An Exploration of the Population and Venues Data of Leeds, UK; Does Leeds Need Another Thai Restaurant?

Applied Data Science Capstone Project
- The Battle of Neighborhoods

Anthony Swain, PhD

March 2, 2020

1. Introduction
1.1 Forward

Many of you who are reading this will already be aware of the context of this post. However, for those who are not aware (I am sharing this as an article on LinkedIn), this report is written in connection with the Applied Data Science Capstone Project for the IBM Data Science Professional Certificate on Coursera [1]. Students of this course are spending the first 8 modules focusing on building up data science skills with the course culminating in a large capstone project where the students are able to exhibit some of the skills they learned throughout the course. It is essential for us to use Foursquare location data in this final project and come up with a creative business case to exhibit our skills.

1.2 Background

I have been working as an expat in Asia since 2006. My background is in molecular biology and as such I have not had much of an opportunity to work on location data before. I was born and raised in the UK, my wife is from Thailand and I live in Japan. Having this mixed background gives me a lot of potential locations to work with, however based on the potential of multiple non-English languages encroaching on this project, I decided to pick my hometown; Leeds, UK, as the topic of this project.

Leeds is a city in the North of England located about 170 miles North of Central London [2]. According to a 2011 census, the population of Leeds is around 2.6 million [3] and is likely to be on the rise. I did my undergraduate degree at one of the five universities in Leeds [4] and found the city to be a very active town that is popular amongst students. As a student I my friends and I would often try to visit different bars and restaurants in the area, but hindsight has led me to realize that there may have been something missing in the city…Thai cuisine. Which brings me to the business problem.

1.3 Business Problem

In the event that I move back to the UK I would want to move back to my hometown with my family. In such a scenario, it would be feasible for me to open a business in the form of a restaurant. My wife is an amazing cook and I think her talent for cooking could really bring in repeat customers. Although there were almost no Thai restaurants in Leeds when I left in 2006, the landscape may have changed a lot. This data science project should be able to 1) Give an overview of the current

restaurant and bar landscape in Leeds, 2) Understand the current competition with regards to Thai restaurants in Leeds, 3) Pick out some possibilities for restaurant locations with minimum competition and an appropriate customer base.

1.4 Interest

This data science project should be interesting for anyone who is interested in the current restaurant landscape of Leeds. It could also be of interest to anyone who is realistically interested in opening up a Thai restaurant in Leeds (in a non-hypothetical sense). Also, given the structure of how this report is written, it may be of interest to anyone who wants to perform a similar analysis in other cities in the UK or around the world.

2. Data

In this project there are three main sources of data used. The first source of is the Wikipedia page for the Leeds Postcode area [5]. Wikipedia pages are not generally sources that I would use academically but the Wikipedia page uses the Office of National Statistics [6] to get this information (the presentation of the data on in the Wikipedia source is in an accessible format which is appropriate for this course). The Information obtained from the Wikipedia page will be used as the base for the data frame using the postcode district and area (coverage) as the main points. The next piece of data is taken from the Office of National Statistics who performed a census to gather data on the population of the UK [7]. The most recent data that is available is from 2011 but this should give a rough picture for the purpose of this experiment. The next source of data is from the website freemaptools.com which has a .csv file directory of all of the postcodes in the UK and their respective latitudes and longitudes [8]. This data is used to complete a base data frame to connect to the post codes and area names of the initial data frame based on the Wikipedia table. The final main data source is Foursquare [9]. Foursquare is an application we have been using through this course and is a technology company which built a massive library of location data. The Foursquare library is used in conjunction with the location data I got from the first two sources in order to perform data analysis of the venues in the Leeds area.

3. Methodology

Most of the data mentioned in the data section was taken directly from the download links with the exception of the census data which required the usage of GitHub as a repository.

