Introduction

For this project I am creating a hypothetical client which is interested in opening a bar in the city of Bristol. In order to his business become successful or even a void failure at first, it is important to know the most suitable location to start such business. Areas with the highest concentration of bars might indicate the perfect spot as people might commute to such locations looking for those environments.

Business Problem

The question to answer is: Which is the most suitable location in Bristol to invest in a Bar? Well with the help of unsupervised machine learning algorithm known as k- mean clustering, we will try to answer this question and provide suggestions to our client regarding the best location.

Target Audience

The client who wants the most suitable location to open his bar.

Data

The data bellow was used to solve this problem:

1. A list of neighbourhoods in Bristol with its corresponding coordinates.
2. List of all venue existent in each neighbourhood, this data will be fetched from foursquare API.

Source of data

1. The neighbourhood data was scrapped via Wikipedia: https://en.wikipedia.org/wiki/BS_postcode_area
2. Latitude and longitude data were acquired downloaded and filtered according the purpose of this project: https://www.doogal.co.uk/postcodedownloads.php
3. Foursquare API was used to get all venues related to each neighbourhood.

Methodology

Both the latitude and longitude of each neighbourhood were acquired via download. This data contained postcode of all England, so for the sake of this project Excel was used to filtered the data which then was uploaded into google drive so that could be accessible directly from jupyter notebook. The neighbourhoods were extracted from Wikipedia, this was accomplished using the pandas library which made much easier the process compared to another known technics such as beautifulsoup. After merging both data in the notebook, the foursaquare API was used to pull a list of 100 venues for each neighbourhood within a radius of 500 m. The API provided the names, categories, latitudes and longitudes of those venues. Then I grouped the venues by neighbourhood and took the mean frequency of each type venue. At this point I narrowed down the investigation to the occurrence of Bars in each neighbourhood, this was then fed to K-mean clustering algorithm. The neighbourhood were clustered into 3 groups based on their occurrence of "Bars".
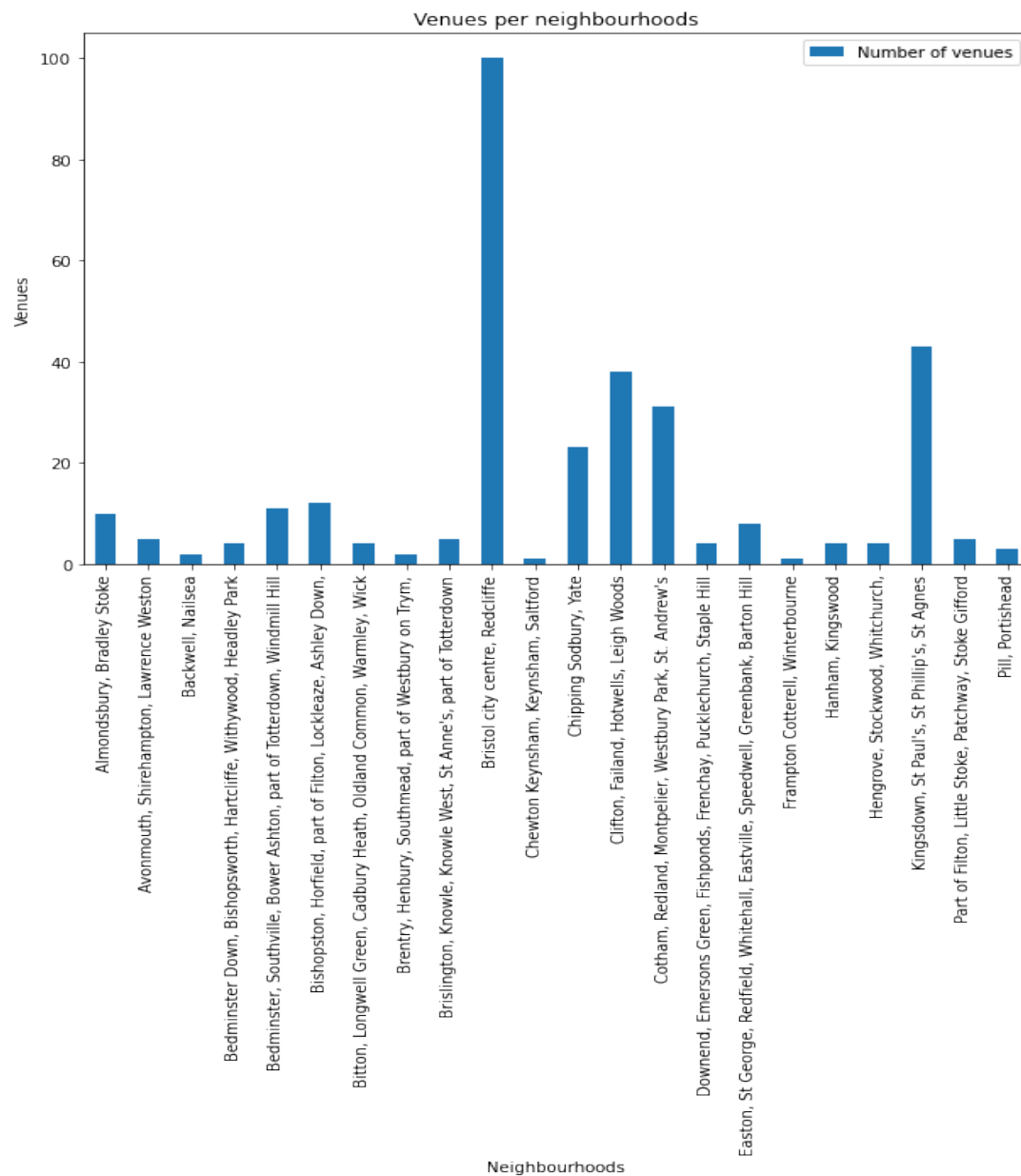
Results

