Clustering Bristol Neighbourhoods with Anti-Social Behaviour and Noise Problems

Visit Jupyter Notebook on GitHub for data and machine learning code.

Haseeb Rahman
IBM Data Science Professional Certificate
Applied Data Science Capstone Project

Introduction/Business Problem:

Bristol is the largest city in the South West of England with a population of over 460,000 and 34 neighbourhoods. Bristol is one of the most popular tourist destinations whose economy is built on the creative media, electronics, aerospace industries and city centre docks as symbols of heritage and culture. Bristol has a unique blend of pubs, restaurants, and Cafés that are just as famous for street food as fine dining.

This project attempts to cluster the neighbourhoods in Bristol based on the venues around their centroids and identify neighbourhoods that consider anti-social behaviour and noise from pubs, clubs and entertainment are problems locally. The analysis provided by this project can be useful for city planners, security agencies, charities, business investors, house buyers, etc., to understand the neighbourhoods that share common characteristics for planning and effective decision making.

Data:

The data required to answer the question was extracted from multiple sources.

1. A list of neighbourhoods was scrapped from Wikipedia by using html5lib library.

	Rank	Ward	Population	Address
0	1	Lawrence Hill	18942	Lawrence Hill, Bristol, United Kingdom
1	2	Ashley	16225	Ashley, Bristol, United Kingdom
2	3	Cabot	15940	Cabot, Bristol, United Kingdom
3	4	Bishopston	13871	Bishopston, Bristol, United Kingdom
4	5	Easton	13541	Easton, Bristol, United Kingdom

2. The information on Anti-Social Behaviour and noise indicators were downloaded from Open Data Bristol, an online data source maintained by Bristol City Council. The excel file was uploaded using pandas.

	Ward	Noise Problem	Anti-Social Behaviour
0	Ashley	29.391346	42.393488
1	Avonmouth	12.609684	44.049799
2	Bedminster	17.130083	35.448836
3	Bishopston	16.436375	11.607648
4	Bishopsworth	10.415815	32.944118

3. Neighbourhoods Location coordinates were added to the main data-frame using geopy library.

	Rank	Ward	Population	Address	Coordinates	latitude	longitude
0	1	Lawrence Hill	18942	Lawrence Hill, Bristol, United Kingdom	(Lawrence Hill, Church Road, Newtown, Lawrence...	51.458580	-2.564170
1	2	Ashley	16225	Ashley, Bristol, United Kingdom	(Ashley Down, Bristol, City of Bristol, South ...	51.478287	-2.585152
2	3	Cabot	15940	Cabot, Bristol, United Kingdom	(Cabot Circus, Broadmead, Bristol, City of Bri...	51.458382	-2.585432
3	4	Bishopston	13871	Bishopston, Bristol, United Kingdom	(Bishopston, Bristol, City of Bristol, South W...	51.478305	-2.595453
4	5	Easton	13541	Easton, Bristol, United Kingdom	(Easton, Bristol, City of Bristol, South West ...	51.464703	-2.567684

4. Venues data was imported from Foursquare location data services using API. The data consisted of venue names, categories and their coordinates.

	name	categories	lat	lng
0	Small Street Espresso	Coffee Shop	51.454766	-2.594400
1	Swoon	Ice Cream Shop	51.452660	-2.600179
2	The Christmas Steps	Bar	51.455979	-2.596378
3	Small Bar	Pub	51.451791	-2.594948
4	Eatchu	Japanese Restaurant	51.453325	-2.594141

Methodology:

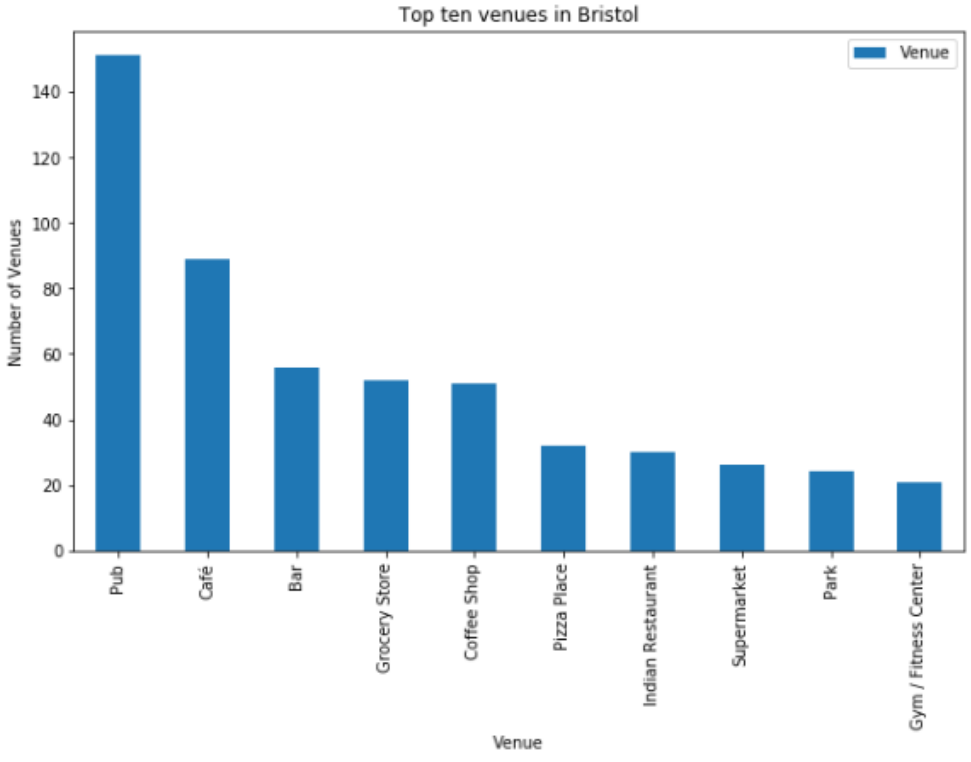
Foursquare location data services were used to explore a maximum of 100 venues within a 1000 meter radius from the centre point of each neighbourhood.

Then an unsupervised machine learning technique called k-means clustering was applied to construct clusters with similar venue locations in each cluster. The best 'k' for the k-means clustering technique was selected by performing 'Silhouette Coefficient' evaluation method on the data frame.

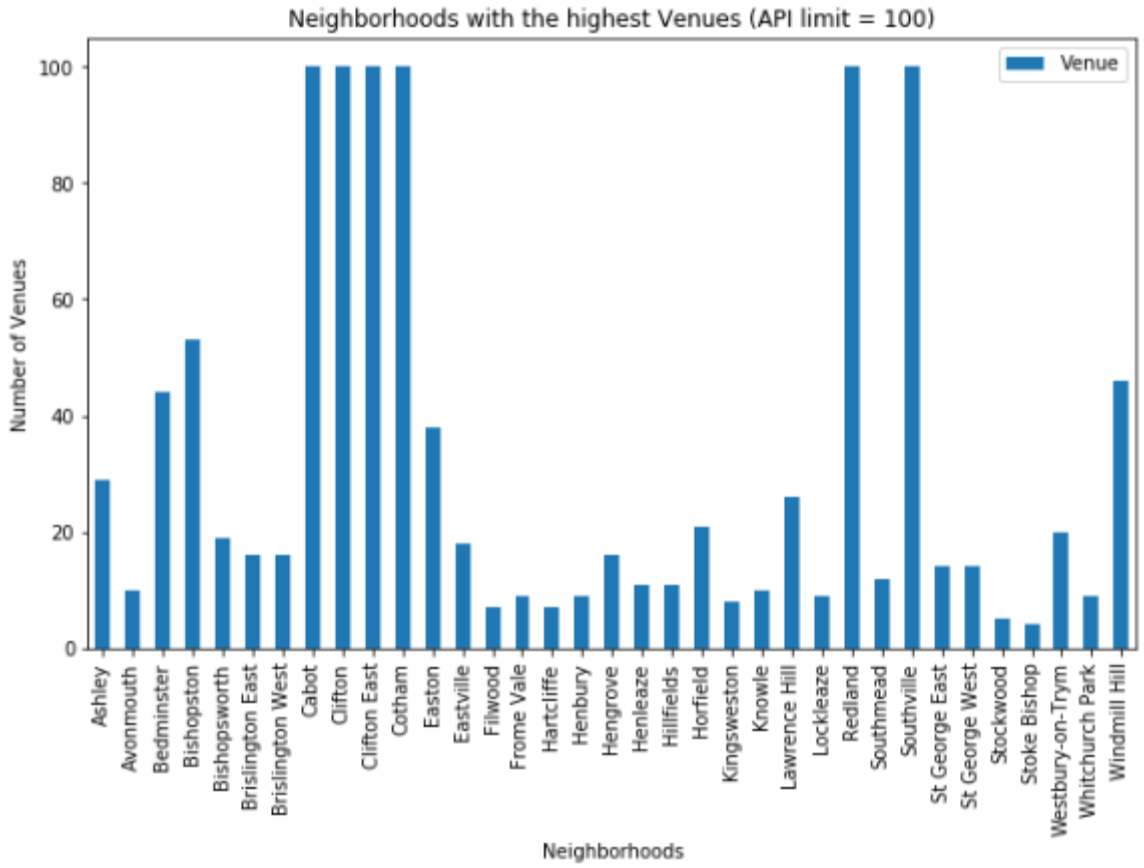
The answers to the project questions were generated by evaluating the mean values of the responses given by the people in the relevant cluster. the response was collected by 'Open Data Bristol'.

