Capstone Project - The Battle of the Boroughs

Richard Dudbridge

Code Jupyter Notebook

1 Introduction

1.1 Background

2020 has deprived the world of many things, including the UK's most popular and time-honoured social venue: the pub. In an effort to achieve some escapism from the shackles of 2020, this report will explore the best boroughs of London to open a new pub.

With pubs closing in the UK at a rate of one every 12 hours, it's very important to make sure that the location chosen gives every chance for the success of the pub [1]. As such, a wide range of factors have been considered and statitisical analysis undertaken to understand the ideal borough to open a new pub in London.

1.2 Problem Statement

When choosing the ideal location for a new pub, there are so many factors to consider and so much data that the decision is very difficult to make.

1.3 Interested Stakeholders

This report will provide a framework for stakeholders interested in opening a new pub in London an objective way to decide the ideal location.

2 Data Acquisition and Preparation

2.1 Data Sources

Data has been sources from the following locations:

Table 1. Data sources

Data Title	Data Description	Data Source
London Age, income, and business survival rates	CSV file containing many different features for the demographics and profile of London boroughs.	London Datastore csv [1]
London pubs	CSV files containing lists of all the London pubs with various descriptive features for each pub.	London Datastore csv and GetTheData csv [1] [1]

London rental costs	CSV file containing rental costs for London boroughs.	London Datastore csv [1]
London venues	API that connects to a database of London venues with many different descriptive features for each venue.	FourSquare API [1]
London borough centre point coordinates	Wikipedia page containing the coordinates for all London boroughs	Wikipedia London Boroughs [1]
London borough boundary data	CSV file containing GIS boundary files for London	London Datastore csv [1]

2.2 Data Preparation

All data has been combined into one master data frame which has been used for creating plots and statistical modelling. For details of data preparation please see the GitHub code.

3 Methodology

3.1 Pub density

Pub density has been calculated by dividing the borough area by the number of pubs. A blank map has been created using Geopandas and Matplotlib. Shape files from the London Datastore were used to give the Borough shapes which were then layered on top of a blank map of London.

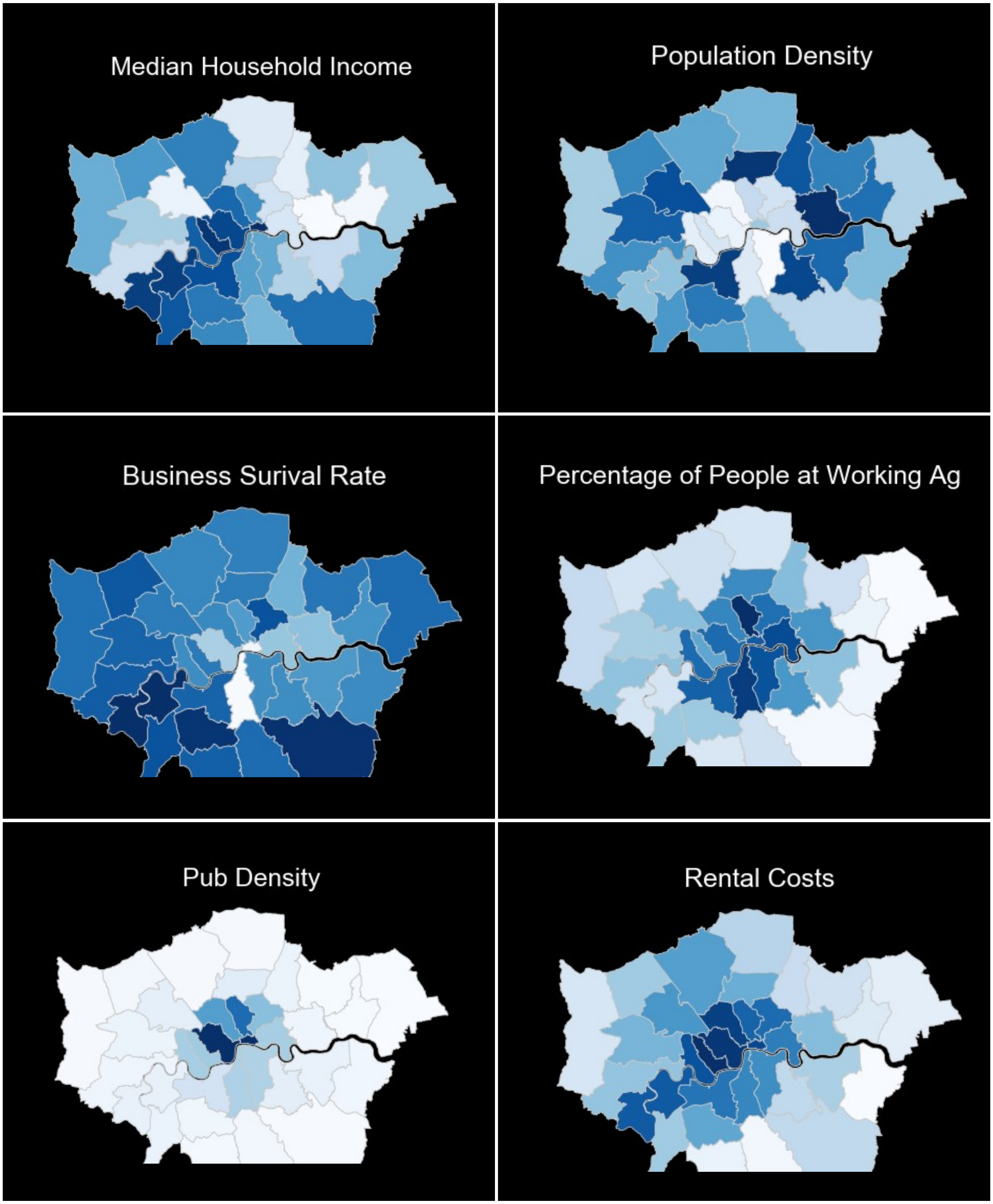
Westminster has a pub density of 80 per km which is 4 times larger than the next closest borough (City of London has 20 pubs per km). When Westminster's pub density is plotted on a choropleth map it is so large that the disparity between other boroughs cannot be seen. Therefore, the pub density of Westminster is temporarily set to that of the next largest borough, City of London.

