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

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.

2. Data

2.1 Data Sources

Brooklyn Neighborhood Data was retrieved from a JSON File accessible here. This dataset contains a lot of information on New York; however, this project only extracted the following:

- Borough
- Neighborhood
- Latitude
- Longitude

[3]:

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

Dataset size: (70, 4)

Coventry Neighborhood Data was scrapped from the Wikipedia page 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

[5]:

	PostalCode	Borough	Neighborhood	Longitude	Latitude
0	CV1	Coventry	Coventry City Centre	-1.522541	52.406551
1	CV1	Coventry	Gosford Green	-1.522541	52.406551
2	CV1	Coventry	Hillfields	-1.522541	52.406551
3	CV1	Coventry	Spon End	-1.522541	52.406551
4	CV1	Coventry	Coventry University	-1.522541	52.406551

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:

[11]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar

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:

[14]:

	Neighborhood	Yoga Studio	Accessories Store	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	...	Vegetarian / Vegan Restaurant	Video Game Store
0	Bath Beach	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.019608
1	Bay Ridge	0.012346	0.0	0.0	0.037037	0.0	0.0	0.0	0.0	0.0	...	0.0	0.012346
2	Bedford Stuyvesant	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000
3	Bensonhurst	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000
4	Bergen Beach	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000

5 rows × 293 columns

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

[17]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	Pizza Place	Pharmacy	Chinese Restaurant	Donut Shop	Italian Restaurant	Bubble Tea Shop	Peruvian Restaurant	Gas Station	Fast Food Restaurant	Sports Bar
1	Bay Ridge	Spa	Italian Restaurant	Pizza Place	Greek Restaurant	American Restaurant	Bar	Bagel Shop	Chinese Restaurant	Café	Playground
2	Bedford Stuyvesant	Coffee Shop	Café	Pizza Place	Deli / Bodega	Bar	Gourmet Shop	Thrift / Vintage Store	Juice Bar	Community Center	Basketball Court
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Flower Shop	Donut Shop	Ice Cream Shop	Sushi Restaurant	Shabu-Shabu Restaurant	Supermarket	Spa	Smoke Shop
4	Bergen Beach	Harbor / Marina	Baseball Field	Athletics & Sports	Park	Hockey Field	Playground	Factory	Falafel Restaurant	Farm	Farmers Market

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.

[23]:

	Neighborhood	Yoga Studio	Accessories Store	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	...	Wings Joint	Women's Store	Arcad
0	Bath Beach	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0
1	Bay Ridge	0.012346	0.0	0.0	0.037037	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0
2	Bedford Stuyvesant	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0
3	Bensonhurst	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0
4	Bergen Beach	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0

