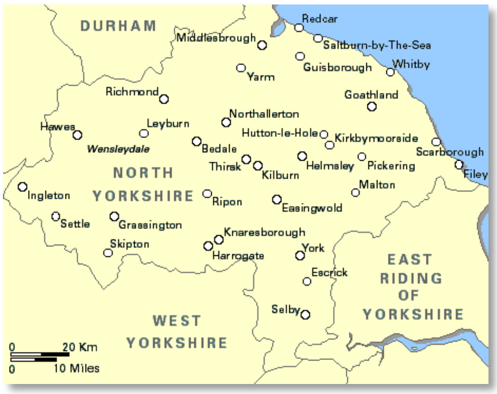
North Yorkshire Cycling:
Identifying the Best Town for a New Bike Shop

INTRODUCTION:



Cycling in North Yorkshire is a hugely popular sport and attracts both locals and enthusiasts from all over the world. Last September the World Championships were held in the spa town of Harrogate, whilst the county has previously hosted the prestigious Tour de France Grande Depart. The Tour de Yorkshire takes place every year and is one of the most loved events in the Yorkshire's sporting calendar. With large events like this inspiring people to cycle there is a huge demand for bikes and cycling shops. Therefore, for this project I take the approach that I am starting my own cycling business and want to open my first shop. I have decided that North Yorkshire is the place to be but am open to starting my business in any of the towns within the county.

Recently, the global pandemic has resulted in an uptake of people cycling as many are avoiding public transport and are now travelling on two wheels. This also presents an opportunity to start a cycling business, as there is definitely sufficient demand in the market for bicycles and related services.

Considering the above, I want to find the optimal location to place my bike shop. I want to choose a location that is not already saturated with shops, which has the opportunity to offer a new service. However, I also want there to be sufficient demand and do not want to place a shop in a location where I shan't be able to make any sales, even if there are currently no bike shops there.

DATA:

For this project I will need to utilise several data sources. Firstly, I will need to scrape the populations of the North Yorkshire towns from Wikipedia, to understand the number of people living in each and the estimated potential market size. I will need to use a location service to acquire the longitude and latitude of the towns so that I can input both into the foursquare API.

I will then use the foursquare API to access several pieces of information. The first aspect I will look at is how many bike shops there are in each town. It will also be interesting to see how this relates to the populations of each town. Next I will also choose to look at the number of sports shops, as many of these will sell bikes and equipment/kit. I also want to look at the current transport options. I want to see if there is a relationship between public transport options, the number of bike shops, and the populations.

Using this data, I want to cluster the towns to see if I can find groupings and relationships. I hope that one specific cluster will stand out as being the one that suggests where I should place my shop. My hypothesis is that the required cluster will have few bike shops per population. It will be interesting to see how any correlations with public transport can be related to my findings.

METHODOLOGY:

BeautifulSoup The first step of my project was to scrape the populations of the North Yorkshire towns. I did this using the BeautifulSoup Package. I was able to create a DataFrame with the 20 largest towns, as shown on the Wikipedia page, https://en.wikipedia.org/wiki/North_Yorkshire. I will choose one of these towns to be the location of my bike shop. All code for this project can be found on my GitHub page. The link to my notebook is:

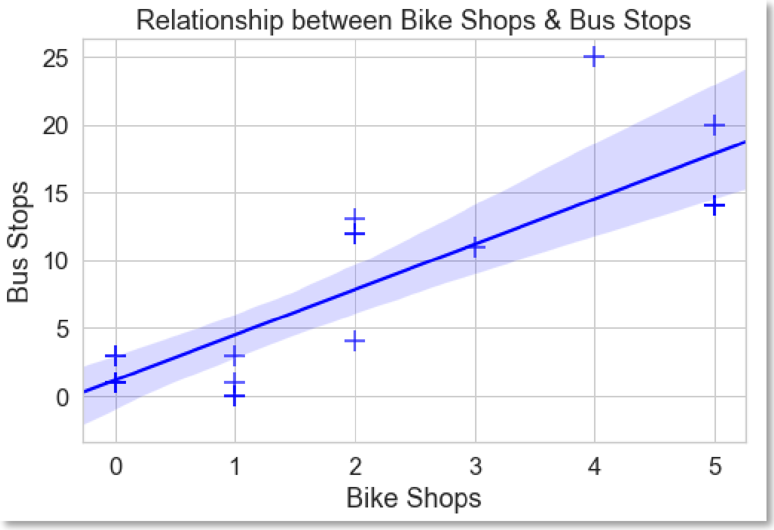
https://github.com/MattHipkin/Coursera_Capstone/blob/master/Final%20Project%20-%20North%20Yorkshire%20Cycling.ipynb

