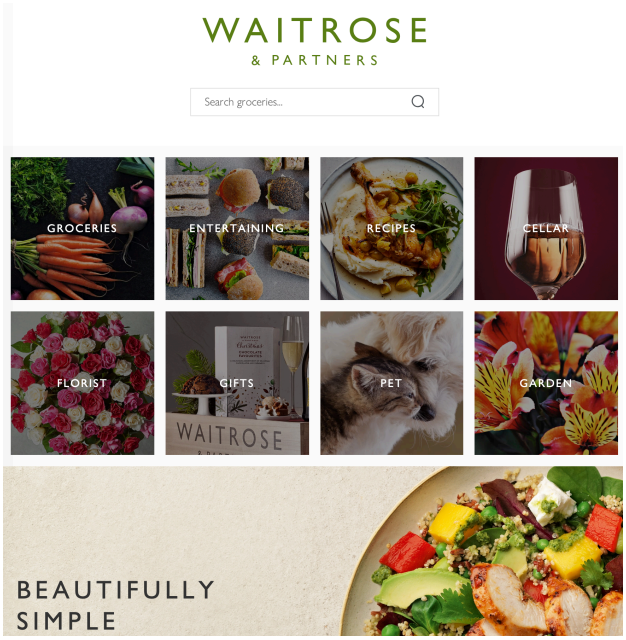
Where will the next Waitrose be built?

Kevin Brown

Disclaimer: Waitrose are unaware of this work. I have no connection with Waitrose or John Lewis.

1.0 Introduction

It is well known that supermarkets choose their locations to fit with their target market – and the location of the supermarket is vital to the success of the shop. The 'low cost' supermarkets will target different areas to 'premium' supermarkets. But what makes the ideal location for supermarkets? Is it possible to look at where supermarkets are located, understand the neighbourhoods of which they are a part, and then find similar neighbourhoods across the country where they could be placed?



Waitrose is a supermarket in the UK. Compared with other supermarkets, it is perceived to be more expensive, associated with the more well-off and linked to Conservative-run councils - and can even increase house prices, more than other supermarkets.

In this project, I will aim to find the location of all current Waitrose stores and find the characteristics of the area to find which ones are most important, by:

- Finding the name and postcode of all stores
- Finding the lat/lon of all stores
- Finding the political constituency of all stores
- Finding the deprivation index of all stores
- Finding the local amenities for each of the stores
- Exploring if there is any correlation between the data (compared with National data) for Waitrose stores, and trying to find similar areas in the UK where new stores could be built.

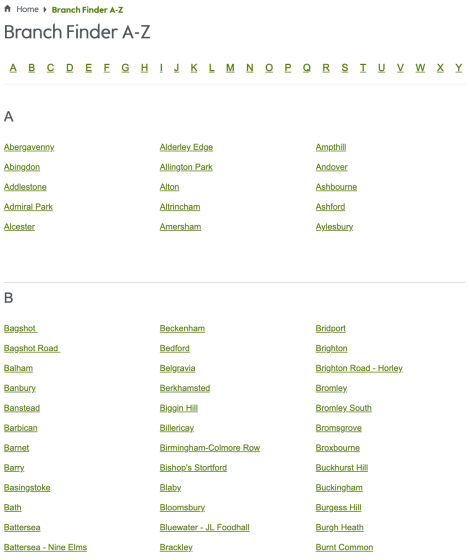
In order to predict where the best places are for Waitrose stores to be built I will:

- Filter the list of postcodes by any correlating factors that are found, to potential neighbourhoods of interest
- Looking at the local amenities for those areas, use K-means to determine which of the non-waitrose areas are most closely aligned to the existing waitrose areas, and present some suggested locations.

2.0 Data

The data required will come from a variety of sources, and be useful in different ways:

Store Data

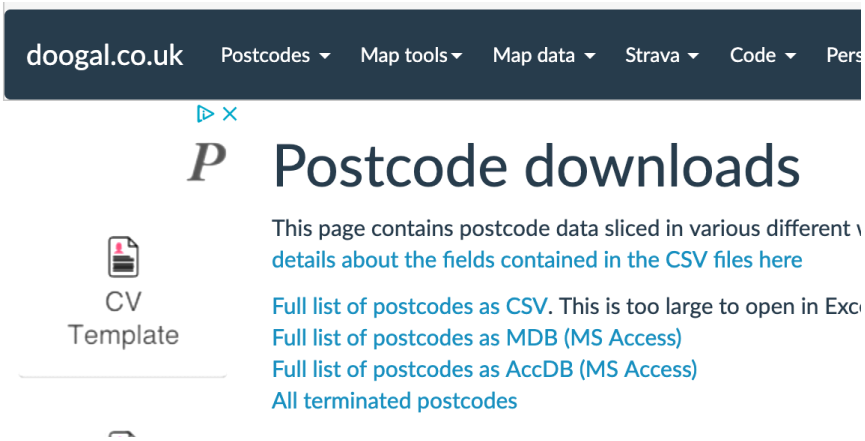


Waitrose A-Z store finder has a list of stores, which I can scrape, in order to get the full list of stores, and the URL of that store's web page.