5 rows × 301 columns

2.3 Data Cleaning

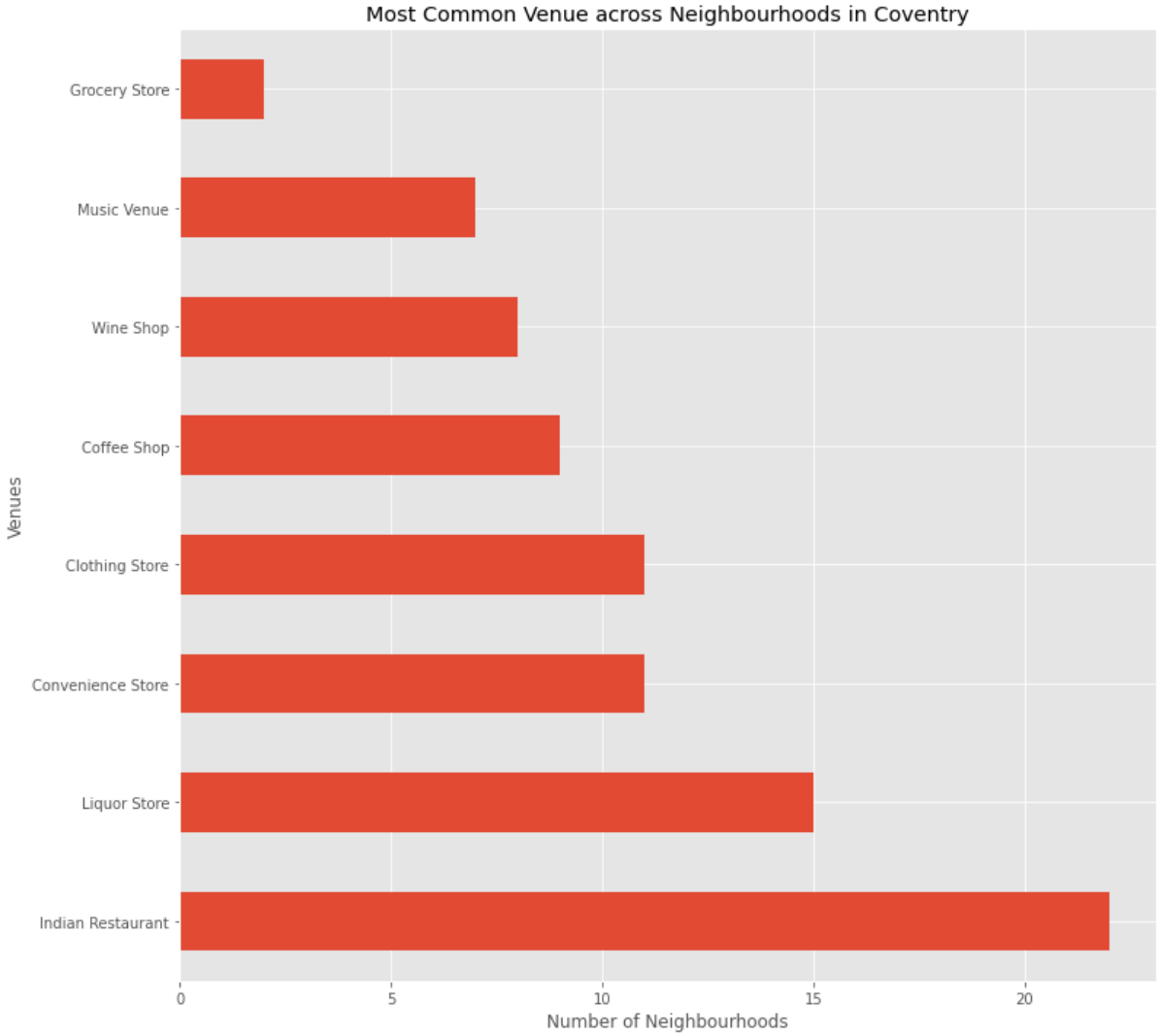
After Data transformation was completed, I handled missing data. Some of this resulted from merging the two groups (as not all venues types are present in both **Brooklyn & Coventry,** there is bound to be missing data). Since I already performed one-hot encoding and the category of data with missing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

