# **Final Report: Identifying the optimal location for a distillery in North London**

## Introduction

## The purpose of this project is to identify where to optimally locate a distillery in North London - to ensure that it is financially successful. The target audience is a company which is looking to build a distillery in North London to serve North London pubs/bars. Distilleries provide their produce to pubs and bars. I need therefore to identify where in North London there is a high concentration of pubs/bars i.e. where pubs/bars are the most popular category of venue.

The problem to be solved: There are numerous pubs and bars in North London, however in a number of neighbourhoods, they are experiencing reduced footfall - which is resulting in pub closures. It is therefore not straightforward to identify where to locate a distillery which can serve parts of North London where pubs are still valued by the local community. If the distillery is located in an area which serves pubs that are not popular venues, then the distillery is unlikely to be successful.

## Data

In order to solve this problem, I will require data from two sources

:

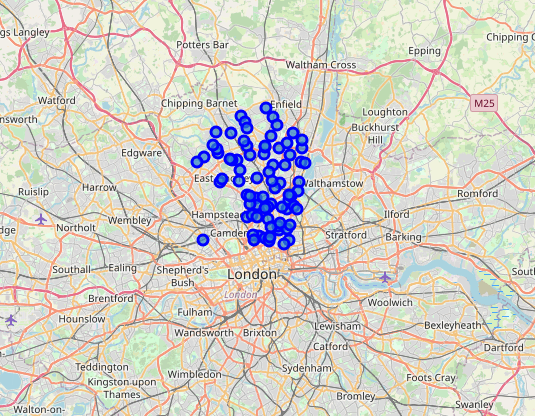
Firstly, I will need to obtain a list of North London postal codes together with their longitude, latitude, Borough and Neighbourhood. As there are a very large number of postal codes in this area, I will select a range of postal codes from different parts of North London e.g. North1, North2 etc, and identify venues on Foursquare which are within a large radius of these post codes. This will ensure that I do not exceed Foursquare's data limits. I will convert this data into a CSV file that can be easily uploaded into Python.

Secondly, I will input this data into Foursquare. I will perform a search on Foursquare to identify venues within the given radius of each postal code - i.e. a "get nearby venues" request. For each neighbourhood (which broadly corresponds to a postal code), I will identify the most popular category of venue. This will be done by assigning a dummy variable to each category of venue and sorting the venues from most to least popular for each neighbourhood. This will enable me to identify for which neighbourhoods’ pubs/bars are the most or one of the most popular category of venue. Neighbourhoods will subsequently be grouped through K-means clustering in order to identify the optimal location for the distillery.

## Methodology

Postal code data was sourced from a publicly available database, which also included details of the longitude and latitude of the postal codes, together with their Borough and Neighbourhood. This was converted into a CSV file and uploaded onto Python.

A plot of the postal codes on folium is shown below:



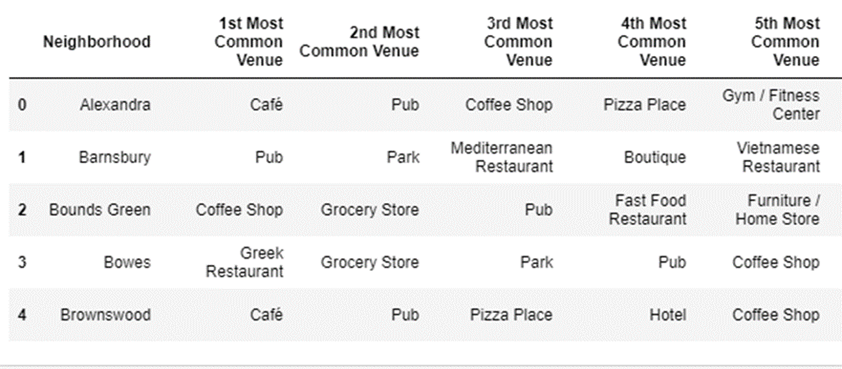
Foursquare location data: Foursquare enables users to identify venues within a certain radius of a location - through the use of longitude and latitude data. I input North London Postal Code data into Foursquare in order to identify the most popular categories of venues within a certain radius of each postal code. This was performed using a “get venues nearby” request.

Dummy variables were assigned to each venue category in order to help identify which categories of venue were the most popular for each postal code.

The results of the Foursquare search were grouped by neighbourhood (which aligns with the Postal Codes).

For each neighbourhood, the most popular category of venue was identified.

The table below displays the output of the analysis of Foursquare locational data. In particular, for each area, the most popular categories of venue are identified.



At this point, the data was ready to be clustered. Clustering would help to identify similar neighbourhoods based on the most popular categories of venue within each neighbourhood.

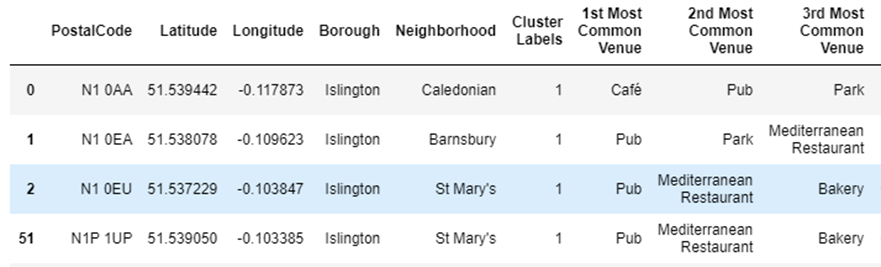
K-Means clustering is a form of machine learning which enables data points to be grouped together based on their characteristics. In this case, we would be grouping neighbourhood data into clusters based on the popularity of different categories of venue.