Each Waitrose store has its own web page. I'll need to use the URL to visit the web page for each of the stores, and scrape the postcode and store details.

The features from the A-Z store page will be store name, and URL.
The features required from the store details will primarily be Postcode – but I will capture the Phone number and county, in case it's needed for display purposes later.

Postcode Data



There is a python library call postcodes, which I explored for mapping postcodes to Lat/Lon, however it relied on a web service which is no longer active, so the python library does not work.

Therefore, I will use the site doogal.co.uk to download a csv mapping of postcodes to Lat/lon, Parliamentary Constituency, Deprivation index or crime index for that postcode

As described, the features required from the Postcode csv will be looked up using the Postcode of the store, and will consist of the Latitude, Longitude, Constituency ID, Index of Multiple Deprivation. The field description page describes how the deprivation rank for the postcode, is a number where 1 is the most deprived. The range of values is as follows:

1 - 32844 = England

1 - 1909 = Wales

1 - 6976 = Scotland

1 - 890 = Northern Ireland

Political Data

Having retrieved the Constituency ID for the postcode, I will need the last election results (2017) to give me the current political party for the constituency under which the Waitrose store currently sits. This is available from the UK Parliament, as a downloadable csv.

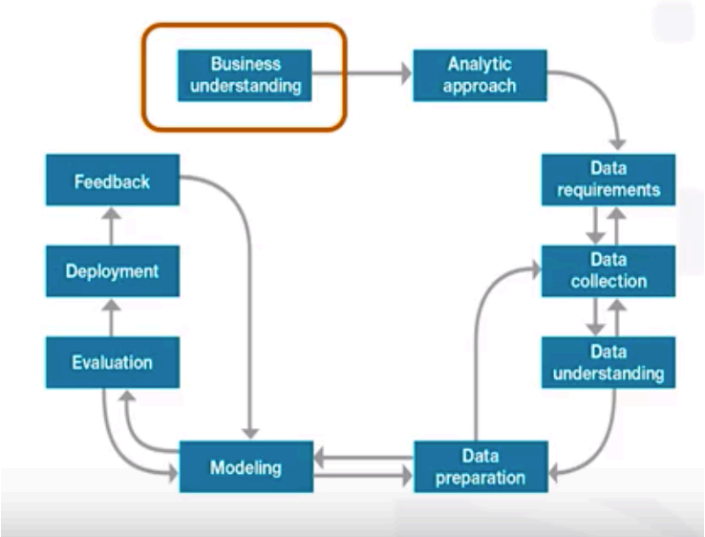
Neighbourhood / Venue data

From experience, it is usual for Waitrose to be located in towns with lots of local amenities, rather than in a trading estate, or in the countryside etc., Therefore, we'll review the profile of existing Waitrose stores to see what venues they have around them, and then compare them to some proposed sites, to see which are most similar. The top local amenities around the supermarket, or any UK location are available from Foursquare.

3.0 Methodology

The approach to this project was to based on the methodology outlined in the 'Data Science Methodology' course on Coursera, and consisted of several stages. The business understanding and data requirements are addressed in the first two sections of this report, and the Evaluation and Feedback sections are covered in sections 4, 5 and 6

Analytic Approach



The problem chosen was to predict the location of a new Waitrose Store. In order to to that, we need to understand in more detail the location of existing stores. Therefore, there are two main analytic approaches used:

- Diagnostic (statistical analysis) which will help us understand where Waitrose have already placed their stores, and why they have placed them there.
- Predictive – which will help us to forecast where they might place their stores in the future.

Data Collection, Understanding and Preparation.

As described by the diagram, data collection, understanding and preparation was done iteratively:

Iteration 1 – Store details

In the first part of this iteration, I used BeautifulSoup to extract the Store name from the Waitrose web page. The only other data that was available about the store, was the URL to the store details page. At the end of this stage, my dataframe consisted of Store name and URL.

	Store name	Store URL
0	Abergavenny	https://www.waitrose.com//content/waitrose/en/bf_home/bf/683.html
1	Abingdon	https://www.waitrose.com//content/waitrose/en/bf_home/bf/211.html

Data cleaning needed to be done. Some Store URLs were absolute, and some relative – so the domain needed to be added. In addition some URLs started /content/waitrose/en/bf_home – and some started /bf_home. This did not matter however as both worked OK as URLs.