3.2 Choropleth Maps

To visualise the data, a series of choropleth maps has been plotted. A blank map of London boroughs was produced using Geopandas. Six maps were initially plotted and can be seen below in Figure 1.

Figure 1 shows there is some clear patterns emerging in the borough location. For instance, the % of people at working age is comparatively very high in the central boroughs. However, a single view of these patterns would make the trend much clearer. Therefore, in the next section, a decision matrix is formed to create a single score for each borough which will be plotted on a choropleth map.

Figure 1. Choropleth maps of London boroughs for chosen features



Median Household Income

Population Density

Business Surival Rate

Percentage of People at Working Ag

Pub Density

Rental Costs

3.3 Creation of Decision Matrix

To combine all of the independent variables on to one map, a decision matrix has been created. All data was normalized using the following formula:

$$\mathbf{X_{new} = (X - X_{min}) / (X_{max} - X_{min})}$$

Rental cost has been made negative as high rent negatively impacts new business prospects. A weighting was then applied to each variable to give importance to each variable. For instance, business survival rate was seen as the most important variable, therefore a weighting of ‘4’ was given. Conversely, pub density was seen as the least important variable and was therefore given a weighting of ‘0.5’. This weighting is subjective and can be varied depending on what stakeholders believe are the most important variables.

The results of each variable were then summarised to give a ‘Pub Score Total’.

