The Battle of Neighbourhoods

Introduction:

As of late, Machine Learning (ML) calculations are generally utilised in the investigation of information rather than customary measurements. ML calculations bring favourable circumstances since they offer answers for issues identified with the huge amounts of information and set less requirements than conventional measurements. Specifically, unaided learning calculations are utilised to discover designs in information as far as similitude between tests. Contingent upon the example inside the information, various calculations are utilised. For non-convex data it is utilised Density-Based Spatial Clustering (DBScan). Then again, for convex data it is used a well known algorithm as K-Means.

Foursquare is where individuals remark and rank food destinations, espresso locales, shopping centres and stops. For example, how about we feel that a Foursquare client needed to move from New York city, USA to the city of Toronto, Canada. Foursquare area information alongside a bunching calculation can recommend an area so as to assist this client with living in Toronto in a comparative spot. The local that will be proposed, won't be an irregular recommendation, yet rather will be a spot for his pleasure. Hence, past information from New York and Toronto will be utilised to anticipate a decent living neighbourhood for him.

Data:

For this project the Foursquare API will be used. A list of neighbourhoods in New York and Toronto is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

- New York neighbourhoods: https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json
- Toronto neighbourhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The data downloaded are the areas situated in New York and Toronto. In addition, their particular directions are consolidated. Just Manhattan neighbourhoods and districts that contain the string "Toronto" are considered. A Foursquare API GET request is sent so as to acquire the surroundings within 500m. The data is formatted utilising the 'Get Dummies' method with the categories of every place. The, mean is calculated grouping all neighbourhoods.

The likenesses will be resolved dependent on the recurrence of the categories found in the areas. These similarities discovered are a solid pointer for a client and can assist him in deciding whether to move in a specific neighbourhood close to Toronto or not.

Methodology:

- **Feature Extraction**

For feature selection, the 'Get Dummies' method is utilised as far as classifications so that the feature can be converted into a number which will be used in our ML model, as the Machine learning algorithm can only take numbers for developing a model.

