**Capstone Project - The Battle of Neighborhoods**

**RESOLVING EMPLOYEE RELOCATION**

**Company:** WXYZ Company, Brooklyn, NY, USA.

**Presentation by:**

Samuel Nnitiwe Theophilus

**GITHUB REPO:**

<https://github.com/nnitiwe-dev/Coursera_Capstone/>

**LinkedIn Profile:** <https://www.linkedin.com/in/samuelnnitiwetheophilus/>

Table of Contents

[1. Introduction 3](#_Toc67007872)

[1.1 Background 3](#_Toc67007873)

[1.2 Problem Definition 3](#_Toc67007874)

[2. Data 4](#_Toc67007875)

[2.1 Data Sources 4](#_Toc67007876)

[2.2 Data Transformation 5](#_Toc67007877)

[2.3 Data Cleaning 6](#_Toc67007878)

[3. Methodology 7](#_Toc67007879)

[3.1 Exploratory Data Analysis 7](#_Toc67007880)

[4. Results 8](#_Toc67007881)

[4.1 Clustering 8](#_Toc67007882)

[4.2 Visualizing Results 9](#_Toc67007883)

[5. Discussion 10](#_Toc67007884)

[6. Conclusion 10](#_Toc67007885)

# **Introduction**

## 1.1 Background

A Company's best resource and its largest costs are the people who bring creativity, productivity and ultimately profitability to a company- It's Employees. A good talent management program can improve an employer's competitiveness, but it does not ensure that the talent is located where it is most needed.

In a situation where an Organization wishes to expand its branch to a new location, there will be a need to deploy staff. While it is possible for the Company to hire new talents, using current capable employees who are already familiar with the company structure and operations is the best decision for the organization in terms of Cost & overhead time required to staff to adapt and handle its operations.

The Company will need to find a way to relocate some keys employees for continued career development or to bring their knowledge to different subsidiaries or locations. These moves can be a daunting task for the company and a high-stress situation for the employee. If a relocation is not handled successfully, it threatens the employer's ability to retain the employee—and it risks losing someone the employer has devoted time and money to develop and move.

## 1.2 Problem Definition

The **WXYZ Company** has been operating in ***Brooklyn, New York, USA*** for the past 5 years. This year, the board made the decision to open an office in ***Coventry, England*** and would like to select some of its existing employees to fill some managerial roles at its new branch.

This Data Science project aims to compare the neighborhoods in Brooklyn, New York (Current company location) with the neighborhoods in Coventry (New Branch) and **create Clusters of similar neighborhoods**.

This will help the company to:

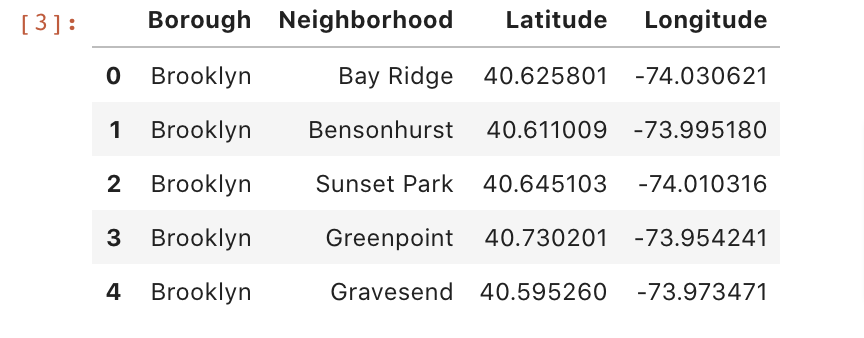
1. Identify Employees who would have a smoother transition to the new branch (by identifying if their current residential address matches a cluster in the new location).
2. Identify Locations to consider as recommendation for employees who agree to relocate.

