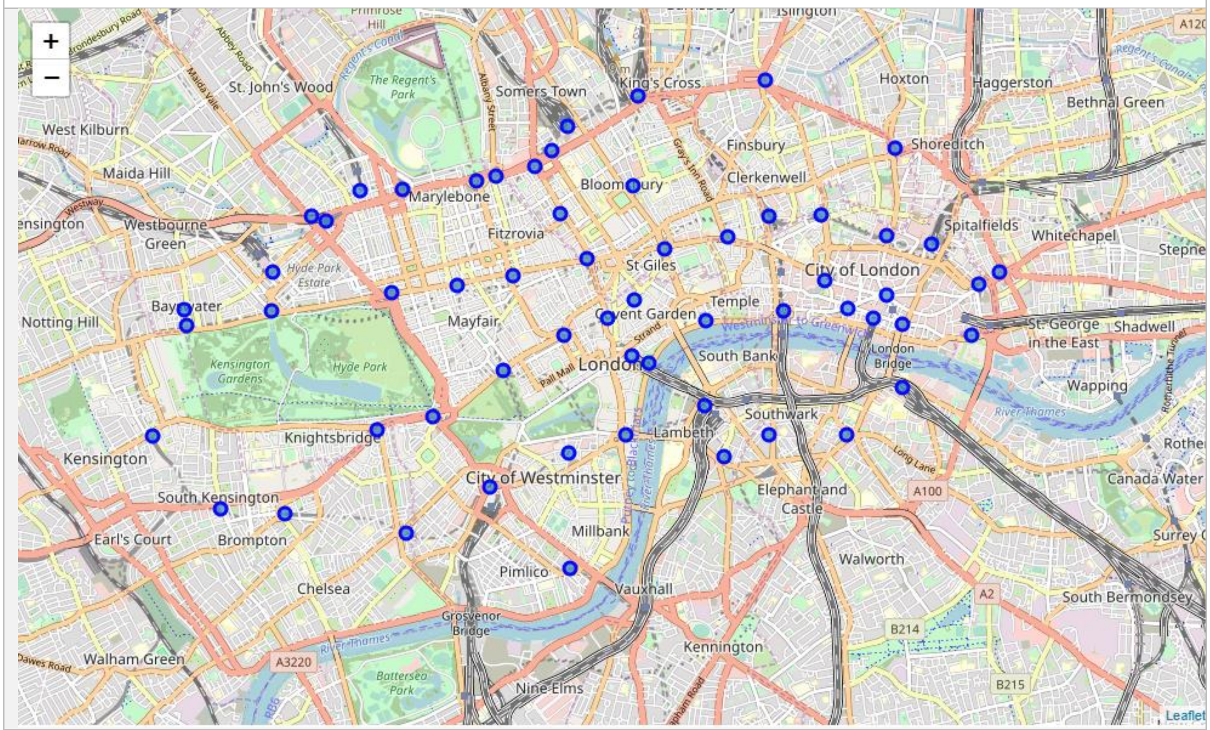
We combined the table with the stations geo-locations with the table containing information on the station exits into one table. We filter only stations within Zone 1 and we count the exits over the full day to get a proxy of the “street” traffic outside the tube station location. We obtain a total of 59 stations.