The initial data frame from which the construction of the dataset began was from the Wikipedia page for the Leeds Postcode area. The read html pandas' function was used to read the data frame directly from the Wikipedia page. From this data frame, the 'Non-geographic' postcodes were dropped along with the columns for 'post town' and 'local authority area' since they are not needed in this report. The column name coverage made little sense in this report so it was changed to 'Area' for the purpose of data interpretation. Below is the head of the data frame for the Postcode district and area:

|	Postcode district	Area
0	LS1	Leeds city centre
1	LS2	Leeds city centre, Woodhouse
2	LS3	Burley, Kirkstall, Woodhouse
3	LS4	Burley, Kirkstall
4	LS5	Hawksworth, Kirkstall

The population csv file was saved as an Excel file then uploaded to my GitHub page. An extension to read excel files was then downloaded and the population data for the whole of the UK was imported into the notebook. In this analysis I decided to keep everything simple by looking at the total population rather than dividing the population into males/females etc so the data was cleaned which produced a data frame containing the UK population by "geography." The geography label was kept until such time that the data frame could be optimized. Below is the head of the population data frame:

	geography	Population
0	AL1	36180
1	AL10	35900
2	AL2	23604
3	AL3	28462
4	AL4	27268

The final part of the master data frame was taken from the freemaptools.com website. The csv file was directly imported into the notebook. The columns were renamed for formatting purposes and all of the previous data frames were merged together to produce one master data frame. The master data frame contained columns with either duplicate data or data which is not needed for analysis and therefore they were dropped from the data frame. The complete data frame containing the Postcode district, Area, Population, Latitude and Longitude is shown below:

	Postcode district	Area	Population	Latitude	Longitude
0	LS1	Leeds city centre	5759	53.79674	-1.54754
1	LS2	Leeds city centre, Woodhouse	10120	53.80123	-1.54597
2	LS3	Burley, Kirkstall, Woodhouse	2876	53.80128	-1.55964
3	LS4	Burley, Kirkstall	9072	53.80944	-1.58082
4	LS5	Hawksworth, Kirkstall	8507	53.81851	-1.60199
5	LS6	Beckett Park, Burley, Headingley, Hyde Park, M...	48277	53.81928	-1.56704
6	LS7	Beck Hill, Buslingthorpe, Chapel Allerton, Cha...	28266	53.81828	-1.53971
7	LS8	Fearnville, Gipton, Gledhow, Harehills, Oakwoo...	48286	53.82445	-1.50926
8	LS9	Burmantofts, Cross Green, East End Park, Gipto...	36406	53.79928	-1.50803
9	LS10	Belle Isle, Hunslet, Leeds city centre, Middle...	35534	53.76162	-1.53123
10	LS11	Leeds city centre, Beeston, Beeston Hill, Cott...	34795	53.77609	-1.55535
11	LS12	Armley, Farnley, New Farnley, Wortley	40949	53.79062	-1.59618
12	LS13	Bramley, Gamble Hill, Moorside, Rodley, Swinnow	35021	53.81052	-1.63396
13	LS14	Killingbeck, Seacroft, Scarcroft, Swarcliffe, ...	34484	53.82782	-1.45604
14	LS15	Austhorpe, Barwick-in-Elmet, Colton, Cross Gat...	34991	53.80619	-1.44417
15	LS16	Adel, Bramhope, Cookridge, Eccup, Far Headingl...	36457	53.85086	-1.60169
16	LS17	Alwoodley, Bardsey, East Keswick, Eccup, Harew...	43426	53.85958	-1.52709
17	LS18	Horsforth	18895	53.84020	-1.64130
18	LS19	Carlton, Rawdon, Yeadon	19835	53.85920	-1.68220
19	LS20	Guiseley, Hawksworth	12663	53.87365	-1.71251
20	LS21	Arthington, Otley, Pool, Fewston, Blubberhouses	17729	53.91132	-1.67876
21	LS22	Collingham, Linton, Wetherby	15791	53.93179	-1.39375
22	LS23	Boston Spa, Bramham, Clifford, Thorp Arch, Walton	9207	53.90372	-1.35458
23	LS24	Saxton, Stutton, Ulleskelf, Church Fenton, Oxt...	11849	53.86987	-1.25168
24	LS25	Aberford, Garforth, Hillam, Kippax, Ledsham, M...	39917	53.78584	-1.33059
25	LS26	Great Preston, Methley, Mickletown, Oulton, Ro...	30021	53.75227	-1.44791
26	LS27	Churwell, Gildersome, Morley	35695	53.74839	-1.60266
27	LS28	Bagley, Calverley, Farsley, Fulneck, Pudsey, S...	37896	53.80163	-1.66897
28	LS29	Addingham, Ben Rhydding, Burley in Wharfedale,...	31456	53.92058	-1.80356

