Capstone Project

The Battle of the Neighborhoods - Week 5 Report

Ricky To

4th July 2020

1.0 Introduction and Problem

As part of life, people move around for many reasons, for career prospects, to be closer to better schools, for culture or to be closer to parents. There could be intercity migration or international migration, but both cases will happen.

Prior to moving to any city, people will research the potential home and try to make sense of it all. There are multiple features of a potential city or neighbourhood all of which define the character or type of place a neighbourhood is and most importantly whether it is a suitable location to settle in. However, it can be a daunting task to collect all of this information and a very time consuming process especially if the potential location is an unknown location.

Doing a search online for a good neighbourhood will only garner results that are opinionated and possibly be subjective as everyone has their own preferences. A young person may want to be close to places associated with nightlife such as restaurants and bars, whereas a young couple may want to be closer to their place of work and within the reach of all the necessary amenities such as gyms or cafes.

In order to truly know whether a neighbourhood is suitable to settle in, we need hard data. Data such as house prices, schools, crime, weather etc. Being able to get this data will better

inform the potential resident of the things they may value more.

This project will focus on Mr Willis, Mr Willis is married to his Wife Mrs Willis for 9 years and they have one Daughter who is currently in Primary School. They are currently overseas residing in Singapore as expatriates and have been doing so for the past 5 years, however they wish to return to the United Kingdom but at the same time wish to move to one of the major UK cities. One of the major cities that they are targeting is Manchester, however aside from having visited Manchester in the past, they have never spent a great deal of time there and do not know anyone there to assist.

They want to move to a location that is quiet, within budget and has restaurants and schools nearby.

What we want to do is to help Mr Willis and his family determine which of the boroughs in Greater Manchester can be called their future home through the collection and analysis of publicly available data on the web.

Other Target Audiences

- This model can also be used by others who may wish to move to Manchester, and can be reconfigured to analyse other Cities around the world.
- Data Scientists, who want to implement the Exploratory Data Analysis techniques to obtain necessary data and analyze it.

This project can be used by the user at the time of rental apartment or buy house in a locality based on the distribution of various facilities available around the borough. As an example, this project would compare Manchester's boroughs and analyses the venues in each of those boroughs based on the number of visits by people in each of those places.

The project will also use K-mean clustering unsupervised machine learning algorithm to cluster the venues based on the place category such as restaurants, park, coffee shop, gym, clubs etc. This would give a better understanding of the similarities and dissimilarities between the various neighborhoods.

1.1 City Background

Manchester is a city and metropolitan borough in Greater Manchester, England. It is historically and traditionally a part of the county of Lancashire. It has a population of 547,627 as of 2018 (making it the fifth most populous English district). It lies within the United Kingdom's second-most populous urban area, with a population of 2.5 million and third most populous metropolitan area, with a population of 3.3 million. It is fringed by the Cheshire Plain to the south, the Pennines to the north and east, and an arc of towns with which it forms a continuous conurbation. The local authority for the city is Manchester City Council.

2.0 Data

In this project, I will be using the following datasets to help solve Mr Willis' problem.

The main criterion for the Willis Family in selecting a suitable borough will be based on:

- Crime and Safety
- Number of Restaurants
- Schools
- House within Budget

2.1 Key Libraries being utilised

Python packages and Dependencies:

• Pandas - Library for Data Analysis

• NumPy – Library to handle data in a vectorized manner

• Geopy – To retrieve Location Data

• Requests – Library to handle http requests

• Matplotlib – Python Plotting Module

• Sklearn – Python machine learning Library

• Folium – Map rendering Library

• Beautifulsoup - Webscraping tool

2.2 General

Firstly we need to find out how many boroughs are in Greater Manchester and their locations. This information will be obtained by Webscraping Wikipedia and then converted into a dataframe using Pandas.

- List of Manchester Boroughs - https://en.wikipedia.org/wiki/List_of_places_in_Greater_Manchester ,and Manchester PostCodes - https://simple.wikipedia.org/wiki/M_postcode_area#:~:text=The%20M%20postcode%20area%2C%20also,Manchester%20boroughs%20except%20for%20Stockport.

The Coordinates (Latitude and Longitude) of each Borough will be extracted from Google Maps manually as Wikipedia does not contain the coordinates of each of the boroughs.

We will also obtain the Demographics of each borough to gauge their size and extract any other information from this that may be of use to the Willis family.

- Demographics of Manchester -
  https://www.citypopulation.de/en/uk/greatermanchester/ or
  https://en.wikipedia.org/wiki/List_of_Greater_Manchester_settlements_by_populat
  ion

#	Settlement	Population[1]	
		Census 2001	Census 2011
1	Manchester‡ (see notes)	405,300	510,746
2	Bolton	182,980	194,189
3	Sale	122,990	134,022
4	Rochdale	102,080	107,926
5	Stockport	106,170	105,878
6	Salford	88,460	103,886
7	Wigan	98,480	103,608
8	Oldham	103,544	96,555
9	Bury	74,540	77,211
10	Leigh/Atherton	65,980	70,542

2.3 Crime and Safety

In order to have an understanding of the safety levels of each borough we shall extract the recorded crime rates in Greater Manchester, this information will be extracted from the UK police database - https://data.police.uk/ and this information will be in the form of a csv file which will be converted into a dataframe using the Pandas library.

We will then later convert this information into a visualisation of the information either using a Chart or a folum map showing the clusters of crime in their respective areas.

