A Tale of Two Cities: Battle of the Neighborhoods Capstone Project Report

Oxford, UK versus Atlanta, USA (Round 1)

Author: Theresa K Foster

Date: 15 May 2020

The link to my Jupyter Notebook with the coding for this data analysis is available below:

A Tale of Two Cities: Battle of the Neighborhoods Capstone Project Jupyter Notebook

1. Introduction

Oxford is a city in central southern England with a population size of around 155,000 people. The city is known for its University, which was established in the 12th century, but is also a hub for manufacturing, publishing and science based industries and research, as well as education and tourism. Atlanta is the capitol of the US state of Georgia and is the most populous city in the state with an estimated 498,044 residents. Atlanta is a culturally and economically diverse city with dominant economic sectors including aerospace, transportation, professional and business services, media and medical operations, and information technology.

The aim of this project is to explore the neighborhoods in both cities and group them by common nearby venues. This will assist anyone visiting or relocating between the cities to consider which areas are most similar to their current neighborhood and therefore might offer their preferred range of amenities. This information is very useful when moving to an unknown city and will help narrow down the list of areas to search for a new home, thus speeding up the relocation process and avoiding overly long and potentially pricey stays in hotels or other temporary living arrangements. Alternatively for those visiting between the cities, this information could be useful in deciding the best location for a vacation rental or hotel booking, based on the interests and priorities of the traveler(s).

2. Data

The following data sources were used to complete this project:

1. Oxford postcode data from Doogal.co.uk updated 2020
2. Atlanta zip code and neighborhood data from local real estate company (The Keen Team) 2020
3. Cross referenced Atlanta zip code and neighborhood data from US Map Guide 2020
4. US Longitude and Latitude data by zip code from Open Data Soft.com 2020
5. Foursquare API

2.1. Oxford, UK Neighborhood Data Sourcing and Cleaning

The data set (1) for Oxford was the most complete and included postal code data, ward (neighborhood) names, and the corresponding latitude and longitude coordinates for all OX postcodes, which covers the entire county of Oxfordshire. The data was in the form of a downloadable excel spreadsheet, which I then cleaned and formatted to include only Oxford city postcodes, ward (neighborhood) names and map coordinates. Finally, I reduced the list of wards by removing duplicate values so that there would only be one occurrence of each neighborhood and corresponding data. It should be noted that this method randomly dropped duplicates so the remaining full postal codes corresponding to each neighborhood were one of many possible options. Different post code choices would have had slightly differing latitude and longitude coordinates. This may have affected the resulting venue data sourced from Foursquare and skewed the results. I then uploaded this data set to my Jupyter notebook and used the <insert to code> function to transform it into a pandas data frame.

Out[3]:

	Postal Code	Neighborhood	Latitude	Longitude
0	OX3 0FR	Barton and Sandhills	51.770563	-1.208585
1	OX4 6AA	Blackbird Leys	51.722543	-1.204506
2	OX1 1AA	Carfax	51.750231	-1.267367
3	OX3 0BD	Churchill	51.753788	-1.230251
4	OX4 2BH	Cowley	51.730350	-1.206266

2.2 Atlanta, USA Neighborhood Data Sourcing and Cleaning

The data sets (2)(3) used to source a list of Atlanta neighborhoods and corresponding postal codes (zip codes) were simply lists from an Atlanta real estate website and a US map guide website respectively. I manually copied and input this data into an excel spreadsheet and added any differences between the data (missing or additional neighborhoods or zip codes) to ensure a more complete breakdown. Unfortunately data available from local city government sources was not in the required format so I could not use more authoritative sources. Therefore the breakdown of neighborhoods to zip codes in this data set should be taken as advisory only and may differ between data sets. Initially I was going to use Geopy Nominatum to find the map coordinates for each zip code. However the results were wildly inaccurate. As an alternative I found and downloaded a spreadsheet of all US zip codes and corresponding latitude and longitude coordinates (4) from the Open Soft Data website. I manually filtered this excel spreadsheet to list only Atlanta zip codes and map coordinates. I uploaded both excel sheets to my Jupyter notebook using the <insert to code> function to transform them into Pandas data frames, dropping any unnecessary columns. Finally, I combined the separate Atlanta data sets using a Pandas join function on the common column value of zip codes.