Once the data frame was constructed it was necessary to visualize the data in a way that could confirm that the postcode areas were appropriately marked. I imported geocoders and the Folium library to make a map of Leeds with the areas marked as shown below:
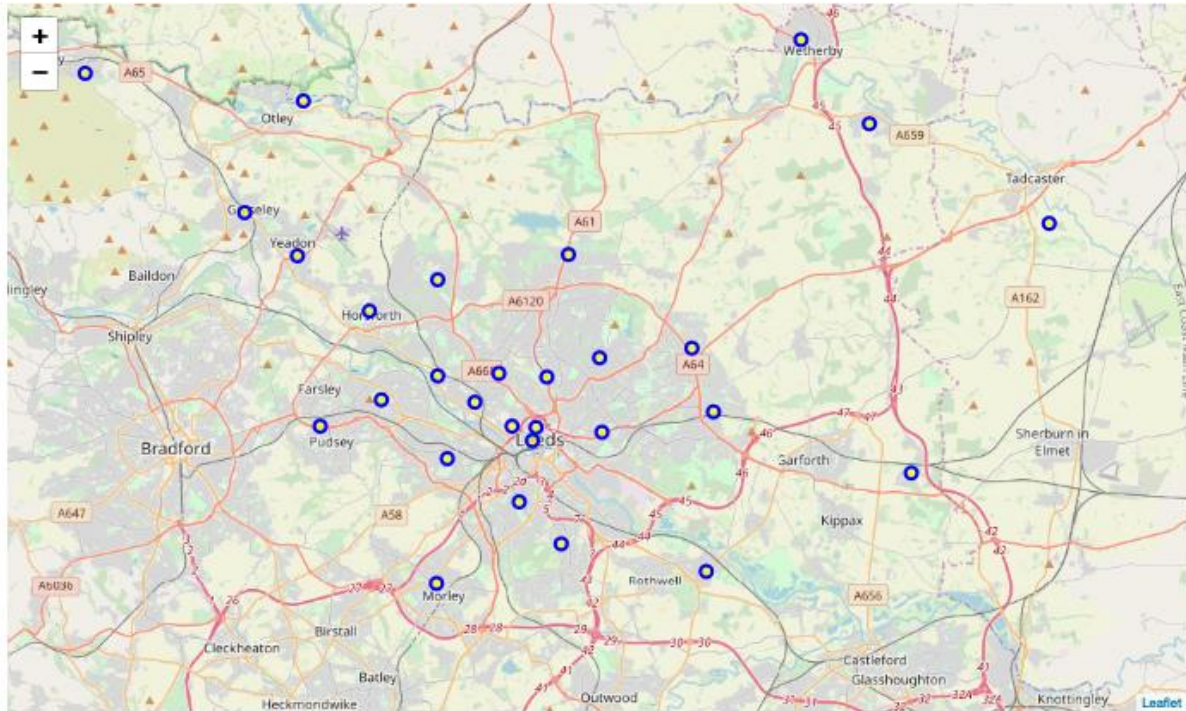
From here I used the Foursquare API to leverage venue data within a 500-meter radius of each area using the latitude and longitude data from the above data frame. The limit was set to 1000 to make sure that all of the venues were captured. Below is the head of the list of venue names:

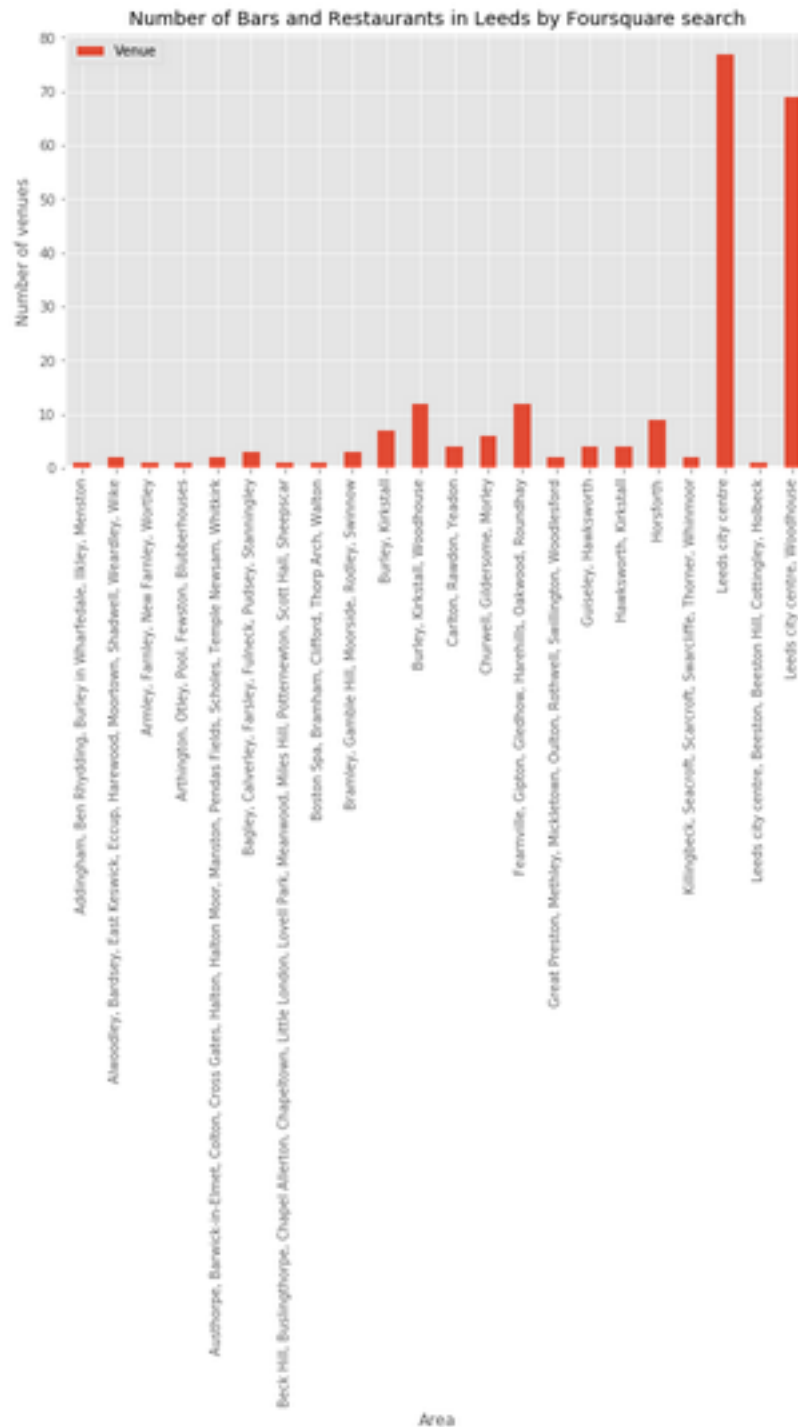|	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Leeds city centre	53.79674	-1.54754	Bundobust	53.795415	-1.545622	Bar
1	Leeds city centre	53.79674	-1.54754	La Bottega Milanese	53.797963	-1.548039	Coffee Shop
2	Leeds city centre	53.79674	-1.54754	MOD Pizza	53.797440	-1.545289	Pizza Place
3	Leeds city centre	53.79674	-1.54754	Trinity Kitchen	53.797316	-1.545640	Restaurant
4	Leeds city centre	53.79674	-1.54754	The Head Of Steam	53.795189	-1.545859	Pub

A total of 355 venues were found by Foursquare with 104 unique categories identified. I used one hot encoding to more easily display the venue categories and found that there were many 'Venue Categories' that were not of interest to this analysis since they were not related to eating or drinking establishments. The data frame included information for venues such as 'bus stop', 'platform' and 'veterinarian', so I dropped 61 unneeded categories from the venue data frame leaving us with 224 venues over 43 unique categories.

From here a mean and frequency of the bars and restaurants were taken from each area to produce a list of top locations in each area. All of the data had now been collected and was ready for visualization.