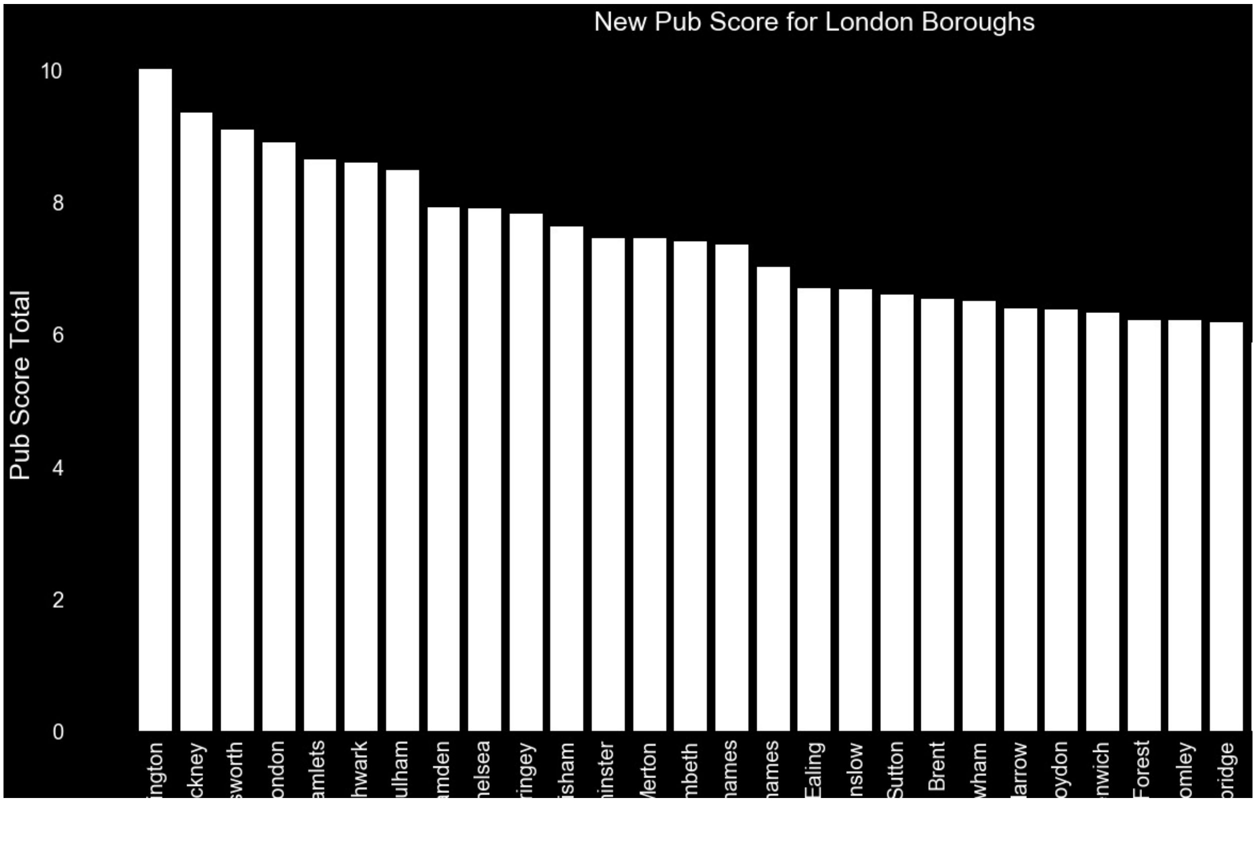
Finally, all ‘Pub Total Scores’ were normalized so that each score was between 0 and 10.

The decision matrix method is summarised below:

1. Normalize all independent variables between 0 and 1
2. Change any negatively correlated variables to negative
3. Weight all variables to give importance
4. Sum all variables together to give ‘Pub Score Total’ for each borough
5. Normalize ‘Pub Score Total’ so that each score is between 0 and 10.