If a group of neighbourhoods can be identified where pubs are amongst the most popular category of venue, then that will help with identifying where to locate the distillery.

In particular, the distillery would need to be located in a location which is in close proximity to the neighbourhoods where pubs are the most popular or one of the most popular categories of venue.

Through trial and error, it was identified that two clusters would yield a suitable split of neighbourhoods between those where pubs were popular venues versus those where pubs were less popular.

The table below displays the output of the K-Means clustering of the geolocational data.

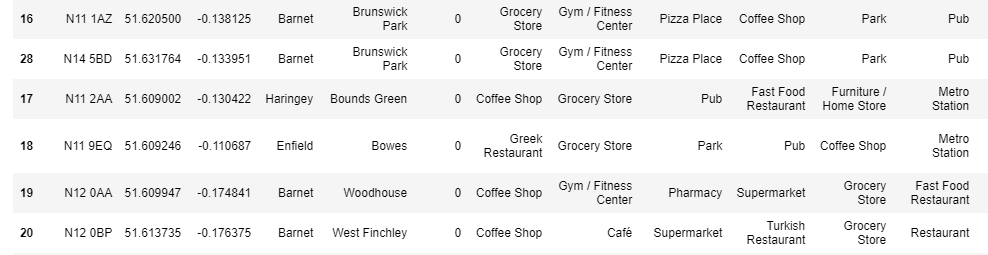


Once a cluster could be identified which encompassed postal codes where pubs were amongst the most popular venues, the median latitude and longitude of the cluster postal codes would be used to determine the optimal location for the distillery.

## Results

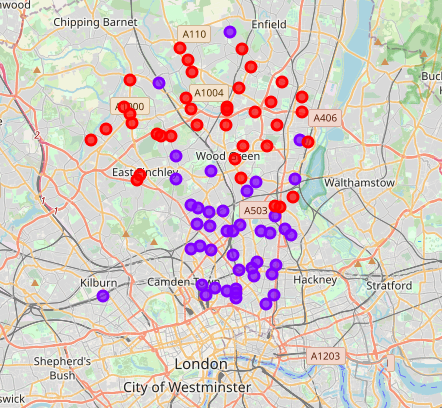
Below is a sample of the cluster 1 postal codes. In cluster 1 neighbourhoods, pubs are amongst the most popular categories of venue. Note that the clustering is not completely perfect (note the results for 52 and 54 below). It is necessary to examine the cluster 0 results for completeness.



For completeness, the output of a sample of cluster 0 postal codes is set out below. In cluster 0 neighbourhoods, pubs are not popular venues. Instead grocery stores, gyms, restaurants or coffee shops tend to dominate. This helps to confirm that cluster 1 is the appropriate cluster to use going forwards.

The folium plot below demonstrates the postal codes categorised by cluster. The purple dots represent postal codes which fall within cluster 1 – where pubs are the or one of the most popular category of venue. The red dots represent postal codes which fall within cluster 0 – where pubs are amongst the least popular category of venue.

Interestingly, cluster 1 postal codes are typically located closer to central London, where people will normally spend their time for recreation e.g. meeting friends.



In order to find the optimal location for the distillery, the median latitude and longitude of cluster 1 postal codes was determined. The result has been plotted below – with one map including a close up of the area. Therefore, the optimal location for the distillery, to ensure that it is in close proximity to areas where pubs are prevalent, is the Finsbury Park area of North London.

## 

## Discussion

Cluster 1 postal codes are typically located closer to central London, where people will normally spend their time for recreation e.g. meeting friends.

However, some cluster 1 postal codes are located far from central London. A commercial decision will therefore need to made as to whether these more distant postal codes should be served by the distillery. If not, then the analysis performed should be re-run and the clustering should exclude those postal codes as outliers.

Based on the analysis undertaken and the results of the clustering, the optimal location for the distillery is Finsbury Park, a neighbourhood in North London which is at the centre of the neighbourhoods where pubs are the, or one of the most popular categories of venue.

## Conclusion

The project has successfully identified, through the use of geolocational data and K-Means Clustering, the optimal location for a distillery in North London.

Based on the analysis, Finsbury Park is the optimal location for the distillery.

Locating the distillery in Finsbury Park will help to ensure that the distillery is located in close proximity to areas where pubs are the most or one of the most popular categories of venue – i.e. there are numerous buyers for the distillery’s products.

An additional analysis which would help to validate that Finsbury Park is the optimal location for the distillery, would be to use Foursquare’s geo-locational data in order to identify trending venues, for example, on a Friday or Saturday evening.

The analysis performed in this project could be re-run in order to classify postal codes based on trending venues instead of where pubs are amongst the most popular category of venue. This would help to confirm that the pubs in the cluster 1 neighbourhoods (where pubs are common) are in fact popular with visitors. There is a possibility that whilst this analysis has identified areas where pubs are common, without performing an additional analysis of trending venues, there is a risk that the distillery is located in the centre of postal codes which have a large number of pubs which aren’t popular with visitors and therefore are just surviving, instead of identifying areas where there are fewer but more popular pub venues, with a greater footfall and therefore greater demand for the distillery’s products.