Table 4: Geo-location and number of exits per Station

	ID	Latitude	Longitude	Station Name	Station Number	Zone	Borough	Time Period	Number exiting
28	145	51.5308	-0.1238	King's Cross St. Pancras	625.0	1.0	Camden	Total day	147949
56	279	51.5036	-0.1143	Waterloo	747.0	1.0	Lambeth	Total day	147683
40	192	51.5150	-0.1415	Oxford Circus	669.0	1.0	Westminster	Total day	138502
4	13	51.5133	-0.0886	Bank / Monument	513.0	1.0	City of London	Total day	136749
5	166	51.5108	-0.0863	Bank / Monument	513.0	1.0	City of London	Total day	136749
54	273	51.4965	-0.1447	Victoria	741.0	1.0	Westminster	Total day	127176
33	156	51.5178	-0.0823	Liverpool Street	634.0	1.0	City of London	Total day	113087
34	157	51.5052	-0.0864	London Bridge	635.0	1.0	Southwark	Total day	108319
41	193	51.5154	-0.1755	Paddington	670.0	1.0	Westminster	Total day	81691
24	107	51.5067	-0.1428	Green Park	590.0	1.0	Westminster	Total day	69611
18	89	51.5282	-0.1337	Euston	574.0	1.0	Camden	Total day	64272
52	259	51.5165	-0.1310	Tottenham Court Road	728.0	1.0	Westminster	Total day	63613

We used python **folium** library to visualize geographic details of central London tube stations and its boroughs and we created a map of London with tube stations superimposed on top. We used latitude and longitude values to get the visual as in Chart 1 below.

Chart 1: Central London tube stations locations



Exploring the neighbourhoods around tube stations and locating existing coffee shops

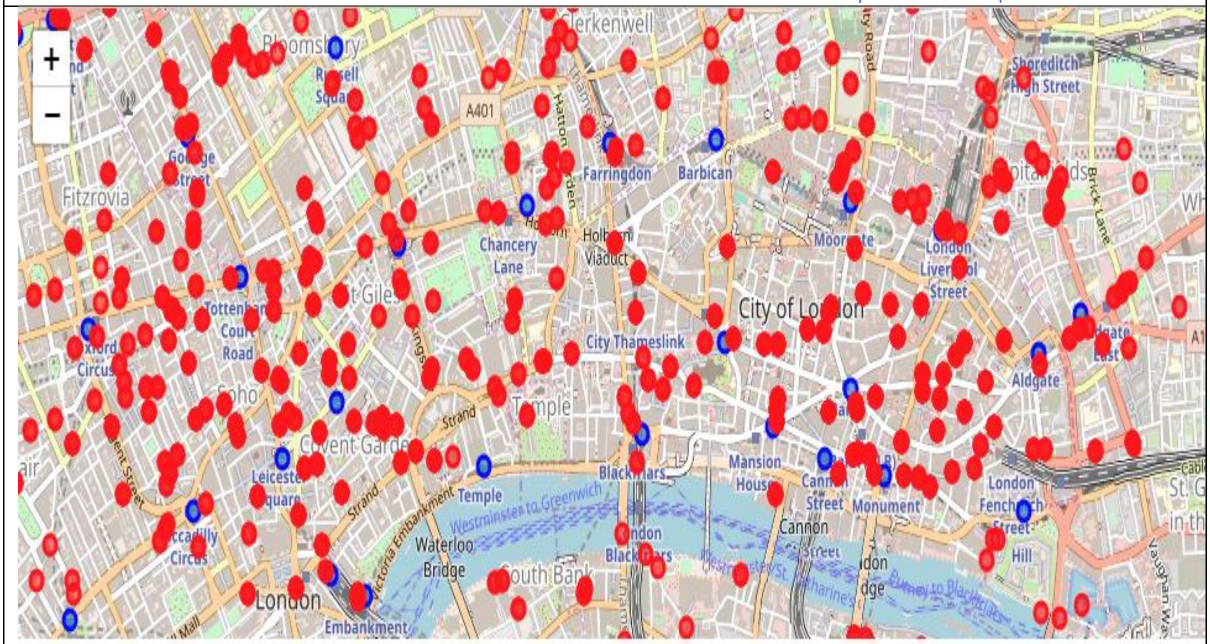
We explore the neighbourhoods around tube stations in London to locate existing coffee shops. The location of existing coffee shops situated in the proximity of each underground station is obtained using the Foursquare API. We utilized the Foursquare API to explore the presence of coffee shops within the area designated by the tube stations locations. We designed a query that limits to 100 the number of coffee shops within a radius of 500 meter from each tube station; we use the latitude and longitude information from the tube station data. Using the criteria set above, we obtained from Foursquare API a list of 1473 coffee shops locations. The table below gives an illustration of some of the coffee shops located around the 500-meter radius from Aldgate tube station.