4. Results

Foursquare found 355 venues in the Leeds area distributed across all of the areas. This distribution was visualized in the bar graph below:

Number of venues in Leeds by Foursquare search

The areas designated as 'Leeds city centre' and 'Leeds city centre, Woodhouse' were the areas that had the greatest number of venues with counts of 100 and 96 respectively.

As I mentioned in the methods section, there was a large amount of venues which was not relevant to this analysis. After removing the venues that were not useful for analysis I created a new data frame and made another bar graph (no pun intended) of the drinking establishments and restaurants by area:

Number of Bars and Restaurants in Leeds by Foursquare search

The shape of the above chart is very similar to the previous one despite the removal of the non-essential categories. This information indicates that the areas that have the greatest concentration of both all types of venues and restaurants are in the 'Leeds city centre' and 'Leeds city centre, Woodhouse' areas.

Next, I decided to create a visualization of the bars and restaurants by frequency of occurrence in Leeds. Over the course of my career I have used pie charts to display this kind of data so I made a pie chart of the venue data in Leeds as shown below:

Distribution of Bars and Restaurants in Leeds, UK
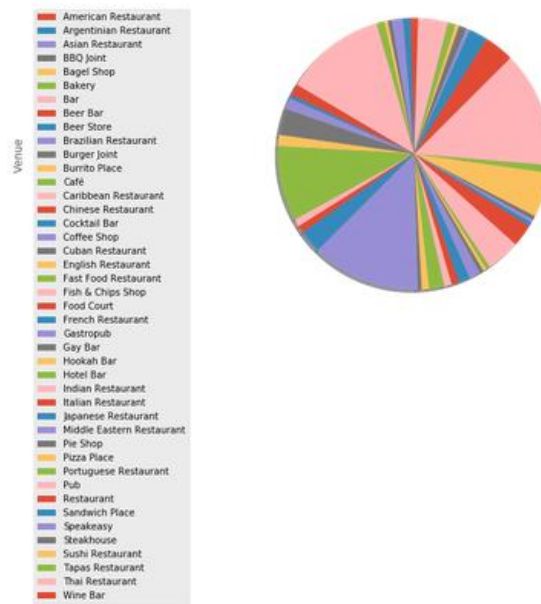
This visualization is not easy for anyone to interpret due to the large number of categories. In the IBM Data Science Professional Certificate course, it was mentioned in the Data Visualization with Python module that pie charts are sometimes very difficult to interpret and may not be the best for visualization. I wanted to see this for myself and demonstrate why this is not a good model which is why I plotted the above chart.

In order to make the visualization more user friendly I used another bar chart to plot the venues in a way that can be interpreted:

Distribution of Bars and Restaurants in Leeds, UK
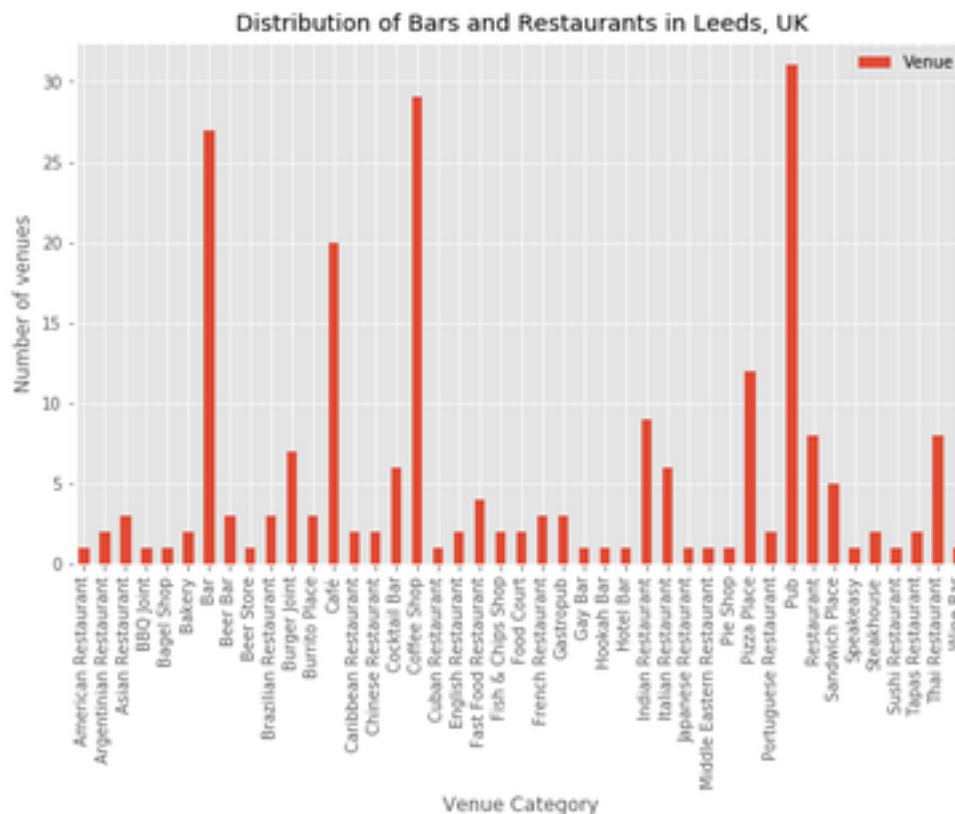
In the bar chart above, we can see that the top categories in Leeds as a whole are 'Pub', 'Coffee shop', 'Bar' and 'Café'. After this we get into less obvious categories but the next most common venues are 'Pizza Place', 'Indian Restaurant' and 'Thai Restaurant' (with 'Thai Restaurant' being a venue of major interest in this analysis).