Out[8]:

	Postal Codes	Neighborhood	Latitude	Longitude
0	30303	Central Business District, Fairlee Poplar	33.752856	-84.39013
1	30305	Buckhead, Garden Hills, Haynes Manor, Peachtre...	33.830054	-84.38472
2	30306	Virginia Highlands, Morningside/Lenox Park, Po...	33.786755	-84.35149
3	30307	Candler Park ,Edgewood, Inman Park, Lake Clair...	33.768205	-84.33786
4	30308	Midtown, Old Fourth Ward	33.771755	-84.38065

The resulting dataset for Oxford had 24 neighborhoods and the dataset for Atlanta had 28 neighborhoods.

```
In [9]: merged_ATL.shape
Out[9]: (28, 4)

In [10]: df_data_OXF.shape
Out[10]: (24, 4)
```

2.3 Final List of Neighborhoods Used for this Project

Oxford (24)	Atlanta (28)
Barton and Sandhills	Central Business District, Fairlee Poplar
Blackbird Leys	Buckhead, Garden Hills, Haynes Manor, Peachtree Battle, Peachtree Hills, Tuxedo Park
Carfax	Virginia Highlands, Morningside/Lenox Park, Poncey-Highland, Druid Hills
Churchill	Candler Park ,Edgewood, Inman Park, Lake Claire, Little Five Points
Cowley	Midtown, Old Fourth Ward
Cowley Marsh	Midtown, Ansley Park, Brookwood Hills, Loring Heights
Headington	Adair Park, Capitol View, Oakland City, West End
Headington Hill and Northway	Cascade
Hinksey Park	Grant Park
Holywell	Castlebury Hill
Iffley Fields	Vines City, Mozely Park
Jericho and Osney	Peoplestown - Lakewood
Littlemore	Cabbagetown, East Atlanta Village, Ormewood Park
Lye Valley	Kirkwood, Edgewood, East Lake
Marston	Home Park, Collier Hills, Underwood Hills
North	Brookhaven, Dunwoody
Northfield Brook	Morningside/Lenox Park, Piedmont Heights, Lenox
Quarry and Risinghurst	Lenox
Rose Hill and Iffley	Buckhead
St Clement's	Emory, Toco Hills, Briarcliff
St Margaret's	College Park
St Mary's	Vinings
Summertown	Northlake, Tucker
Wolvercote	Chamblee
	North Buckhead, Chastain Park
	East Point
	Briarcliff Woods, Oak Grove, Northlake
	Hapeville

3. Methodology

Before sourcing the venue data, I completed initial visual analysis of the neighborhood data for both cities to view the layout of the neighborhoods on a map. This was to ensure the coordinates were initially generally correct and to see the spread of the neighborhoods across each city, as they vary significantly in geographical size.

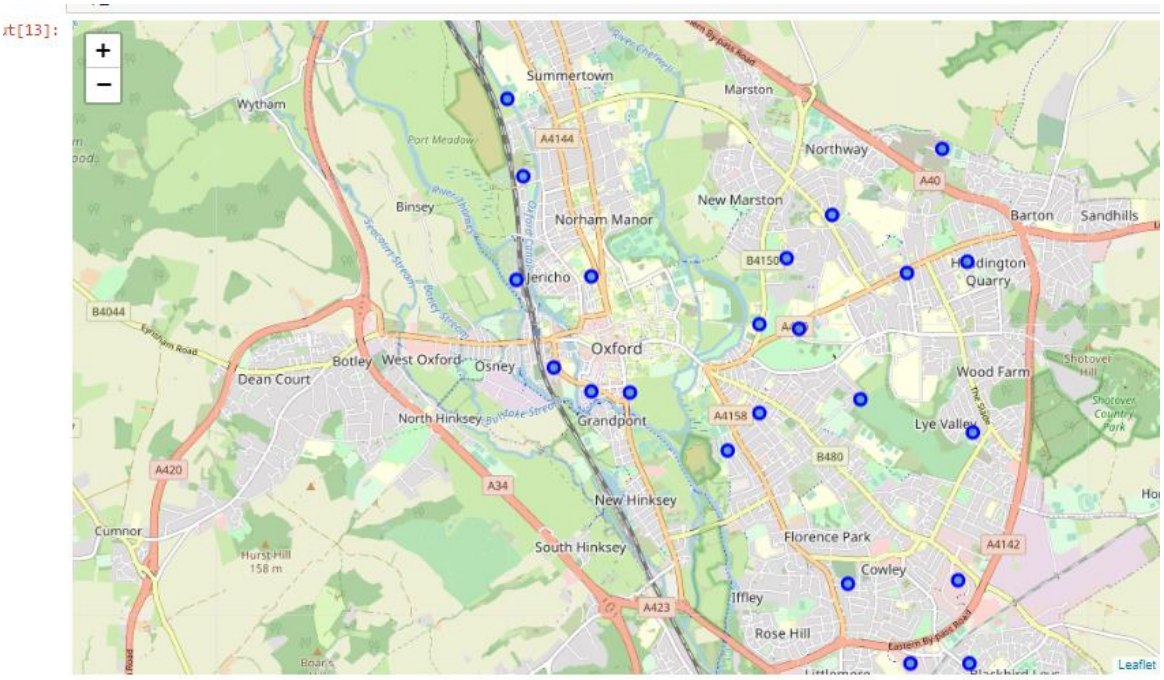
Using the Nominatum tool in Geopy, I calculated the latitude and longitude coordinates of both cities.

```
In [12]: address = 'Oxford, UK'

geolocator = Nominatim(user_agent="Oxf_explorer")
location = geolocator.geocode(address)
latitude_OXF = location.latitude
longitude_OXF = location.longitude
print('The geograpical coordinates of Oxford, UK are {}, {}.'.format(latitude_OXF, longitude_OXF))

The geograpical coordinates of Oxford, UK are 51.7520131, -1.2578499.
```