2.4 Number of Restaurants

We will use the Foursquare API gather information in relation to venues and places in and around the Greater Manchester boroughs. Foursquare is a social location service that allows users to explore the world around them.

Foursquare has other information such as reviews and photos giving a more layered view of the venues in a certain area. By gathering this information we will be able to get a more

rounded view of what amenities and venues are available in a certain borough such as restaurants and cinemas.

HTTP requests would be made to this Foursquare API server using the location of Manchester's boroughs neighborhoods to pull the location information (Latitude and Longitude).

2.5 House within Budget

We would also need to be realistic and consider that the cost of a property must be taken into consideration, as the best neighbourhoods often come at a price. There is a fine balance between a reasonably priced home and a nice neighbourhood to live in.

We shall therefore extract the median household prices paid in Greater Manchester from a csv extracted from https://mappinggm.org.uk/metadata/ which has recorded prices paid in all areas of Greater Manchester across a number of years, however we will use the latest prices paid in 2020 to analyse.

2.6 Schools

Finding a good school for the Willis family is also a key requirement. The information source that we will be dealing with is a csv file downloaded from the gov.uk website:

- https://www.compare-school-performance.service.gov.uk/download-data?currentstep=year&regiontype=beforeStep&la=&downloadYear=2018-2019

3.0 Methodology

3.1 Data preparation

I used Wikipedia to obtain the boroughs of Greater Manchester and then searched the Latitude and Longitudes separately as they were not available in the original list from Wikipedia. There are a number of smaller neighborhoods that were not identified in Wikipedia and therefore cannot be used in this Analysis.

The Population data was also extracted from Wikipedia. Although, both the population data and the borough list are from Wikipedia, the boroughs listed in each list did not align and due to around 40% of the boroughs not having population data available this feature had to be disregarded.

The crime data was widely available and comprehensive enough to work with, it also included very specific locations including streets but unfortunately as the Overall list of boroughs extracted from Wikipedia was shorter it meant the smaller neighbourhoods were merged into the larger boroughs.

The property price paid data was also comprehensive and was only limited by the borough list's high level nature found on Wikipedia.

Once information on location and population of the boroughs were gathered, they were merged into one dataframe for further use later as part of our analysis.

3.2 Feature Selection

After data cleaning, there were 20 features in the dataset that were deemed to assist us in determining the best Borough for the Willis Family to move to. All other features were removed in the Data Cleaning stage. Each of the features that have been retained are independent and are required to answer each of the Criterion that has been laid out in the beginning.

3.3 Exploratory Data Analysis

Now that all the dataframes have been set up extensive comparative analysis of the boroughs can begin to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.

The primary display formats that will be utilised are folium maps to visualise the collated information on a map of Manchester and the supported using matplotlib charts to display information that could be seen to be better represented as.

3.31 Borough Locations

Tool used: BeautifulSoup, Pandas, Geopy, and Folium

In order to extract the information on Wikipedia we will use Beautifulsoup to web scrape the information contained in the tables.

```
[ ]  import requests
     import lxml
     from bs4 import BeautifulSoup
```

```
# Obtaining data from Wikipedia
source = requests.get('https://en.wikipedia.org/wiki/List_of_Greater_Manchester_settlements_by_population').text
soup = BeautifulSoup(source, 'lxml')
soup.encode("utf-8-sig")
```

As a starting point it is useful to visualize the locations of each borough. For that, the map of Manchester was created with the borough locations superimposed on top.
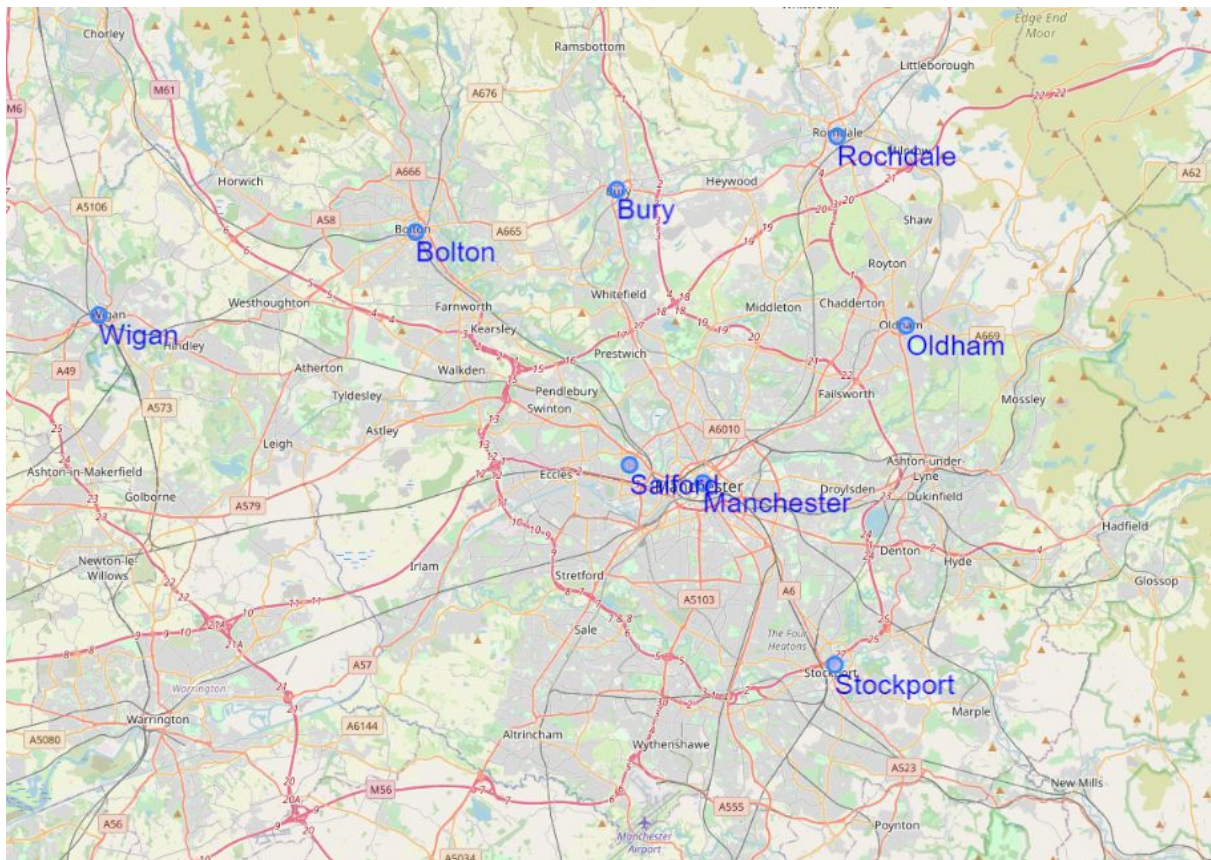
3.32 Crime and Safety