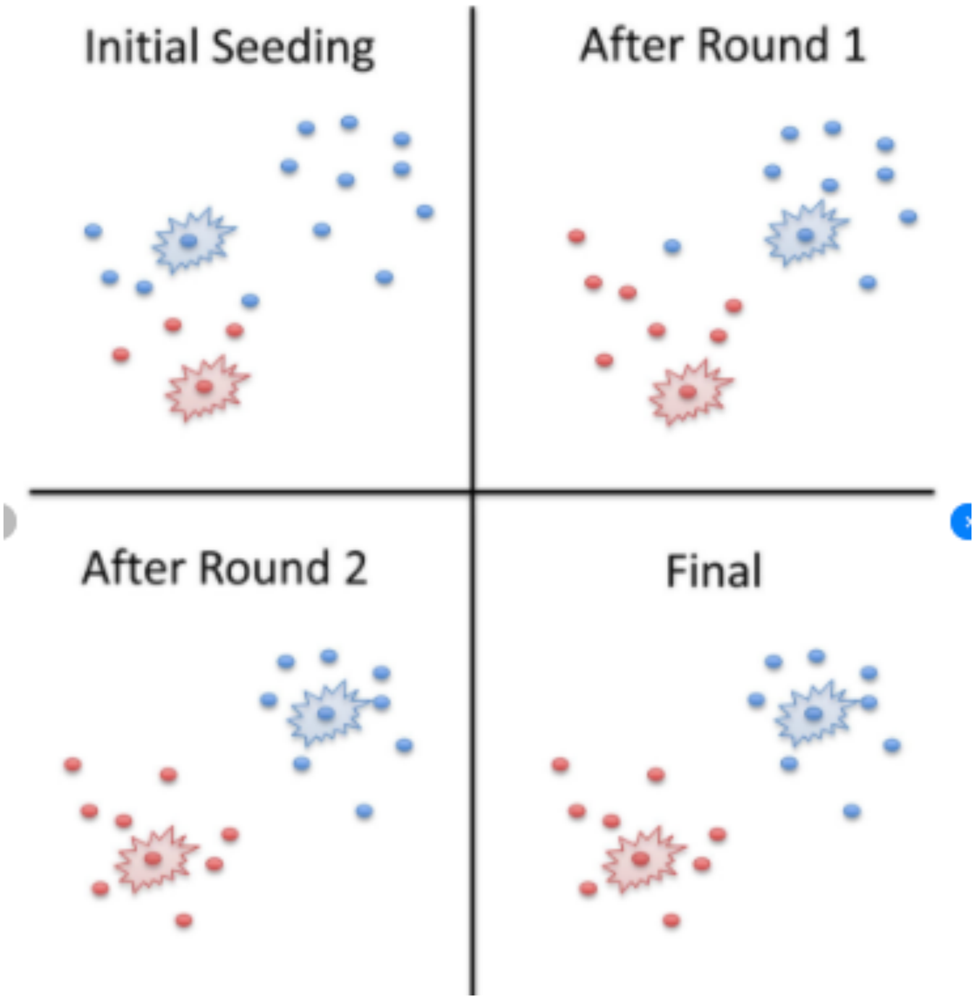
- **Unsupervised Learning**

For unsupervised learning to find similarities between neighbourhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similarity approach to found patterns.

- **K-Means:**

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimise the overlapping of each cluster.

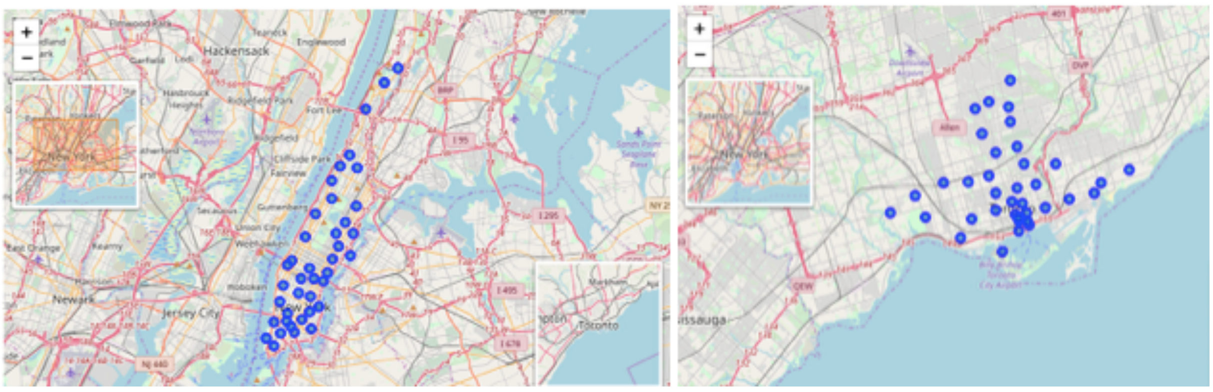
In technical terms its main objective is to reduce the data between the data points in each cluster from the data points of that cluster and increase the distance from other clusters. So that we can differentiate between clusters.



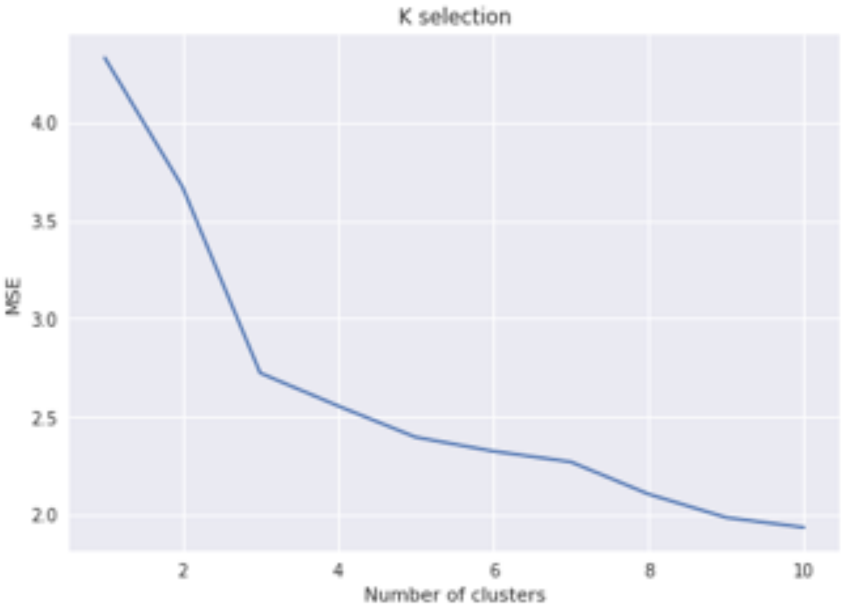
As we can see above in K-means we will introduce centroids in iteration to the data points so as to determine the group of differentiate between them.

