Finding the Optimal Location for Opening a New Restaurant in South West London

IBM Data Science Capstone Project

Ken Young, August 2020

Contents

1. Introduction and Business Problem

1.1. Introduction

Our client is a successful restaurant business with a long established presence in the West End and City in Central London.

They have built a reputation for excellent service in the casual dining market, offering quality innovative menus blending dishes from East and South East Asia.

Their recent addition of fixed price menus appealing to time-sensitive customers has proven popular. This is valued, for example, by office workers on lunch breaks and diners heading to the cinema or theatre.

Our client is looking to expand by opening a new restaurant in South West London.

From the client's initial research, which involves limited detailed analysis, there are a number of locations that are potentially suitable.

1.2. Business Problem

In the restaurant business, location is one of the key factors in determining whether a new venture will be profitable.

The objective of this project is to identify, using data science techniques and the Foursquare location API, optimal locations for the proposed restaurant in South West London.

We will analyse and compare neighbourhoods with the following factors which we believe are critical in identifying a suitable location:-

1.2.1. Existing restaurants particularly potential competitors in the vicinity.
1.2.2. Spending power of local population.
1.2.3. Proximity to transportation hubs.
1.2.4. Proximity to entertainment venues such as cinemas.

1.3. Audience

The findings of this project will be presented to the client's senior management and we believe our analytical approach will identify suitable locations that will support the profitability and long term competitive advantage of the venture

2. Data

2.1. Data Sources

To solve our business problem, we will use the following data:-

2.1.1. **List of neighbourhoods in South West London.**
This defines the scope of the project.
As agreed with our client, this definition includes the London Boroughs of Merton, Richmond Upon Thames and Wandsworth.
We will use the following Wikipedia page to extract a list of neighbourhoods in the desired Borough list.
https://en.wikipedia.org/wiki/List_of_areas_of_London

2.1.2. **Latitude and Longitude data for these neighbourhoods.**
This will be obtained from the Python Geocoder package and is necessary to plot maps and obtain venue data.

2.1.3. **Venue data, particularly cinemas and competitor restaurants.**
We will use the Foursquare API and we will be using this data to perform clustering on neighbourhoods.

2.1.4. **Locations of railway stations.**
Locations close to railway stations, in particular tube stations, will generally have heavier footfall and potential customers.
We will use Foursquare API to obtain railway station locations.

2.1.5. **Median household income for each neighbourhood.**
As the restaurant is in the casual dining sector, information on potential disposable income is an important consideration.
We will use the Excel spreadsheet obtained from the following website.
https://data.london.gov.uk/dataset/household-income-estimates-small-areas

2.2. Data Uploads and Cleaning

A list of neighbourhoods and their boroughs were scraped from the Wikipedia website using the Python library Beautiful Soup. A pandas dataframe was then created to store our data, reducing the list to only include neighbourhoods from the three boroughs required.
Latitude and Longitude coordinates for each neighbourhood were extracted using the Python library Geocoder.
Median Household Income data per neighbourhood was then extracted from the London Government website, converting the file into csv format prior to uploading. Generally the neighbourhoods from the spreadsheet matched the names from the Wikipedia page, therefore no adjustments were required.
The database created for Income was then merged into the main pandas dataframe.
Column headers were updated to ensure consistency across the project.
Further data cleaning checks were conducted where it was found that one item for Income was missing. This was amended in the dataframe as the missing item was in the original csv file.
The finalised database ready for exploratory data analysis is now ready:-

	Neighbourhood	Borough	Latitude	Longitude	Median Household Inc
0	Balham	Wandsworth	51.4456	-0.150364	53420.0
1	Barnes	Richmond upon Thames	51.4719	-0.238744	55450.0
2	Battersea	Wandsworth	51.4708	-0.172214	55380.0

Figure 1: First 3 lines of 'neighbourhoods' dataframe

The dataframe contains 39 Neighbourhoods and has 5 columns of data.

2.3. Exploratory Data Analysis

Before we decide on our methodology for solving our business problem, we first of all want to explore our data and obtain a level of confidence that it is sufficient for our purpose.
We will also use this stage to assess our initial choice of variables and refine them where necessary.

2.3.1. **Visualising our List of Neighbourhoods on a Map**

We obtain the coordinates of a reasonably central neighbourhood in our dataset, using geoPy. This frames our geographical area of interest – South West London.
The neighbourhood locations are correctly placed and are therefore fit for purpose.

