Opening A New Bar in Leeds (UK) – A Location Analysis

Lewis Lawton

June 2020

Introduction

The city of Leeds is located in the north of England, 272 km north west of the capital city London. In fact, it is the largest city in the county of West Yorkshire and has become the largest legal and financial centre outside of the capital. It first began as a small borough in the 13th Century, but by the 18th Century it had become a major hub for the production and trading of many goods, such as wool. By the mid-20th Century; Leeds had expanded and absorbed the surrounding villages, leading the city to sit amongst the fourth-most populous urban areas in England – with a population of roughly 2.6 million people[1].

Known nowadays for its wide array of food and drink and in such a highly populous area; there are, of course, many bars to be found in the city of Leeds. The city is also served by 5 universities, ultimately lowering the average age of its residents compared to the rest of Yorkshire or England as a whole. Roughly 10% of the residents of Leeds are between the age of 20 - 24, many of which are likely to go out to drink regularly.

Leeds is constantly growing. With a bustling city centre, a young demographic and vibrant nightlife, it is a great place in which to set up a new business.

Business Problem

A local business owner is looking to open a new bar in the city of Leeds, the issue is that they are unsure on a specific location in which to open. They also do not know exactly what type of bar to open (e.g. simply a bar, speakeasy, traditional pub, etc.). They wish to open as soon as possible and have spoken to several stakeholders concerning this task. A stakeholder has suggested using location data to decide upon a suitable location to open the bar. As such, they have reached out to a Data Scientist in order to complete this task.

Leveraging the Foursquare location data API, this report will endeavour to highlight potential areas where it may be profitable to open a bar. It may be the case that a location with few bars surrounding it may be the best fit, however more factors are to be considered. For example, many people do not

stay in one bar for an entire night out and as such it may be more profitable to open a bar in a busier area in order to increase foot traffic and subsequent profit. It completely depends on the type of establishment that is to be opened, so this report will attempt to answer this question. A clustering approach will be taken to segment the different areas around Leeds in order to locate an optimal location to open this business.

The report will also aim to inform any future strategy for similar business owners considering Leeds as a potential area to set up a new and developing business. Potential implications of the appropriateness and use of location data for similar projects in the future will be discussed and highlighted.

Data

Packages Used

All statistical, exploratory and clustering analysis was completed within a Jupyter Notebook running Python (version 3.6). Web scraping was undertaken using the **Beautiful Soup** library. The **Pandas** library was utilised to display and store tabular data, **NumPy** allowed specific statistics to be run and for specialised manipulation of the data to take place.

Geopy, **Matplotlib** and **Folium** were also implemented in the report in order to extract and visualise results; assisting in the communication of significant findings. Matplotlib built any graphs used and Folium assisted in displaying specific points on maps.

The **SciKit-Learn** library handled the machine learnings of the project, specifically the clustering algorithm. The algorithm utilised was **K-Means Clustering**. This clustered our dataset and was displayed in tabular form using Pandas, as well as in graphical form using Folium.

Data Sources

A web scrape of the Wikipedia article containing the postcodes (LS) for the Leeds area forms the basis of the investigation. The link to the Wikipedia article can be viewed here[2]. The table contained within this article also lists the specific neighbourhood(s) assigned to each postcode within the Leeds area.

The latitude and longitude for each postcode was obtained in .csv format from the freemaptools.com website, linked here[3]. This data was inspected, cleaned, and the relevant features and data were obtained.

The Foursquare Places API[4] was utilised to determine nearby venues located within each postcode and neighbourhood, in order to determine the frequency and distance of similar businesses. This data was collected for each neighbourhood within the postcodes and formed the basis of map plots to help visualise the data. The data from Foursquare also allowed detailed data tables to be populated for use in the segmentation of different areas around Leeds.

Methodology

Data Wrangling

The data scraped from the Wikipedia article containing the postcodes and he neighbourhoods which exist within them was first loaded into Python. This table was tricky to work with as table rows were formatted as table headers, so two different for loops were utilised to gather the correct information. In each for loop, the data within the table was appended to an empty list, then the lists were combined to make a Pandas DataFrame (named *df*).

Removal of the "\n" characters from the table rows took place and erroneous data were excluded as part of the above process. For example, 3 postcodes were removed from analysis as they did not relate to any neighbourhoods and weren't geographically relevant. Our completed DataFrame (*leeds_df*) contained **29 rows** of data, one for each postcode found within the Leeds area.