# **Data**

## 2.1 Data Sources

**Brooklyn Neighborhood Data** was retrieved from a JSON File accessible [here](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json). This dataset contains a lot of information on New York; however, this project only extracted the following:

* Borough
* Neighborhood
* Latitude
* Longitude



**Dataset size**: (70, 4)

**Coventry Neighborhood Data** was scrapped from the [Wikipedia page](https://en.wikipedia.org/wiki/CV_postcode_area) which contains a list of PostalCodes and respective Neighborhood information. This data did not have the latitude and longitude information, so the geocoder python library was used to to extract the coordinates.

* Postal Code
* Borough
* Neighborhood
* Latitude
* Longitude

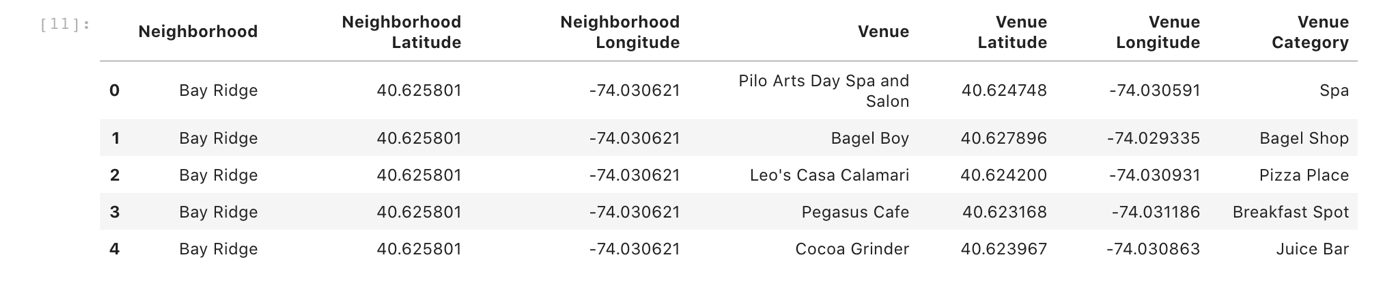


**Dataset size**: (86, 5)-Before drop; (86,4)- After dropping “PostalCode”

The dataset presented above is not enough to help us generate good clusters.

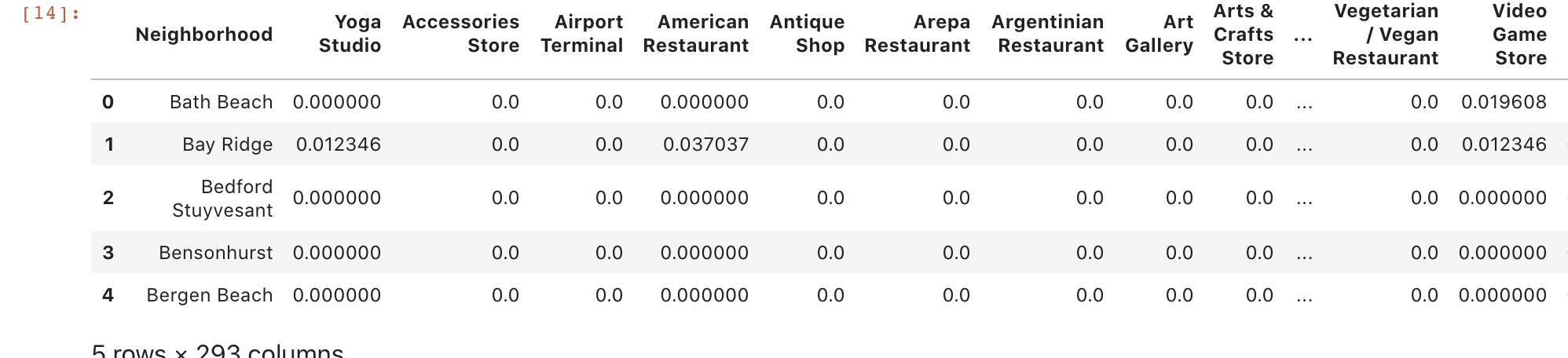
It is logical to decide that the places (Gym, Restaurant, etc.) people visit influences the type of environments they will be comfortable in.