Figure 2: Map of South West London showing locations of neighbourhoods

2.3.2. **Explore Neighbourhoods with Foursquare Venue Data**

The Foursquare data is extracted in json file format, hence we structure it into a pandas dataframe. We combine the neighbourhood coordinates with venue coordinates and obtain the following, which we can then analyse.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Balham	51.445645	-0.150364	We Brought Beer	51.444324	-0.150656	Beer Store
1	Balham	51.445645	-0.150364	M1LK	51.444450	-0.150913	Coffee Shop
2	Balham	51.445645	-0.150364	Franco Manca	51.443616	-0.149959	Pizza Place
3	Balham	51.445645	-0.150364	Brickwood Coffee & Bread	51.444509	-0.151127	Coffee Shop
4	Balham	51.445645	-0.150364	Ciullosteria	51.447144	-0.148981	Italian Restaurant

Figure 3: First 5 lines of dataframe combining Neighbourhoods with Foursquare data

Most neighbourhoods have significantly less than 100 venues within a 1km radius though a small number have much more. We will limit our data gathering to these criteria for simplicity.
We will use one-hot encoding to prepare the data for subsequent machine learning algorithms.

2.3.3. **Distribution of Competitor Restaurants**

Using the Foursquare explore function, we can obtain numbers of restaurants by venue category.
From our analysis we observe for example a Japanese restaurant may be classified in Foursquare as 'Japanese' or 'Asian'.
For this project our assumption is that any restaurant categorised as Asian, Thai, Japanese, Vietnamese, Korean and Japanese restaurant categories are potential competitors.
This list has been cross referenced back to Foursquare's category hierarchy, which is available on their developer website.

	Neighbourhood	Asian Restaurants	Thai Restaurants	Japanese Restaurants	Vietnamese Restaurants	Korean Restaurants	Chinese Restaurants
0	Balham	1	0	1	0	0	0
1	Barnes	0	1	0	0	0	0
2	Battersea	1	1	2	2	0	1
3	Castelnau	1	3	2	2	0	0
4	Colliers Wood	0	1	0	0	0	0

Grouped to:-

	Neighbourhood	Total Asian Restaurants
0	Balham	2
1	Barnes	1
2	Battersea	7
3	Castelnau	8
4	Colliers Wood	1

Figure 4: First 5 lines of dataframe combining total Asian restaurants

With this grouping of restaurant venues complete, we are now confident we have a list of potential competitors by neighbourhood.
We can now visualise a distribution of Asian restaurants by neighbourhood.

Figure 5: Asian Restaurants by Neighbourhood

We can now visualise that some neighbourhoods have many Asian restaurants whereas others have none. It may prove difficult to establish a new restaurant in an area that already has many competitors operating.

It is worth noting that some neighbourhoods are very close to the River Thames, for example Castlenau. In these cases a 1km radius could pick up restaurants from a neighbourhood in a different borough. For Castlenau that could be Hammersmith on the north side of the Thames. Our working assumption is that isn't a problem for our analysis but it is something to be aware of.

2.3.4. **Distribution of Median Household Income**

Using our data uploaded from the London Government website, we can visualise median household income by neighbourhood.

