CAPTSTONE PROJECT: BATTLE OF THE NEIGHBORHOODS

I.   Purpose:

This document provides the details of my final peer reviewed assignment for the IBM Data Science Professional Certificate program – Coursera Capstone.

II.   Introduction:

Glasgow is the most populous city in Scotland, and the third most populous city in the United Kingdom, as of the 2017 estimated city population of 621,020. Historically part of Lanarkshire, the city now forms the Glasgow City council area, one of the 32 council areas of Scotland;Glasgow is situated on the River Clyde in the country's West Central Lowlands. It is the fifth most visited city in the UK.

Glasgow has a population density of 3,400 people per square kilometer, which makes it the most densely populated city in Scotland. The larger Greater Glasgow area has an estimated population of 1.2 million, while the region surrounding the conurbation has about 2.8 million residents. This represents about 42% of the population of Scotland. Additionally, the city proper takes 175 square kilometers of area (approximately 68 square miles), while the metro area expands out to 3,338 square kilometers (about 1,289 square miles).

Coffee is the most popular drink worldwide with around two billion cups consumed every day. In the UK, we now drink approximately 95 million cups of coffee per day. The coffee industry creates over 210,000 UK jobs. The Gross Value-Added contribution from the UK coffee industry to the economy is estimated to be £9.1 billion, whilst output contribution, including indirect and induced multiplier impacts, of £17.7 billion in 2017.

In this project, we will atempt to use FourSquare and K-Means clustering to find the optimal location for opening a new cafe.

III.   Data Acquisition:

This demonstration will make use of the following data sources:

- Greater Glasgow & Clyde areas and their size and population density: https://en.wikipedia.org/wiki/List_of_places_in_Glasgow
- Glasgow Top Venue Recommendations from FourSquare API (FourSquare website: www.foursquare.com)

I will be using the FourSquare API to explore areas in Glasgow. The Foursquare explore function will be used to get the location of coffee shops in each neighborhood, and then use this feature to group the neighborhoods into clusters. The following information are retrieved on the first query. Using these results, we can determine which neighborhood has the most coffee shops, where the most frequented coffee shop locations are and use this to find the best location for opening a new business.

IV.   Methodology:

To get the required data, we first extracted the data concerning the areas of Glasgow – such as population density, coordinates and size.

	District	Population	Area (km^2)	Density (/km^2)
1	Govanhill	9,725	0.86	11,308
2	Pollokshields	9,738	1.59	6,125
3	Partick	8,884	0.85	10,452
4	Hillhead	6,275	0.96	6,536
5	Govan	5,860	1.63	3,595
6	Gorbals	6,030	0.83	7,265
7	Shawlands	7,015	0.52	13,490
8	Langside	4,425	0.46	9,620
Σ	Total	57,952	7.7	7,526

We can see from the table above that Shawlands has the highest population density, followed by Govanhill and Partick. So we can look to these three districts first. A quick google search will show us that Partick would be a better choice, as its close to

the more popular west end living location and also close to the University of Glasgow, making it more likely that a great number of students will live in the area, and it is known that students are avid consumers of coffee.

	District	Population	Area (km²)	Density (/km²)	District_Coord	Latitude	Longitude
0	Govanhill	9,725	0.86	11,308	(55.8363741, -4.2581531)	55.836374	-4.258153
1	Pollokshields	9,738	1.59	6,125	(55.8422663, -4.2849973)	55.842266	-4.284997
2	Partick	8,884	0.85	10,452	(55.8699211, -4.3094365)	55.869921	-4.309437
3	Hillhead	6,275	0.96	6,536	(55.8752091, -4.293281)	55.875209	-4.293281
4	Govan	5,860	1.63	3,595	(55.860879, -4.3185273)	55.860879	-4.318527
5	Gorbals	6,030	0.83	7,265	(55.851813, -4.2531625)	55.851813	-4.253163
6	Shawlands	7,015	0.52	13,490	(55.8292301, -4.2924584)	55.829230	-4.292458

I then used the Nominatim module to find the coordinates of each district, which will be used with the FourSquare API to get information on the venues in each of the respective districts. I also added two separate columns, each containing the Latitude and Longitude of these coordinates, for ease of analysis.

This data will be used in the second part of this project, where I will employ the FourSquare API to generate information on the venues in each district, see which district prefers coffee shops over everything else and use the population density and frequency of visits to find the best spot to open a coffee shop.