Tool used: Pandas, sklearn, Folium and Matplotlib

Unfortunately there was not enough population data to cover all boroughs in this list in order to normalise the criminal data that has been extracted for Manchester. However instead we are able to visually inspect the cluster of crimes around each Borough as well as group them by Borough in order to get a good overview of the safety environment.

Bar chart - cases by borough
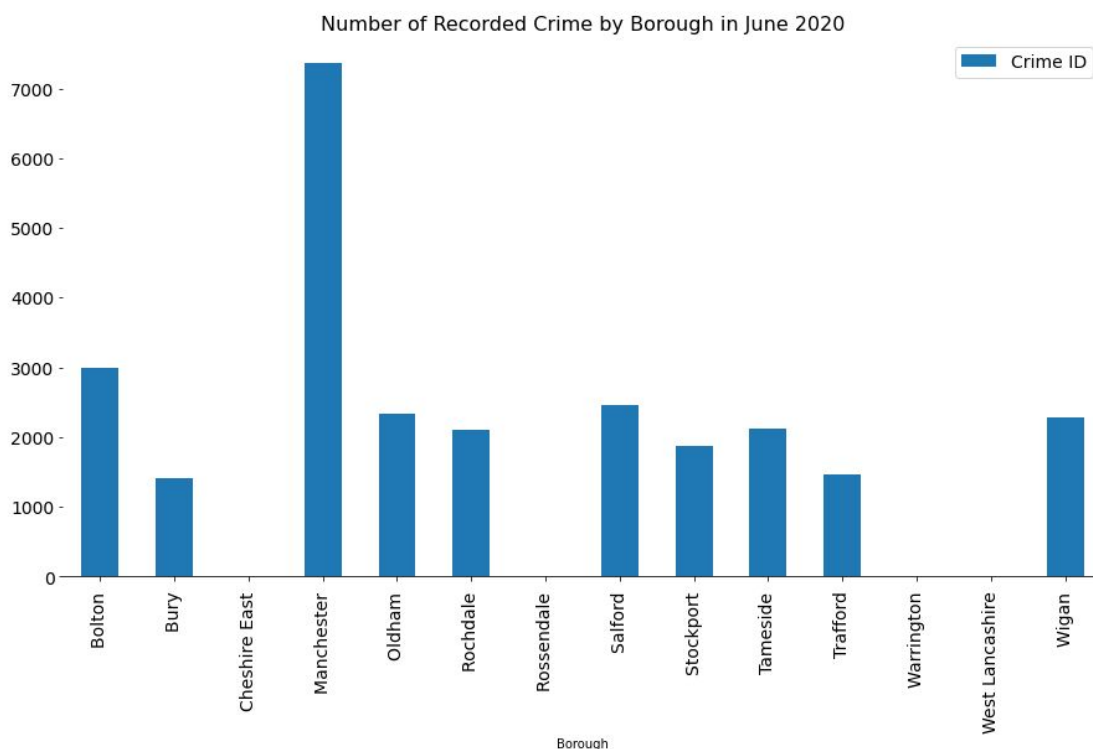
This is a bar chart displaying boroughs for the month of June 2020 for all crime incidents reported. Manchester has the biggest number of reported crime, which is at the heart of the City, which is then followed by Bolton. The Areas with the least are Cheshire East, Rossendale, Warrington and West Lancashire, however their figures are nominal and could therefore not be visualised in this graph. But Since these Boroughs are not part of our main list of Boroughs we will consider them as outliers, it will therefore mean Bury, Trafford and Stockport having the lowest number of crime incidents.
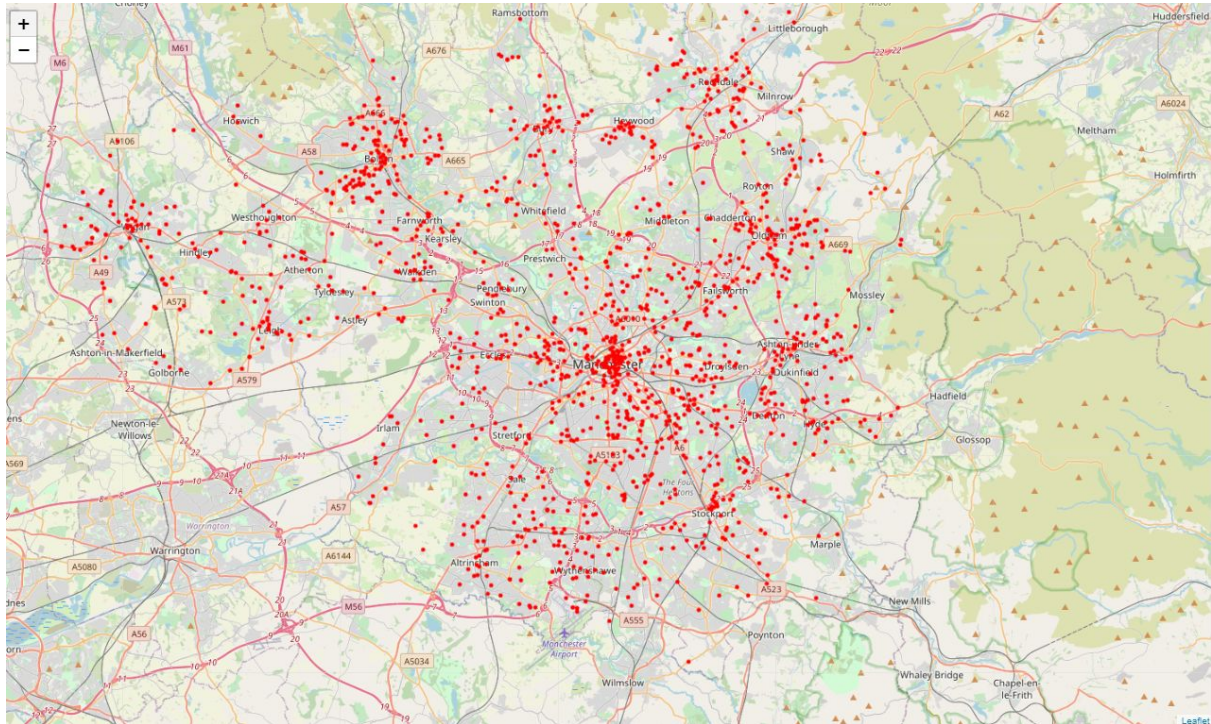
In order to get a visualised perspective on the density of the Crime rate locations this chart is also created that maps out all the June Crime Statistics onto the Map of Greater Manchester. Due to the sheer number of crime statistics we were unable to plot them all onto the map, however a sample size of 5% was taken in order to generate the map using sklearn to split.

```
from sklearn.model_selection import train_test_split

train, test = train_test_split(crime, test_size=0.05)
```