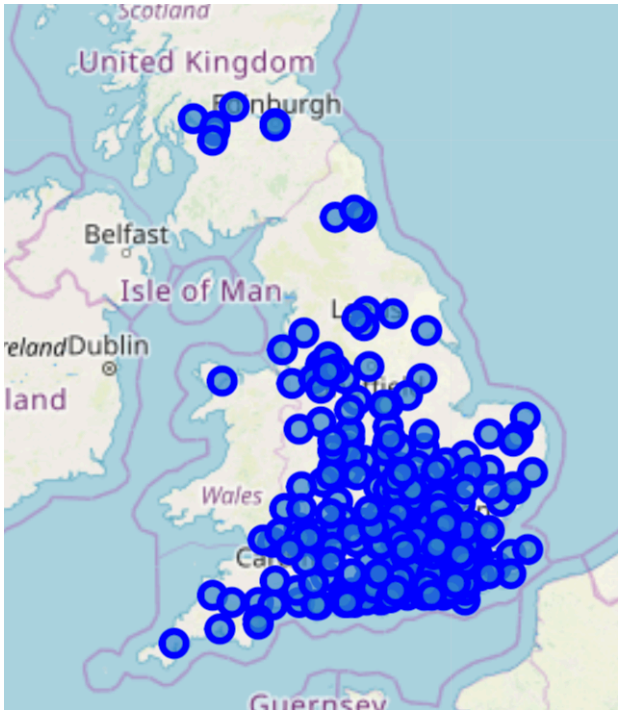
I then used the URL obtained in the first iteration to scrape the Store page, to pull back the Postcode, phone number and county. At this point I found that the Peterborough store, which was on the A-Z list, had been closed, so I removed this from my list. At the end of this stage, I had a dataframe with the extra fields, which, to save lots of web-scraping in future, I saved into its own csv file:

Store name	Store URL	Area	Postcode	Phone
Abergavenny	https://www.waitrose.com//content/waitrose/en/...	Gwent	NP79LL	0187385118
Abingdon	https://www.waitrose.com//content/waitrose/en/...	Oxfordshire	OX143HL	0123553500

Iteration 2 – Lat/Lon

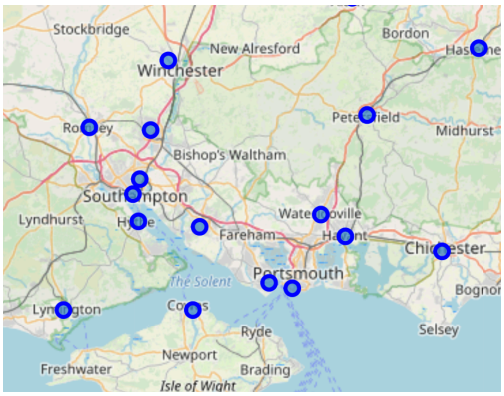
In this iteration, I needed to use the Postcode, to get Lat/Lon for each store and add columns to dataframe. I loaded the csv of Postcodes into a dataframe, and reviewed the columns. I considered if “Distance to train station”, “Police Force name”, “Ward name” etc. would be useful. The only features selected for use were Latitude, Longitude, Constituency and Index of Multiple Deprivation (see Section 2 for details). Very little data cleaning was required to join the postcode dataset with the Store data set:

Store name	Store URL	Area	Postcode	Phone	Postcode district	Constituency Code	Index of Multiple Deprivation	Latitude	Longitude
Abergavenny	https://www.waitrose.com//content/waitrose/en/...	Gwent	NP79LL	0187385118	NP7	W07000054	1042.0	51.818294	-3.028245
Abingdon	https://www.waitrose.com//content/waitrose/en/...	Oxfordshire	OX143HL	0123553500	OX14	E14000874	23835.0	51.672083	-1.279705

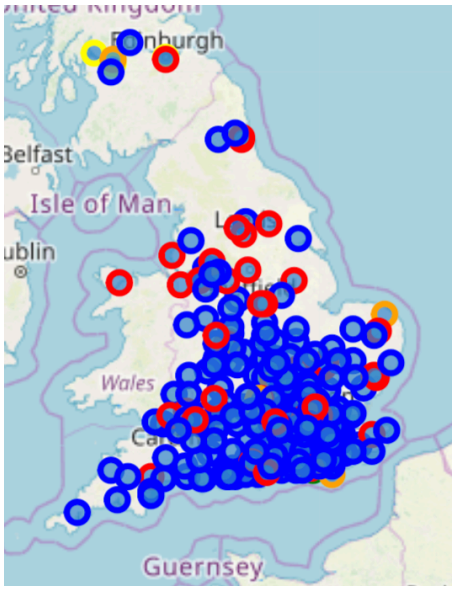


To help understand the data, I plotted the 350 locations of the stores on a map. This shows that there are very few Wales, Scotland or Northern Ireland.