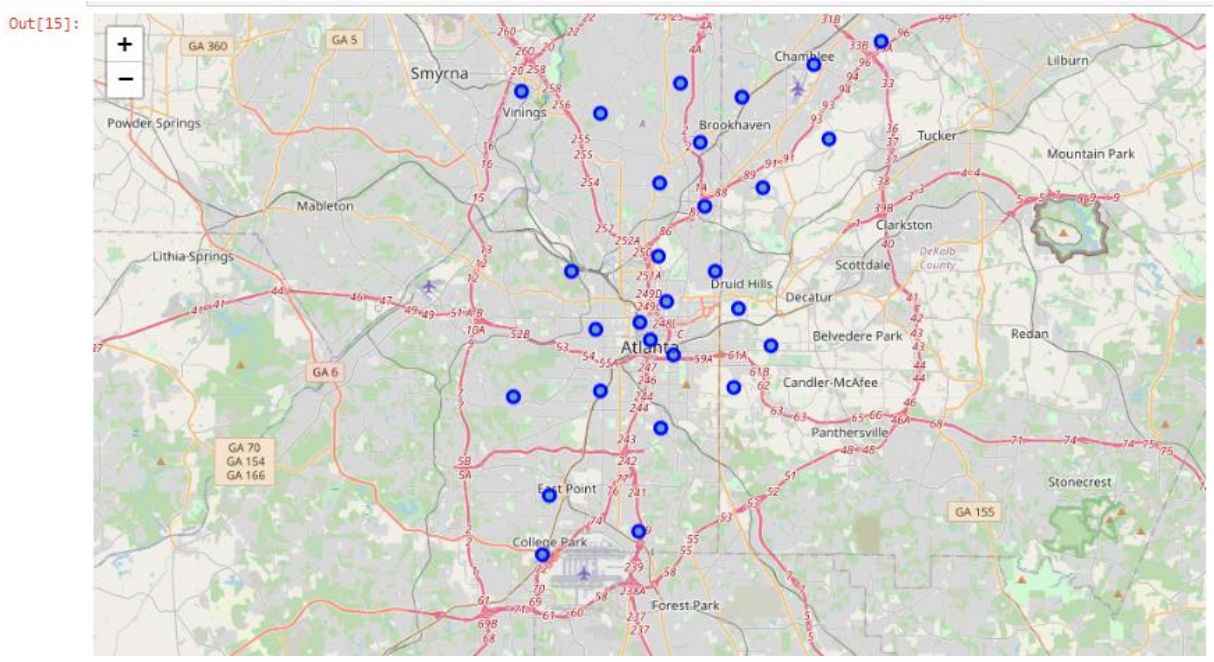
```
In [14]: address = 'Atlanta, GA'

geolocator = Nominatim(user_agent="Atl_explorer")
location = geolocator.geocode(address)
latitude_ATL = location.latitude
longitude_ATL = location.longitude
print('The geograpical coordinates of Atlanta, GA are {}, {}.'.format(latitude_ATL, longitude_ATL))

The geograpical coordinates of Atlanta, GA are 33.7490987, -84.3901849.
```

I then used Folium to create maps of the two cities using the above generated coordinates. Finally, I was able to code markers onto each city map of the corresponding neighborhood coordinates using the data from the previously created data frames.

3.1 Oxford, UK Neighborhoods Map



3.2 Atlanta, USA Neighborhoods Map



3.3 Foursquare API: Venue Data

Using Foursquare, I was able to generate a list of venues by category in each neighborhood based on the corresponding map coordinates in the data sets for both cities. I set the radius to 500 and limited the venue results to 100 per neighborhood or set of coordinates. I then transformed this venue data into Pandas data frames (see below example of Oxford neighborhood venue data generated using Foursquare API.) The process was repeated for Atlanta neighborhoods.