This is why I made the decision to use **popular venues** (places) in this project. This information was extracted using the **Foursquare API** to neighborhoods in Brooklyn, New York and Coventry. I explored the most **common venue categories** in each **neighborhood**, and then used this feature to generate the data structure as shown below:



## 2.2 Data Transformation

After preparing the venues data retrieved from the Foursquare API, the data which was a combination of numeric and categorical values needed to be transformed so that the clustering algorithm can take it as input. The rows (records) of each neighborhood was grouped by taking the mean of the frequency of occurrence of each venue category (e.g. Gym, Spa, Bar) to generate data as shown below:

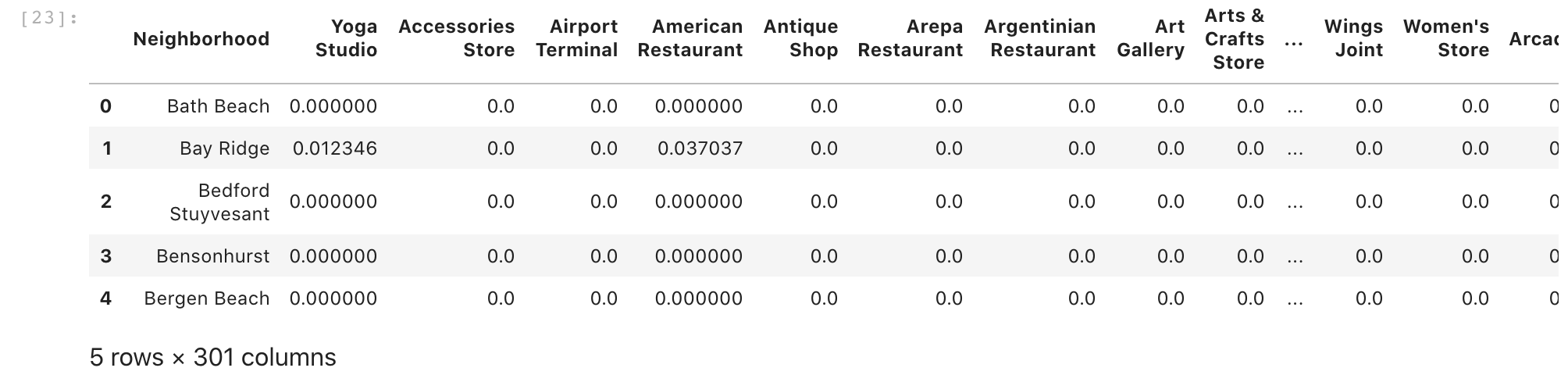


Next I used the mean of frequency values identify the top 10 Venues for every neighborhood.



From the table it can be seen that each neighborhood has 10 top venues.

Finally, the two tables were merged into one new table and One-hot encoding was applied on the transformed data to convert the data to numeric form.