The .csv file containing the postcode outcodes (the first part of the postcode) was subsequently loaded into the notebook and simply merged with our existing *leeds_df* DataFrame. This was completed using an inner join on the postcode column, so that only the relevant rows were joined. The .csv file contained **3004** rows of data, as it contained every outcode in England. Once joined, this was reduced to just the **29** relevant rows that were required for our analysis.

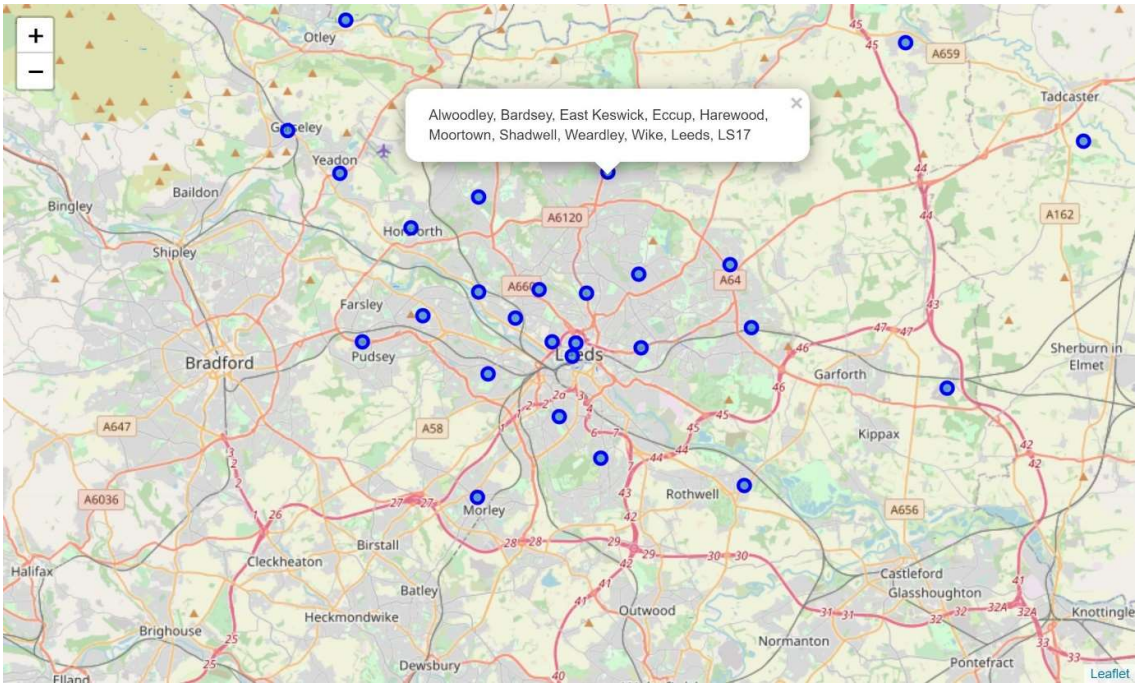
Figure 1 shows the head of the *leeds_df* DataFrame, detailing the columns **PostCode**, **Town**, **Neighbourhood**, **latitude** and **longitude**:

Out[8]:

	PostCode	Town	Neighborhood	latitude	longitude
0	LS1	Leeds	Leeds city centre	53.79674	-1.54754
1	LS2	Leeds	Leeds city centre, Woodhouse	53.80123	-1.54597
2	LS3	Leeds	Burley, Kirkstall, Woodhouse	53.80128	-1.55964
3	LS4	Leeds	Burley, Kirkstall	53.80944	-1.58082
4	LS5	Leeds	Hawksworth, Kirkstall	53.81851	-1.60199
5	LS6	Leeds	Beckett Park, Burley, Headingley, Hyde Park, M...	53.81928	-1.56704
6	LS7	Leeds	Beck Hill, Buslingthorpe, Chapel Allerton, Cha...	53.81828	-1.53971
7	LS8	Leeds	Fearnville, Gipton, Gledhow, Harehills, Oakwoo...	53.82445	-1.50926

The data in this DataFrame was then used to populate a Folium map, so that the locations of centre of each postcode could be visualised. To complete this task, a for loop was used to iterate through each row of the data and assign labels to them which could be applied to the map. The latitude and longitude values were then passed as circle markers in Folium and the map displayed.

