Identifying Suitable Cycle Shop Sites in the City of London

Prepared for: IBM Coursera Data Science Capstone
Prepared by: Moaeed Sajid, Data Scientist
23 August 2020
Proposal number: 123-4567

INTRODUCTION

At the start of May 2020 during the Covid-19 pandemic, Grant Shapps the UK transport secretary announced he was bringing forward publication of the National Cycling Plan to help double cycling and increase walking by 2025. As part of this plan Londons boroughs and Transport for London (TFL) have been allocated £5 million in the phase 1 and an indicated £20 million for the second phase[1]

Later that month TFL announced it's new cycling route map as part of it's Streetspace for London project[2]. This is part of a plan by Londons Mayor to…

- Make it easier and safer for people to keep up social distancing

- Help people walk and cycle more often

- Avoid a sharp increase in car use. If people switch even a fraction of their previous journeys to cars, essential deliveries and emergency services will be gridlocked

- Keep London's air as clean as possible to protect everyone's health and to reduce carbon emissions

The UK government also announced a Fix Your Bike voucher scheme[3] aiming to encourage more people in England to embrace cycling as an alternative to private cars. Anyone with an unused cycle in need of repair can avail of the voucher to get them and their bike back on the road.

Since the Covid pandemic the public are more reluctant to share public transport and as obesity has been highlighted as a key factor in Covid deaths more people have been encouraged to take up cycling.

In fact the Bicycle Associations market data service crunched some numbers revealing that between April and June this year there was an year on year increase of 63%. Sales were strongest for bikes below £1000 and e-bikes.

To build on this popularity and the increase of cycle routes in London, this report is being prepared for potential cycle shop/repair businesses to help identify suitable sites in London to set up a new business.

---

[1] https://www.gov.uk/government/publications/emergency-active-travel-fund-local-transport-authority-allocations/emergency-active-travel-fund-total-indicative-allocations

[2] https://tfl.gov.uk/travel-information/improvements-and-projects/streetspace-for-london?cid=streetspace

[3] https://fixyourbikevoucherscheme.est.org.uk/

DATA SOURCES

I will be using a varying number of data sources for this project. To make sure this work is reproducible we will automate and code as much of this as possible. The data sources that we will be using are as follows…

Foursquare location Data

This project requires that I use Foursquare location data. In this case I will connect to FourSquare using their own API protocol to download a list of cycle shops for each borough. This will require me to identify the category ID for bike shops which are available at https://developer.foursquare.com/docs/build-with-foursquare/categories/

We will be collecting…

1. Shop Name
2. Latitude
3. Longitude

As there is a limit of 50 results with the API I have, I will likely break up the searches by the boroughs boundaries.

London Borough boundaries

In order to produce meaningful maps I will need mapping data for the boundaries of each of Londons boroughs. I have found a JSON file that has a section for each borough and a list of co-ordinates defining it's boundaries. This is available from http://skrange.github.io/data.html

London Boroughs

Wikipedia contains a list of London Boroughs at https://en.wikipedia.org/wiki/List_of_London_boroughs

List of boroughs and local authorities [edit]

Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Nr. in map
Barking and Dagenham [note 1]			Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194,352	51.5607°N 0.1557°E	25
Barnet			Barnet London Borough Council	Conservative	Barnet House, 2 Bristol Avenue, Colindale	33.49	369,088	51.6252°N 0.1517°W	31
Bexley			Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236,687	51.4549°N 0.1505°E	23
Brent			Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317,264	51.5588°N 0.2817°W	12
Bromley			Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317,899	51.4039°N 0.0198°E	20
Camden	✓		Camden London Borough Council	Labour	Camden Town Hall, Judd Street	8.40	229,719	51.5290°N 0.1255°W	11
Croydon			Croydon London Borough Council	Labour	Bernard Weatherill House, Mint Walk	33.41	372,752	51.3714°N 0.0977°W	19
Ealing			Ealing London Borough Council	Labour	Perceval House, 14-16 Uxbridge Road	21.44	342,494	51.5130°N 0.3089°W	13
Enfield			Enfield London Borough Council	Labour	Civic Centre, Silver Street	31.74	320,524	51.6538°N 0.0799°W	30
Greenwich [note 3]	✓ [note 5]	Royal	Greenwich London Borough Council	Labour	Woolwich Town Hall, Wellington Street	18.28	264,008	51.4892°N 0.0648°E	22
Hackney	✓		Hackney London Borough Council	Labour	Hackney Town Hall, Mare Street	7.36	257,379	51.5450°N 0.0553°W	9
Hammersmith and Fulham [note 4]	✓		Hammersmith and Fulham London Borough Council	Labour	Town Hall, King Street	6.33	178,685	51.4927°N 0.2339°W	4
Haringey	[note 3]		Haringey London Borough Council	Labour	Civic Centre, High Road	11.42	263,386	51.6000°N 0.1119°W	29
Harrow			Harrow London Borough Council	Labour	Civic Centre, Station Road	19.49	243,372	51.5898°N 0.3346°W	32
Havering			Havering London Borough Council	Conservative (council NOC)	Town Hall, Main Road	43.35	242,080	51.5812°N 0.1837°E	24

Screenshot from Wikipedia

It is my intention to use the following information from this Wikipedia page in the report…

1. Borough
2. Area
3. Population
4. Co-ordinates

Cycle Routes

TFL have an open data portal at https://cycling.data.tfl.gov.uk/ with a JSON file with the latest cycle routes. This dataset has the name of each route, wether is is open, in progress or planned for the future and a list of co-ordinates mapping the route.

I will use this data to map how much of each route sits within the boundary of each borough.

METHODOLOGY

I will be using a varying number of data sources for this project. To make sure this work is reproducible we will automate and code as much of this as possible. The data sources that we will be using are as follows…

Data Preparation

The majority of time for any data science project is to prepare and clean your data. In this instance the following was performed…

The Wikipedia co-ordinates had a lot of hidden data that was not showing on the Wikipedia page. This included the '/ufeff' text that meant we were not able to extract a clean set of co-ordinates. This was removed and we extracted only the set of co-ordinates shown on the Wikipedia page

As we had the area and co-ordinates of each borough, we calculated the radius (converted to metres). This was used later on in Foursquare to try and narrow down our search of cycle shops to only those within the borough

In order to see the popularity of cycle routes within each borough we downloaded a JSON of all the cycle routes. A calculation was then performed on each co-ordinate to see if it sat within each of the 32 boroughs boundaries. This allowed us to build a data frame showing the density of cycle routes within each borough and whether they were existing, being built or in a planning phase.

Once we had the density of the cycle routes we divided this by the number of shops Wikipedia had given us for each borough giving us a route/shop ratio. We would expect with more cycle paths more shops would be required.

Exploratory Data Analysis

As part of the data analysis we produce box and whisker plots. These are a great way of determining if there are any outliers in our data. For example, in the area sizes for the boroughs, three were selected as outliers. These three, Bromley, Hillingdon and Havering are larger boroughs on the edge of the city.

As these seemed to have a relatively large population I calculated a population density by dividing the population in each borough by it's area size. Once we did this the three boroughs were no longer outliers.

