**Data Science Project: Battle of London Boroughs**

**Finding suitable locations to open a car wash in London, UK**

**February 2021**

### Table of Contents

* Introduction
* Data
* Methodology
* Analysis
* Results
* Conclusion
* References

### **Introduction**

Imagine for a moment that an investor is looking to open a car wash in London, United Kingdom. The investor crunched all of the numbers and on paper it is looking like a good business idea. However, the question now stands about the location. What would be the best place to open a new car wash to maximise return on investment? There is already thriving competition present in the city, and the investor wants to find out which one of Greater London's 32 Boroughs would be the optimum place.

To answer this question, the investor turns to the field of **Data Science** in order to explore all of the Boroughs from a deep analytical point of view. The main objective is to explore what each Borough contains, and what each Borough lacks. A perfect outcome would be to find an area that 1) isn't too crowded with competition, 2) has a lot of passing traffic and 3) relatively large population, and then to open a new car wash in that area.

### **Data**

To answer the main question of this report, data will be collected from several sources. First, a list of all 32 Greater London Boroughs (excluding city of London as it is predominantly a business district), including latitude and longitude positions and the population of each Borough which can be scraped from the Wikipedia page.

Next, it is necessary to identify the number and locations of existing car washes in London. Foursquare API is going to be used to help find and extract this information. The API allows one to search for businesses and venues around a given geographic point (like the Borough information from Wikipedia) and find out how many and what types of businesses are in that area. This data will then be extrapolated and prepared for further analysis.

In addition, vehicle traffic information (volume) of each Borough will be used from UK's Department of Transport. Data figures give the total volume of traffic across each local authority for the whole year. This will be used in order to assess how much 'potential' traffic there is in each Borough for a given car wash.

For further sources of data please refer to the References section.

### **Methodology**

There are several algorithms that can be used to cluster location data. For this particular exercise, K-means clustering is going to be used to help group car wash locations and analyse the best potential candidate.

K-means aims to partition the observations into a predefined number of clusters (k) in which each point belongs to the cluster with the nearest mean. It starts by randomly selecting k centroids and assigning the points to the closest cluster, then it updates each centroid with the mean of all points in the cluster through iterating the data. This algorithm is convenient when you need to get a precise number of groups, and it’s more appropriate for a small number of even clusters.

In order to define the right k, Elbow Method is going to be used by plotting the variance as a function of the number of clusters and picking the k that flattens the curve.

The main open-source Python libraries that are used are the following:

**Data Analysis and Manipulation**

* Pandas
* SciPy
* NumPy

**Data Visualisation**

* Matplotlib
* Seaborn
* Folium

**Machine Learning**

* Scikit-Learn

### **Analysis**

First, the abovementioned packages will be imported. Then we will scrap the list of boroughs and clean the data. Once we have a list of boroughs, we can use Nominatim to extract latitude and longitude for each area and plot it in Folium (Fig. 1):

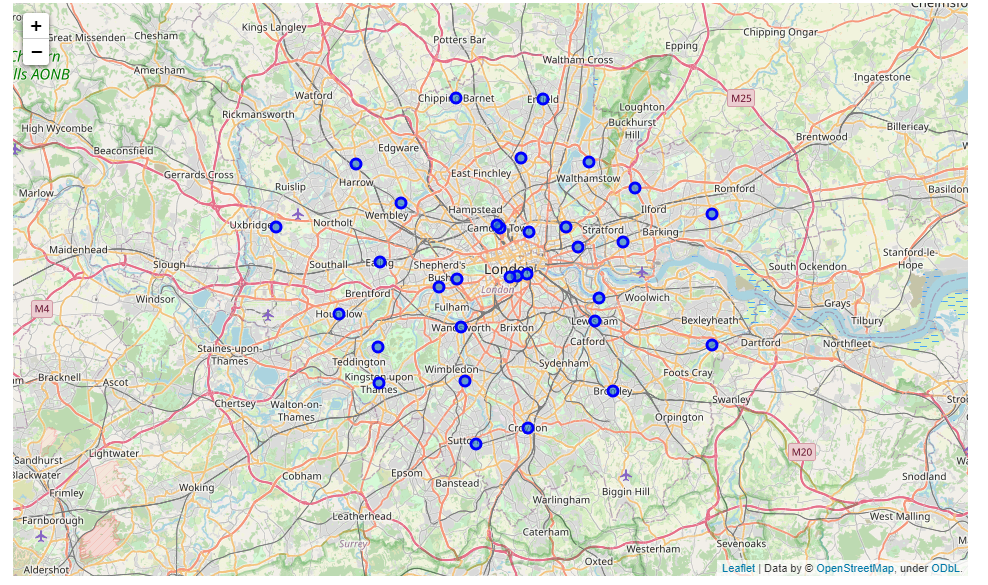


Fig. 1

We can then proceed to extracting existing business location through Foursquare API, cleaning the data and plotting it on top of London borough map (Fig. 2):