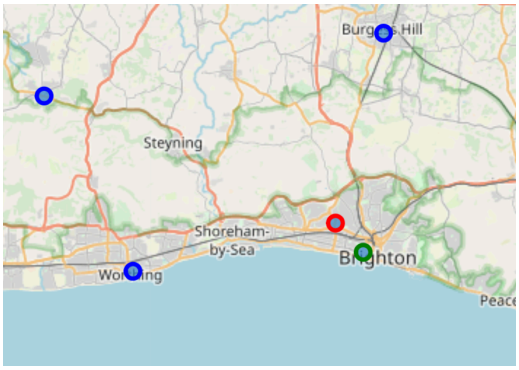
On zooming in more closely to my local area, I can confirm the stores are where they should be.



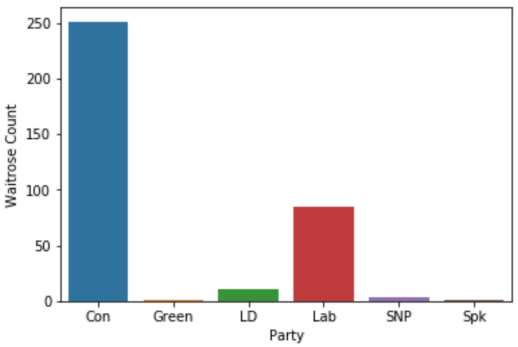
Iteration 3 – Political Data



While the map above is useful, it doesn't tell us much about the stores. This iteration pulls in data from UK Parliament about the party represented in each constituency. Again the dataset was pretty complete, so no real data cleansing issues to merge this with the store data. We could then understand the data a bit more by showing the stores where the colour represented the politics of the area.



The predominant colour on the map is blue (Conservative). The chart confirms this in numbers, showing that there are far more stores in Conservative areas than other areas. However, this does not confirm a bias toward Conservative areas, because it may be that there are just more Conservative areas across the country.

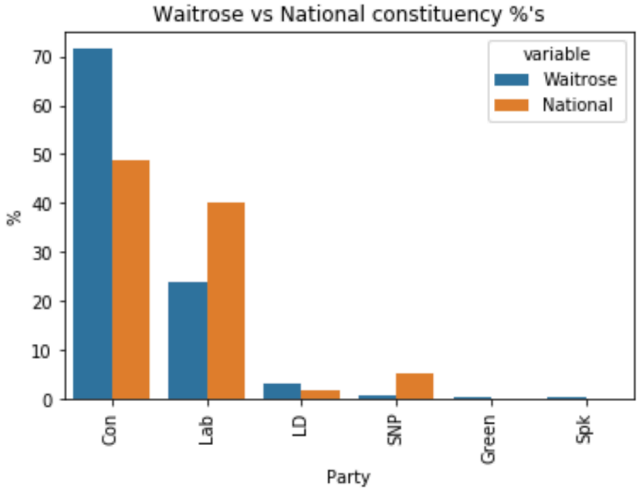


Iteration 4 – Political Data for the Country

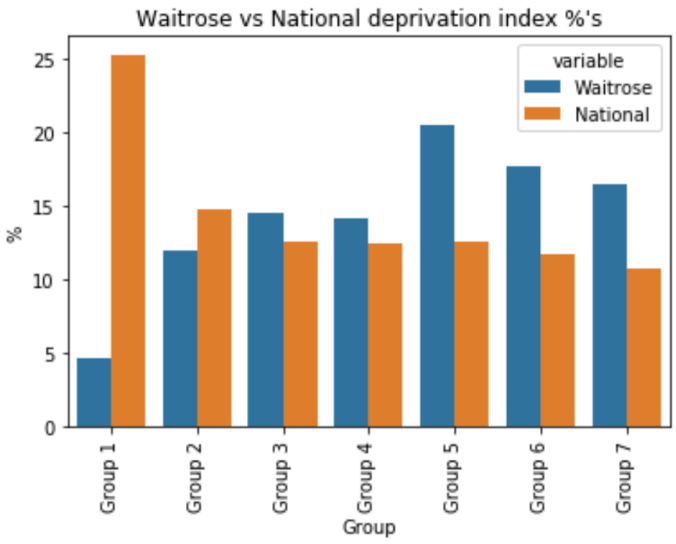
We therefore need to obtain the political results for the whole country, and determine whether these results are in proportion, or if there is a bias. Once the raw numbers were obtained, the percentage of total neighborhoods (and percentage of total stores) was calculated so that the data could be compared. For this project, I used Seaborn charting library. This requires use of the pandas melt function to convert the dataframe from short format (left) to long format (middle). From the chart, it can be seen that there is definitely a skew towards conservative-held areas.