9]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barton and Sandhills	51.770563	-1.208585	bus stop barton village road	51.768160	-1.203738	Bus Stop
1	Barton and Sandhills	51.770563	-1.208585	Booty	51.766914	-1.210675	Boutique
2	Blackbird Leys	51.722543	-1.204506	Soundworks	51.722242	-1.201936	Music Venue
3	Blackbird Leys	51.722543	-1.204506	The Blackbird	51.721002	-1.202126	Pub
4	Blackbird Leys	51.722543	-1.204506	Delteys Supermarket	51.720504	-1.202680	Convenience Store

8]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Central Business District, Fairlee Poplar	33.752856	-84.39013	Walgreens	33.754345	-84.389484	Pharmacy
1	Central Business District, Fairlee Poplar	33.752856	-84.39013	Ebrik Coffee Room	33.753897	-84.388782	Coffee Shop
2	Central Business District, Fairlee Poplar	33.752856	-84.39013	The Masquerade	33.751720	-84.389739	Music Venue
3	Central Business District, Fairlee Poplar	33.752856	-84.39013	Weekday Cafe	33.754278	-84.388892	Burger Joint
4	Central Business District, Fairlee Poplar	33.752856	-84.39013	Blossom Tree	33.755496	-84.389006	Korean Restaurant

Finally, I created a new data frame for each city listing the top 10 most common venues in each neighborhood based on frequency.

3.4 Oxford Top 10 Venues by Neighborhood

5]:

	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	Barton and Sandhills	Bus Stop	Boutique	Wine Bar	Dessert Shop	Discount Store	Donut Shop	Electronics Store	Farmers Market	Fast Food Restaurant	Field
1	Blackbird Leys	Bus Stop	Pub	Recreation Center	Music Venue	Convenience Store	Wine Bar	Department Store	Dessert Shop	Discount Store	Donut Shop
2	Carfax	Pub	Bar	Nightclub	Hotel	Indian Restaurant	Pizza Place	Sandwich Place	Portuguese Restaurant	Bakery	Hostel
3	Churchill	Coffee Shop	Gym / Fitness Center	Park	Café	Deli / Bodega	Wine Bar	Fish & Chips Shop	Dessert Shop	Discount Store	Donut Shop
4	Cowley	Hotel	Clothing Store	Fast Food Restaurant	Electronics Store	Sporting Goods Shop	Market	Pub	Supermarket	Grocery Store	Gym / Fitness Center

3.5 Atlanta Top 10 Venues by Neighborhood

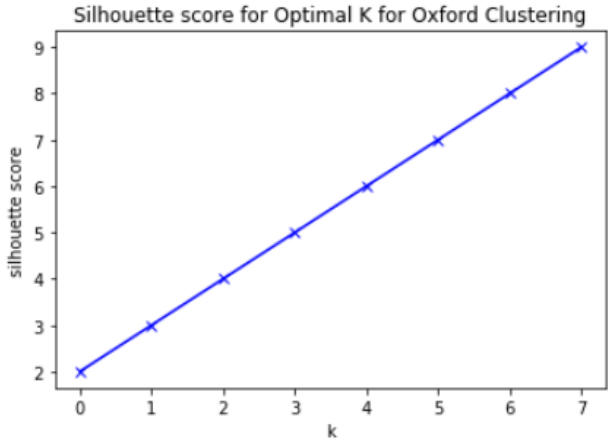
4]:

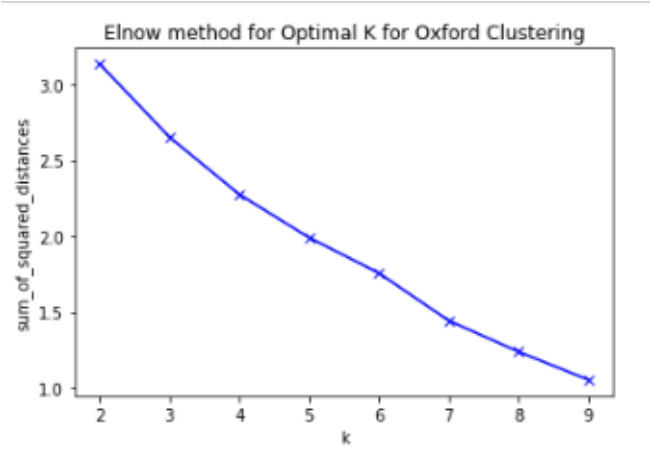
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adair Park, Capitol View, Oakland City, West End	Brewery	Trail	Thrift / Vintage Store	Liquor Store	Beer Store	Market	Food	Paper / Office Supplies Store	Boutique	Juice Bar
1	Briarcliff Woods, Oak Grove, Northlake	Lake	Zoo Exhibit	Fast Food Restaurant	Fried Chicken Joint	Fountain	Football Stadium	Food Service	Food Court	Food	Flower Shop
2	Brookhaven, Dunwoody	American Restaurant	Gym / Fitness Center	Ice Cream Shop	Sandwich Place	Frozen Yogurt Shop	Mexican Restaurant	Irish Pub	Pizza Place	Soup Place	Shopping Mall
3	Buckhead	Gym / Fitness Center	Fast Food Restaurant	Fried Chicken Joint	Fountain	Football Stadium	Food Service	Food Court	Food	Flower Shop	Fish Market
4	Buckhead, Garden Hills, Haynes Manor, Peachtre...	Italian Restaurant	Café	Basketball Court	Shipping Store	Mobile Phone Shop	Sandwich Place	Salon / Barbershop	Farmers Market	Sushi Restaurant	Cosmetics Shop