The results of the decision matrix can be seen in the bar chart below

Figure 2. New Pub Score by Borough

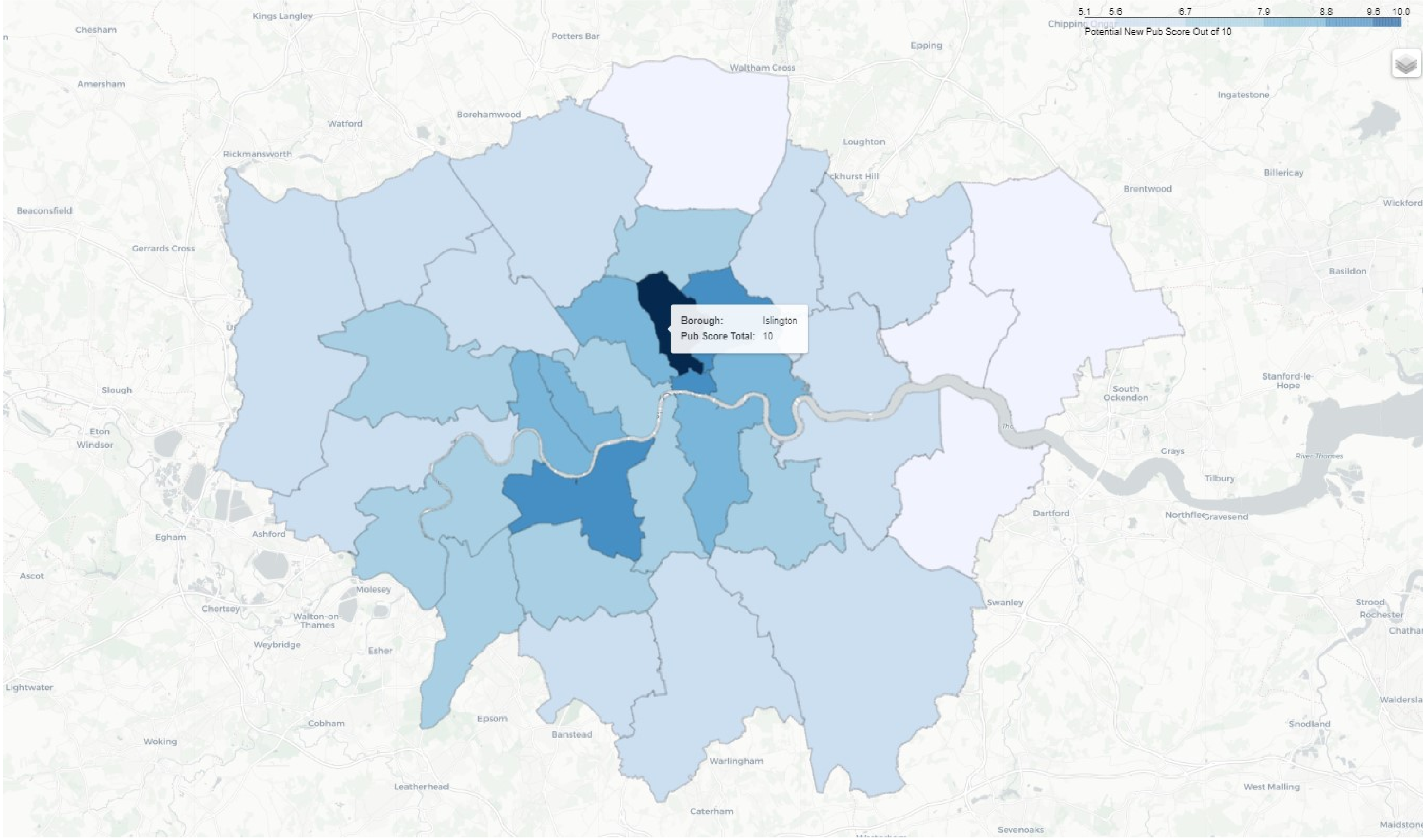


The results of the decision matrix were then plotted on a choropleth map. However, this time Folium was used instead of Geopandas so that the map was interactive. Figure 3 below shows a screenshot of the map. When the map is viewed via Jupyter Notebook, you can hover the cursor over each borough and each borough becomes highlighted and shows the borough pub score from the decision matrix. The basic method for creating the interactive Folium map is outlined below:

1. Master data frame is updated to include pub score data
2. Create blank interactive map of London using Folium
3. Create scale for map
4. Add choropleth layer to the map that shows the pub score of each borough through the intensity of the colour
5. Add highlight tooltip to the map

The results of the interactive choropleth map can be seen below in figure .3.