I also produced a folium map showing a visual representation of the data that is to be used. This was a good way to get a visual representation of the data we are expecting to use for this report.
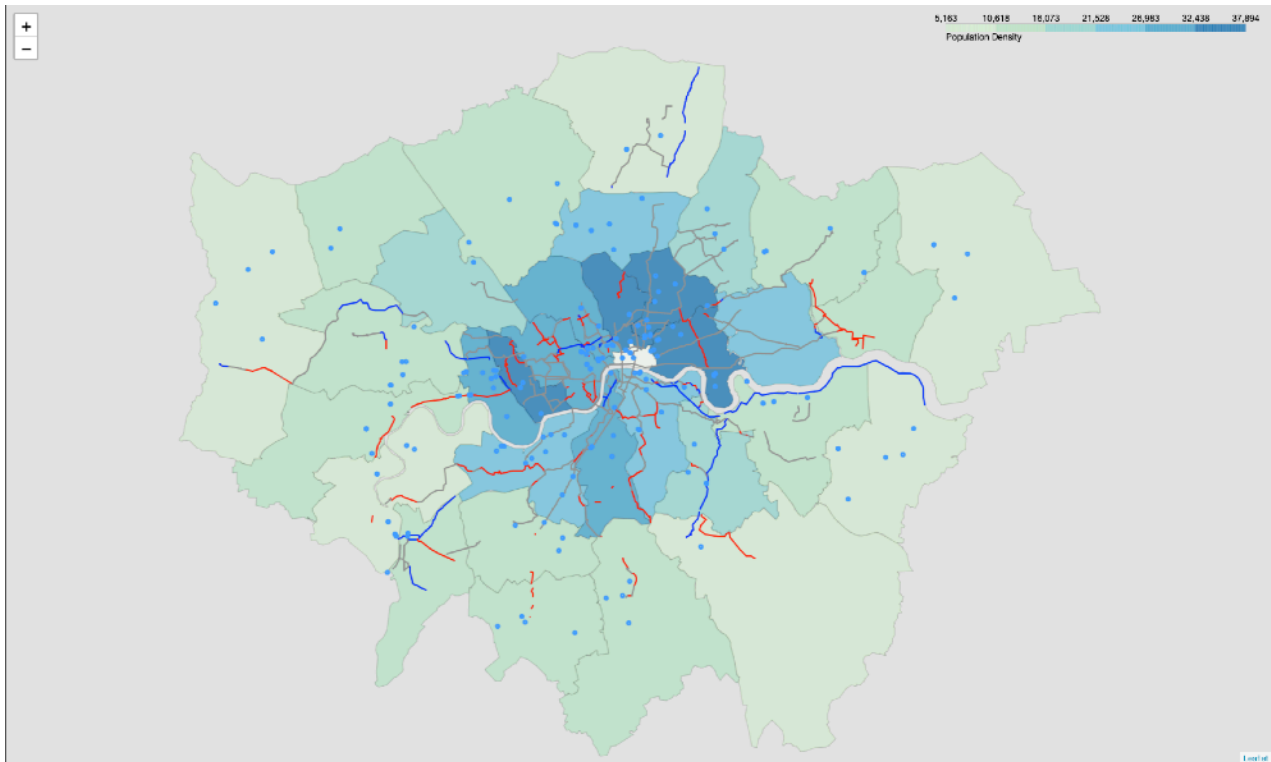
EDA Map generated using Leaflet in Python

This map shows the following....

1. Population Density - the darker areas show more densely populated boroughs. This shows that brought closed to the centre of the city are not only shorter but also much more densely populated.

2. Current Cycle Shops - All the blue dots quickly show all those businesses listed as cycle shops within the FourSquare database.

3. Cycle Routes - These have been split into three groups
    -   Grey Routes - Are those cycle routes that are already in place
    -   Blue Routes - Are routes that are currently being rolled out
    -   Red Routes - These are routes that have been planned for the future but where work on them has not yet begun.

Finally as part of the exploratory data analysis I produced a correlation matrix. The purpose of this matrix is to allow us to identify and variables that have a strong correlation with each other.