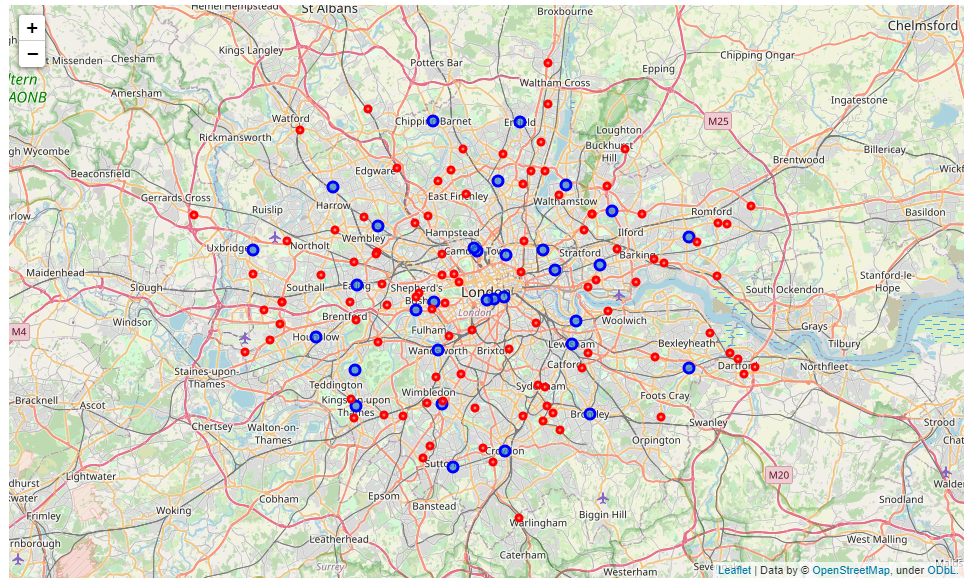


Fig. 2

Produce a heat map (Fig. 3):

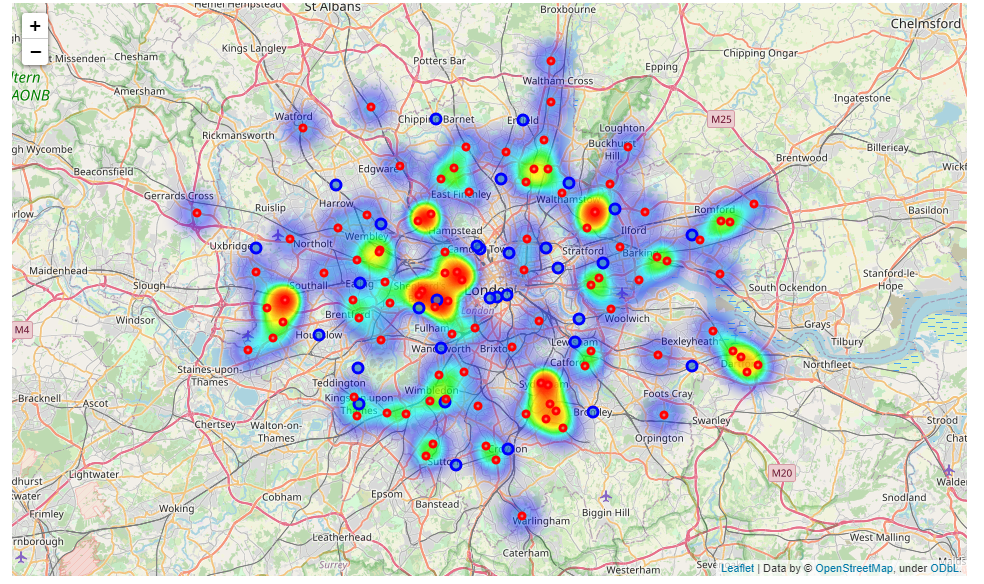


Fig. 3

And finally, we can analyse each borough (Fig. 4):