Exploratory Analysis:

The top 10 venues in Bristol with Pubs, Cafes and Bars leading the numbers.

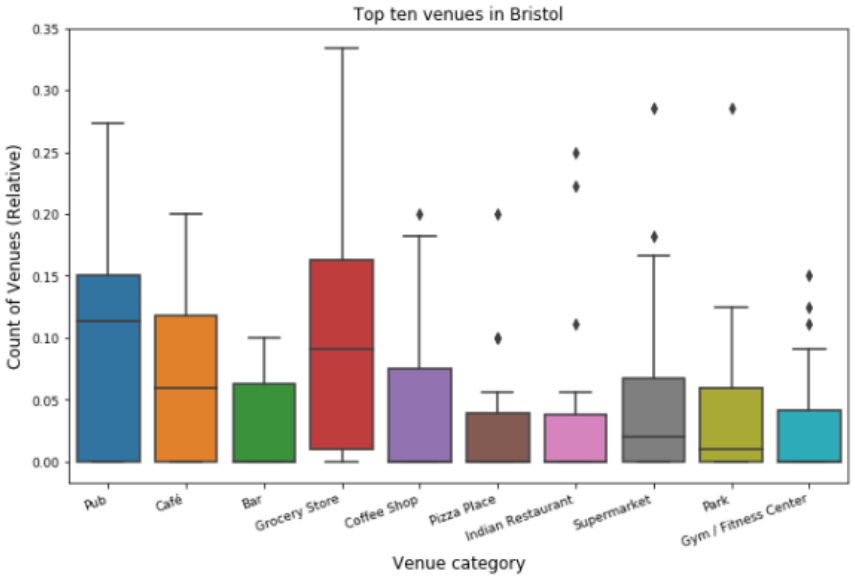


The neighbourhoods with the highest number of venues, the data is restricted by the maximum limit of 100 venues.



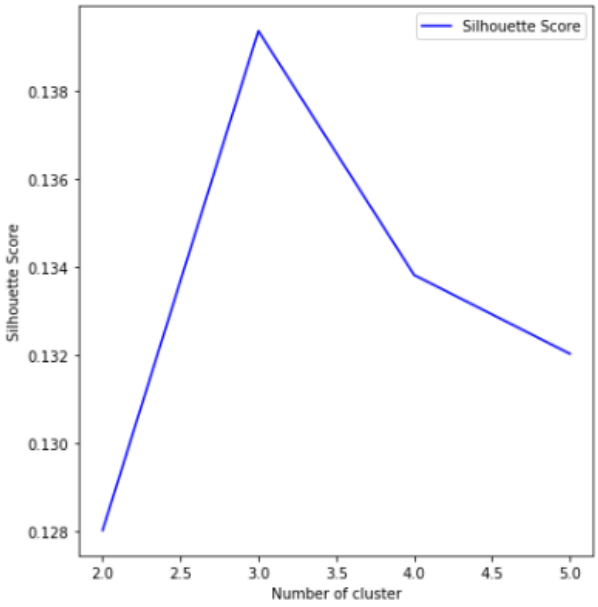
Data Preparation:

At this stage, onehot technique was used to normalize the data required for processing k-means clustering technique. The scaled diagram below reflects relative spreads and outliers of each category of the top ten venues.