- **Results**

Firstly, we will plot geographical to get the location.



Secondly, the cluster algorithm is implemented. For this purpose, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 and iterating until 10. This chart is shown in the image below.

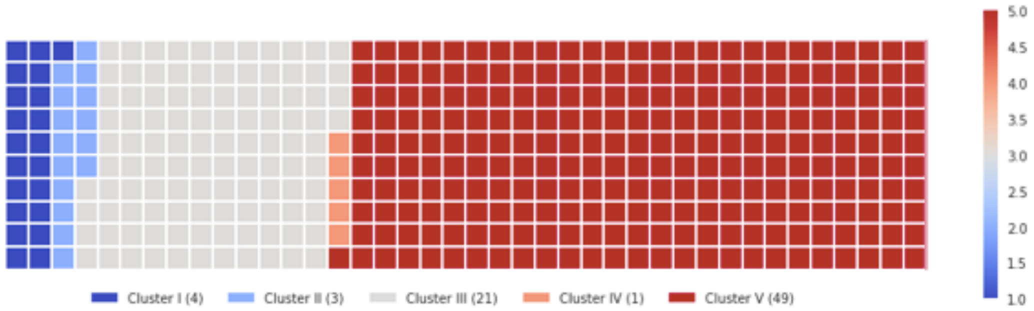


Again, For visualisation, the geographical data is plotted but with colours with each colour representing the cluster for which that neighbourhood belongs. The map is shown below:

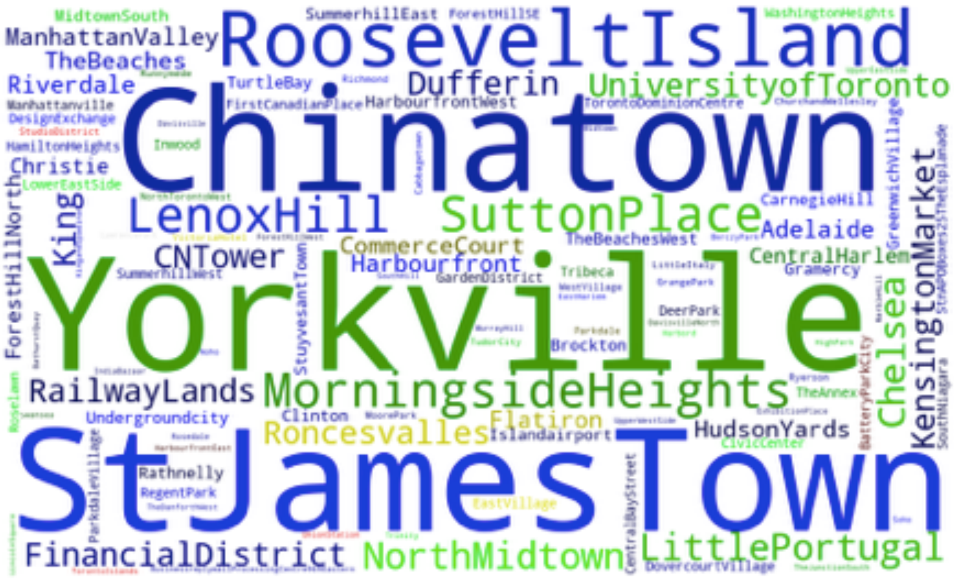
Hackensack
Newark
Jersey City
Mississauga

In this map, it is clear our algorithm is not segmenting the areas. This implies it isn't accurate that geolocation of neighbourhoods is related to the classifications of the neighbourhoods around every area. However, it is possible to see which neighbourhoods inside Manhattan, New York are like the areas in Toronto. Those areas that are similar among them have a place with a similar group. Henceforth, they have a similar shading in the picture above.