3.6 Finding the best K for K-Means Clustering

K-Means is one of the most common methods of unsupervised machine learning for clustering. Using one hot encoding and mean frequency on the new data frames, I was able to then apply algorithms from the SciKit Learn library to calculate the best K value for K-means clustering of the neighborhoods in each city. I initially used the Silhouette method but the results were inconclusive. I therefore tried the Elbow method (sum of squared distances) and achieved slightly better results in both cases. I used Matplotlib to plot the results.

3.7 Finding K for Oxford Clustering

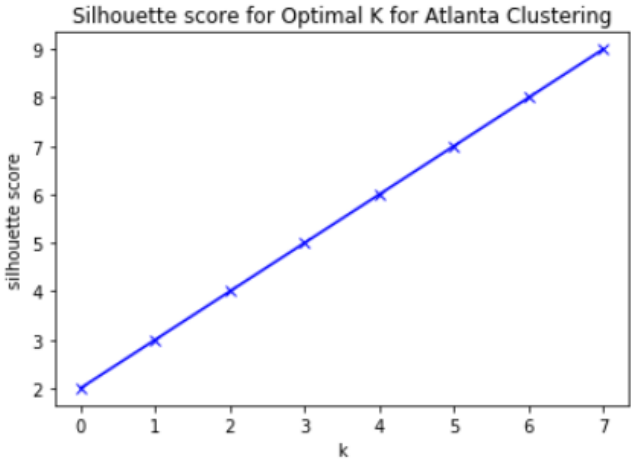


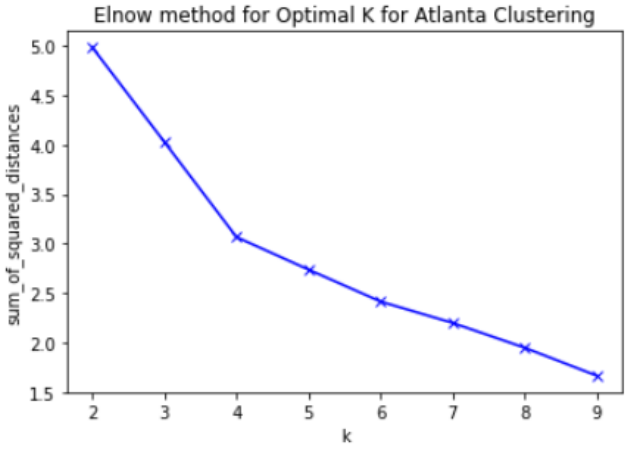


I determined the best K would be either 5 or 6 for the Oxford venue data. However after implementing both, it was clear the neighborhood clustering stopped at 5.

3.8 Finding K for Atlanta Clustering

I determined the best K could be 4 for the Atlanta venue data. However I felt that was a bit low for clustering 28





neighborhoods and wanted there to be at least as many clusters in Atlanta as in Oxford. I implemented clustering using 5 and 6 and ultimately choose 6 as a good option for K in this case.

3.9 K-Means Clustering Neighborhoods

Using the K-means algorithm, I clustered the neighborhoods in both cities and merged this data with the Top 10 Venue data frames. I also cleaned the data to ensure the clusters were integers and not floats, as otherwise they would not show up properly on the maps using Folium.