3.33 House within Budget

Tool used: pandas and Matplotlib

After extracting the information into a Dataframe using Pandas we grouped them into Boroughs and generated the Median and Mean prices paid for each location.

Dataframe of Property Prices Paid by Borough

	Borough	Population	Latitude	Longitude	Median-PricePaid	Mean-PricePaid
0	Manchester	510,746	53.480759	-2.242631	165000.0	226208.930079
1	Bolton	194,189	53.576866	-2.428219	128000.0	165499.729730
2	Rochdale	107,926	53.613600	-2.156100	125000.0	148055.183024
3	Stockport	105,878	53.410600	-2.157500	230050.0	282065.896332
4	Salford	103,886	53.487500	-2.290100	150010.0	193393.930185
5	Wigan	103,608	53.545100	-2.632500	125000.0	172105.265248
6	Oldham	96,555	53.540932	-2.111366	125000.0	146423.039216
7	Bury	77,211	53.593460	-2.298540	160000.0	181432.515850

Barchart - Property Price Paid by Borough

Median and Mean Price Paid for a Property by Borough in 2020
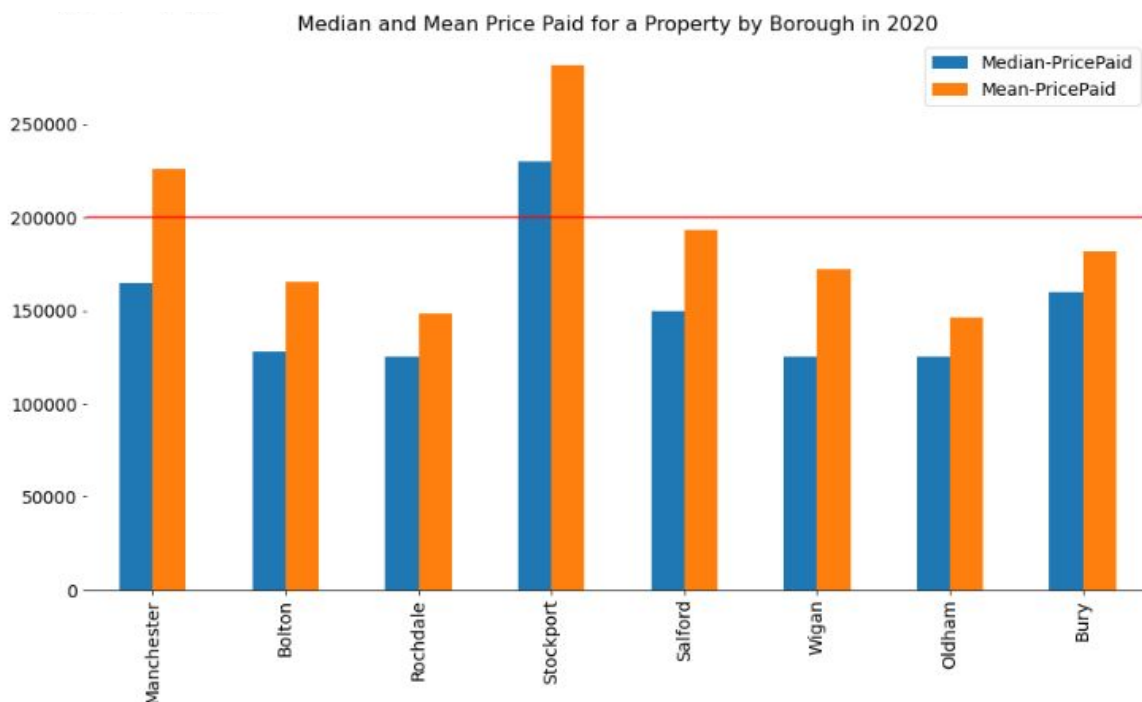
We have set the threshold line to be 200,000GBP as this is the budget that Mr. Willis has to purchase a house. This would mean that properties in Stockport would be out of the range, however the remaining Boroughs are still available to them, with the exception of Manchester if you consider then Mean Price Paid instead of the Median Price Paid.