## 2.3 Data Cleaning

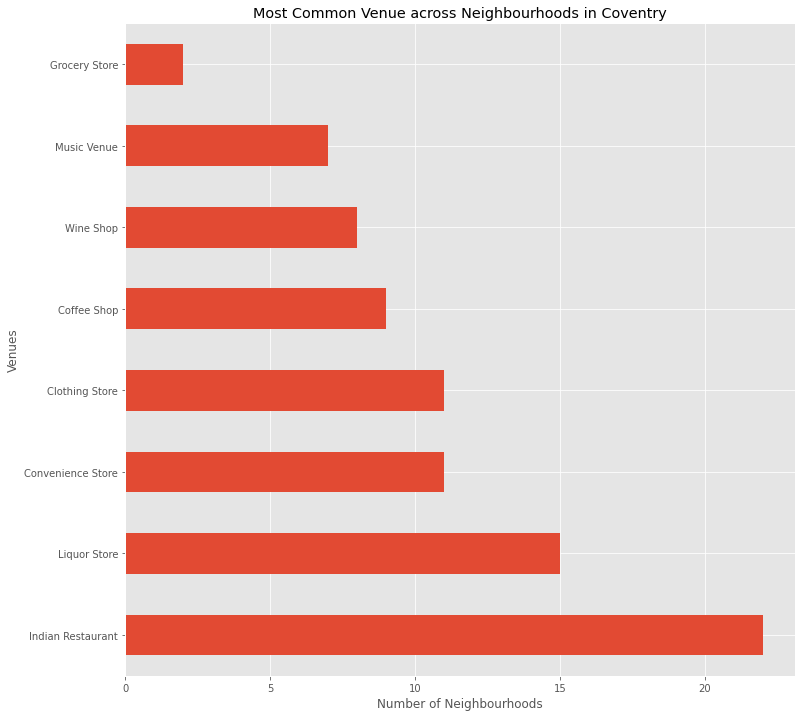
After Data transformation was completed, I handled missing data. Some of this resulted from merging the two groups (as not all venues types are present in both **Brooklyn & Coventry,** there is bound to be missing data). Since I already performed one-hot encoding and the category of data with missing values were venue related data, I filled missing data with 0.

Data scraped from the Wikipedia page were also handled by dropping rows with NaN values. were combined into one table.

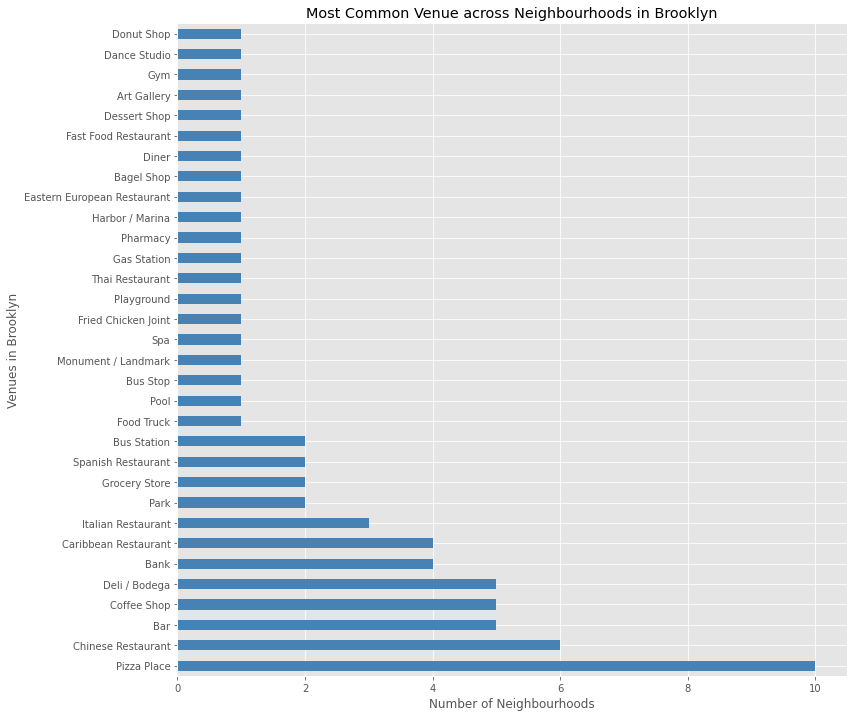
# **Methodology**

## 3.1 Exploratory Data Analysis

Since the commons venues were going to be the key features in the clustering model training, I decided to visualize the top venues in both Brooklyn and Coventry.



From this horizontal bar chart it is clear that Indian Restaurant and Liquor Store are the most common venues, which implies that Indians are probably quite comfortable in Coventry and might not have an issue relocating to Coventry from Brooklyn.