Figure 2 illustrates the map created showing postcode centres, also showcasing one of the popups for the LS17 postcode:



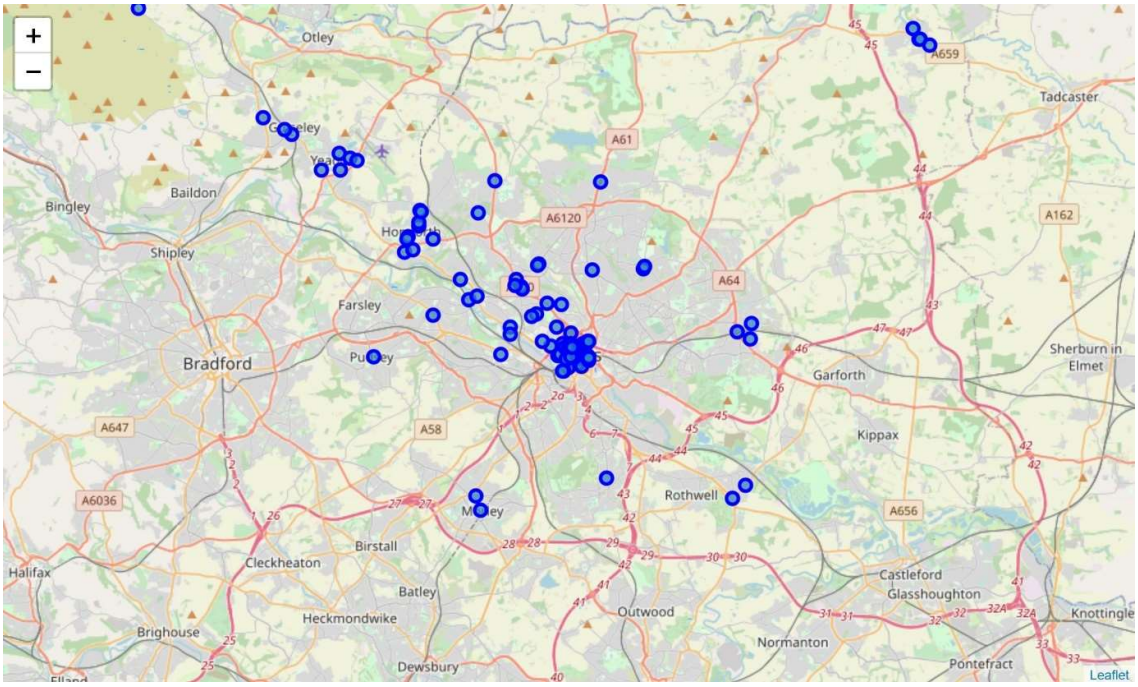
Using Location Data

The **Foursquare Places API** was used to gather nearby venue and location data around the city of Leeds. A function was defined in order to collect the data from venues within a **1 km** radius of each

postcode centre. This function iterates through each postcode in our DataFrame and appends the results to a new one, *leeds_venues*, which stored the results. The resulting *leeds_venues* DataFrame contained **666** rows of data and many different types of venue were found. It was clear that most of these venues were situated closer to the city centre, as would be expected. This DataFrame contained the data that was to be used for the subsequent clustering analysis.

The rows only containing pubs, bars and similar venues were filtered from this DataFrame and stored into a new one (*leeds_bars*). **133** venues in the *leeds_bars* DataFrame fit the criteria, however when duplicates were removed (due to overlap across different postcodes) this number dropped to **98** unique venues. The *leeds_bars* DataFrame was used to show the spread of different drinking establishments across Leeds, based on their location data.

Figure 3 illustrates the venues returned, placed on an interactive map using Folium:



Clustering Algorithm

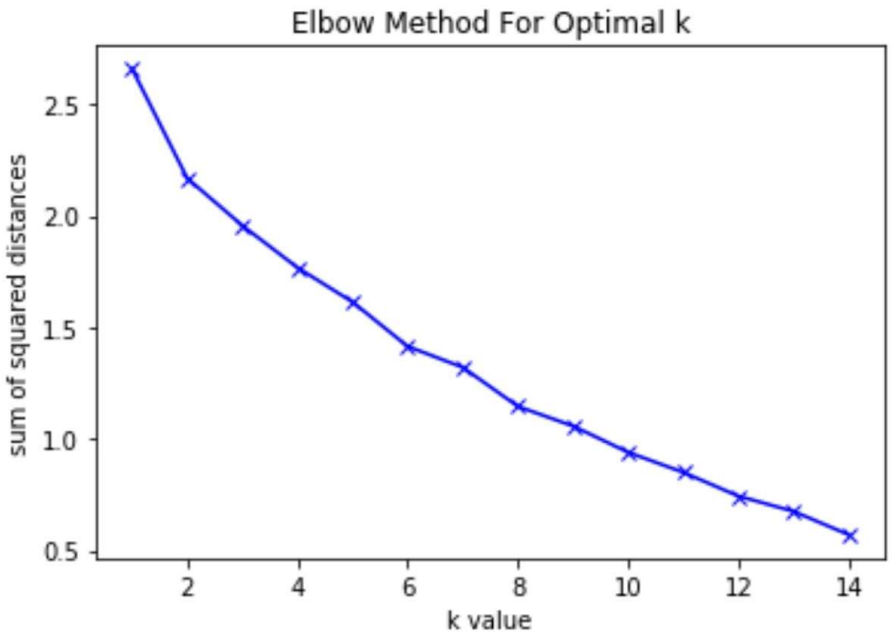
As aforementioned, the DataFrame *leeds_venues* contained the data which was used for our clustering analysis. Before this could take place, however, we first needed to remove any strings from the data so the algorithm could run. To do this, **One-Hot encoding** was used to represent the string values as integers. One-Hot encoding allows us to transform non-ordinal variables. It places a 1 in one row for each variable and 0 in all the others, for all columns in the dataset. This allows our machine learning algorithm to recognise the difference between values using integers instead of strings.

The “Neighbourhood” column of the *leeds_onehot* DataFrame was then dropped, so that the clustering algorithm could be run on the data. This was represented by the DataFrame *leeds_grouped_cluster*.

The **K-means clustering algorithm** from the Sci-Kit Learn library was utilised for this analysis. The sum of squared distances of samples to the nearest cluster centre were first calculated for multiple values of K in order to ascertain the appropriate number for K for this study. Sum of squared differences were calculated as the *inertia_* value from the K-means library. The result of this was plotted using the Matplotlib library and the “elbow method” implemented to decide upon the K value.

The elbow method attempts to show the optimal value to use for K by plotting the sum of squared differences against the number of clusters used in the current pass of the algorithm on a simple line plot. The “elbow” respresents a bend in the line, which itself illustrates the algorithm’s ability to converge properly.

Figure 4 shows a simple line graph to decide upon the number for K, using the elbow method:

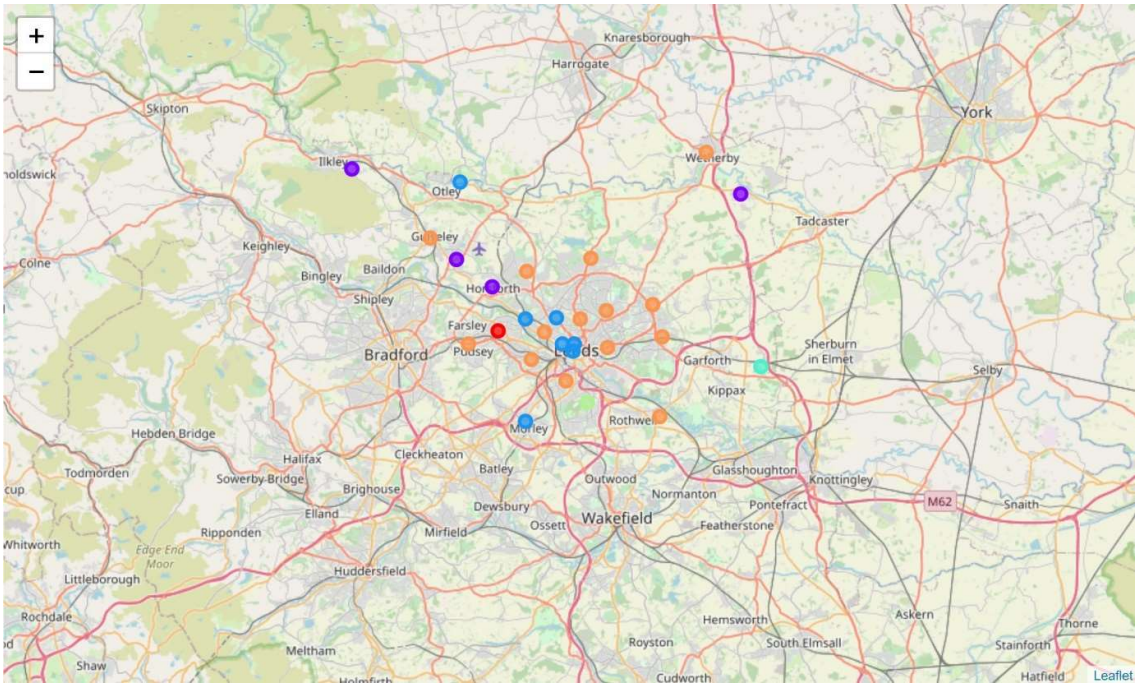


From the above plot, although not conclusive, it was decided upon that **6** was the appropriate number to use for K – that 6 clusters were to be created by our algorithm on the venues data.

Once the algorithm had been run on the data, the labels of each cluster were inserted into the DataFrame alongside their respective postcode and neighbourhood. It was also apparent that no

results were returned for any venues around **LS24 – Tadcaster**, so this row was removed. With our cluster labels assigned to our postcodes, the data was plotted on one final map, so that the way in which the data was clustered could be viewed.

Figure 5 shows the different colour-coded clusters on the map of Leeds:



Clusters within the final DataFrame were also visualised in tabular form, to examine the most common venue types within each, as well as which postcodes they belonged to.

Figure 6 shows a table for the **cluster 1** values and the most popular venues within these postcodes:

	PostCode	Town	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
17	LS18	Leeds	1	Pub	Coffee Shop	Pizza Place	Bar	Cricket Ground	Sandwich Place	Park	Café	English Restaurant	Fast Food Restaurant
18	LS19	Leeds	1	Pub	Lake	Cricket Ground	Fish & Chips Shop	Supermarket	Gas Station	Sports Club	Hotel	Gym	Golf Course
22	LS23	Wetherby	1	Pub	Bar	Convenience Store	Cricket Ground	Chinese Restaurant	Music Venue	Tea Room	Discount Store	Fast Food Restaurant	Farm
28	LS29	Ilkley	1	Pub	Train Station	Scenic Lookout	Trail	Gastropub	Construction & Landscaping	Convenience Store	Cricket Ground	Cuban Restaurant	Fast Food Restaurant

Results

Locations of Existing Businesses

The first section of this report focused on locating the existing businesses which are listed as some form of pub or bar on Foursquare. Using the Places API, it was possible to highlight **98** such businesses around the area of Leeds. The API call specified a radius of 1000 (equalling a radius of 1 kilometre from the centre of each postcode) and a limit of 200 results for each postcode.

There was no increase in the number of results when a larger limit parameter was specified. With a larger radius parameter, the results only increased due to overlapping of the postcodes finding the same businesses and listing them multiple times. Removing duplicate values always resulted in the API finding and returning **98 venues**.

The vast majority of the venues returned were within the centre of Leeds, specifically within postcodes **LS1**, **LS2** and **LS3**. This is illustrated in figure 3, which shows many results towards the centre of the map (which itself is centred on the latitude and longitude values for central Leeds). This aids in establishing a picture of where may be profitable to open a bar in these areas, presumably with a higher density of bars receiving higher foot traffic and profit.

Clustering Analysis

The K-means clustering algorithm used in the report was able to effectively cluster the different areas around Leeds based upon the types of venues found within them. Through the use of the "elbow method", the optimal number of clusters for the data was found to be **6**. The clusters labelled 0, 3 and 4 contained few instances of bars/pubs, owing to the fact they clustered differently from clusters 1, 2 and 5.

Specifically, the most frequent venue in cluster 1 was a pub for all four postcodes found within it. The postcodes included in the cluster were more rural areas, compared to some of the other clusters. Cluster 2 featured many of the central postcodes, where different types of bars were very frequent in the results. Cluster 5 contained the greatest number of postcodes within it but was slightly less conclusive in terms of the number of pubs/bars located within them. This cluster was a bit more generalised and represented more of a middle ground compared to the other, bigger clusters.