3.34 Number of Restaurants

Tool used: Foursquare, Folium and sklearn

The information on venues, venue types and their locations were extracted using the Foursquare API and then put into a dataframe using pandas.

Foursquare API search feature was used to collect the nearby places of the neighborhoods. The number of places per neighborhood parameter was set to 100 and the radius parameter would be set to 2000m as the boroughs are quite large.

Extracting this information can gave us a broad view on whether a particular neighbourhood is primarily residential or more of a mixed use borough with restaurants.

After extracting a list of venues in the Boroughs and this list was further filtered down to the number of Restaurants per Borough.

```
#Get top 100 venues in 3000m radius of the center of each Borough
LIMIT = 100
venues = getNearbyVenues(names=McrCoord['Borough'],
                         latitudes=McrCoord['Latitude'],
                         longitudes=McrCoord['Longitude']
                        )
```

```
Manchester
Bolton
Rochdale
Stockport
Salford
Wigan
Oldham
Bury
```

```
print(venues.shape)
venues.head()
```

(639, 7)

	BoroughName	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Manchester	53.480759	-2.242631	Café Beermoth	53.481359	-2.241902	Bar
1	Manchester	53.480759	-2.242631	Royal Exchange Theatre	53.482555	-2.244770	Theater
2	Manchester	53.480759	-2.242631	Manchester Art Gallery	53.478882	-2.241817	Art Gallery
3	Manchester	53.480759	-2.242631	Northern Soul • Grilled Cheese	53.483074	-2.238303	Sandwich Place
4	Manchester	53.480759	-2.242631	Hotel Gotham	53.480767	-2.242235	Hotel

Restaurant locations in Greater Manchester using Folium Map were generated.
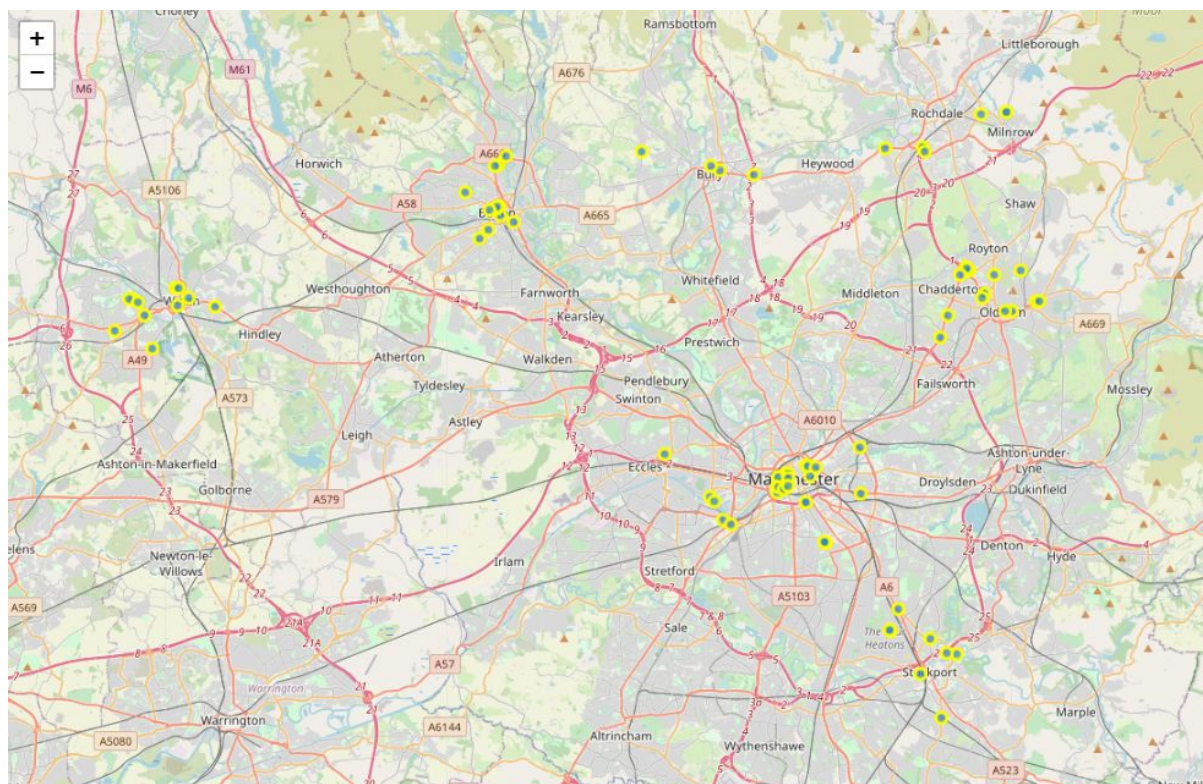
Here is the merged Dataframe for the Venue Categories