Figure 3. Interactive Choropleth Map of Pub Score by Borough



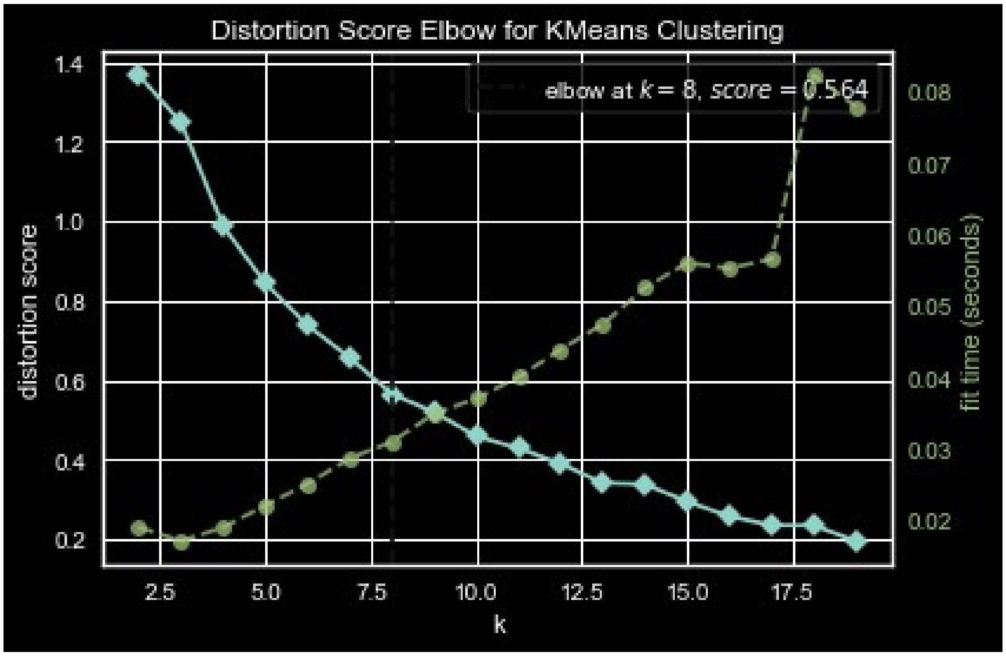
3.4 K Means clustering

To further explore which borough is most suitable for a new pub, the boroughs have been clustered together based on venue types using the machine learning algorithm, K-means clustering. K- means groups similar Boroughs together in k number of clusters. The aim is to see which clusters currently contain pubs as the most common venue.

The Foursquare API has been used to extract venue data for all boroughs. The process for clustering the boroughs by venue is outlined below.

1. Generate unique Foursquare URL using personal Foursquare account. This URL is limited to 100 venues per borough as the Foursquare account used is not a premium account.
2. Make get request to Foursquare API using unique URL. This produces a json file of venue data.
3. Run loop on json file to filter only the useful data.
4. Convert json file to data frame.
5. Use one hot encoding to make all data numerical so it can used for K-means
6. Run K-means clustering on numerical data
7. Validate K number using K-Elbow visualiser. K- Elbow visualiser plots a line graph which shows that increasing K centroids decreases distortion (sum of squares). This is expected behaviour as the more centre points there are, the closer each data point is to a centre point. The visualiser also shows that increasing centre points increases algorithm time. The ideal K point is the cross section between time to run algorithm and distortion. A K values of 10 is given as the optimal K value.