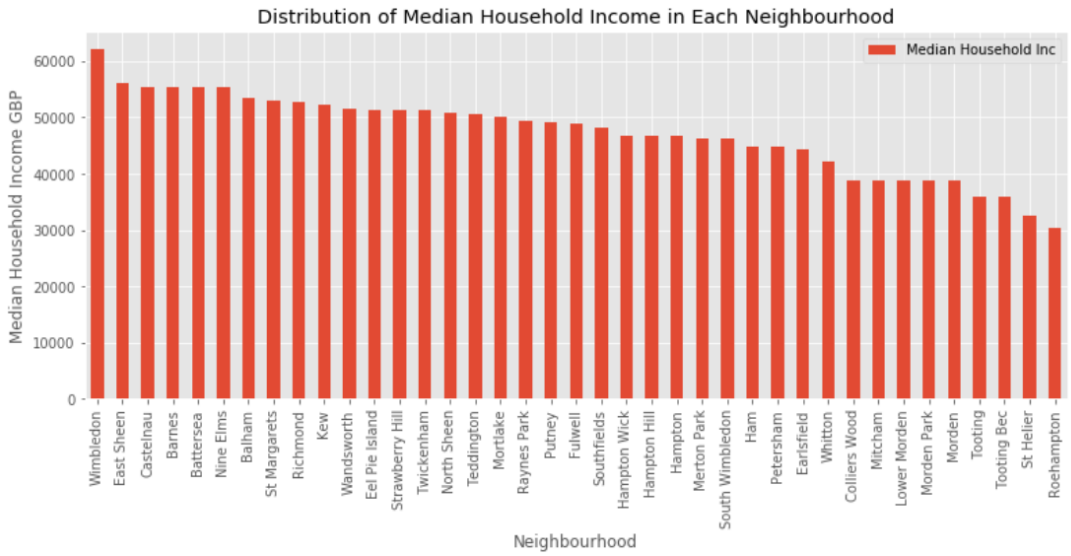

Figure 6: Median Household Income by Neighbourhood

For the casual dining market which our proposed restaurant operates in, we'd anticipate operating in an area where potential customers have sufficient disposable income.
Our assumption is that areas with higher median income will have more money to spend on restaurants and are therefore more favourable.

2.3.5. **Analysis of areas with high footfall such as train stations and entertainment venues**

Using the Foursquare explore function, we can obtain numbers of train or other metro stations and venues such as cinemas, theatres and tourist attractions.
From our analysis we observe that almost all of our neighbourhoods have at least one train station. We'd expect this as suburban London is in general well served by transportation links to central London.
We observe that there are some neighbourhoods with cinemas or theatres and there are a few neighbourhoods with major tourist attractions such as Kew Gardens.

Our assumption here is that the inclusion of train stations, cinemas and other entertainment venues are not particularly useful as key variables for the next stage of this project.
Some neighbourhoods will be outliers in terms that they don't have train stations, and some will be outliers in terms that they do have significant entertainment venues. We will certainly make reference to this in our discussion and conclusions, but for now we will concentrate on the key variables competitor restaurants and household income.

3. Methodology

3.1. Overview

In this project we will focus our efforts on identifying neighbourhoods of South West London that have higher household income and some existing potential competitor restaurants.

Our logic is that higher household income areas are more likely to have disposable income to spend in restaurants.
Using median household income doesn't of course tell us anything about distribution of income, but we believe it is a good starting point to build up a picture of the local customer base.

For competitor restaurants we believe that a large number of existing venues may prove very difficult for a new restaurant to get a foothold in the market.
At the opposite end, a lack of competitors may mean there is little demand in the area. Therefore our ideal is somewhere in the middle.

In our Data section we also specified data gathering on train stations and other venues likely to create significant footfall such as cinemas.
From exploratory data analysis, it is clear that most parts of South West London are well served by train stations. It is probably more of interest to identify the few areas which don't have close access to a train station. Therefore we believe there is little value in conducting data analysis including train station data at this stage.
For cinemas etc., much of South West London is suburban with centres of commerce of various sizes in each neighbourhood. Although our client gets a sizeable amount of customers in its existing central London restaurants from pre/post cinema/theatre traffic, there aren't many venues of this nature in our target area. Our assumption is that customers will likely be a mix of locals who live/work in the area and those who commute by train to central London.

Following on from our exploratory data analysis, our next step is to prepare our dataset for machine learning.

3.2. Machine Learning

We would now like to identify groups of neighbourhoods that are alike, using our criteria of competitor Asian restaurants and median household income. We can then analyse these groups to identify a list of potentially suitable neighbourhoods for our new restaurant.

We will use the clustering method to obtain these neighbourhood groupings.
k-means clustering will be utilised as it is straightforward and its results are relatively easy to interpret.

3.2.1. **Identifying number of clusters to use in the model**

First of all we need to normalise our dataset. We use Standard Scaler to achieve this.
Before we can fit the normalised values for restaurants and income into our cluster model, we have to pre-assign the number of clusters, k, to be used. We will use the Squared Error method to assess the performance of each potential number of clusters.

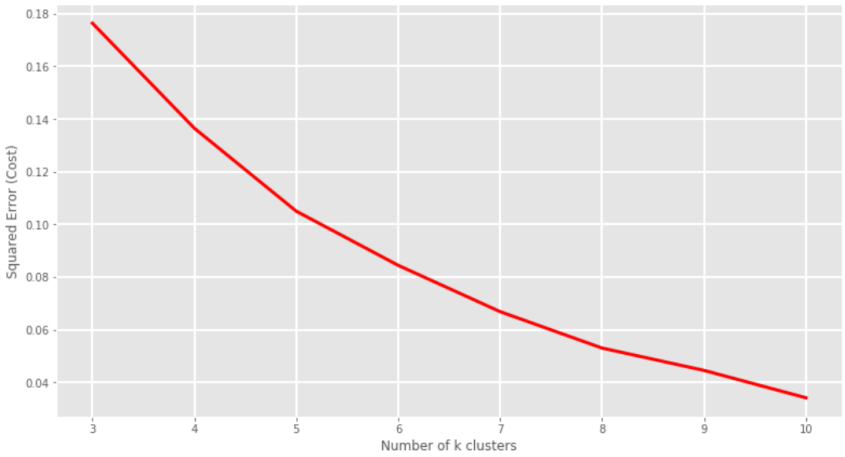

Figure 7: Squared Error for each number of k clusters

The curve appears less steep from k=5 and k=7. From trial and error, we establish that k=6 appears to be the best fit.