In the picture beneath it is discovered the extent of the areas allotted to each cluster. Thus a waffle diagram is executed. There are two significant clusters and two minor clusters. Besides, there is a cluster that has just a single neighbourhood. This area is Lawrence Park from Toronto. This neighbourhood has no similitude with New York City.

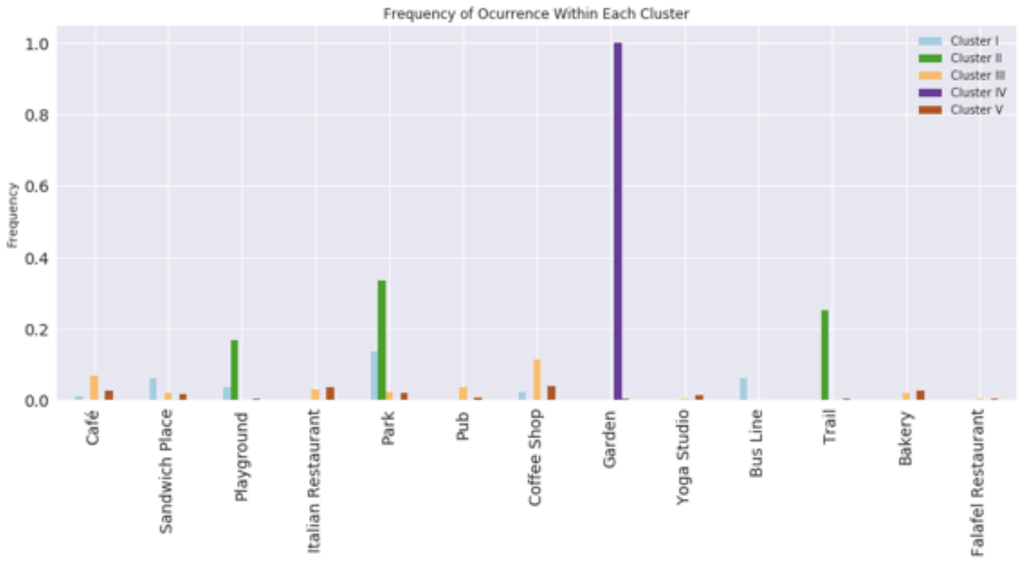


For depicting the name of neighbourhoods, a word cloud is mentioned below, which can quickly portray the true picture. With the help of coloured words, an individual attempting to find a comparative neighbourhood in Toronto can find it by searching for the areas with a similar shading. Here we can see that Chinatown, Roosevelt Island, St James Town, etc have a place are comparative among them. Then again, York Ville, Sutton Place, North Midtown, etc are not the same as those we referenced before however again they are comparable among them.