If 2 variables are closely correlated with each other then it is better to not use both of them in your model. By using correlated variables you increase their weight on the model for no gain, as well decreasing the effect of other variables and losing out on their importance on your model.
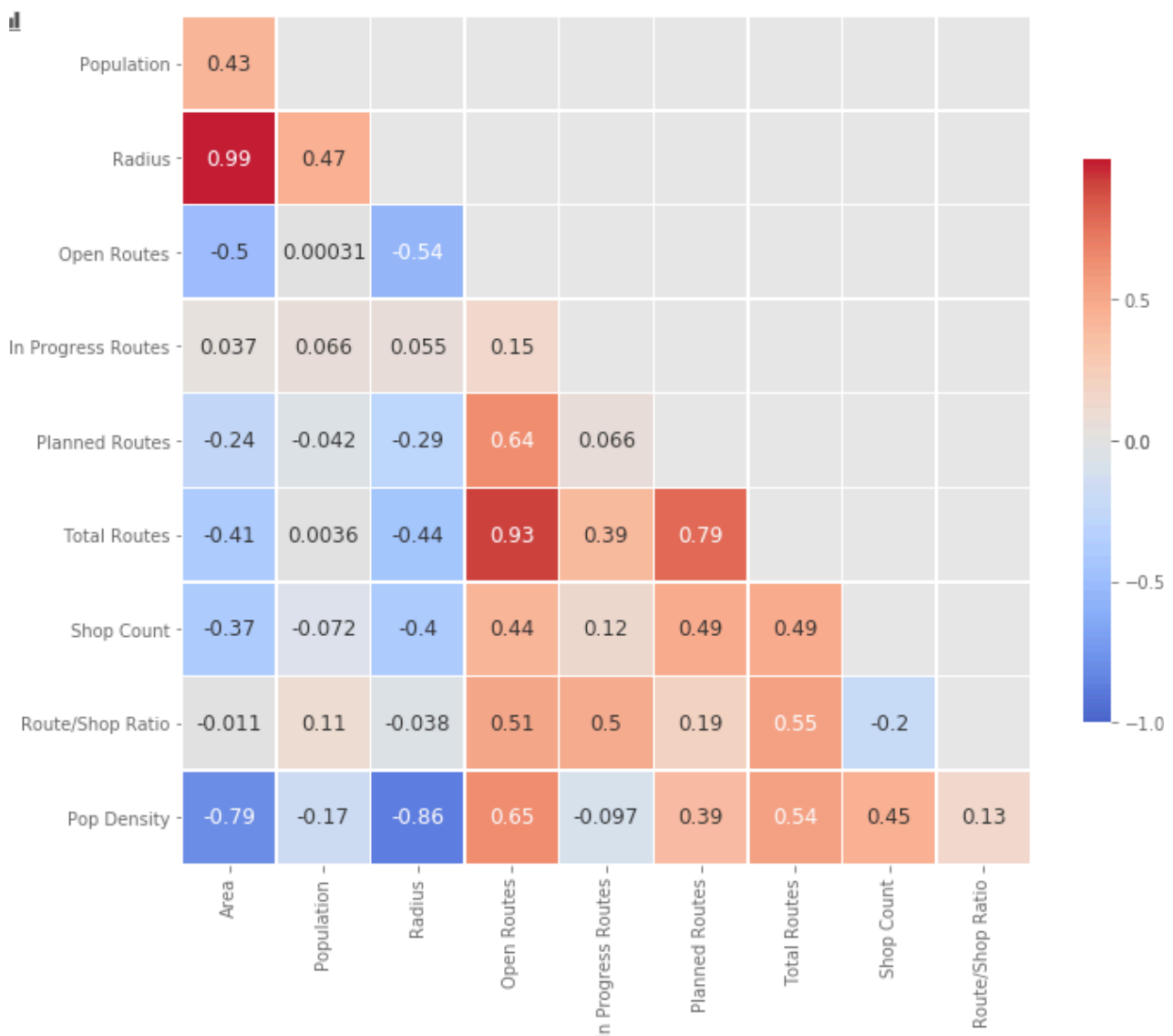
During our model we will look to avoid choosing any values that are highly correlated, including Area with Radius (0.99), Total Routes with open Routes (0.93) or Planned Routes (0.79), Pop Density with Radius (-0.86), and Area with Pop Density (-0.79).

Machine Model

We used K-Means algorithms to identify similar groups of boroughs within the city of London.

This was done by creating a new data frame using the variables selected for our model. I ran the model a number of times with different variables to find a model that was interpretable for our project.

K-means calculates the distance between 2 point by using the Euclidean distance, that is the straight line between 2 points. If one of the features has a wide range of values the distance will be governed by this feature. For this reason we look to feature scaling to normalise all the values so that the mean value of all the variables is 0 and a standard deviation of 1.

Once the data was normalised we need to identify the best number of clusters for our data. This is done by running each model a number of times with 2 to 15 clusters.

For each of these models we plotted the sum of squared distances, that is the differences between our data points and centroids. As the number of K approaches the number of data points our SSD approaches 0 but having each borough in it's own cluster with no grouping is not a good solution.