Figure 4 . K Elbow Visualiser



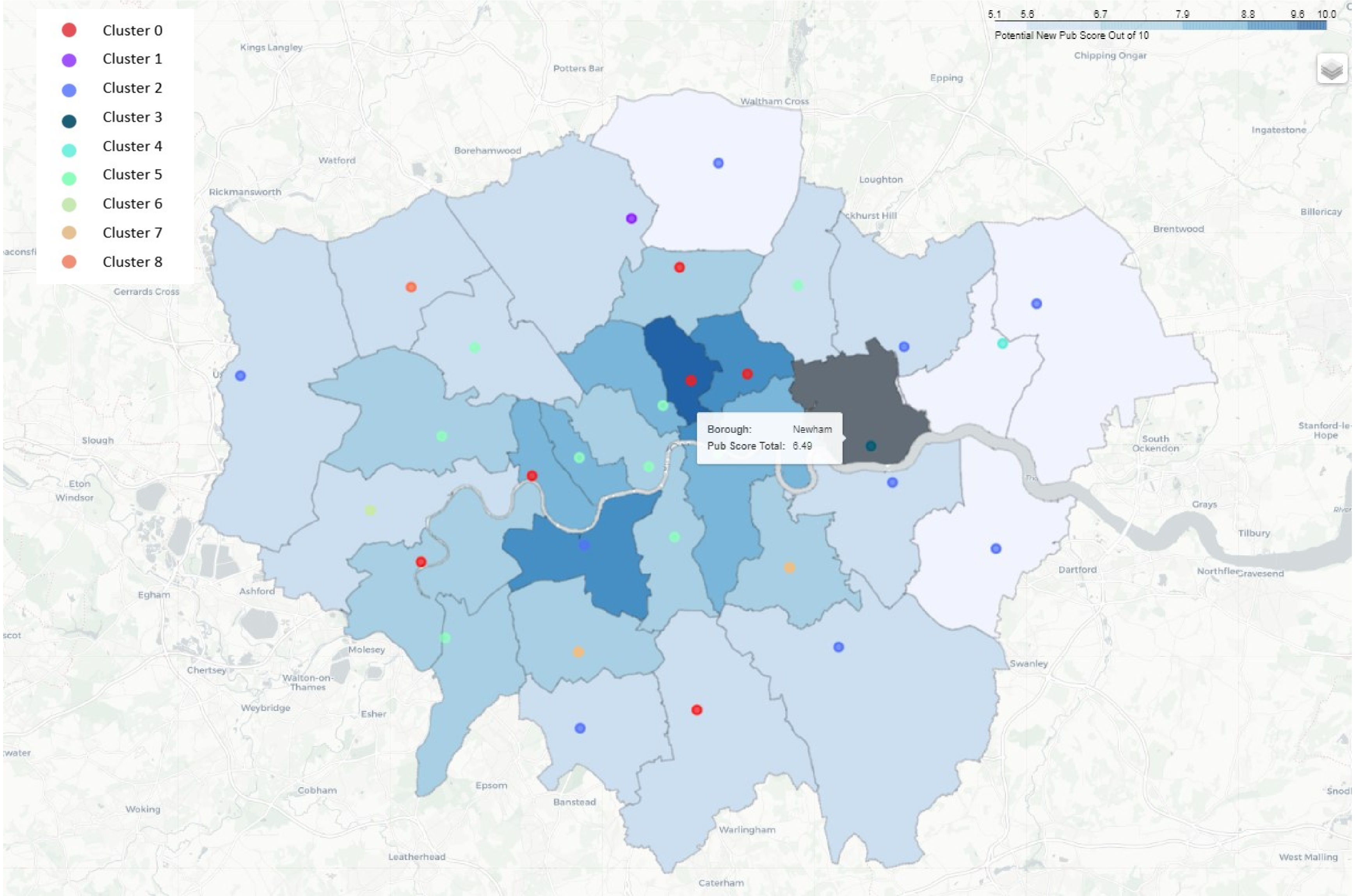
8. Data frame organised to list top 10 venues. K cluster labels added to data frame. See table below for the top 5 venues from cluster 0 boroughs. The top venue for all these boroughs is pub (apart from Haringey).

Table 2. Top 5 venues from cluster boroughs

Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Croydon	0	Pub	Coffee Shop	Portuguese Restaurant	Gym / Fitness Centre	Donut Shop
Hackney	0	Pub	Coffee Shop	Bakery	Brewery	Cocktail Bar
Hammersmith and Fulham	0	Pub	Café	Hotel	Indian Restaurant	Italian Restaurant
Haringey	0	Café	Pub	Turkish Restaurant	Fast Food Restaurant	Movie Theatre
Islington	0	Pub	Cocktail Bar	Music Venue	Burger Joint	Ice Cream Shop
Richmond upon Thames	0	Pub	Coffee Shop	Italian Restaurant	Pharmacy	Grocery Store

9. The clusters were then plotted as a layer on the interactive choropleth map to see if there is any correlation between clusters and the pub scores generated in the previous section. The results of this plot can be seen below.