	Borough	Mean-ReadingScore	Mean-MathScore	Population	Latitude	Longitude	Median-PricePaid	Mean-PricePaid	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
0	Bolton	102.720000	104.090000	194,189	53.576866	-2.428219	128000.0	165499.729730	2	Pub	Clothing Store	Fast Food Restaurant	Supermarket	Pizza Place	Coffee Shop	Grocery Store
1	Bury	105.200000	105.175000	77,211	53.593460	-2.298540	160000.0	181432.515850	3	Pub	Discount Store	Coffee Shop	Clothing Store	Hotel	Department Store	Grocery Store
2	Manchester	102.535836	102.774744	510,746	53.480759	-2.242631	165000.0	226208.930079	1	Coffee Shop	Pub	Indian Restaurant	Beer Bar	Hotel	Café	Park
3	Oldham	103.812500	104.437500	96,555	53.540932	-2.111366	125000.0	146423.039216	2	Supermarket	Fast Food Restaurant	Indian Restaurant	Grocery Store	Clothing Store	Pub	Discount Store
4	Rochdale	102.534884	103.953488	107,926	53.613600	-2.156100	125000.0	148055.183024	0	Supermarket	Pub	Grocery Store	Gastropub	Pizza Place	Sandwich Place	Clothing Store
5	Salford	103.172414	103.793103	103,886	53.487500	-2.290100	150010.0	193393.930185	1	Coffee Shop	Pub	Bar	Hotel	Gym / Fitness Center	Indian Restaurant	Park
6	Stockport	104.800000	104.785714	105,878	53.410600	-2.157500	230050.0	282065.896332	4	Pub	Grocery Store	Coffee Shop	Supermarket	Pharmacy	Clothing Store	Pizza Place
7	Wigan	102.594595	103.202703	103,608	53.545100	-2.632500	125000.0	172105.265248	2	Pub	Coffee Shop	Grocery Store	Clothing Store	Sandwich Place	Fast Food Restaurant	Discount Store

Clusters and k-means

Using the sklearn k means clustering package we a start to categorise the boroughs by the cluster types. We have given different colour codes depending on the type of Cluster it is, whether it is a Borough that is focused on restaurants or other amenities.
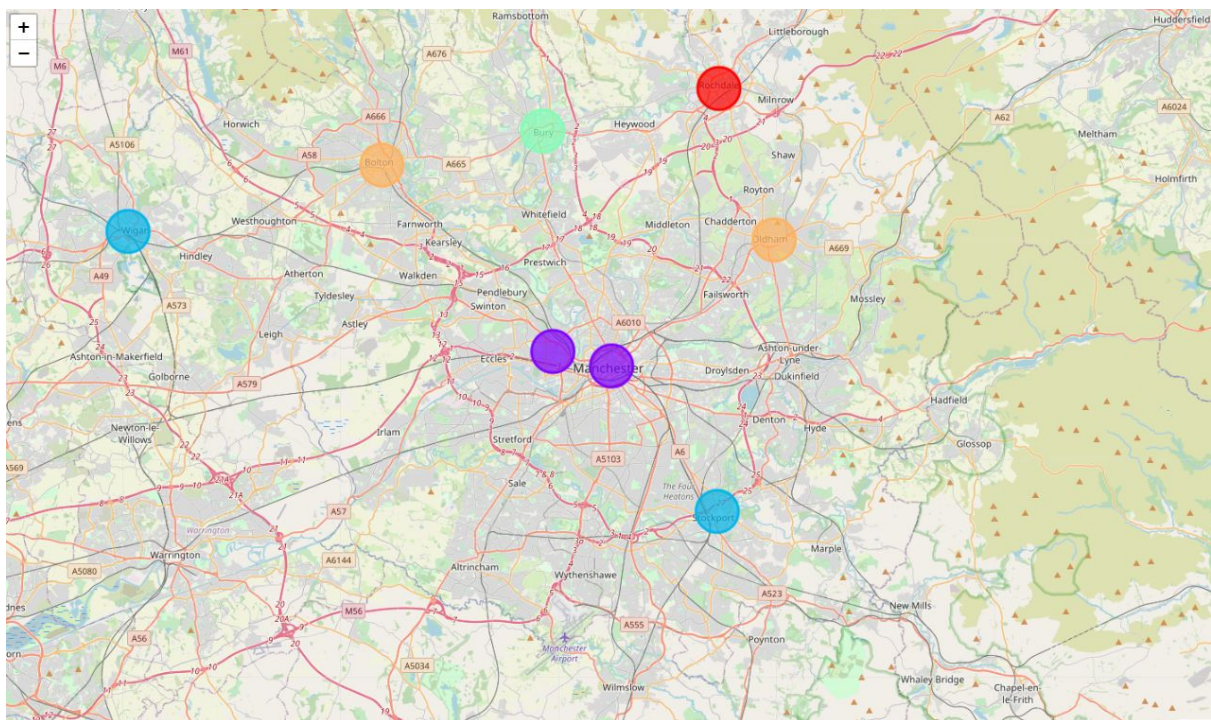
- Pub and Grocery Stores (Green) - Cluster 0
- Bars, Restaurants, Hotels (Red) - Cluster 1
- Coffee shops, Restaurants, Grocery stores (Purple) - Cluster 2
- Pubs, Small Stores and Shops (Orange) - Cluster 3
- Grocery Stores, Restaurants and Shops (Blue) - Cluster 4

The differences are minor as each of the Boroughs appear to be generally of mixed use, however some boroughs have some tendencies for certain amenities over others. Of all the locations it will be

the Red, Purple and Blue Clusters that will be of interest to the Willis Family as they are interested in an area with good accessibility to Restaurants.