Fig. 1 – Number of venues per neighbourhood

From the results obtained in fig. 1 the second and third neighborhood with the highest numbers of venues are the ones with the most bars.
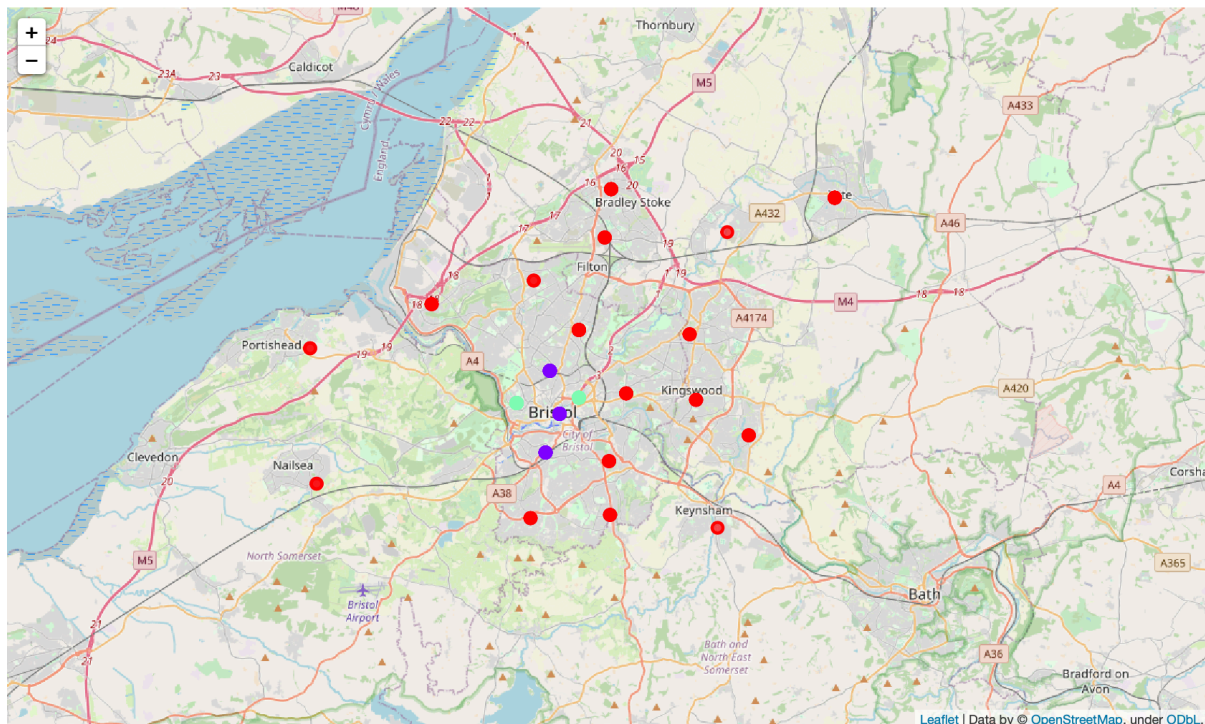
Fig. 2 – Cluster 0 is red, cluster 1 is purple and cluster 2 is green.

The results from the clustering algorithm were divided into 3 cluster based on the occurrence of Bars in each neighbourhood. In addition, each cluster were labelled as the following:

- Cluster 0: With very low occurrence of Bars
- Cluster 1: With medium occurrence of Bars
- Cluster 2: With High occurrence of Bars

Discussion

The cluster with the most bars is the ones in green which account with at least five establishments per neighbourhood. As this neighbourhoods are located close to the city center, it might indicate that most people concentrate in these areas to spend their leisure time. Therefore, cluster 2 represent the neighbourhoods most suitable to invest in a bar. Also, neighbours Clifton, Failand, Hotwells and Leigh which belong to cluster 2 could be further explored for better consideration as they are a bit decentralized compared to the other. You might argue that the competition will likely kill the investment, but if the business idea is promising clients will be attracted. However, the neighbourhoods in cluster 1 could be considerate due to high rental prices in areas around the city center (cluster 2).

Conclusion

This project gives us a glance on how machine learning can be used to create recommendations and help individuals make decisions. From the whole analyse cluster 2 indicate the neighbourhoods of choice to implement the business idea, and in the future additional data such as rent prices can be added for a better recommendation.