**Battle of the Neighbourhoods**

**Introduction/Business Problem:**

The problem that I am solving lies in the fact that we travel all the time throughout the world in different cities and different countries. In doing that we usually book hotels in advance. Over the years this business has expanded in its influence, but what about the adventure seeking **solo travellers** or the **group of backpackers**who travel without online reservations. There is no instant solution for those. Some hotels can be frustrating, and this could put you off them, sometimes hotels or you can also cut your reservation. All these problems can be solved by this solution**. It tracks your current location and gives you a list of available hotels in close proximity and gives you the direction.** What if we could leave the responsibility to a machine and algorithm? This would save us a lot of time. The hospitality industry is multi-billion-dollar industry and one of the biggest in the world, by implementing this platform, the industry could expand their influence furthermore and appeal to a wider array of consumers

## Data:

There are two primary data sets we will be using for this program.

1. The New York location data in JSON format (As provided previously). The dataset contains all the information about the boroughs and the neighbourhoods present in New York. Apart from the names, the co-ordinates of the neighbourhoods are the primary data that we can use from this dataset.
2. The other part of the data comes from the foursquare API. We use two types of queries for fetching the data from the foursquare API.
   * First type of query is ‘explore’ that is used to fetch the venues present in a 1.5 km radius of the neighbourhood of our target.
   * Second type of query that I am using is “venues”. This query is used to check the details about the venues that are hotels and get those details about them since we are interested in only those venues that are hotels. We check those venues by using their venue id.

#### Machine Learning:

To use most of this data, machine learning will be required to find relevant data clusters. This can be found using trends in ratings, likes and tips. By organising these into neighbourhoods we could identify similar clusters and their hotels.

## User Interface:

Here we will find the data the user will need to personally decide on the hotel. We can show the user details like the location and distance away.

Using columns excluding ***'Neighbourhood Latitude’, ‘Neighbourhood Longitude’, ‘Venue Id', 'Venue Latitude’, ‘Venue Longitude’, ‘Venue Category'***from the **Manhattan hotels** data frame and ***'Rating'*** column from the **manhattan\_hotels\_data** data frame we create the user interface for the Manhattan Hotel data called user\_sees. Like below : |Neighbourhood|Hotel Name |Distance |Rating| |-----------:|:---------:|:-------:|:----| |Chinatown |Hotel 50 Bowery NYC| 214m |8.9| |Chinatown |Crosby Street Hotel| 866m |9.3| |Manhattanville| Aloft Harlem | 1003m|8.1|

## Backend Statistical Analysis

Here we shall cluster the hotels into categories of similar feature. m**anhattan\_hotels\_data** has many columns and KNN will help to organise. The ones we shall use for clustering are ***Like Counter, Tip Counter, Rating*** and we can use these to create the data frame **manhattan\_hotels\_cluster**. This is then used geographically to show if the similarities may also be linked to location. The same is done for staten.

**Result:**

The resulting data frame with the added clusters showed us good insight into which hotels would be recommended to a user of the app.

* 1st Cluster: The like and tips are at a high rate but are not as high as the third cluster
* 2nd Cluster: The number of likes and tips are lower than both the other clusters. The ratings are around average
* 3rd Cluster: The ratings of these hotels are the highest of the three category’s

The same results apply for the staten dataframe

**Discussion:**

We can check the hotel dataframes for Manhattan and Staten Island and see that there are more options of good hotels in Manhattan borough than that in Staten Island. One observation that we can make on the Staten Island dataframe is that some of the hotels do not have Ratings and they are shown in the dataframe as **Not rated yet**. We had to remove those hotels from the dataframe for clustering since Rating is one of our clustering parameters. By checking the Rating and other features like, **Number of Likes** and **Number of Tips** we can straightaway tell that Manhattan has many more good hotels than Staten Island has. Not only that, the number of hotels in Staten Island is lesser than that in Manhattan. This can lead to unreliable reviews.

**Conclusion:**

We have gone through a process of accessing data from the Foursquare API to create our dataframe so that we can show the necessary data to the users of our app. All they need to know is the **Name of the Neighbourhood, Name of the hotel, Rating, Distance from your current location**. Other information that we have collected are used for our own analysis. We have used the venue id for using the **Venues** feature of the Foursquare API. Then we got the Rating, Number of Likes, Dislike Flag, Number of tips etc. Since none of the hotels were disliked that column was not used for clustering. Other three columns were used for clustering. Since clustering was used for grouping similarly featured hotels distinguished them. We also learnt from the foursquare data that Manhattan has more and better options than there is in Staten Island.