This graph in Brooklyn shows that Pizza Place and Chinese Restaurant are the most common venues among neighborhoods. This could mean that relocating the Chinese or Italian employee might not be the best idea.

# **Results**

## 4.1 Clustering

Clustering is a Machine Learning technique that involves the grouping of data points. Data points that are in the same group should have similar properties, while data points in different groups should have highly dissimilar properties. As discussed at the introduction, the purpose of this project was **to find neighborhoods in Coventry that have similarities with neighborhoods in Brooklyn** so that employee will have smooth transition when they relocate to Coventry at the Company’s new branch office.

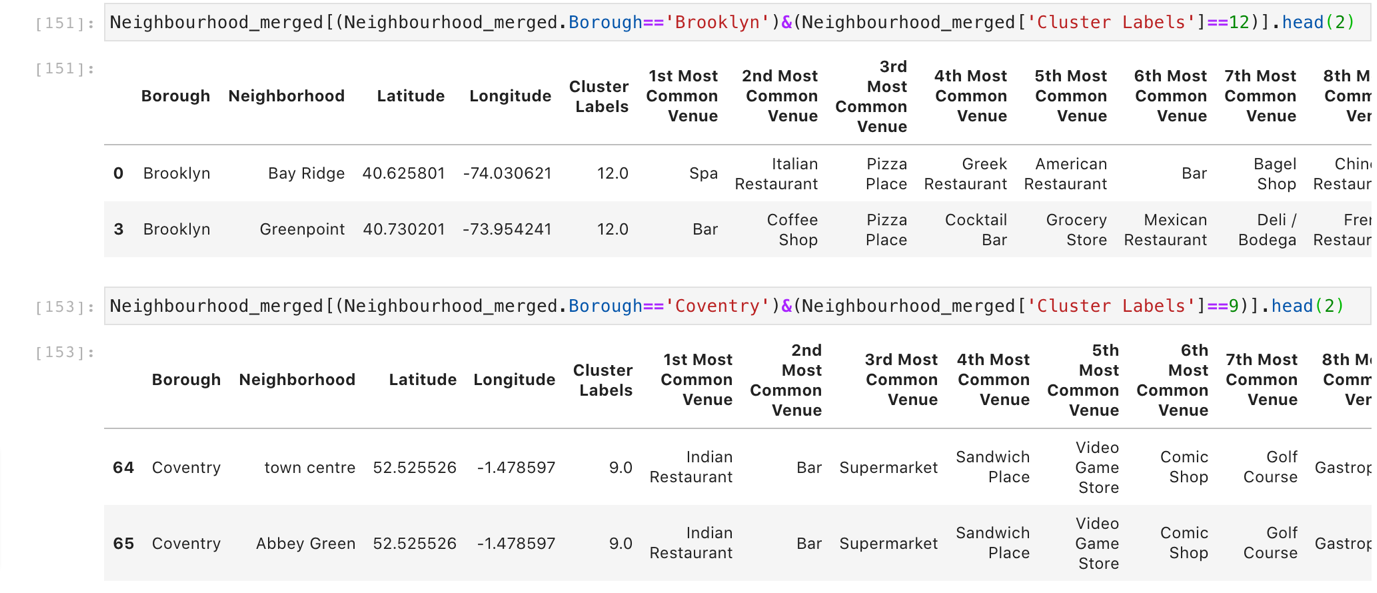
k-means clustering was the clustering algorithm used in this project. It is a method of vector quantization, that aims to partition observations into **k clusters** in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The Elbow method was used to pick an appropriate value of K and 15 was selected.

The size of the dataset used for clustering was 156 by 300.