	Party	Waitrose	National
0	Con	71.5	48.8
1	Green	0.3	0.2
2	LD	3.1	1.8
3	Lab	23.9	40.3
4	SNP	0.9	5.4
5	Spk	0.3	0.2

	Party	variable	value
0	Con	Waitrose	71.5
3	Lab	Waitrose	23.9
2	LD	Waitrose	3.1
4	SNP	Waitrose	0.9
1	Green	Waitrose	0.3
5	Spk	Waitrose	0.3
6	Con	National	48.8
9	Lab	National	40.3
10	SNP	National	5.4
8	LD	National	1.8
7	Green	National	0.2
11	Spk	National	0.2



Iteration 5 – Deprivation Index



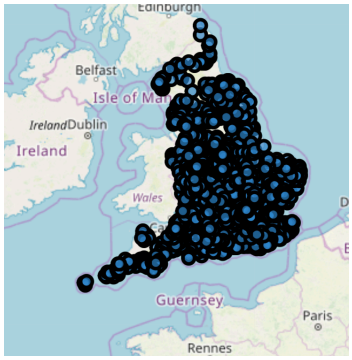
The deprivation rank is given for each postcode, where 1 is the most deprived, and the least deprived is 32844 (in England).

I again wanted to compare the locations for Waitrose stores with the national picture. I therefore binned the deprivation index into 7 groups, with Group 1 being most deprived.

I then repeated the charting methodology above – the resulting chart shows the percentage of stores per deprivation group (blue), and the percentage of postcodes per deprivation group (orange).

From this chart, we can see that there are far fewer stores in the most deprived areas (Group 1) and proportionally more stores in Groups 5, 6 and 7. From this we can conclude there is a bias (intended or not) towards placing Waitrose in the least deprived areas.

Modelling



At this point, we make our first entry into Modelling, though we return to data collection soon, then back to modelling later. We now understand enough about the Waitrose stores, to suggest a broad set of initial locations for new stores: Those postcodes which have a conservative constituent, and those which have a deprivation index that fits into Group 5, 6 or 7.

I also decided that predicting per postcode was too specific (given there is usually one or two postcodes per road, and added the postal district (often an area of 1 or 2 square km), from the postcodes CSV, into our store data.

This analytical based modelling gives us our first representation of potential sites- but even with postal districts, there are almost 1200 of these locations, so we need to collect more data about the area, and use these features in order to better predict locations.

Data Collection, Understanding and Preparation.

Iteration 6 – Foursquare data for proposed sites

This iteration involved using the Foursquare API, for each lat / lon of our proposed sites, to collect information about the facilities (or 'venues') in each area - before then clustering these and fitting the Waitrose stores into the clusters. The cluster where most Waitrose stores fit (ie most similar to existing neighborhoods which contain Waitrose) is where we should look to build new stores.

To start, we collected the data from Foursquare. After the API failed to respond once or twice towards the end of my 1200 requests, I resorted to collecting small batches of data, and appending them together to form a dataframe, which I saved as a CSV, to avoid the same difficulties again later:

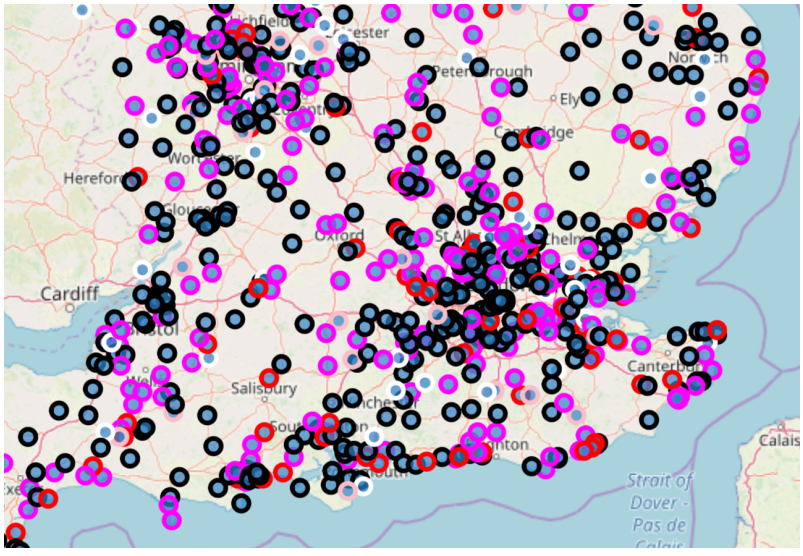
	Unnamed: 0	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	AL1	51.748475	-0.320944	Med Grill	51.752137	-0.320707	Mediterranean Restaurant
1	1	AL1	51.748475	-0.320944	The Crown	51.751278	-0.323276	Pub
2	2	AL1	51.748475	-0.320944	Lebanese Kitchen	51.751080	-0.323537	Lebanese Restaurant
3	3	AL1	51.748475	-0.320944	Chilli Raj	51.751223	-0.323385	Indian Restaurant