Oxford Clustered Neighborhoods Pandas Data frame

[34]:

	Postal Code	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 Most Common Venue	9th Con
0	OX3 0FR	Barton and Sandhills	51.770563	-1.208585	0	Bus Stop	Boutique	Wine Bar	Dessert Shop	Discount Store	Donut Shop	Electronics Store	Farmers Market	Fast Resta
1	OX4 6AA	Blackbird Leys	51.722543	-1.204506	4	Bus Stop	Pub	Recreation Center	Music Venue	Convenience Store	Wine Bar	Department Store	Dessert Shop	Dis
2	OX1 1AA	Carfax	51.750231	-1.267367	4	Pub	Bar	Nightclub	Hotel	Indian Restaurant	Pizza Place	Sandwich Place	Portuguese Restaurant	B
3	OX3 0BD	Churchill	51.753788	-1.230251	3	Coffee Shop	Gym / Fitness Center	Park	Café	Deli / Bodega	Wine Bar	Fish & Chips Shop	Dessert Shop	Dis
4	OX4 2BH	Cowley	51.730350	-1.206266	4	Hotel	Clothing Store	Fast Food Restaurant	Electronics Store	Sporting Goods Shop	Market	Pub	Supermarket	Gr

Atlanta Clustered Neighborhoods Pandas Data frame

t[50]:

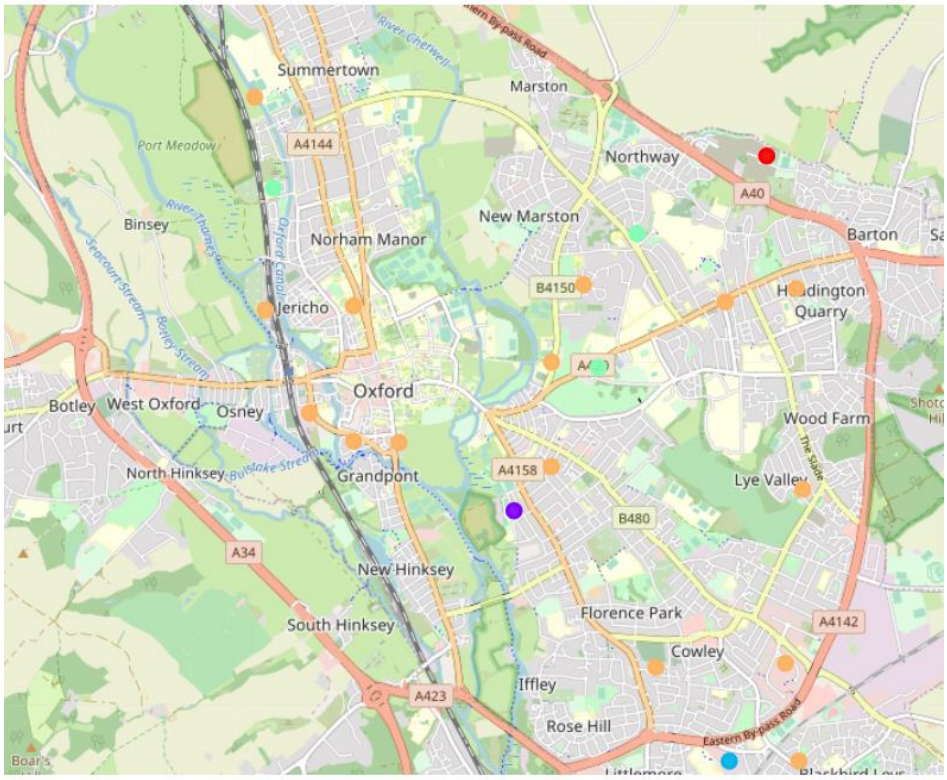
	Postal Codes	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Con
0	30303	Central Business District, Fairlee Poplar	33.752856	-84.39013	0	Sandwich Place	Mexican Restaurant	Theater	Smoothie Shop	Korean Restaurant	Burger Joint	Poke Place	Deli / Bodega	
1	30305	Buckhead, Garden Hills, Haynes Manor, Peachtre...	33.830054	-84.38472	0	Italian Restaurant	Café	Basketball Court	Shipping Store	Mobile Phone Shop	Sandwich Place	Salon / Barbershop	Farmers Market	Resta
2	30306	Virginia Highlands, Morningside/Lenox Park, Po...	33.786755	-84.35149	0	Doctor's Office	Thai Restaurant	Pet Store	Deli / Bodega	Coffee Shop	Café	Burger Joint	Massage Studio	Bou
3	30307	Candler Park ,Edgewood, Inman Park, Lake Clair...	33.768205	-84.33786	0	Playground	Athletics & Sports	Church	Golf Course	Tennis Court	Outdoor Sculpture	Park	Basketball Court	F Resta
4	30308	Midtown, Old Fourth Ward	33.771755	-84.38065	0	Southern / Soul Food Restaurant	Hotel	Gay Bar	Pizza Place	Donut Shop	Hookah Bar	New American Restaurant	Nail Salon	C Resta

4. Results and Discussion

4a. Mapping the Neighborhoods by Clusters

Using Folium once again and the new data frame including the top 10 venues in each neighborhood and the Cluster labels, I mapped out the neighborhoods in both cities. The neighborhoods are color coded by cluster to show the cluster groupings visually.