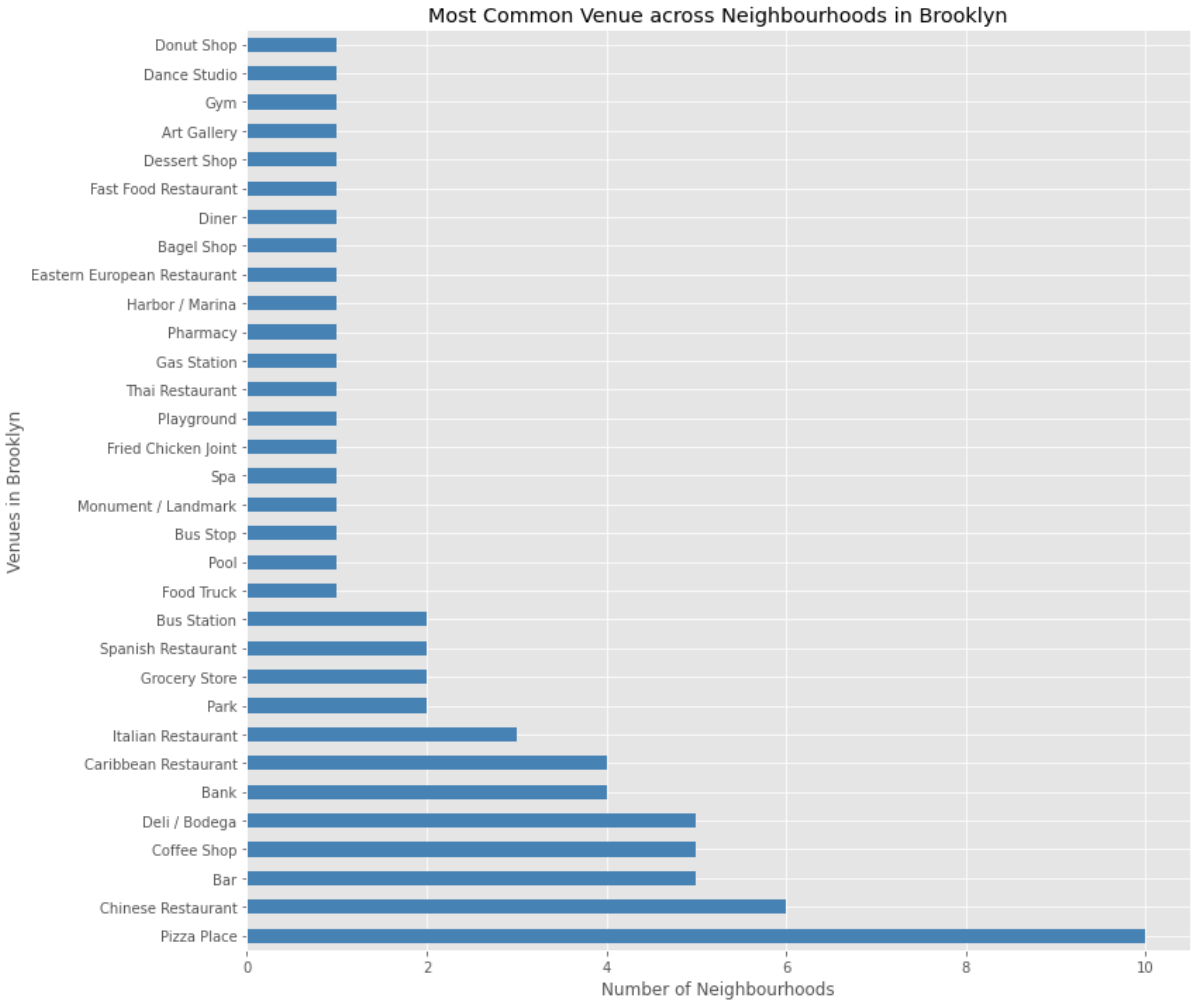
3. Methodology

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.

4. 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.

```
[151]: Neighbourhood_merged[(Neighbourhood_merged.Borough=='Brooklyn')&(Neighbourhood_merged['Cluster Labels']==12)].head(2)
```

[151]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th M Comm Ver
0	Brooklyn	Bay Ridge	40.625801	-74.030621	12.0	Spa	Italian Restaurant	Pizza Place	Greek Restaurant	American Restaurant	Bar	Bagel Shop	Chin Restaur
3	Brooklyn	Greenpoint	40.730201	-73.954241	12.0	Bar	Coffee Shop	Pizza Place	Cocktail Bar	Grocery Store	Mexican Restaurant	Deli / Bodega	Fre Restaur