3.2.2. **k-means clustering**

The k-means clustering algorithm is now run with the assumption that we will have 6 clusters of neighbourhoods. The cluster labels from 0 to 5 are added back into our database.

	Cluster Label	Neighbourhood	Borough	Latitude	Longitude	Median Household Inc	No. of Asian Restaurants
0	0	Balham	Wandsworth	51.445645	-0.150364	53420.0	-0.183218
1	0	Barnes	Richmond upon Thames	51.471896	-0.238744	55450.0	-0.559298
2	5	Battersea	Wandsworth	51.470793	-0.172214	55380.0	1.697179
3	5	Castelnau	Richmond upon Thames	51.487429	-0.231228	55450.0	2.073258
4	1	Colliers Wood	Merton	51.418275	-0.177863	38780.0	-0.559298

Figure 8: First 5 lines of dataframe including cluster labels

We are now ready to plot these results on a map and discuss each cluster.

4. Results and Discussion

4.1. Results

We observe the following results from our clustering algorithm, plotted on a map of SW London.

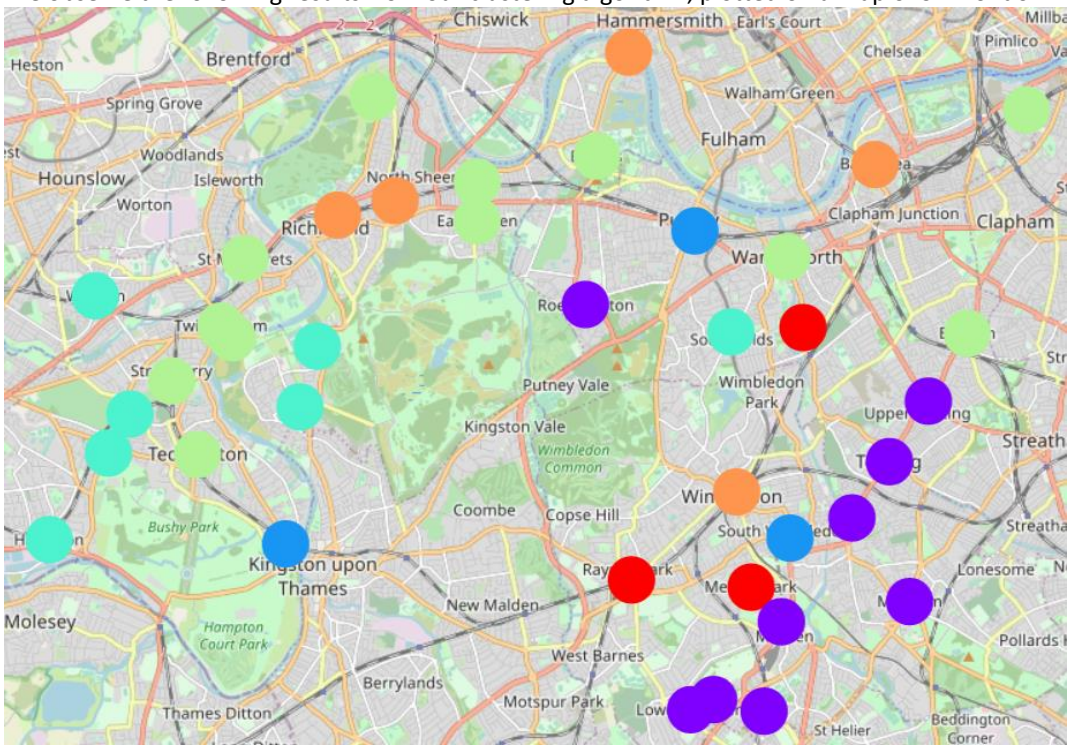

Figure 9: Map of clusters

Cluster 4 (green on map): Higher Income, Mid number of competitors (our optimum cluster)