The next step was to use the OpenCage Geocoding API to retrieve the longitude and latitude for each town I retrieved from the Wikipedia page. This service gives you a free trial with a restricted number of calls.

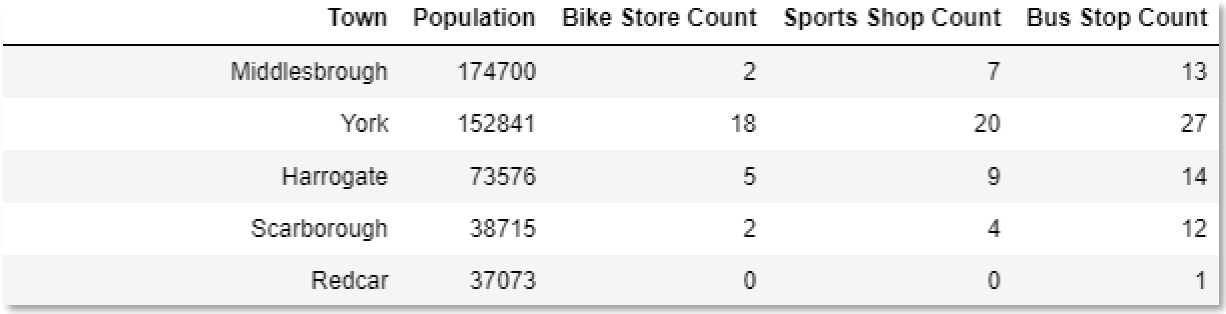
Having acquired the top 20 towns by population in North Yorkshire and their respective Latitude and Longitudes, I used the foursquare API to retrieve the number of bike stores, the number of sports shops, and the number of bus stops for each town.

I chose to retrieve the number of bus stops per town, as my way of adding public transport into the investigation. My reasoning is that in the smaller towns of North Yorkshire the main method of public transport is by bus. Compared to larger cities there are fewer rail networks and these are mainly used for long distance and between city travel. Due to the pandemic there may be a large volume of people who no longer want to travel on public transport and may feel that traveling by bike in the open air is safer. I have also assumed that where there are less bus stops, people may travel by car or already cycle so there may be less of a market in these locations.

As I am interested to see if there is a relationship between Bike Shops and Public Transport, I used Linear Regression to see if there is a relationship between the number of bike shops and the number of bus stops. The resulting R2 score was 0.61 which shows that the model is a good fit and that a relationship can be modelled. I improved this even further by trimming York out of the data, as it is a large outlier compared to the other towns. The resulting R2 value rose to 0.71. The resulting regression plot is as follows:



The final table that I produced, having scraped the data and pulled the information I needed from both the OpenCage and foursquare APIs can be seen below *(showing first 5 rows)*:

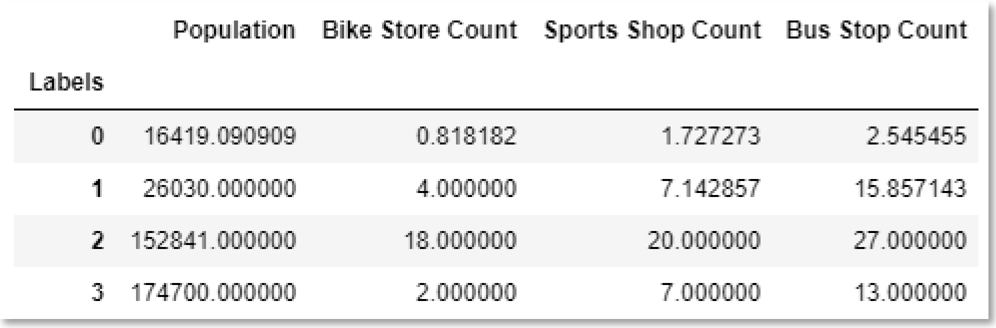


Town	Population	Bike Store Count	Sports Shop Count	Bus Stop Count
Middlesbrough	174700	2	7	13
York	152841	18	20	27
Harrogate	73576	5	9	14
Scarborough	38715	2	4	12
Redcar	37073	0	0	1

Using this data, I carried out k-means clustering to group the towns and to apply a label to each. I chose a cluster size of 4 and normalised the data so that the clustering was not skewed by the large differences in population sizes. I used 12 rotations, starting the algorithm from 12 different random starting points and selecting the result with the lowest error.

RESULTS:

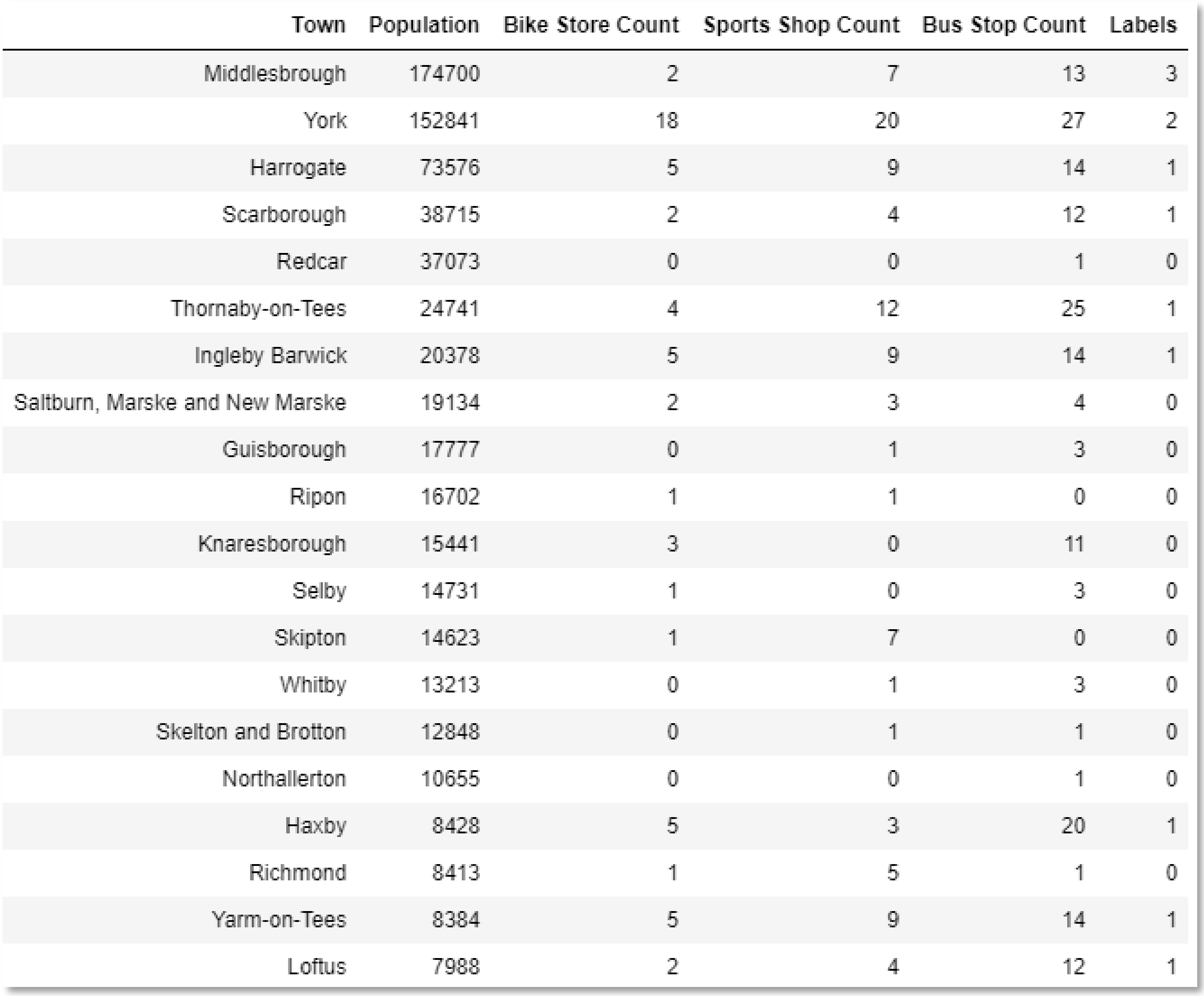
The result of my clustering is four clearly defined groups. The below table shows the label and the average (mean) value for each of the columns:



	Population	Bike Store Count	Sports Shop Count	Bus Stop Count
Labels				
0	16419.090909	0.818182	1.727273	2.545455
1	26030.000000	4.000000	7.142857	15.857143
2	152841.000000	18.000000	20.000000	27.000000
3	174700.000000	2.000000	7.000000	13.000000

0: Small Population: Few Bike Stores, Sports Shops and Bus Stops.
1: Small Population: Medium number of Bike Stores, Sports Shops and Bus Stops.
2: Large Population: Many Bike Stores, Sports Shops and Bus Stops.
3: Large population. Few Bike Stores, Sports Shops. Many Bus Stops

Below is the table with each town assigned its label:



Town	Population	Bike Store Count	Sports Shop Count	Bus Stop Count	Labels
Middlesbrough	174700	2	7	13	3
York	152841	18	20	27	2
Harrogate	73576	5	9	14	1
Scarborough	38715	2	4	12	1
Redcar	37073	0	0	1	0
Thornaby-on-Tees	24741	4	12	25	1
Ingleby Barwick	20378	5	9	14	1
Saltburn, Marske and New Marske	19134	2	3	4	0
Guisborough	17777	0	1	3	0
Ripon	16702	1	1	0	0
Knaresborough	15441	3	0	11	0
Selby	14731	1	0	3	0
Skipton	14623	1	7	0	0
Whitby	13213	0	1	3	0
Skelton and Brotton	12848	0	1	1	0
Northallerton	10655	0	0	1	0
Haxby	8428	5	3	20	1
Richmond	8413	1	5	1	0
Yarm-on-Tees	8384	5	9	14	1
Loftus	7988	2	4	12	1

Discussion:

The ideal place to locate a bike shop would be a town where there is a high population and thus ample market size and opportunity for selling bikes and cycling products. This location would not have too many bike stores or sports shops located there already. The town would also have a good public transport network; during this current pandemic, if residents heavily rely upon public transport, then there is more opportunity to sell to those wanting to cycle instead. This description matches to label 3 from my clustering. Considering that there is a linear relationship between the number of bike shops and number of bus stops, it seems that Middlesbrough is an outlier to this trend. Therefore, I further think that there is the option to increase the number of stores and that there is currently untapped potential.

We may also want to consider other towns, for example those that do not currently have many bike shops present. Therefore, we could turn to those towns classified by cluster 0. However, there is a risk here because the population sizes of these towns are small; there may not be an adequate market size in order to fuel a sustainable business.

One aspect that I have not considered which could further expand this investigation, is the willingness of a person to spend cash on a bike, especially considering that the country is entering a recession. I could also consider aspects such as GDP and disposable income. In addition, cycling may be much more popular in some town than others.

Conclusion:

To conclude I would choose **Middlesbrough** as my destination of choice for the new bike shop. It is clear from the results that this town has the most potential for starting a new business. It is a large town that stands out from my investigation as currently having very little providers of bikes and services. To improve my findings, I could expand the number of venues I retrieved from the foursquare API. I also think that if I were to apply my investigation over a larger area, I would also find more suitable towns to start a business in. For example, there may be a town just over the border in another county that I have not considered which is actually geographically very close. This may also improve the accuracy of my k-means clustering as I would have a larger dataset to utilise.