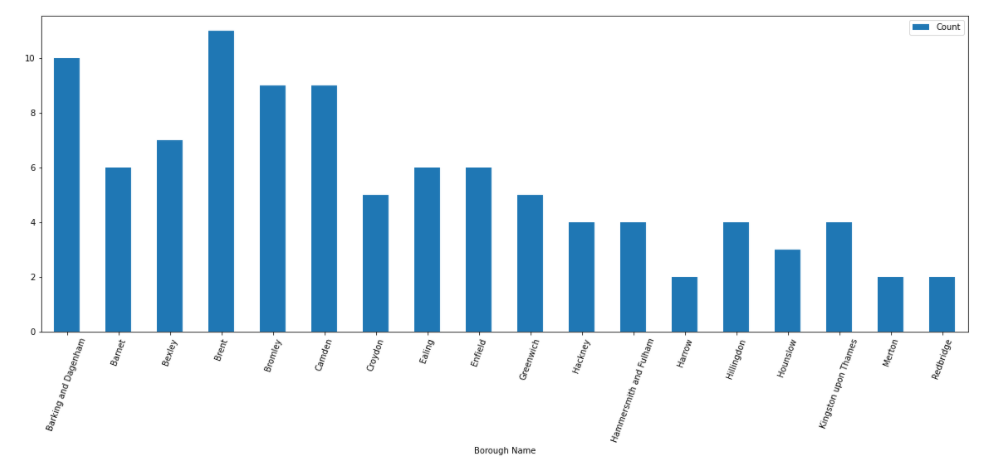


Fig. 4

Finally, to complete all the data, we can extract vehicle traffic volume information from Department for Transport. The final table contains 99 existing car wash venues and is presented below (Table 1):



Table 1

We can produce histograms for population of each borough as well as its traffic volume to get an idea of how wide the data range is (Fig. 5 & Fig. 6):