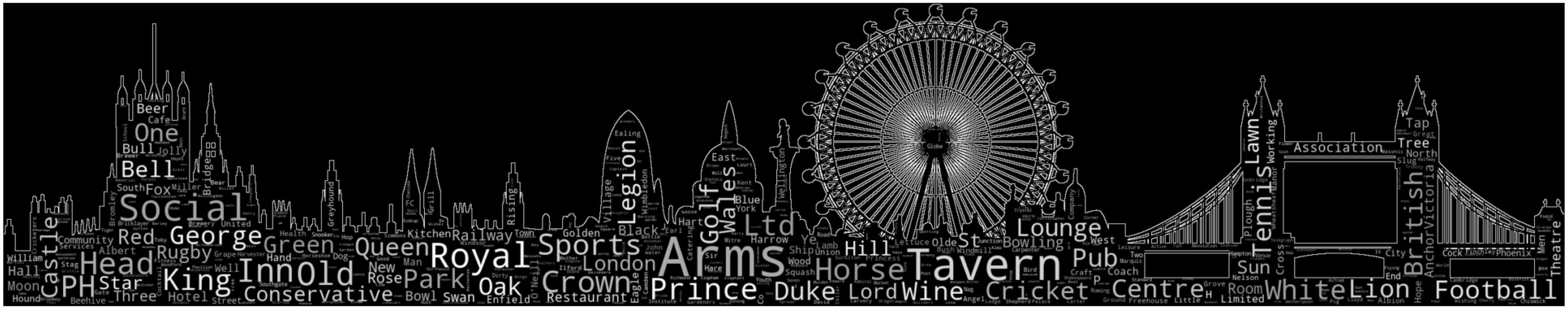
Figure 5. Interactive Choropleth Map of Pub Score by Borough Plus K-Means Labels of Boroughs Clustered by Venue Types



3.5 Word Map of London Pub Names

Finally, to help inspire a name for the new pub, a word has been created showing the most common pub names in London, in the shape of the London skyline. The pub names of London were extracted into a single list and passed through the word cloud using a mask png in the shape of the London skyline

Figure 6. Word map of London Skyline using London Pub Names



4 Results

4.1 Results and Discussion

Figure 5, the interactive Choropleth Map of Pub Score by Borough, shows a clear pattern of boroughs scoring highly in central and south west London. This indicates that newly opened pubs would perform well in these areas. This makes sense as these areas are the youngest, most affluent, and most densely populated areas of London.

The borough that scored most highly was Islington, which achieved a score of 10/10. This is because Islington has high scores for all the independent variables considered in the decision matrix, including a top score for population density, and working age %. Islington is a residential area very popular with young affluent people, which is the target demographic for pubs. Islington is also one of the most restaurant-dense areas in the UK, suggesting it is an area with good capacity for supporting new pubs. Finally, if these stats weren't enough to convince you that Islington is a top borough to open a pub, then it's worth noting that Islington is home to Europe's 3rd longest escalator as well as the past Labour leader, Jeremy Corbyn.

Hackney, Wandsworth, City of London, and Tower Hamlets also featured in the top 5 pub scores. These boroughs also have famously bustling restaurant and pub scenes and would be well suited to a new pub. Though it should be noted that the City of London has very expensive rent and low business survival rate. Moreover, given that less people will be travelling to central London as a result of Covid-19, it would be wise to avoid opening a pub in a central location.

The cluster that consistently has pubs as the top venue is cluster 0. Cluster 0 has some cross over with the boroughs that have the highest pub score, including Islington and Hackney. This strengthens the argument that Islington and Hackney are the best boroughs to open a new pub.

Finally, the London pub word cloud shows that the most common pub names include Tavern, Arms and Royal. If you wanted use a very generic name for the new pub then why not go for the 'Royal Tavern Arms'.

4.2 Recommendations

Given the results of this report, the best borough to open a new pub would be Islington or Hackney.

However, this framework is very generalised and does not cater for different types of pub. For instance, if you want to open a pub that appeals to young people, then you would weight the 'working age %' variable as more important and you might want to look at the type of other pubs in the area as you would want to open the pub in a 'hip' area.

Furthermore, boroughs in London are huge places. An important next step would be to break down the most popular boroughs into their respective areas and rate these areas in terms of suitability for opening a new pub.

Finally, it's worth noting that most of this data is from 2019 and may need refreshing to understand whether the results are the same in a post-Covid world.

5 Conclusion

The purpose of this study was to understand the best borough to open a new pub in London. Through the creation of a decision matrix and K-means clustering, it was found that the best two boroughs to open a new pub in London are Islington and Hackney.