# BIRMINGHAM NEIGHBOURHOODS

## Applied Data Science Capstone Report

### The Battle of Neighbourhoods

Opening a new Restaurant in Birmingham, England

Shoaib Ur Rehman Khan

December 15, 2019

# 1. Introduction:

## 1.1 Background:

Birmingham is a major city in England's West Midlands region, with multiple Industrial Revolution-era landmarks that speak to its 18th-century history as a manufacturing powerhouse. It's also home to a network of canals, many of which radiate from Sherborne Wharf and are now lined with trendy cafes and bars. Birmingham is blessed with some top notch restaurants and more keep arriving and it can be hard to know which one is best for you or where to book a table. In a city with several fine dining eateries as well as places serving cuisines covering over 30 countries all over the world, it's safe to say they've got a decent choice of places to eat.

## 1.2 Business Problem:

The objective of this capstone project is to analyze and select the best locations in the city of Birmingham, England to open new restaurant. Using data science methodology and machine learning clustering this project aims to provide solutions to answer the business problem which is: In the city of Birmingham, England, if someone is looking to open a restaurant, where would you recommend that they open it?

## 1.3 Target Audience:

The objective is to locate and recommend to the management which neighborhood of Birmingham city will be best choice to start a restaurant. The Management also expects to understand the rationale of the recommendations made. This would interest anyone who wants to start a new restaurant in Birmingham.

# 2. Data:

## 2.1 Data to Solve the Problem:

- List of all neighbourhod of Birmingham, England.
- Geographical coordinates of the neighborhoods which is, longitude and latitude and we will need this to plot the map and venue data.

- Venue data, we will obtain this from foursquare API and it will related to restaurants. We will use this data to perform clustering.

### 2.2 Data Sources:

Here is the Wikipedia page contains list of neighborhoods of Birmingham with postal codes and the approximate coverage of the postcode districts with the list of authority area of all the neighborhoods (https://en.wikipedia.org/wiki/B_postcode_area).

Here we will use web scraping technique to extract data from Wikipedia page, then we will get geographical coordinates from python geocoder package which will give and latitude and longitude coordinates of the neighborhood. After that we will analyze data with folium package to see the venues in map and then we use the Foursquare API to obtain the venue data to explore neighborhoods in Birmingham and the Foursquare API will help us to solve the business problem.

This project will make use of many Data Science techniques like,

- Web Scraping (Wikipedia)
- Working with Foursquare API
- Data Cleaning
- Data Wrangling
- K-means Clustering
- Plotting Map (Folium)

Which we will discuss on Methodology section.

## 3. Methodology:

### 3.1 Exploratory Data Analysis:

First we need to obtain the list of neighbourhoods of Birmingham, England which is available in given Wikipedia page in Data Source Section. We will then do web scraping using python requests and beautifulsoup package to extract the list of neighborhoods data. But, this is just the list of neighborhoods, we need to geographical coordinates and for this we will use geocodor package that will allow us to convert address into geographical coordinates to get the latitude and

longitude of each neighbourhood. After getting the complete data then we will analyze it into pandas dataframe and visualize the neighborhoods in map from folium package.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods. We will then extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Restaurant" data, we will filter the "Restaurant" as venue category for the neighbourhoods.

### 3.2 Machine Learning (k-means Clustering):

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Restaurant". The results will allow us to identify which neighbourhoods have higher concentration of restaurant while which neighbourhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurants.