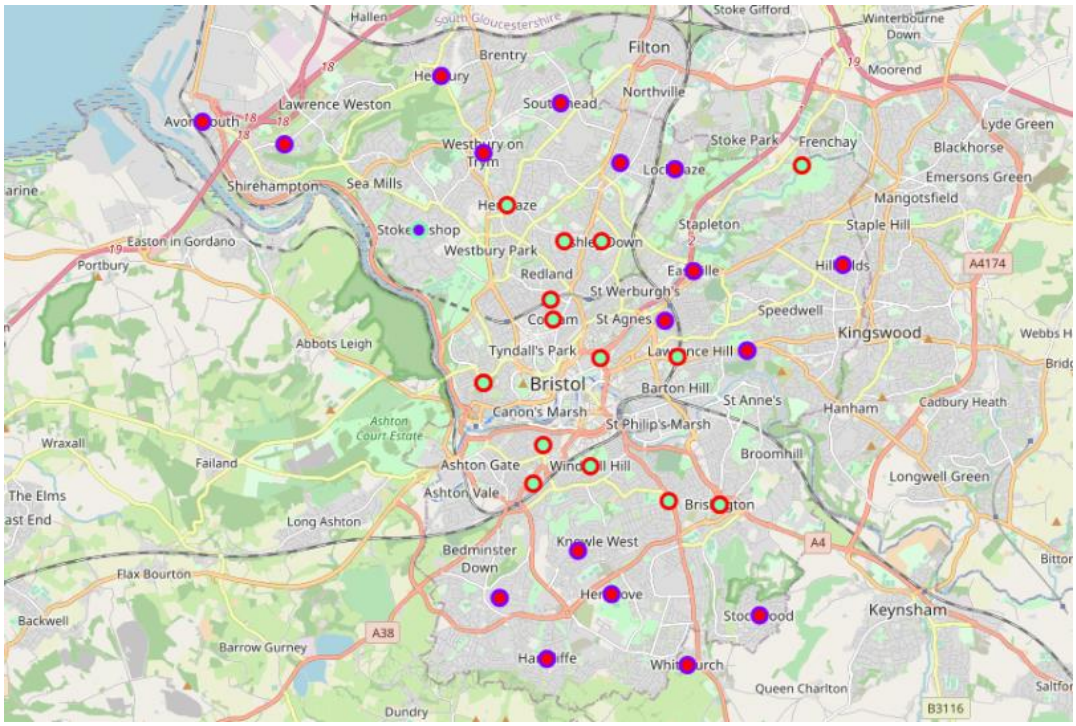


Clustering:

We performed k-means machine learning technique to cluster the neighbourhoods. The optimum ‘k’ was 3 where ‘Silhouette Score’ was the highest.



Each cluster was indicated with a colour coded circle. The cluster 0 was denoted by a red circle filled with green, the cluster 1 was denoted by a dark blue circle filled with red, while cluster 2 was denoted by sky blue circle filled with blue.



1. Cluster 0: the cluster consisted of 16 neighbourhoods dominated by pubs, cafes and coffee shops. The city centre and its environs are the main areas included in this cluster.

2. Cluster 1: the cluster consisted of 18 neighbourhoods dominated by grocery stores and a variety of other locations.
3. Cluster 2: the cluster consisted of just one location with grocery stores and Indian restaurant.

Discussions/Results:

1. ***Noise from pubs, clubs and entertainment is a problem locally***: The mean value of the observations in cluster 0 was 17.23 compared to 11.72 and 9.81 of cluster 1 and cluster 2 respectively. The cluster 0 was dominated by pubs and also reflected the highest mean value of observations of the people who considered the noise was a problem. Thus, we could say that the people living in the cluster with the highest number of pubs were more concerned with the noise compared with the people in other clusters.
2. ***Anti-social behaviour is a problem locally***: The mean value of the observations in cluster 0 was 34.33 compared to 32.33 and 11.62 of cluster 1 and cluster 2 respectively. The people in cluster 0 were slightly more concerned with the anti-social behaviour compared to other areas, the cluster 0 was dominated by pubs. However, the anti-social behaviour problem seems to be unrelated to the pubs as people in most of the neighbourhoods consider this a common problem.

Conclusion:

The venue data provided by Foursquare location services is subjective and based on the observations of its users. However, this Foursquare data is the most accurate available at the moment and can provide accurate insights if used in combination with other relevant sources.

---xxx---