Clusters 0, 3 and 4 featured only one postcode each within their clusters. Pubs and bars were listed infrequently in these clusters, showing a low amount of them situated in these areas of Leeds. None of

the postcodes included in these clusters can be considered central Leeds, representing more rural areas.

Discussion

Existing Business Locations

As is to be expected, the location data collected through Foursquare revealed a higher density of venues towards the centre of Leeds. This was true for all venue types, but also specifically for drinking establishments such as bars and pubs. A more central location within Leeds almost guarantees higher foot traffic and potential earnings, however this needs elucidating with further data collection and market research.

A more central location in a city as large as Leeds usually means higher costs for upkeep, for example paying rent to the building owners or more competition with rival businesses. These factors should be considered when looking at the difference between opening a new business within a city or within a rural location and are beyond the scope of this report.

By plotting the locations of the establishments returned from Foursquare, this report has highlighted potential “hotspots” which can help inform a decision on where to open. These areas may be an appropriate place to open based on the fact they receive a lot of customers; however quieter areas are to be considered.

It appears that from the location data, a higher number of bars exist towards the centre of the city when compared with other types of drinking establishments. Similarly, more pubs can be viewed on the outskirts of the city compared to different types of bars. To this end, I would recommend that the owner take this into account when making their decision on whether to open a bar or a pub – or similar. A bar would suit the centre of the city much more, as customers are more likely to visit multiple bars on a day / night out in the city compared to a pub. A pub usually represents a larger place for more people to spend more time within, meaning that it suits the rural areas of Leeds more compared to a bar. The homely the establishment, the longer people are likely to spend drinking there – increasing the profit to be made.

If the business owner is looking to open up a more specific style of bar or pub (such as a cask-ale bar, wine bar, speakeasy, etc.), I would certainly recommend a more central location within the first three postcodes of the city for that venture. This is due to the higher chance that more people will visit the bar, helping them to establish a foot hold within the business quickly.

Once more, it would be beneficial to have data pertaining to the financial information of existing bars within the city to help inform a decision. It would also be beneficial for the specific data science analysis if there was an appropriate .geojson file available for use which detailed the geographical boundaries of the Leeds postcodes, as an appropriate one could not be located for this report. A .geojson file would aid in creating a Chloropleth map using the Folium library, in order to truly highlight the "hotspots" around the city in terms of density of bars.

Clustering Analysis

Our K-means clustering algorithm was able to effectively cluster the areas around Leeds based on the number of popular venues located within each postcode. Our visual representation of the clusters on the map of Leeds shows geographically disparate postcodes being clustered together – demonstrating the algorithm's ability to cluster the data effectively based off the venue data alone.

The optimal number for K was found to be 6, following an analysis of the sum of squared differences to each cluster centre and the elbow method. As viewed in Figure 4, the "bend" in the data to highlight convergence was not as explicit as we would have hoped. In order to combat this, the silhouette method may have been employed instead. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters and increases in value as the algorithm converges[5]. A combination of these two validation methods may have been used to find an optimal value for K for the report and is certainly something which should be considered for future analyses in similar studies.

Another potential reason why the elbow method struggled to highlight the optimal K value may be due to the data itself. There may not have been enough data points to accurately depict the multiple clusters required in the analysis and as such was not as robust a measure as it could be. For this specific type of report, it may be worth experimenting with the DBSCAN algorithm to cluster the postcodes. This may be more appropriate due to its ability to cluster high density vs. low density within a dataset, capable of creating various clustered shapes as it does so.

We can view that the clusters labelled 0, 3 and 4 contained few instances of bars/pubs, owing to the fact they clustered differently from clusters 1, 2 and 5 which contained many. Clusters 0, 3 and 4 also contained only one postcode in their cluster - potentially highlighting that starting up a bar in these areas would not be profitable.

Specific Clusters

Cluster 0 featured one instance of a pub in the top ten venues, but most other venues included landmarks, sports clubs and train stations. There were also no bars found in the top ten most popular venues for this postcode. The location for this postcode geographically is the second-most western area of Leeds and far from the city centre, supporting the notion that bars are much more frequent in the centre.

Cluster 3 was similar in scope to cluster 0, in that there was only one instance of any type of pub/bar in the top ten most popular venues – a wine bar. Featuring mostly sports clubs, restaurants and some types of shops; it clusters differently from the rest and represents an area on the far-eastern side of Leeds.