## 4. Result:

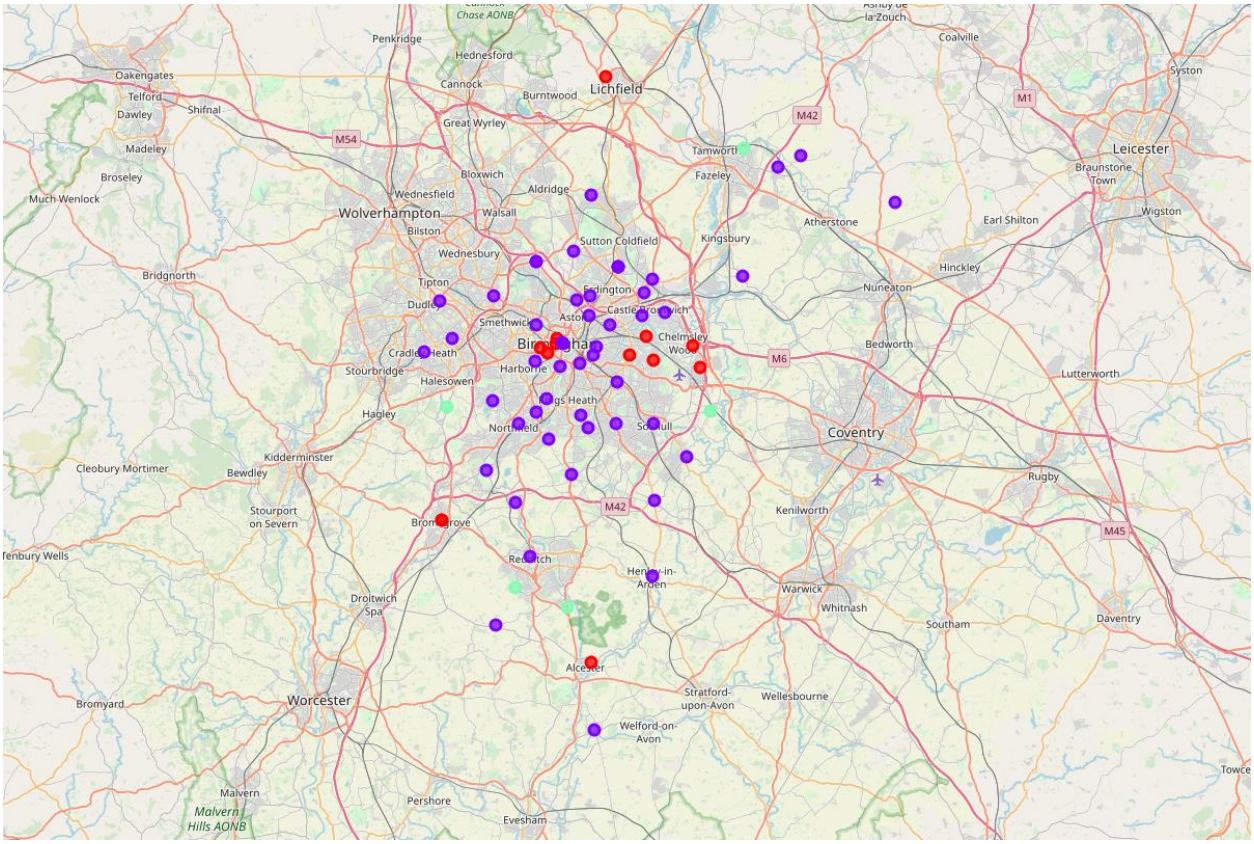
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Restaurants":

Cluster 0: Neighbourhoods with moderate number of restaurants

Cluster 1: Neighbourhoods with low number to no existence of restaurants

Cluster 2: Neighbourhoods with high concentration of restaurants

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.




# 5. Discussion:

As the observation noted from the map in result section, we can see most of the restaurant are found in the center of the Birmingham City, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to totally no restaurants in the neighborhoods. Therefore, this project recommends property developers to capitalize on these findings to open new restaurant in neighborhoods in cluster 1 with little to no competition and cluster 0 have a little competition where you can also open restaurant but must avoid cluster 2. We have shown you above the total restaurants by locations so anyone can see this dataset and can setup their restaurant as they want but if we see the suburb areas have very few restaurants so it may be the best place for them to open restaurant.

# 6. Recommendations:

In this project, we only consider one factor i.e. frequency of occurrence of restaurants, there are other factors such as population and income of residents that could influence the location decision of a new restaurants. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could revise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurants. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

# 7. Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. Interested person regarding the best locations to open a new restaurants. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.