	Cluster Label	Neighbourhood	Borough	Latitude	Longitude	Median Household Inc	No. of Asian Restaurants
0	4	Balham	Wandsworth	51.445645	-0.150364	53420.0	-0.199482
1	4	Barnes	Richmond upon Thames	51.471896	-0.238744	55450.0	-0.588472
6	4	East Sheen	Richmond upon Thames	51.462371	-0.267094	56090.0	-0.977462
7	4	Eel Pie Island	Richmond upon Thames	51.445126	-0.325408	51380.0	-0.199482
13	4	Kew	Richmond upon Thames	51.480663	-0.291929	52130.0	-0.588472
19	4	Mortlake	Richmond upon Thames	51.468069	-0.266940	50100.0	-0.588472
20	4	Nine Elms	Wandsworth	51.478743	-0.136263	55380.0	-0.199482
30	4	St Margarets	Richmond upon Thames	51.456709	-0.322412	52870.0	0.189508
31	4	Strawberry Hill	Richmond upon Thames	51.438592	-0.339937	51380.0	-0.199482
32	4	Teddington	Richmond upon Thames	51.427784	-0.333653	50600.0	-0.588472
35	4	Twickenham	Richmond upon Thames	51.446744	-0.328189	51350.0	-0.199482
36	4	Wandsworth	Wandsworth	51.457027	-0.193261	51420.0	-0.199482

Cluster 0 (red on map): Mid income, Mid number of competitors
Cluster 1 (purple on map): Lower income, Lower number of competitors
Cluster 2 (dark blue on map): Mid income, Higher number of competitors
Cluster 3 (light blue on map): Mid income, Lower number of competitors
Cluster 5 (orange on map): Higher income, Higher number of competitors

4.2. Discussion

To reduce our potential list of suitable neighbourhoods significantly, we can clearly eliminate cluster 1 according to our criteria.
Areas with lower disposable income and few (or lack of) competitors are not of interest to us.

The cluster most likely to interest us is cluster 4 with higher income and a mid number of competitors. These neighbourhoods are more likely to have households with the disposable income to spend in our client's restaurant. Also the presence of some Asian restaurants indicates there is already demand, but not too many restaurants likely to create a barrier to entry.

Looking at the other clusters, cluster 5 has the highest income and a higher number of competitors. The higher number of competitors may suggest difficulty in opening yet another restaurant offering Asian cuisine.

Clusters 0, 2 and 3 don't offer our optimal blend of income and competitors and are therefore not candidates for recommended areas.

From exploratory data analysis we gathered data on other venues likely to create footfall near our potential restaurant. Neighbourhoods such as Balham and Wandsworth are well connected by train to central London with high populations of young professionals. Kew is connected by London Underground train and also has the major tourist draw of Kew Gardens. Neighbourhoods such as Twickenham and Teddington are much further out from central London but are still well connected by train and are popular with commuters and families. Each of these neighbourhoods will likely have a different mix of potential customers.

It is also worth noting that the neighbourhood Nine Elms is a large redevelopment zone with many new apartments being built, a new tube station soon to open, and major employers such as the US Embassy have recently relocated there. This could indicate potential for future growth.

4.3. Recommendation

Further to the above, we recommend our client considers neighbourhoods in cluster 4 for further consideration as they offer the optimum blend of potential customers' disposable income and number of existing competitor restaurants.
We'd recommend as a next step, further analysis of potential customers in the neighbourhoods listed in cluster 4. One potential source of useful information could be from official demographics data. The client's existing restaurants in central London have a mix of after work dinner, pre/post theatre and tourist traffic. The further we move into suburban South West London, the more likely this mix will change. We'd recommend the client considers this when undertaking further research of neighbourhoods.

5. Conclusion

The objective of this project was to identify areas in three South West London boroughs - Merton, Richmond, Wandsworth - potentially suitable for the opening of a new restaurant offering East and South East Asian cuisine. We selected median household income and number of existing competitor restaurants as key criteria in making this assessment. We used Foursquare to gather venue information for each neighbourhood in the three boroughs.

We then used a clustering algorithm to narrow down areas of specific interest based on our criteria, with a preference for areas that have higher household income and medium number of competitors.

Looking at our preferred list of neighbourhoods (cluster 4); areas closer to central London such as Nine Elms, Balham and Wandsworth (Town) are likely to have a different potential mix of customers to areas further out into South West London such as Mortlake, Teddington and Twickenham. We'd recommend further analysis taking into account demographics of these areas.

The final decision on the optimal restaurant location and whether to consider further demographic analysis will be made by the client.

There are many additional factors involved in choosing a suitable location such as real estate availability, rental prices and business rates, proximity to areas with high footfall. We believe the analysis conducted here provides a solid starting point listing neighbourhoods worthy of further investigation.