Although we can get a lot of information from the above bar chart, it is useful to see this data in a list. I arranged a data frame which shows the top 5 most common venues in each area of Leeds.

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Addingham, Ben Rhydding, Burley in Wharfedale,...	Gastropub	Wine Bar	Fish & Chips Shop	English Restaurant	Cuban Restaurant
1	Alwoodley, Bardsey, East Keswick, Eccup, Harew...	Pub	Italian Restaurant	Wine Bar	Burger Joint	Cuban Restaurant
2	Armley, Farnley, New Farnley, Wortley	Pub	Wine Bar	Burger Joint	English Restaurant	Cuban Restaurant
3	Arthington, Otley, Pool, Fewston, Blubberhouses	Café	Wine Bar	Burger Joint	English Restaurant	Cuban Restaurant
4	Austhorpe, Barwick-in-Elmet, Colton, Cross Gat...	Gastropub	Pub	Wine Bar	Burger Joint	Cuban Restaurant
5	Bagley, Calverley, Farsley, Fulneck, Pudsey, S...	Fast Food Restaurant	Indian Restaurant	Café	Fish & Chips Shop	English Restaurant
6	Beck Hill, Buslingthorpe, Chapel Allerton, Cha...	Burger Joint	Wine Bar	English Restaurant	Cuban Restaurant	Coffee Shop
7	Boston Spa, Bramham, Clifford, Thorp Arch, Walton	Restaurant	Wine Bar	Burger Joint	English Restaurant	Cuban Restaurant
8	Bramley, Gamble Hill, Moorside, Rodley, Swinnow	Fast Food Restaurant	Coffee Shop	Pub	Burger Joint	English Restaurant
9	Burley, Kirkstall	Cuban Restaurant	Coffee Shop	Sandwich Place	Hookah Bar	Pub
10	Burley, Kirkstall, Woodhouse	Pub	Café	American Restaurant	Asian Restaurant	Bagel Shop
11	Carlton, Rawdon, Yeadon	Pub	Asian Restaurant	Bakery	Wine Bar	Burrito Place
12	Churwell, Gildersome, Morley	Fish & Chips Shop	Bakery	Bar	Pub	Pizza Place
13	Fearnville, Gipton, Gledhow, Harehills, Oakwoo...	Café	Bar	Italian Restaurant	Fish & Chips Shop	Coffee Shop
14	Great Preston, Methley, Mickletown, Oulton, Ro...	Bar	Pizza Place	Wine Bar	Burger Joint	English Restaurant
15	Guiseley, Hawksworth	Coffee Shop	Pub	Pizza Place	Fast Food Restaurant	Wine Bar
16	Hawksworth, Kirkstall	Coffee Shop	Pub	Wine Bar	Burger Joint	English Restaurant
17	Horsforth	Pub	Coffee Shop	Pizza Place	Gastropub	Bar
18	Killingbeck, Seacroft, Scarcroft, Swarcliffe, ...	Fast Food Restaurant	Pub	Burger Joint	English Restaurant	Cuban Restaurant
19	Leeds city centre	Bar	Coffee Shop	Pub	Indian Restaurant	Thai Restaurant
20	Leeds city centre, Beeston, Beeston Hill, Cott...	Pizza Place	Wine Bar	Burger Joint	English Restaurant	Cuban Restaurant
21	Leeds city centre, Woodhouse	Coffee Shop	Bar	Café	Pub	Thai Restaurant