Map of Oxford Neighborhoods (Color Coded by Cluster)



4b. Labelling and Initial Analysis by Cluster: Oxford

Cluster 4 (Orange): Pubs, Shopping Mall, Restaurants, Museums and Bars
This cluster is the largest by a significant margin and includes 17 of the 24 Oxford Neighborhoods. This could be due to a number of factors including the range of venue types returned by Foursquare. As mentioned in the data section of this report, the venue list generated relies on the latitude and longitude coordinates provided for each neighborhood. If these coordinates are not the optimal choice then the venue data may be inaccurate and this could have skewed the cluster results.

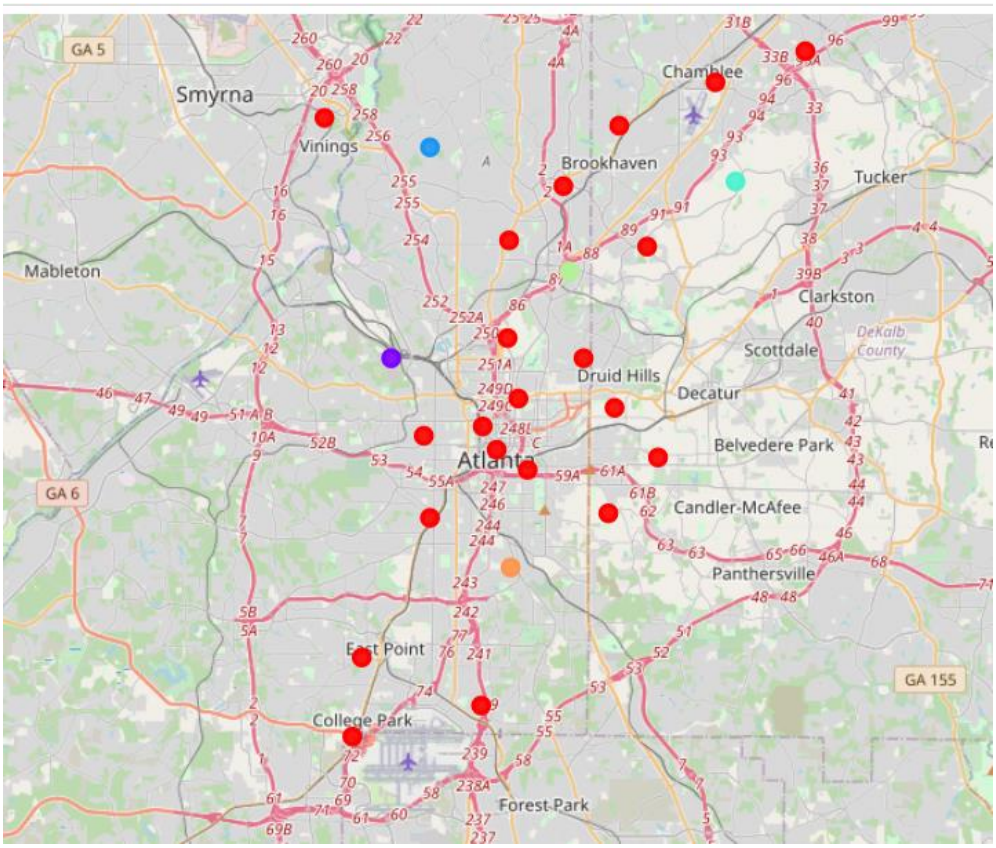
Cluster 3 (Light Green): Cafes and Parks
This cluster is the second largest with 3 neighborhoods.

Cluster 2 (Light Blue): Small Shops and Food

Cluster 1 (Purple): Pubs and Gyms

Cluster 0 (Red): Bus Transport, Boutiques and Food
The remaining clusters were assigned one neighborhood each. It may be that these areas did not have enough venues to properly cluster them or there were very distinctive venues. However looking at the top three venues listed for clusters 0, 1 and 2, this does not seem likely. It is also possible they are heavily residential or zoned for business.