For this reason the SSD for each K is plotted on a chart and we look to choose a point where diminishing returns are no longer worth the additional cost in computational power or leading to overfitting of the data where noise begins to affect our model. This method is known as the elbow method and we look for the elbow to choose the bet value of K.
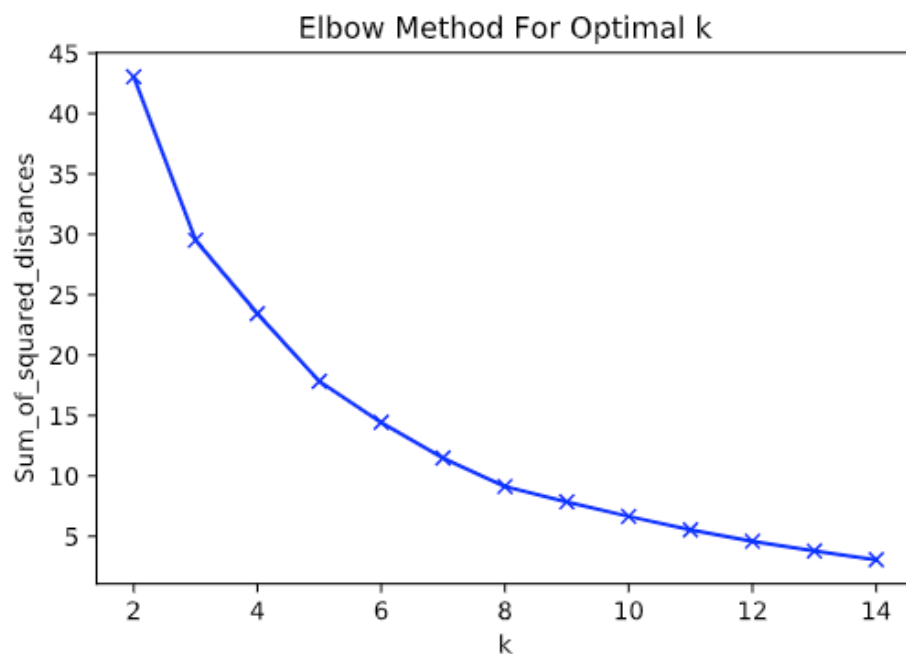
In this case and along with the calculated silhouette score we chose a value for K of 7

RESULTS

Results

With our 7 clusters this report identifies 7 boroughs where a new cycle business would work well. These are split as