The above data frame shows that 'Thai Restaurant' is a top 5 venue category in the 'Leeds city centre' and 'Leeds city centre, Woodhouse' areas. I checked the count of Thai Restaurants in Leeds as a whole and 8 venues were returned. I created another data frame containing the details of the Thai Restaurants in Leeds and found that all of the Thai Restaurants in the Leeds area are distributed exclusively in the 'Leeds city centre' and 'Leeds city centre, Woodhouse' areas.

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	Leeds city centre	53.79674	-1.54754	Sukhothai	53.799071	-1.547604	Thai Restaurant
61	Leeds city centre	53.79674	-1.54754	Chaophraya	53.794334	-1.544047	Thai Restaurant
73	Leeds city centre	53.79674	-1.54754	my Thai	53.797178	-1.553922	Thai Restaurant
76	Leeds city centre	53.79674	-1.54754	Mommy Thai	53.796024	-1.541822	Thai Restaurant
102	Leeds city centre, Woodhouse	53.80123	-1.54597	My Thai	53.801343	-1.543007	Thai Restaurant
121	Leeds city centre, Woodhouse	53.80123	-1.54597	Sukhothai	53.799071	-1.547604	Thai Restaurant
146	Leeds city centre, Woodhouse	53.80123	-1.54597	Zaap Thai	53.800286	-1.539767	Thai Restaurant
158	Leeds city centre, Woodhouse	53.80123	-1.54597	Thai Aroy Dee	53.799753	-1.539604	Thai Restaurant

Using this information, I removed areas in Leeds that have Thai Restaurants from the master data frame and sorted the data frame by population.

	Postcode district	Area	Population	Latitude	Longitude
7	LS8	Fearnville, Gipton, Gledhow, Harehills, Oakwoo...	48286	53.82445	-1.50926
5	LS6	Beckett Park, Burley, Headingley, Hyde Park, M...	48277	53.81928	-1.56704
16	LS17	Alwoodley, Bardsey, East Keswick, Eccup, Harew...	43426	53.85958	-1.52709
11	LS12	Armley, Farnley, New Farnley, Wortley	40949	53.79062	-1.59618
24	LS25	Aberford, Garforth, Hillam, Kippax, Ledsham, M...	39917	53.78584	-1.33059
27	LS28	Bagley, Calverley, Farsley, Fulneck, Pudsey, S...	37896	53.80163	-1.66897
15	LS16	Adel, Bramhope, Cookridge, Eccup, Far Headingl...	36457	53.85086	-1.60169
8	LS9	Burmantofts, Cross Green, East End Park, Gipto...	36406	53.79928	-1.50803
26	LS27	Churwell, Gildersome, Morley	35695	53.74839	-1.60266
9	LS10	Belle Isle, Hunslet, Leeds city centre, Middle...	35534	53.76162	-1.53123
12	LS13	Bramley, Gamble Hill, Moorside, Rodley, Swinnow	35021	53.81052	-1.63396
14	LS15	Austhorpe, Barwick-in-Elmet, Colton, Cross Gat...	34991	53.80619	-1.44417
10	LS11	Leeds city centre, Beeston, Beeston Hill, Cott...	34795	53.77609	-1.55535
13	LS14	Killingbeck, Seacroft, Scarcroft, Swarcliffe, ...	34484	53.82782	-1.45604
28	LS29	Addingham, Ben Rhydding, Burley in Wharfedale,...	31456	53.92058	-1.80356
25	LS26	Great Preston, Methley, Mickletown, Oulton, Ro...	30021	53.75227	-1.44791
6	LS7	Beck Hill, Buslingthorpe, Chapel Allerton, Cha...	28266	53.81828	-1.53971
18	LS19	Carlton, Rawdon, Yeadon	19835	53.85920	-1.68220
17	LS18	Horsforth	18895	53.84020	-1.64130
20	LS21	Arthington, Otley, Pool, Fewston, Blubberhouses	17729	53.91132	-1.67876
21	LS22	Collingham, Linton, Wetherby	15791	53.93179	-1.39375
19	LS20	Guiseley, Hawksworth	12663	53.87365	-1.71251
23	LS24	Saxton, Stutton, Ulleskelf, Church Fenton, Oxt...	11849	53.86987	-1.25168
22	LS23	Boston Spa, Bramham, Clifford, Thorp Arch, Walton	9207	53.90372	-1.35458
3	LS4	Burley, Kirkstall	9072	53.80944	-1.58082
4	LS5	Hawksworth, Kirkstall	8507	53.81851	-1.60199
2	LS3	Burley, Kirkstall, Woodhouse	2876	53.80128	-1.55964