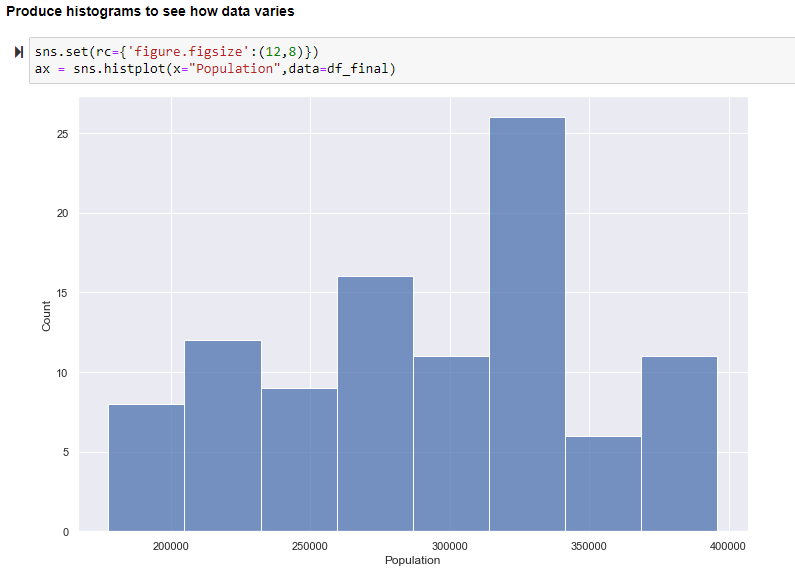


Fig. 5

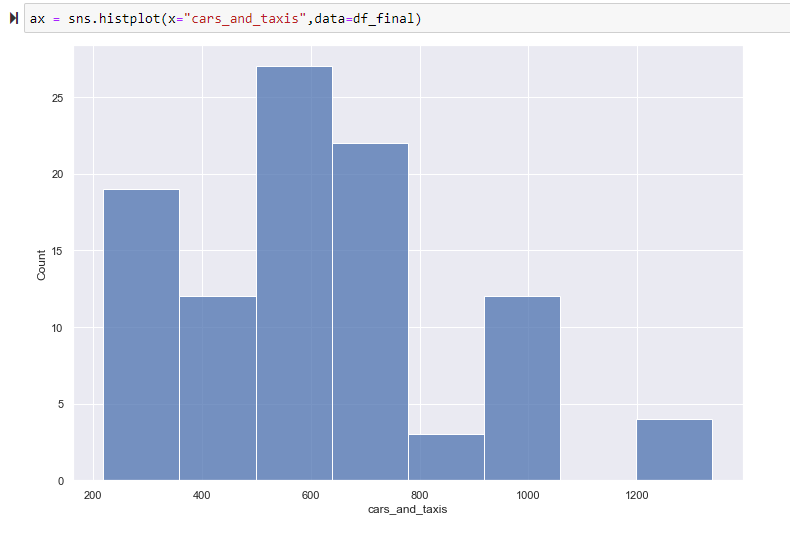


Fig. 6

We can proceed to cluster the locations. There are several algorithms that can be used to cluster data.

K-Means aims to partition the observations into a predefined number of clusters (k) in which each point belongs to the cluster with the nearest mean. It starts by randomly selecting k centroids and assigning the points to the closest cluster, then it updates each centroid with the mean of all points in the cluster. This algorithm is convenient when you need the get a precise number of groups, and it’s more appropriate for a small number of even clusters.  
  
Here, in order to define the right k, I shall use the Elbow Method (Fig. 7): plotting the variance as a function of the number of clusters and picking the k that flats the curve.

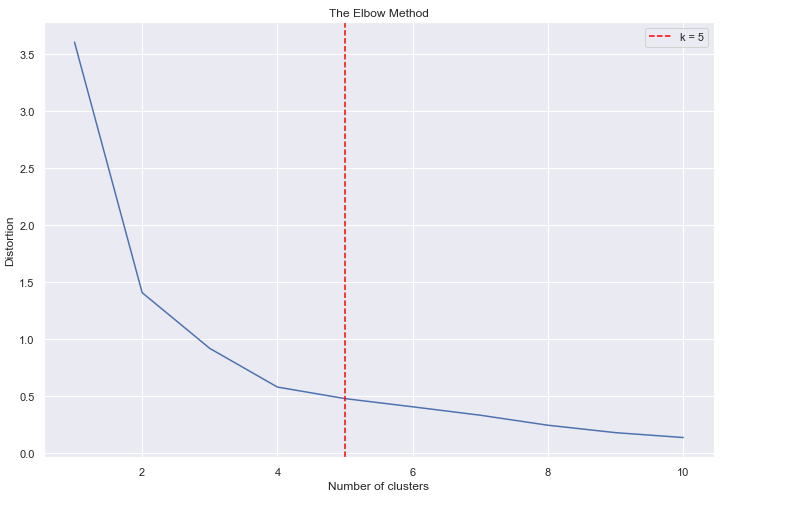


Fig. 7

It appears that k = 5 is the optimum number of clusters that needs to be used. We can then re-train the model using the correct number of clusters and proceed to analyse the results.

First, we can Plot a scatter plot with all of existing Car Washes allocated to individual clusters with their centroid points (Fig. 8):

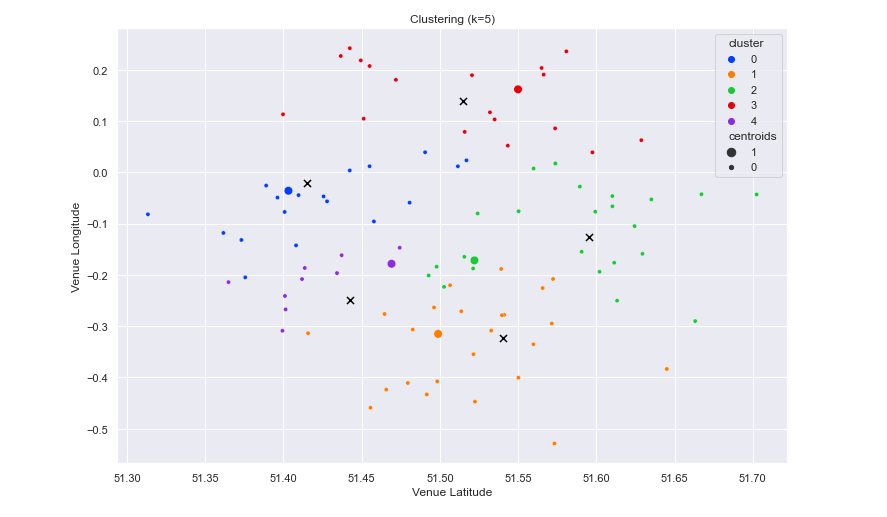


Fig. 8

And plot a map with existing clustered Car Washes with a bubble function. The bubble size represents volume of traffic in a particular area (Fig. 9).