Cluster 4 was the final cluster to contain one postcode (LS10), featuring more restaurants and shops – as well as a stadium. Geographically, this is just south of the centre of Leeds and near to the Elland Road stadium - home of Leeds United Football Club. Once more, this cluster featured specifically a pub within the top ten venues of the area, further highlighting the high density of bars in central Leeds.

Cluster 1 is clustered mainly on pubs, with all the postcodes included featuring a pub as the most common venue. The geographical location of these postcodes is either north west or north east of the centre of Leeds, in more rural areas such as Ilkley and Wetherby. As stated before, this follows the trend that pubs are more likely to be located along the outskirts of Leeds and in neighbouring villages. The postcodes also feature a number of bars within this cluster, highlighting that either venture may be a profitable endeavour within these postcodes.

Cluster 2 heavily features a range of different types of bars and pubs within the top ten most common venues within these postcodes. The postcodes included are the most central within the Leeds area, which coincides with the above findings about the higher density of venues towards the city centre. It is safe to say that areas within this cluster will provide the most competition in terms of other venues.

This may be considered a positive and a negative – the positive being the increase in foot-traffic mentioned above. A downside to starting a business within these areas would be the time it would take to establish yourself as a place to visit instead of the many other bars and pubs. It is a decision that must be explored further than the scope of this analysis, especially if considering opening a specific type of venue.

Cluster 5 contained the most postcodes compared to any other cluster within this analysis. Regardless, there are many bars and pubs featured in the top 10 most common venues for these postcodes and clearly there is a need for them. The locations of each of these postcodes around Leeds vary, featuring some more central and some of the most rural postcodes included in the analysis. The more central locations appear to be the areas of Leeds mostly populated by students, such as LS7 and LS8.

One final observation would be that there was no instance of a speakeasy within the top ten most common venues for any of the postcodes used in this analysis. This may be due to speakeasies being represented by different, more generalised names on Foursquare; or it could genuinely highlight a potential gap in the market within the city of Leeds. Many people prefer a speakeasy compared to a traditional bar, where they can relax and have more of a conversation with others as if in a café, so there should be demand for a venture of this type. I would certainly recommend looking at this as a possibility when opening a new business of this kind in the city of Leeds.

Concluding Remarks

This report was commissioned by a local business owner looking to open up a new bar or pub in the city of Leeds. They were unsure of a place to open and what type of establishment they wanted to settle on. Through the use of the Foursquare Places API and using unsupervised machine learning clustering algorithm (K-means clustering); this question has attempted to be answered. This report, combined with appropriate market research of the area, should hopefully provide an insight into the correct decision to make on this business venture.

It is clear that this report should not be considered the entire picture, but a tool to aid in the decision-making process for this endeavour. To this end, there are a few pieces of advice that can be drawn from this analysis.

Firstly, if the business owner is considering opening a bar, it may be appropriate to do this closer to the centre of Leeds. Due to the higher volume of bars in these areas and the increased foot-traffic, this should be the most profitable path to take. If the business owner is looking to open up a pub, I would certainly suggest more rural areas of Leeds – potentially even the four postcodes included in cluster 1 (LS18, LS19, LS23 and LS29).

Competition is not a bad thing in this circumstance, so I believe it would be unwise to open up an establishment in the postcodes LS10, LS13 and LS25. These areas do not seem to contain many bars

or pubs, however there may be good reason for this. Of course, this observation may be reversed if the business owner conducts any more research on the area or if the model used in this study can be tuned further.

A final observation would be that opening a speakeasy specifically may be a profitable venture. As aforementioned, speakeasies do not appear to be very common in the city of Leeds and so a new venue such as this may work well. This may be even more apparent within the centre of Leeds where I believe it would be more popular.

One factor that would improve this analysis would be the availability of more data for use in a report. As mentioned, a .geojson file of the postcodes around the city of Leeds could be utilised to build a better picture of the current venues and what they may be missing. Lastly, the use of the DBSCAN clustering algorithm may be much more appropriate for a project such as this, which I would recommend exploring if this report leads to further work in this area.

References

1. https://en.wikipedia.org/wiki/Leeds
2. https://en.wikipedia.org/wiki/LS_postcode_area
3. https://www.freemaptools.com/download-uk-postcode-lat-lng.htm
4. https://developer.foursquare.com
5. https://en.wikipedia.org/wiki/Silhouette_(clustering)