5. Discussion

There are 29 geographical areas in Leeds with a population of around 2.6 million people [3]. Using the Foursquare data base I identified that there are 224 venues over 43 unique categories over the whole of the Leeds city. There are 8 Thai restaurants; all of which are in the 'Leeds city centre' and 'Leeds city centre, Woodhouse' areas, which gives a lot of room for people to establish Thai Restaurants in areas that are not in the center of Leeds. The final table in the results section shows a list of areas arranged by population which essentially represents a list of places where no Thai restaurants currently exists which have a large population of people that may be interested in such an establishment.

It should be noted that some of the areas in Foursquare did not bring up much information. Having lived in the Leeds area for most of my life, I believe that there could be a significant amount of data missing from Foursquare. If the Foursquare database were to be updated with more information for each area then a more accurate analysis could be done.

The data used in this analysis could quite easily be adapted to other areas in the UK. The location data and population data for all of the postcode areas of the UK can be leveraged from the data sources used so only the Wikipedia pages for the postcodes would be required to run this kind of analysis on other cities in the UK.

This is my first opportunity to produce anything using Python and Jupyter notebooks so this analysis may be a little simple but I am looking forward to having more opportunities in the future to work with this kind of technology. Thank you very much for taking the time to read this report. I hope you found it useful.

6. Conclusion

If you are going to build a Thai restaurant in Leeds, the completion in the 'Leeds city centre' and 'Leeds city centre, Woodhouse' areas is pretty fierce. Several areas of Leeds with no Thai restaurants were identified in the final table which was separated by population. The feasibility of constructing a Thai restaurant in Leeds will be very much dependent on whether a good location can be acquired at a reasonable cost but ultimately, as long as the city center is avoided there is a chance to fill a currently unoccupied niche in the area.

7. Notebook

The following is a link to the Jupyter notebook I used to create this report:

https://github.com/Hakase102/Coursera-CourseworkAS/blob/master/Capstone%20-%20Anthony%20Swain.ipynb

8. Acknowledgements

I would like to thank IBM, Coursera and the instructors of the IBM Data Science Professional Certificate; Joseph Santarcangelo, Alex Aklson, Rav Ahuja, Saeed Aghabozorgi, and Polong Lin for putting together such an engaging program. It has been a pleasure.

9. References

1. Coursera Course - https://www.coursera.org/professional-certificates/ibm-data-science
2. Leeds definition - https://www.collinsdictionary.com/dictionary/english/leeds
3. Office for National Statistics - http://www.nomisweb.co.uk/articles/747.aspx
4. Leeds Wikipedia - https://en.wikipedia.org/wiki/Leeds
5. Leeds LS Wikipedia Page - https://en.wikipedia.org/wiki/LS_postcode_area
6. Office for National Statistics Source Data - www.ons.gov.uk/ons/guide-method/geography/products/postcode-directories/-nspp-/onspd-user-guide-and-version-notes.zip
7. Freemaptools.com postcode data - https://www.freemaptools.com/download/outcode-postcodes/postcode-outcodes.csv
8. Office of National Statistics Census Data - http://www.nomisweb.co.uk/census/2011/ks101ew
9. Foursquare - https://foursquare.com/