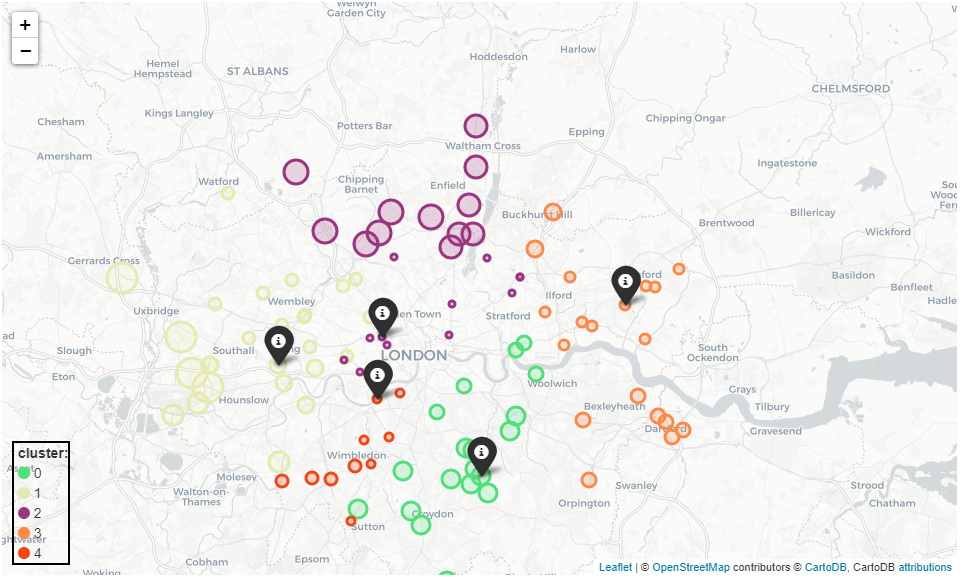


Fig. 9

It seems that the ML algorithm has split the existing facilities in the following manner:

* Cluster 1: South London
* Cluster 2: West London
* Cluster 3: North London
* Cluster 4: East London
* Cluster 5: South-West London

### **Results**

We can now proceed and analyse each cluster by dividing population size and volume of traffic by number of car washes. We are interested in lowest number of these. Final table below (Table 2):

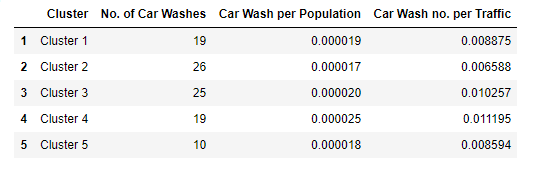


Table 2

### **Discussion**

Cluster no.2 has the lowest number of car wash facilities per population and traffic volume, despite having the highest number of Car Washes. Therefore, for a potential investor starting a new venture this area would be highly desirable. The Boroughs include:

* Brent
* Harrow
* Ealing
* Hillingdon

Indeed, this area benefits from dense population and high amount of vehicle traffic, for example due to Heathrow Airport, therefore it is an attractive location. Further analysis could benefit incorporating a cost consideration i.e., the optimum location from a financial point of view.

### **Conclusion**

This report evaluated the best location to open a new car wash in London using Data Science techniques. Publicly available data was extracted on locations of London boroughs, existing car wash locations and supporting information such as population size and vehicle volume. This data was plotted and visualized and K-Means clustering was used to partition the observations into an optimum number of clusters. The analysis concluded that cluster no.4 which is roughly the area of West London has less competition compared to other areas. Four boroughs were identified that would be the most beneficial.

### **References**

1) Wikipedia (2021), List of London Boroughs, Available at: <<https://en.wikipedia.org/wiki/List_of_London_boroughs>> (Accessed 31/01/2021)

2) Departmment for Transport (2021), Road Traffic Statistics, Available at: <<https://roadtraffic.dft.gov.uk/downloads>> (Accessed 31/01/2021)