3.35 **Schools**

Tool used: Pandas and Matplotlib

The School information in the csv file is converted into a dataframe and merged with the other dataframes, in order to associate the schools to a borough coordinate using Pandas.

	Borough	Mean-ReadingScore	Mean-MathScore	Population	Latitude	Longitude	Median-PricePaid	Mean-PricePaid
0	Bolton	102.720000	102.720000	194,189	53.576866	-2.428219	128000.0	165499.729730
1	Bury	105.200000	105.200000	77,211	53.593460	-2.298540	160000.0	181432.515850
2	Manchester	102.535836	102.535836	510,746	53.480759	-2.242631	165000.0	226208.930079
3	Oldham	103.812500	103.812500	96,555	53.540932	-2.111366	125000.0	146423.039216
4	Rochdale	102.534884	102.534884	107,926	53.613600	-2.156100	125000.0	148055.183024
5	Salford	103.172414	103.172414	103,886	53.487500	-2.290100	150010.0	193393.930185
6	Stockport	104.800000	104.800000	105,878	53.410600	-2.157500	230050.0	282065.896332
7	Wigan	102.594595	102.594595	103,608	53.545100	-2.632500	125000.0	172105.265248

We can then also start making score comparisons between the schools of various boroughs.

Dataframe of the School scores for Key Stage 3 students Greater Manchester.

| Borough	Mean-ReadingScore	Mean-MathScore
Bolton	102.720000	104.090000
Bury	105.200000	105.175000
Manchester	102.535836	102.774744
Oldham	103.812500	104.437500
Rochdale	102.534884	103.953488
Salford	103.172414	103.793103
Stockport	104.800000	104.785714
Wigan	102.594595	103.202703

School Scores by Borough
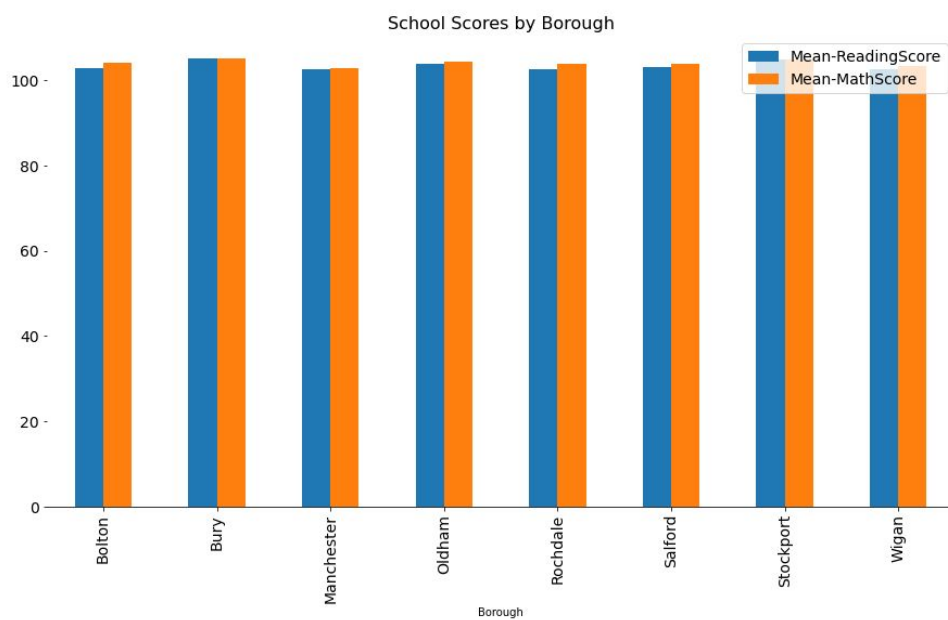
Barchart - school median by borough

As indicated through the data extracted and in the graphs it is apparent that the quality of schools are very much similar with no outstanding performance from each of the boroughs. However the 2 boroughs that were slightly ahead of the rest were Bury and Stockport in both the median and mean results which correlates with each other as well and show little difference in the results.

To see the differences up closer we normalise the results to achieve the below graph.