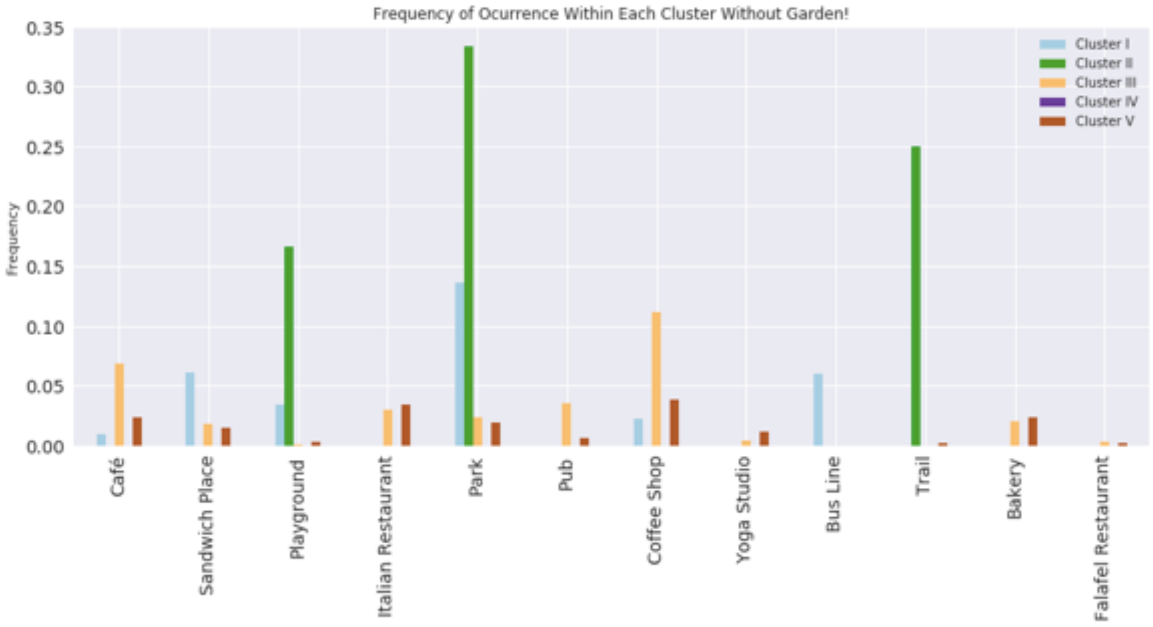


.

The bar chart that is shown below shows the features with higher frequency in the centroids found by the algorithm. In this way we can learn what the algorithm is
finding. The chart depicts a specific classification "garden" with a high recurrence of 1. This class is identified with the IV group, which is the one that has only one
neighbourhood. This group doesn't include additional data. Thus, we can evacuate it and examine different ones.



It can be seen that 1st cluster centres around neighbourhoods that have parks around them, transport lines and sandwich places. Then again 2nd cluster centres around neighbourhoods that have parks, play areas, and trails. 3rd cluster centres around neighbourhoods that have coffeehouses, bars, and Italian cafés around them. The 5th cluster centres around neighbourhoods that have bistros, parks, and bread shops. Please refer the chart below.



- ## Discussion

It is worth taking note that this work is valuable the individuals who live in Manhattan, New York, s in Toronto. Also, there is a cluster with one neighbourhood. In a
result we discovered that this cluster has a recurrence of 1 in garden places. This implies the group isn't dividing data effectively and the centroid is situated in the
specific situation of that area. This area has a high recurrence of parks around. Thus, we can say the Kmeans algorithm is doing an incredible job since there is no
other cluster with comparable neighbourhoods around.

- ## Conclusion

In this project, we had a close look at different neighbourhoods of the two different countries, which includes Manhattan, New York, and the areas close to the centre of Toronto. The data is gathered and the settings around the areas is acquired utilising the Foursquare API. The 'Get dummies method' is utilised for converting categorical information into integers. Then taking into consideration all clusters are divided as per the features.

The K-Means algorithm is used for clustering similar neighbourhoods and the number of centroids are measured by the elbow method to classify each cluster as per features of the data point. Out of 5, there are 2 clusters and 2 minor clusters and one cluster contains just a single neighbourhood that is different from others. The depiction of the cluster is as follows:

Group

I: Neighbourhoods that have parks, transport lines, and sandwich places.

II: Neighbourhoods that have parks, play areas, and trails.

III: Neighbourhoods that have cafés, bars and Italian eateries.

IV: Neighbourhood that has gardens.

V: Neighbourhoods that have cafés, parks, and pastry kitchens.

At long last, anyone who needs to move from Manhattan to Toronto and vice versa can utilize this framework to get a hint regarding what would be the best place for him.