The dataframe was grouped by Neighbourhood and most common venues varied widely, as expected from a (reasonably) diverse set of locations.

	Neighborhood	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
0	AL1	Italian Restaurant	Lebanese Restaurant	Building	Café	Fast Food Restaurant	Grocery Store	Mediterranean Restaurant	Chinese Restaurant	Breakfast Spot	Pub
1	AL10	Indian Restaurant	Coffee Shop	Clothing Store	Sandwich Place	Hotel	Fast Food Restaurant	Bubble Tea Shop	Bookstore	Mobile Phone Shop	Movie Theater
2	AL2	Fish & Chips Shop	Grocery Store	Liquor Store	Park	Exhibit	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field
3	AL4	Fish & Chips Shop	Grocery Store	Chinese Restaurant	Home Service	Gym / Fitness Center	Financial or Legal Service	Factory	Falafel Restaurant	Farm	Farmers Market

One-hot encoding was then used to turn the data into numbers so it could be used by the K-mean clustering algorithm.

Modelling



The K-means clustering was applied to split the data into 5 clusters, which were then plotted on the map. Each cluster had different, sets of venues to each other, but, looking at the data, they were similar within group. (See next page)

It was interesting to hypothesize which of the clusters our existing Waitrose stores would fall in. Would they fall across a wide spectrum, or fall into one category?

Proposed site clusters

Cluster 0 – Convenience stores and food?

1061	51.111206	39.0	15.0	84.0	NaN	21279.0	1.0	0.0	0.679196	0.0	Convenience Store	Pizza Place
1081	51.554608	65.0	26.0	18.0	6.0	24754.0	1.0	0.0	0.632117	0.0	Convenience Store	English Restaurant
1108	51.847845	13.0	5.0	65.0	NaN	29634.0	1.0	0.0	6.507675	0.0	Convenience Store	Food

Cluster 1 – Health and Leisure?

	Latitude	Population	Households	Altitude	London zone	Index of Multiple Deprivation	Quality	User Type	Distance to station	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
1	52.648906	26.0	11.0	72.0	NaN	25758.0	1.0	0.0	2.361065	1.0	Pub	Yoga Studio	Fishing Spot	Falafel Restaurant
5	52.435366	47.0	19.0	117.0	NaN	23973.0	1.0	0.0	1.112000	1.0	Pub	Yoga Studio	Fishing Spot	Falafel Restaurant
9	52.255890	19.0	8.0	128.0	NaN	22936.0	1.0	0.0	5.924595	1.0	Pub	Fish & Chips Shop	Yoga Studio	Fishing Spot
9	51.023754	7.0	3.0	97.0	NaN	21856.0	1.0	0.0	3.118720	1.0	Pub	Yoga Studio	Fishing Spot	Falafel Restaurant

Cluster 2 – Outdoor venues and groceries?

87	53.850979	23.0	10.0	133.0	NaN	23716.0	1.0	0.0	0.886335	2.0	Grocery Store	Pub	Indian Restaurant
90	53.827513	29.0	12.0	194.0	NaN	20850.0	1.0	0.0	1.430700	2.0	Grocery Store	Park	Soccer Field
95	54.071605	11.0	6.0	162.0	NaN	20064.0	1.0	0.0	0.640313	2.0	Grocery Store	Bakery	Food & Drink Shop
104	50.725631	42.0	18.0	8.0	NaN	21734.0	1.0	0.0	1.144200	2.0	Track Stadium	Grocery Store	Liquor Store

Cluster 3 – Services? Miscellaneous?