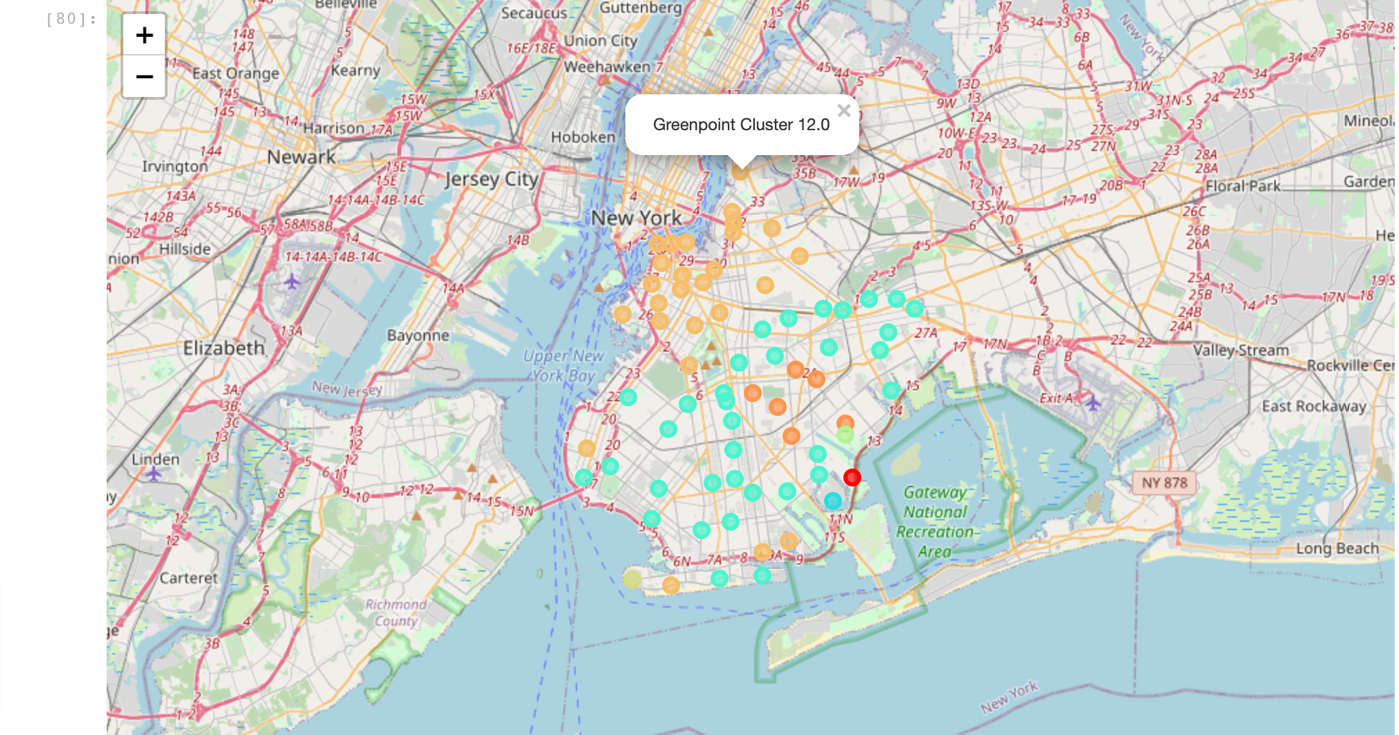
## 4.2 Visualizing Results

Look at the query results of **2 Brooklyn neighborhoods** and **2 Coventry neighborhoods** clustered using K-means. By merely comparing the values in the columns you will be able to understand why these neighborhoods fall in the same clusters.