Table 5: Central London tube stations locations

	Borough	Number exiting	Station Name	address	categories	distance	lat	lng	name	postalCode
0	City of London	16167	Aldgate	126 Whitechapel High St	Coffee Shop	210	51.514966	-0.072658	Costa Coffee	E1 7PU
1	City of London	16167	Aldgate	9 Aldgate High St	Coffee Shop	34	51.513990	-0.075459	Black Sheep Coffee	EC3N 1AH
2	City of London	16167	Aldgate	90 Mansell St	Coffee Shop	413	51.511502	-0.071572	Costa Coffee	E1 8AL
3	City of London	16167	Aldgate	2 Leman St	Coffee Shop	248	51.515066	-0.072126	Black Sheep Coffee	E1 8FA
4	City of London	16167	Aldgate	83 Whitechapel High St	Coffee Shop	402	51.515919	-0.070309	Exmouth Coffee	E1 7QX
5	City of London	16167	Aldgate	133, Whitechapel High Street	Coffee Shop	157	51.514746	-0.073347	Department of Coffee & Social Affairs	NaN
6	City of London	16167	Aldgate	122 Leadenhall St	Coffee Shop	461	51.513736	-0.082098	Black Sheep Coffee	EC3V 4AB
7	City of London	16167	Aldgate	30 St Mary Axe	Coffee Shop	360	51.514643	-0.080671	Notes Coffee Roaster & Wine Bar	EC3A 8EP
8	City of London	16167	Aldgate	NaN	Coffee Shop	320	51.512188	-0.078651	Coffee Society	EC3M 4BR
9	City of London	16167	Aldgate	NaN	Coffee Shop	305	51.514543	-0.071105	Hyde Independent Specialty Coffee Bar	E1 8EN

The map below visualizes a zoomed-in snapshot of geographic details of central London tube stations and coffee shops superimposed on top. We used latitude and longitude values to get the visual as below: Central London tube stations location (blue) and coffee shops (red)s.

Chart 2: Central London tube stations location and nearby coffee shops



Analysis of coffee shops distribution by Boroughs and tube stations

We then analysed the total number of coffee shop for each Borough in central London to find that Westminster is the borough with the highest number of coffee shops, 579 followed by City of London with 354.

Table 6: coffee shops distribution by Boroughs

	Borough	Count CoffeeShops x Borough
8	Westminster	579
1	City of London	354
0	Camden	258
7	Tower Hamlets	66
6	Southwark	58
3	Islington	46
5	Lambeth	46
4	Kensington & Chelsea	38
2	Hackney	28

As a high number of existing coffee shops could be a deterrent in finding the right location for the new outlets, we should consider these number in conjunction with street traffic defined by the total number of people exiting on the street from the nearby tube stations. To put this thought into action we create an index to proxy the potential street traffic available per coffee shops; this "traffic index" is calculated as the ratio between the total number of exits per each tube station ("Number exiting") divided by the number of existing coffee shops within 500-meter radius from the station itself ("CountCoffee Shops Station"). The median traffic-index calculated for all the existing coffee shops within our sample is 1362.5; we set this number as a criterion to select the potential location of our 5 new coffee shops and then we exclude all the station that have a Traffic index below the sample median. Henceforth, we restrict the potential location for the new coffee shops to the 29 stations locations reported in the Table 7.