67	51.111565	13.0	6.0	27.0	NaN	23100.0	1.0	0.0	9.768610	3.0	Park	Pub	Yoga Studio
68	51.090155	16.0	7.0	86.5	NaN	30342.0	1.0	0.0	1.173485	3.0	Home Service	Pub	Thai Restaurant
73	53.824325	16.0	7.0	187.0	NaN	28246.0	1.0	0.0	3.556480	3.0	Construction & Landscaping	Pub	Yoga Studio

Cluster 4 – Restaurants, coffee shops and pubs/bars

	Latitude	Population	Households	Altitude	London zone	Index of Multiple Deprivation	Quality	User Type	Distance to station	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	51.748475	33.0	14.0	97.0	NaN	29152.0	1.0	0.0	0.967912	4.0	Italian Restaurant	Lebanese Restaurant	Building
1	51.759492	45.0	15.0	79.0	NaN	22209.0	1.0	0.0	1.917680	4.0	Indian Restaurant	Coffee Shop	Clothing Store
9	51.733807	21.5	10.0	93.0	NaN	27777.0	1.0	0.0	0.723527	4.0	Train Station	Bar	Breakfast Spot

Iteration 7 – Foursquare data for Waitrose sites

As with the proposed sites, we used the Foursquare API, for each lat / lon of our proposed sites, to collect all the data about venues around the Waitrose sites

Looking at the data, it's clear to see the kinds of venues that Waitrose like to have nearby – Pubs, Coffee Shops, restaurants:

```
waitrose_venues.groupby('Venue Category').count().sort_values(['Neighborhood'], ascending=False)
```

	Unnamed: 0	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category							
Pub	819	819	819	819	819	819	819
Coffee Shop	786	786	786	786	786	786	786
Café	456	456	456	456	456	456	456
Italian Restaurant	389	389	389	389	389	389	389
Hotel	358	358	358	358	358	358	358
Grocery Store	293	293	293	293	293	293	293
Pizza Place	270	270	270	270	270	270	270
Bar	232	232	232	232	232	232	232
Sandwich Place	222	222	222	222	222	222	222

We one-hot encoded the Waitrose venue data in preparation for running kmeans.predict() against the model prepared with our proposed site data.

Some features (venues) obtained with this Waitrose location data set were not in the proposed site data set – and some in the proposed site data set were not in the Waitrose dataset. In order to use our K-means model to predict, these needed to be the same.

Therefore, we added features to the Waitrose location data that were not present, by setting these to 0, and removed some features that were extra. We were now able to fit the data to the model, and view the cluster that each site fitted into.

```
# run k-means clustering
kmeans_waitrose = kmeans_proposed.predict(waitrose_grouped_cluster)
# check cluster labels generated for each row in the dataframe
kmeans_waitrose
```

```
array([4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 0, 4, 4, 4, 4, 4, 4, 4, 4, 4, 0, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4],
      dtype=int32)
```

Most of the Waitrose stores fitted into Cluster 4 (Restaurants, coffee shops and pubs / bars), with some fitting to Cluster 0 (Convenience stores / food). This was a surprising result initially, but looking at the expected venues above, this closely matches the description we gave to cluster 4.

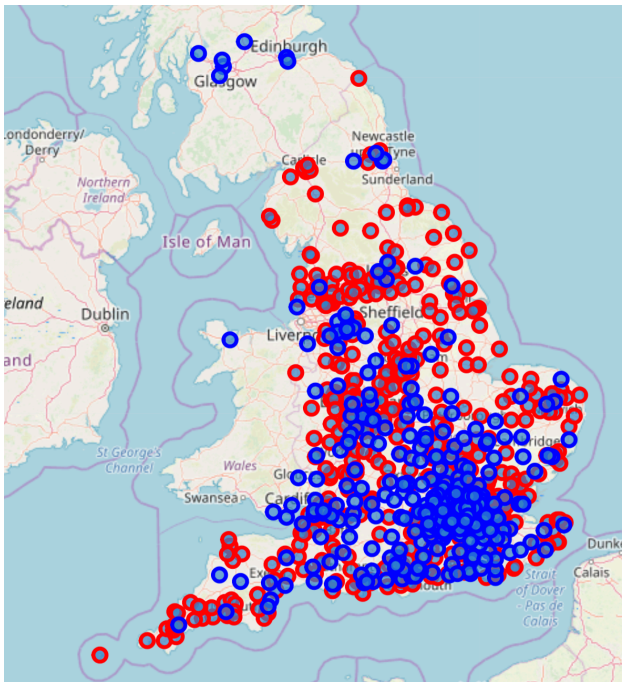
Given this result, we were able to remove clusters 0-3 from the proposed site data set, and focus on cluster 4. I plotted the results onto a map – as shown in the results section.

4.0 Results

A number of conclusions were reached during the project, including:

- Waitrose has around 350 stores across the UK, but these are mainly in England; the store does not feature strongly in Wales, Scotland or Northern Ireland.
- There is a bias towards Conservative run areas
- There is a bias towards areas which are least deprived
- Waitrose stores are often found near pubs, coffee shops and restaurants.