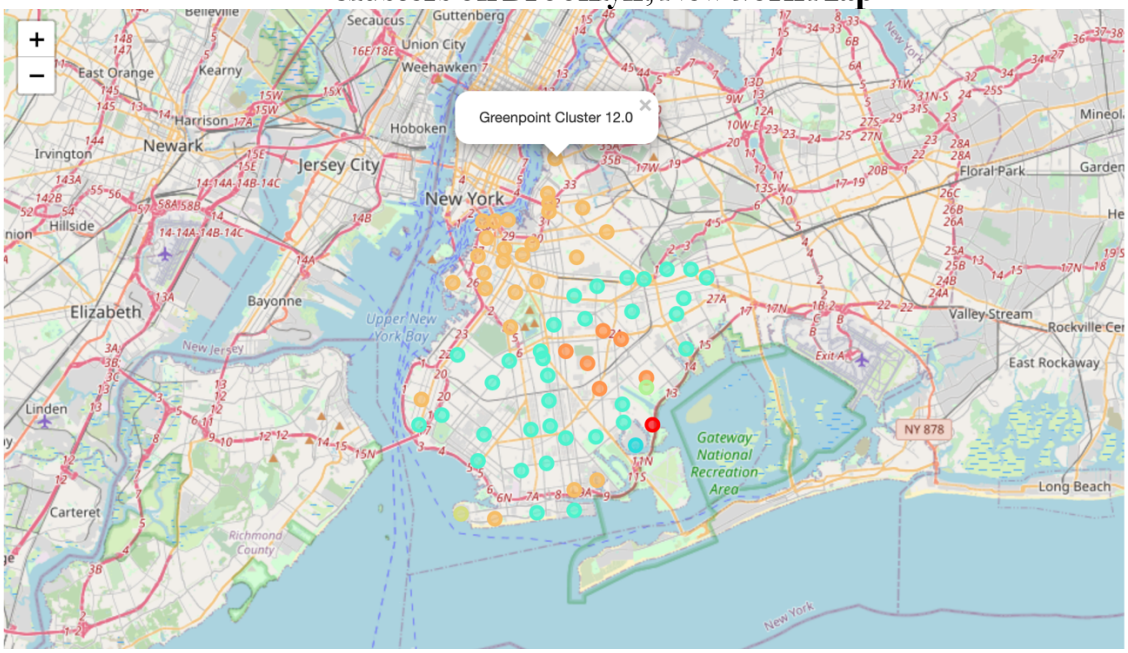
```
[153]: Neighbourhood_merged[(Neighbourhood_merged.Borough=='Coventry')&(Neighbourhood_merged['Cluster Labels']==9)].head(2)
```

[153]:

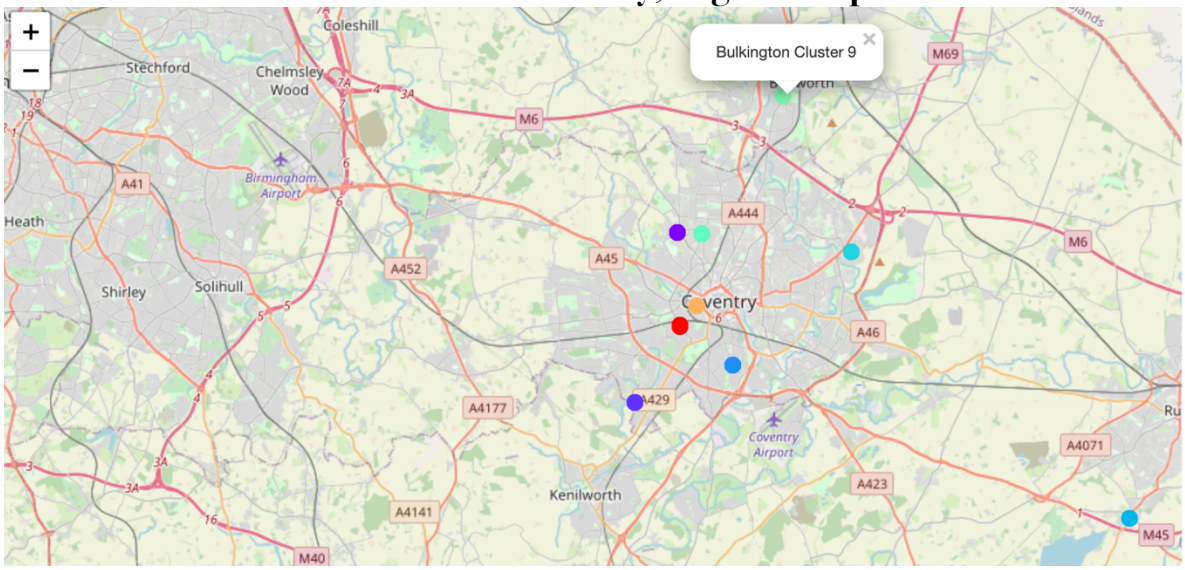
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th M Comm Ver
64	Coventry	town centre	52.525526	-1.478597	9.0	Indian Restaurant	Bar	Supermarket	Sandwich Place	Video Game Store	Comic Shop	Golf Course	Gastrop
65	Coventry	Abbey Green	52.525526	-1.478597	9.0	Indian Restaurant	Bar	Supermarket	Sandwich Place	Video Game Store	Comic Shop	Golf Course	Gastrop

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



5. 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.

6. 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.