Map of Atlanta Neighborhoods (Color Coded by Cluster)



4c. Labelling and Initial Analysis by Cluster: Atlanta

Cluster 0 (Red): Restaurants, Businesses, Tourist Attractions, Hotels, Breweries, Music Venues, Bars

This is by far the largest cluster of neighborhoods and we can see that neighborhoods across all areas of Atlanta have been included in this group. 21 of the 28 neighborhoods in Atlanta have been assigned to this cluster. As with cluster 4 from the Oxford data, it may be that the neighborhoods in this cluster have too wide a range of venue results to be very useful as a measure of similarity. Clustering based on other data or a subsection of the venue data could be required to better categorize these neighborhoods and break them down into smaller and more distinct clusters. It may also be that the radius needs to be changed when generating the venue lists from Foursquare.

Cluster 1 (Purple): Event Venues, Zoo Exhibits, and Fish Market

Cluster 2 (Light Blue): Gyms, Fast Food and Sports Stadiums

Cluster 3 (Teal): Nature/Parks, Zoo and Fast Food

Cluster 4 (Lime green): Residential Apartments, Gay Bars, and Smoke shops

Cluster 5 (Orange): Discount shops, Playgrounds and Southern/Soul Food Restaurants

The remaining clusters have only one neighborhood each. Again this may be due to inaccurate or incomplete venue data or it may be the result of better clustering than the above Cluster 0.

4d. Comparing Neighborhood Clusters Between Cities

For both cities we see a similar results pattern in the clustering of neighborhoods. Both have returned one cluster comprising the majority of the neighborhoods, with the remaining clusters generally having one neighborhood each. The most similar clusters between the two cities are these large clusters, Cluster 0 in Atlanta and Cluster 4 in Oxford. However it is clear that more clustering analysis on the basis of other data beyond nearby venues will be required to more accurately group similar neighborhoods in each city. Even if this is accomplished, the results may still show that there are many neighborhood clusters that do not have direct comparison between these two cities. This could be due to a number of factors, such as the geographical size and layout of the neighborhoods and differences in culture and lifestyle between the US and the UK. Further analysis and investigation is required.

It may also be necessary to better clean the venue data returned by Foursquare API. As we can see below, some of the top venues listed and used in the clustering analysis include uninformative categories such as 'Bus Stop' or 'Miscellaneous Shop' or 'Discount Store'. This may or may not be a significant venue and could be excluded for more statistically significant venues. This is something to consider if this project were to be replicated.

Top Five Venues in each Cluster: Oxford and Atlanta

Cluster Labels	OXF Cluster 0	OXF Cluster 1	OXF Cluster 2	OXF Cluster 3	OXF Cluster 4
1st Most Common Venue	Bus Stop	Pub	Miscellaneous Shop	Coffee Shop	Bus Stop
2nd Most Common Venue	Boutique	College Gym	Wine Bar	Gym / Fitness Center	Pub
3rd Most Common Venue	Wine Bar	Thai Restaurant	Food Truck	Park	Recreation Center
4th Most Common Venue	Dessert Shop	Flower Shop	Department Store	Café	Music Venue
5th Most Common Venue	Discount Store	Deli / Bodega	Dessert Shop	Deli / Bodega	Convenience Store

Cluster Labels	ATL Cluster 0	ATL Cluster 1	ATL Cluster 2	ATL Cluster 3	ATL Cluster 4	ATL Cluster 5
1st Most Common Venue	Sandwich Place	Event Service	Gym / Fitness Center	Lake	Residential Building (Apartment / Condo)	Discount Store
2nd Most Common Venue	Mexican Restaurant	Zoo Exhibit	Fast Food Restaurant	Zoo Exhibit	Gay Bar	Playground
3rd Most Common Venue	Theater	Fish Market	Fried Chicken Joint	Fast Food Restaurant	Smoke Shop	Southern / Soul Food Restaurant
4th Most Common Venue	Smoothie Shop	Frozen Yogurt Shop	Fountain	Fried Chicken Joint	Zoo Exhibit	Dumpling Restaurant
5th Most Common Venue	Korean Restaurant	Fried Chicken Joint	Football Stadium	Fountain	Fish Market	Electronics Store

5. Conclusions

This project has given us some insight into the amenities in the selected neighborhoods in both Oxford and Atlanta, which partially fulfills the intended purpose of the exercise. The information garnered provides a useful, albeit cursory and broad, snapshot of each neighborhood. However based on the results it is clear we need more holistic data to improve the accuracy and usefulness of our neighborhood clustering. If I were to redo this project, I would consider including data on population, cost of living, demographics, schools and transportation. I would also better clean the venue data and ensure that the best map coordinates were being used to represent each neighborhood in order to improve the accuracy of venue results. Finally, I would consider whether factors such as culture or geographical size and spread are impacting the results and how these could be minimized to better standardize the data and subsequent results to ensure more accurate comparison.

Thank you for reading! This project was created for my Coursera capstone course to complete my IBM Professional Certificate in Data Science.