1. **Problem Description**

In 2020, an area of Bristol has been named one of the best places to live in the UK. Having studied in a university in Bristol myself, I come to realize that the city has more than a picturesque harborside and Clifton suspension bridge, it is also number of activities that one can do to relax with friends that make the city one of the best.

The Bristol and Bath area houses over 550,000 people according to data in 2019, with at least 4 universities in the area, the demand for nightlife is high. Nevertheless, Bristol has been named the city with the third-highest pubs' density in the UK following Portsmouth and Liverpool, with 10 pubs per square mile, which means the competition is also very strong in the area. Therefore, in this report, we will be investigating on where the best location would be to set up our pub, by utilizing the machine learning method of K-means clustering.

1. **Data Resources and Preprocessing**

There are three different types of data we need to obtain to conduct our analysis. First, we collected the data on neighbourhoods in the Bristol and Bath area, as well as their postcode. Both sets of information collected from their respective Wikipedia page:

1. Bristol: <https://en.wikipedia.org/wiki/BS_postcode_area>
2. Bath: <https://en.wikipedia.org/wiki/BA_postcode_area>

Secondly, we need the information of each neighbourhood’s latitude and longitude position to plot the location on the map, we will utilize the geocoder Python library to collect these data. Then finally, we would need to collect the data regarding the venue types of each neighbourhood areas, these data will be collected from the foursquare API.

Once we have collected the data from Wikipedia, we will first need to identify if there are any missing data. In fact, upon investigation, some postal areas have not been assign a neighbourhood or local authority areas, these postal codes do not relate to a specific geographical area and are therefore used for postal purposes only. Thus, we will be dropping these codes and from our list of neighbourhoods.

* 1. Initial data inspection

To ensure the data we collected was valid, we have plotted the locations on a map using the folium library with the latitude and latitude we collected. Errors can occur at this stage due to the fact that even in the same country, some towns might have similar names, or the mapping library might give out the wrong data. Fortunately, the areas are in fact located in around areas where we expected, and none of them are out of the scope of the Bristol and Bath area.

Secondly, we looked at the aggregate number of each venues for the area we are interested. As anticipated, pubs are by far the most frequent venue in the area, followed by coffee shops and hotels. While some of the venue data seems misleading, such as where there are only one mobile phone shop in all these area in combine, we will make the assumption that the data is a good approximation of what we need. In fact, there are no other ways for us to obtain better data than what is already in the foursquare API that are publicly available.

Figure 1 – Neighborhoods in the Bristol and Bath Area

Map

Description automatically generated

Figure 2 – Total Numbers of Venue in the Area

Chart, histogram

Description automatically generated

1. Methodology
   1. Unsupervised Learning - K-means Clustering

In this project, we apply the K-means clustering unsupervised learning algorithm to cluster the venues based on their latitude and longitude locations. This allow us to identify neighborhoods with similar traits in terms of venues type and gain an insight to decide which neighborhood(s) are the best for setting up a pub.

To determine the value of K in out k-means clustering algorithm, we will be using the Silhouette Analysis method. After running the analysis, we have found that the optimal number of clusters is five, given that it has the largest silhouette score.

Figure 3 – Optimal number of clusters is five

Table

Description automatically generated with medium confidence

1. Results

Upon diving the neighborhoods into five different clusters, we plotted it onto a folium map. As shown in figure 4 below, most of the areas are labelled in purple, and the rest are mostly in green, followed by one to two areas labeled red and blue.

Map

Description automatically generated

Figure 5 to 10 – Clustering results

Cluster 0

Table

Description automatically generated

Cluster 1

Table

Description automatically generated

(\* only showing the top ten rows)

Cluster 2

A picture containing table

Description automatically generated

Cluster 3A screenshot of a computer

Description automatically generated with medium confidence

Cluster 4

Table

Description automatically generated

1. Discussion

From the clusters above, we can see that cluster 0 consist of two locations that has a high number Dog Runs and Tracks, while cluster 1, the cluster with the largest number of neighborhoods has a mix of areas and top venues. Cluster 2 consist of one area and that is the location BS36, where in the location there is the only sport shops in the Bristol and Bath area. For Cluster 3, this is a cluster with locations that are populated with pubs and other hospitality venues such as bars and hotel. And finally, cluster 4, are areas where that hosts construction and landscaping venues and Zoo exhibit.

While the clustering algorithm provided us with a classification of the areas, it does not implicitly answer our question that which area(s) are best for setting up our Pub. Nevertheless, the strong characteristic of cluster 3, with all the areas consist of high numbers of pubs, it would be a reasonable strategy to start our new pub venue in the area. However, the high number of pubs also suggest that the area will be high level of competition, and better alternative might be to set up a pub in the areas in cluster 1. While in some areas in cluster 1 there are high numbers of pubs, and hospitality venues similar to cluster 3, there are areas where pubs are amongst the lesser common area, and given the similar trait within the cluster, it could be a good idea to set up a pub in areas in cluster 1 where pubs are amongst the less popular venues, such as BS11 shown in the figure above, where pub is only the fourth most popular venue.