Borough	MathSchoolScore	ReadingSchoolScore
Bolton	0.547965	0.069459
Bury	1.000000	1.000000
Manchester	0.000000	0.000357
Oldham	0.692741	0.479385
Rochdale	0.491091	0.000000
Salford	0.424271	0.239213
Stockport	0.837815	0.849913
Wigan	0.178297	0.022405

Normalised School Scores by Borough
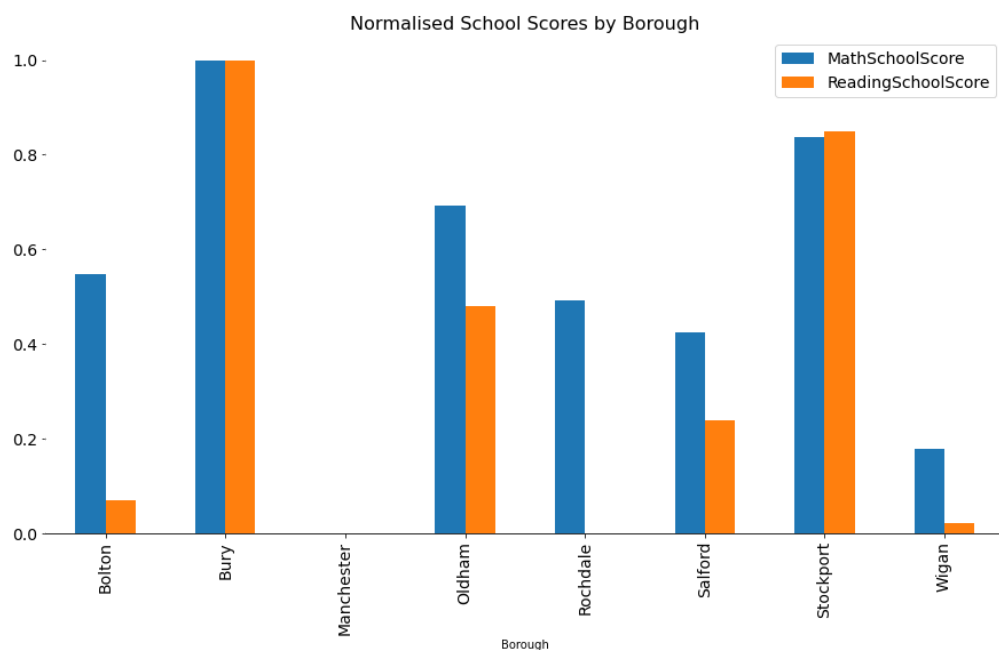
4.0 Results and Recommendations

The publicly made data for Greater Manchester allowed us to delve deeper than perhaps we could have compared to other cities, which is fortunate. Without this data it would have been a difficult exercise to come up with a recommendation.

During the analysis we had worked with 7 Boroughs in Greater Manchester with 5 Clusters generated by K-Means.

In Terms of Crime and Safety the Safest Boroughs identified from the results were **Bury, Trafford and Stockport** if we ignore the outlier results, which are the results closest to the value of 0.

	Borough	Number	NormalizedScore
6	Rossendale	1	0.000000
12	West Lancashire	3	0.000227
11	Warrington	5	0.000453
2	Cheshire East	6	0.000566
10	Trafford	1764	0.199728
1	Bury	1822	0.206299
8	Stockport	2424	0.274499
5	Rochdale	2510	0.284242
9	Tameside	2548	0.288547
7	Salford	2771	0.313810
4	Oldham	2842	0.321853
13	Wigan	2924	0.331143
0	Bolton	3609	0.408746
3	Manchester	8828	1.000000

Considering the other key features using the below dataframe with Normalised data

	Borough	Median-PricePaid	Mean-ReadingScore	Mean-MathScore	Cluster Labels	ReadingSchoolScore	MathSchoolScore
0	Bolton	128000.0	102.720000	104.090000	2	0.069459	0.547965
1	Bury	160000.0	105.200000	105.175000	3	1.000000	1.000000
2	Manchester	165000.0	102.535836	102.774744	1	0.000357	0.000000
3	Oldham	125000.0	103.812500	104.437500	2	0.479385	0.692741
4	Rochdale	125000.0	102.534884	103.953488	0	0.000000	0.491091
5	Salford	150010.0	103.172414	103.793103	1	0.239213	0.424271
6	Stockport	230050.0	104.800000	104.785714	4	0.849913	0.837815
7	Wigan	125000.0	102.594595	103.202703	2	0.022405	0.178297

After reviewing the Cluster makeup it was determined that Clusters 1,2 and 4 are locations with a more stronger presence of Restaurants which leaves us with **Manchester, Bolton, Oldham, Wigan and Stockport**.

200,000 is the Willis' Family budget to buy a house and based on the Median Price paid in the year of 2020 most Boroughs would be affordable with the exception of Stockport.

Finally, with regards to finding an area with a good school for Mr Willis' daughter to attend, the Boroughs with the highest scores were from **Bury, Stockport and Oldham**.

4.1 What could have been done better?

Using Foursquare API, information could be collected for various venues around Manchester however not all venue categories appear to have been available. In this exercise it appeared that Schools could not be easily filtered out from Foursquare unless the actual school name was known, this lead to having to go to other publicly available sources to obtain their information.

One of the other challenges in this was that when using Wikipedia to extract the Borough information there appeared to be inconsistencies between different wiki pages as to what were the main Boroughs, as some lists were longer than others. So clearly the definition of a Borough will vary depending on the source you would use.

5.0 Conclusions

So which area is the best for the Willis Family to move to:

Borough	Crime	Restaurants	Property	School	Total
Manchester		1	1		2
Bolton		1	1		2
Oldham		1	1	1	3
Wigan		1	1		2
Stockport		1		1	2
Salford	1		1		2
Rochdale			1		1
Bury	1		1	1	3
Trafford	1		1		2

Counting up all the scores from the results section there are 2 potential locations that the Willis Family should focus on, Bury and Oldham. However, each location does not offer all 4 requirements and therefore it will be down to which criteria they value more. Oldham has more Restaurants but has slightly more Crime incidents compared to Bury, which has fewer Restaurants but less crime.

Both locations in Greater Manchester are suitable to move to, and it is now down to personal preference for the Willis Family.