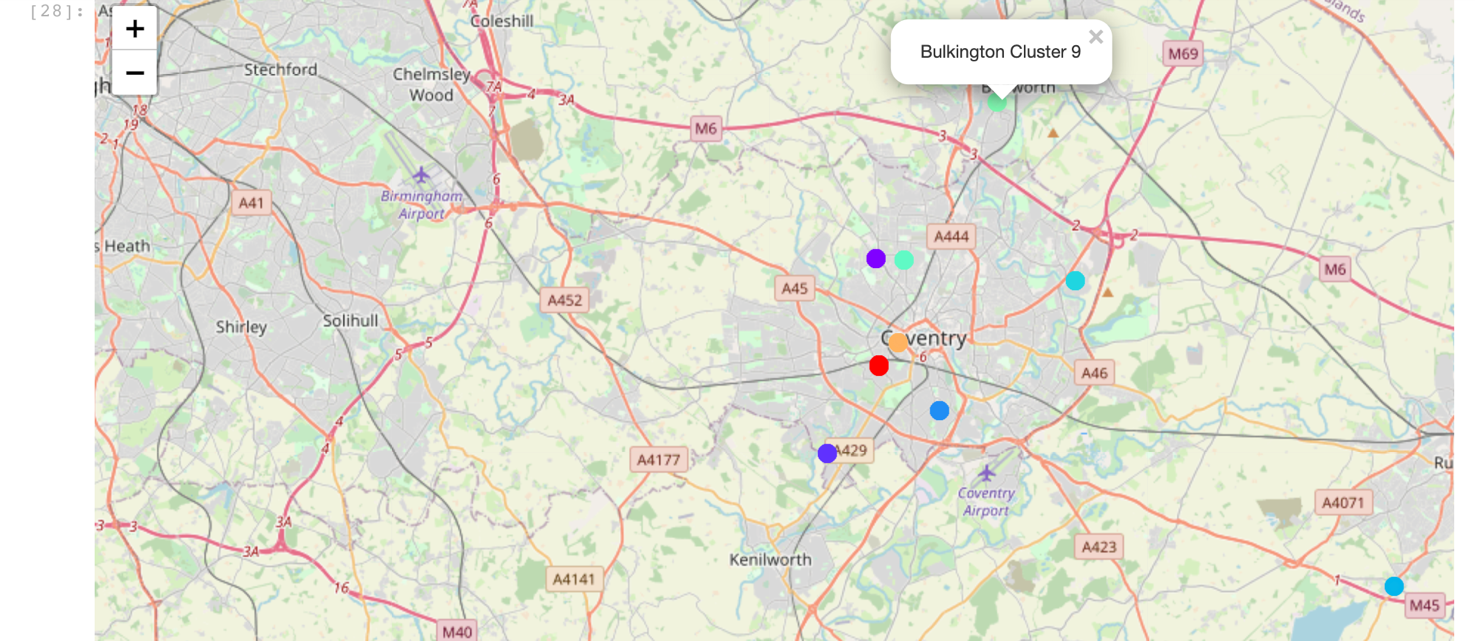


Now lets look at the Clusters as visualized on a Geographical map to have a better idea.

**Clusters on Brooklyn, New York Map**



**Clusters on Coventry, England Map**



# **Discussion**

Since the company is expecting its employees to move to the new area, it’s very beneficial to allow the qualified candidates time to acclimatize so that they can look into the recommendations as modeled in this project and decide the neighborhood they prefer. After researching similar neighborhoods, it is now possible to draft a list of recommendations from this cluster model.

The color codes in the two graphs show that **Cluster 9 -Green (Followed closely by Cluster 12-Orange)** is very common neighborhood type in Brooklyn and appears a little on the Coventry Map. This indicates that employees relocating to Coventry will have the smoothest transition on these clusters.

The organization should put together a portfolio of the new areas may to help employees learn more about what they’re going to be facing. Including information on the most common venues as this will help the employees make an informed decision.

# **Conclusion**

In this study, I extracted Brooklyn and Coventry Neighborhood data along with their most popular places using the Foursquare API. I transformed the data and clustered the data to find similar neighborhoods between Brooklyn and Coventry. All this was down to help the Company to identify residential neighborhoods for its staff that will have the smoothest transitional effect. International relocation is an attractive facet of a career at any company, and is good for the company as a whole and this project helps the company to ease the relocation process.