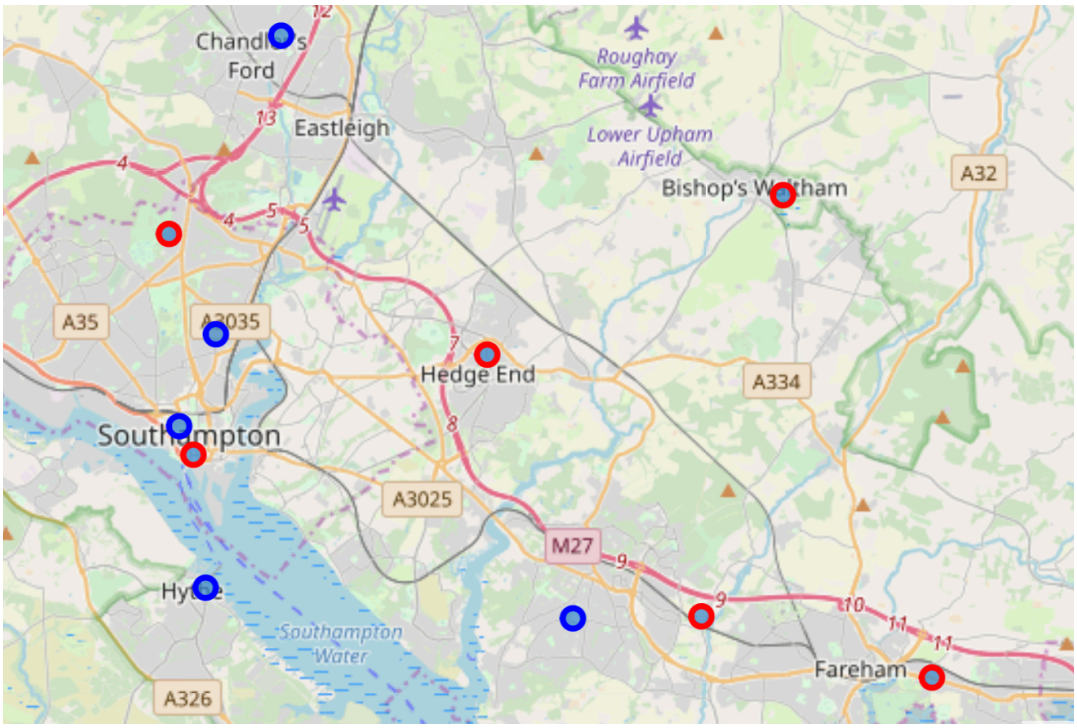
Bishops Waltham



Having then applied the K-means model (created with the proposed locations) to the Waitrose venue data set, I found that almost all of the stores fitted into Cluster 4. I removed cluster 0-3 from the data set and plotted the map of the UK shown here. The existing stores (blue) were placed alongside the suggested stores (the output from our prediction) with our new refined set of predicted sites.

I then zoomed in, to my local area, and found the red dots showing suggested Waitrose locations. I noticed that Bishop's Waltham and Hedge End were proposed sites and was surprised that they didn't already have a Waitrose store, and did seem a good choice.

Other locations would be too close to existing stores – some distance measure from existing stores would also be helpful in refining the model.



5.0 Discussion

There are a number of observations I would make about different aspects of the project

- **Extensions to the project:** There would be a lot more that could be done to refine and check the model.
    - Using more data - correlating with crime statistics and other neighbourhood features
    - Taking into account the financial results of each store (with access to confidential data) so only the best stores could be modelled
    - Knowing which stores have been closed (and therefore have not succeeded, in order to provide a supervised learning model with counter-examples.
    - Taking into account the proximity to other stores, delivery depots, supply chains, size of stores etc., to make the results more compelling.
- **Validation:** With more time, and a real paying customer, it would be possible to spend as long a time as budget allowed refining the model. One aspect that could be more thoroughly undertaken is the validation of results, through cross-checking with different data sets, looking at variance, correlations, p-Values, to help validate results.
- Overall however, the results seemed sensible, other than suggestions which are too close to other stores.

6.0 Conclusion

It would be possible to keep refining the model to get from the 500 current positions down to the best one. However, no business is going to just build a store where the computer tells them to build a store, so the business will always need to have a range of options, not least because the model would never be perfect – there are always factors outside the knowledge of the model at play (market conditions, news events, etc.)

However, in this project I set out to predict some good locations for Waitrose stores, using a range of data science techniques. Through following the Data science methodology described at the start of section 3, I chose statistical methods, and k-means supervised learning as an analytics approach. I researched the data requirements, and iteratively collected, visualised, re-visited the requirements, and then created the first Model. I then returned to data collection, analysis, visualising , and refining the mode. It’s clear that this process can continue, so another key factor is knowing when to stop refining.

Thank you for reviewing this document.