**Introduction**

We have seen previously that different neighbourhoods in New York and Toronto can be clustered together into similar groups. However geographical information were not included, and we have seen cases of clusters which were formed of disjoint neighbourhoods, and separated by other clusters geographically.

In this project, I will look to group neighbourhoods in Sutton, UK into clusters based on both venues in each area as well as their geographical locations. This will be useful to both local business and councils, to determine local town centres, as well as possible catchment areas that will be of interest for each area. For example, when trying to open a new store, business may consider an area which have wide geographical reach, yet currently lacks competitors in the industry it is in.

**Data Used**

1. Ordnance Survey data, which is public information, that contains the latitude and longitude for postal areas. In our project, we will be looking at all the postal sectors in Sutton (all except for last 2 letters of postcode), and the latitude and longitude for each sector. This will be useful to our project from the geographic location point of view.

Sample dataset

|  |  |  |
| --- | --- | --- |
| Postal Sector | Latitude | Longitude |
| SM1 1 | 51.3646 | -0.19466 |

1. Foursquare data, by making calls to its API, we can obtain the venues of interest in each postal sector, such as restaurants, supermarket etc. This will be useful to our project from the business point of view, to help cluster neighbourhoods together based on similar businesses

Sample dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Venue | Venue Latitude | Venue Longitude | Venue Category |
| Holiday Inn - Sutton | 51.361713 | -0.196645 | Hotel |

**Methodology**

**Exploratory Analysis**

The venue categories from Foursquare API has been cleaned in order to group small niche categories together. This is equivalent to factor analysis, however as we have background information on the categories, manual grouping will already suffice our needs. This would help to reduce the noise and correlation within columns, and help to result in better clusters.

One Hot Encoding is then used on these categories, to transform the data as a percentage of occurrences within each category, as categorical variables ca not be used directly during clustering.

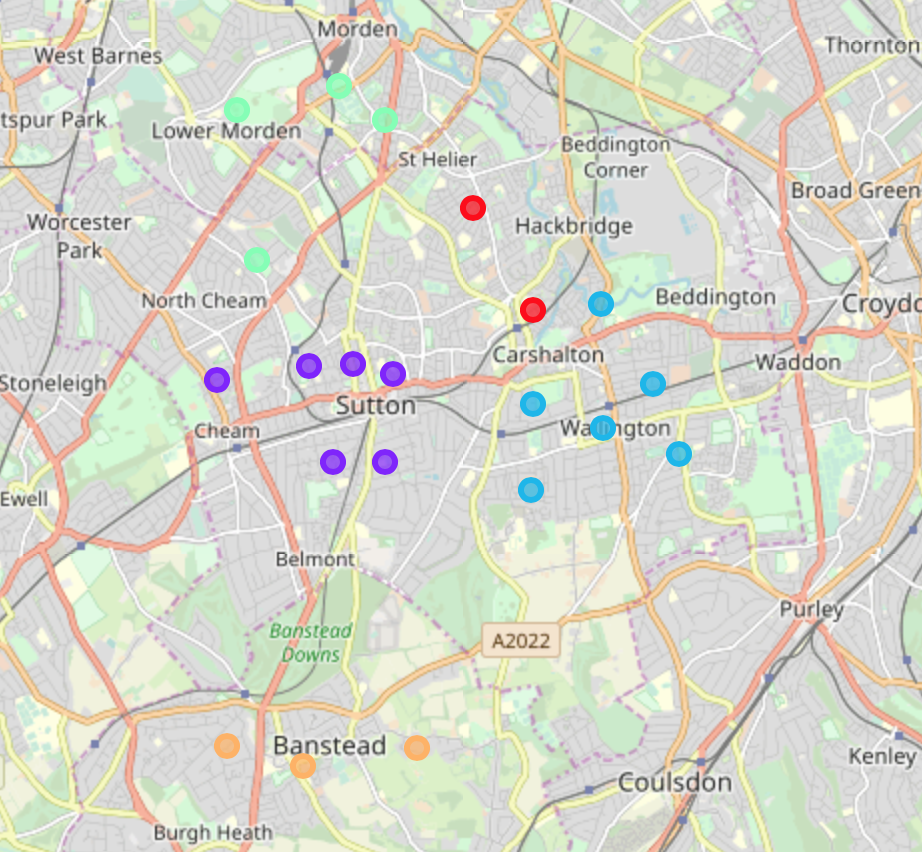
Latitude and longitudes have also been standardised according to how far they are from the mean latitude and longitudes. This will help the clustering to be done at standardised formats, rather than being affected by the different scales of metrics in each variable.

**Machine Learning**

The neighbourhoods will be clustered using K-Means methods. This method starts at random centre cluster points, and each time assign neighbourhoods to the closest cluster. After each iteration, the cluster centre point will be re-calibrated according to the data currently already in each cluster, and each neighbourhood will once again be reassigned to the closest clusters, and so on. This is fitting method to our project, as we will be including geographic data such as latitude and longitude, hence the methodology of assigning the closest together make sense both logically and from business point of view.

**Results and Discussion**

The output of our project produced 5 clusters, map as below



The output is as to our expectation, with each clusters disjoint from each other, and none by cut in halves by other clusters. The clusters also align fairly nicely with the postal districts (all except last 3 letters of the postcode), as shown on Wikipedia <https://en.wikipedia.org/wiki/SM_postcode_area>. This shows that the different postal districts in Sutton are not only distant from each other geographically, but also exhibit slightly demographics of businesses. There are slight blurs around the edges of the clusters, suggesting that nearby area in adjacent district may still have similar businesses.

**Conclusion**

Overall, we have shown that by adding in geo-location information into clustering can help to achieve more sensible results, and more importantly results that can be explained better from business points of views. Example being how they very nicely align with the towns in real-life.



The results from this project in particular will be useful to interested parties. For example, local government and council can use this to classify urban areas and town. Whilst businesses can use this to establish their catchment areas and new store planning.