Table 7: Selected stations

	ID	Latitude	Longitude	Station Name	Station Number	Zone	Number exiting	Borough	CountCoffeeShops Station	CountCoffeeShops Borough	TrafficIndex
0	7	51.5322	-0.1058	Angel	507.0	1.0	30631	Islington	14	46	2187.928571
1	11	51.5226	-0.1571	Baker Street	511.0	1.0	45436	Westminster	19	579	2391.368421
2	13	51.5133	-0.0886	Bank / Monument	513.0	1.0	136749	City of London	88	354	1553.965909
3	28	51.5142	-0.1494	Bond Street	524.0	1.0	61940	Westminster	35	579	1769.714286
4	87	51.5074	-0.1223	Embankment	542.0	1.0	35515	Westminster	18	579	1973.055556
5	89	51.5282	-0.1337	Euston	574.0	1.0	64272	Camden	17	258	3780.705882
6	99	51.4945	-0.1829	Gloucester Road	583.0	1.0	20619	Kensington & Chelsea	6	38	3436.500000
7	107	51.5067	-0.1428	Green Park	590.0	1.0	69611	Westminster	13	579	5354.692308
8	122	51.5009	-0.1925	High Street Kensington	605.0	1.0	20350	Kensington & Chelsea	6	38	3391.666667
9	133	51.5027	-0.1527	Hyde Park Corner	614.0	1.0	8198	Westminster	5	579	1639.600000
10	145	51.5308	-0.1238	King's Cross St. Pancras	625.0	1.0	147949	Camden	29	258	5101.689655
11	146	51.5015	-0.1607	Knightsbridge	626.0	1.0	23486	Kensington & Chelsea	11	38	2135.090909
12	149	51.5119	-0.1756	Lancaster Gate	629.0	1.0	9284	Westminster	6	579	1547.333333
13	156	51.5178	-0.0823	Liverpool Street	634.0	1.0	113087	City of London	48	354	2355.979167
14	157	51.5052	-0.0864	London Bridge	635.0	1.0	108319	Southwark	26	58	4166.115385
15	163	51.5225	-0.1631	Marylebone	641.0	1.0	22220	Westminster	16	579	1388.750000
16	188	51.5263	-0.0873	Old Street	665.0	1.0	42022	Hackney	28	28	1500.785714
17	192	51.5150	-0.1415	Oxford Circus	669.0	1.0	138502	Westminster	49	579	2826.571429
18	193	51.5154	-0.1755	Paddington	670.0	1.0	81691	Westminster	19	579	4299.526316
19	197	51.5098	-0.1342	Piccadilly Circus	674.0	1.0	57840	Westminster	39	579	1483.076923

Result and Discussion

We used python **folium** library to visualize geographic details of the potential locations for the new coffee shops using the 29 stations locations which have a traffic index above median. However, after the data are displayed it is impossible to determine where to ideally open the coffee shops by just looking at the map.

Chart 3: Selected tube locations which have a traffic index above median

Let's set up DOcplex to write and solve an optimization model that will help us determine where to locate the coffee shops in an optimal way. The optimization process has identified Bank/Monument, Westminster, King's Cross St. Pancras, Lancaster Gate and Hyde Park Corner as the optimal location for the new 5 coffee shops. The solution can be analysed by displaying the location of the coffee shops on a map. Displaying the solution Coffee shops are highlighted in red.

Optimization model result: Optimal locations for 5 new coffee shops		
Station proximity	*Latitude*	*Longitude*
Bank / Monument	*51.5133*	*-0.0886*
Westminster	*51.5010*	*-0.1254*
King's Cross St. Pancras	*51.5308*	*-0.1238*
Lancaster Gate	*51.5119*	*-0.1756*
Hyde Park Corner	*51.5027*	*-0.1527*

Chart 4: Optimal solution for the new coffee shops locations

Conclusion

In this study we implemented a model to identify the optimal geo-location of new shops. We used this technique to generate a few most promising locations based on some filtering criteria and an optimization framework. 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). The goal was to optimal locate those new shops in such a way that all the city underground stations are within minimal walking distance.

References

- [Decision Optimization CPLEX Modelling for Python documentation]
- [Watson Studio documentation]