Cluster 1 - Enfield

Cluster 3 - Kensington and Chelsea, Lambeth, Tower Hamlets and Westminster

Cluster 5 - Lewisham and Waltham Forest

Cluster Labels	Borough	Area	Population	Radius	Open Routes	In Progress Routes	Planned Routes	Total Routes	Shop Count	Route/Shop Ratio	Pop Density
0	Southwark	11.14	298464	2899.935777	1795	555	1109	3459	24	144.12	26792.10
0	Brent	16.70	317264	3550.618569	426	5	0	431	3	143.67	18997.84
0	Camden	8.40	229719	2518.172214	753	223	235	1211	26	46.58	27347.50
0	Newham	13.98	318227	3248.623415	0	0	0	0	4	0.00	22763.02
0	Wandsworth	13.23	310516	3160.281047	1638	0	872	2510	12	209.17	23470.60
0	Hammersmith and Fulham	6.33	178685	2185.987325	329	77	283	689	14	49.21	28228.28
0	Haringey	11.42	263386	2936.154041	358	0	0	358	5	71.60	23063.57
1	Enfield	31.74	320524	4894.961110	997	798	0	1795	2	897.50	10098.42
2	Barking and Dagenham	13.93	194352	3242.808799	64	10	23	97	2	48.50	13952.05
2	Sutton	16.93	195914	3574.985326	0	0	101	101	4	25.25	11572.00
2	Richmond upon Thames	22.17	191365	4090.992008	340	3	128	471	7	67.29	8631.71
2	Merton	14.52	203223	3310.770621	388	0	67	455	4	113.75	13996.07
2	Havering	43.35	242080	5720.583583	0	0	0	0	4	0.00	5584.31
2	Hillingdon	44.67	286806	5807.025869	129	36	167	332	4	83.00	6420.55
2	Greenwich	18.28	264008	3714.786685	0	69	0	69	4	17.25	14442.45
2	Barnet	33.49	369088	5028.093653	0	0	0	0	9	0.00	11020.84
2	Bexley	23.38	236687	4201.148547	0	589	0	589	5	117.80	10123.48
2	Harrow	19.49	243372	3835.762456	0	0	0	0	2	0.00	12487.02
2	Croydon	33.41	372752	5022.084576	89	0	174	263	4	65.75	11156.90
3	Westminster	8.29	226841	2501.629847	2388	222	1329	3939	12	328.25	27363.21
3	Tower Hamlets	7.63	272890	2399.982381	1335	0	469	1804	5	360.80	35765.40
3	Kensington and Chelsea	4.68	155594	1879.614173	1248	70	529	1847	4	461.75	33246.58
3	Lambeth	10.36	314242	2796.569792	1590	362	447	2399	7	342.71	30332.24
4	Bromley	57.97	317899	6615.267871	0	286	413	699	2	349.50	5483.85
4	Ealing	21.44	342494	4023.075426	628	565	36	1229	6	204.83	15974.53
4	Kingston upon Thames	14.38	166793	3294.770945	417	891	187	1495	6	249.17	11598.96
4	Hounslow	21.61	262407	4038.993627	0	0	881	881	4	220.25	12142.85
4	Redbridge	21.78	288272	4054.849338	533	0	164	697	4	174.25	13235.63
5	Lewisham	13.57	286180	3200.631744	409	728	137	1274	3	424.67	21089.17
5	Waltham Forest	14.99	265797	3363.927296	1614	0	0	1614	3	538.00	17731.62
6	Islington	5.74	215667	2081.621336	797	13	308	1118	9	124.22	37572.65
6	Hackney	7.36	257379	2357.136276	1429	0	108	1537	9	170.78	34969.97

Discussion

Cluster Labels	Area	Population	Latitude	Longitude	Radius	Open Routes	In Progress Routes	Planned Routes	Total Routes	Shop Count	Route/Shop Ratio	Pop Density
0.0	11.6	273752	51.5212	-0.139643	2928.54	757	122.857	357	1236.86	12.5714	94.9071	24380.4
1.0	31.74	320524	51.6538	-0.0799	4894.96	997	798	0	1795	2	897.5	10098.4
2.0	25.7836	254513	51.4934	-0.111055	4322.64	91.8182	64.2727	60	216.091	4.45455	48.9627	10853.4
3.0	7.74	242392	51.4925	-0.113525	2394.45	1640.25	163.5	693.5	2497.25	7	373.377	31676.9
4.0	27.436	275573	51.4718	-0.17788	4405.39	315.6	348.4	336.2	1000.2	4.4	239.6	11687.2
5.0	14.28	275988	51.518	-0.01715	3282.28	1011.5	364	68.5	1444	3	481.335	19410.4
6.0	6.55	236523	51.5433	-0.07875	2219.38	1113	6.5	208	1327.5	9	147.5	36271.3

Mean Values for each Cluster

Cluster 1 - Enfield

Although not very densely populated Enfield has a lot of new cycle routes coming on line soon. It covers a large area and there are only 2 cycle shops at the moment. In fact we can see that to the North of this borough new routes are being put in and there would be scope for a new cycle business here. I think a business here would lead to slower but more high value leisure cycling sales

Enfield

Cluster Labels	Borough	Area	Population	Radius	Open Routes	In Progress Routes	Planned Routes	Total Routes	Shop Count	Route/Shop Ratio	Pop Density
1	Enfield	31.74	320524	4894.96111	997	798	0	1795	2	897.5	10098.42

Cluster 3 - Kensington and Chelsea, Lambeth, Tower Hamlets, Westminster

Cluster Labels	Borough	Area	Population	Radius	Open Routes	In Progress Routes	Planned Routes	Total Routes	Shop Count	Route/Shop Ratio	Pop Density
3	Kensington and Chelsea	4.68	155594	1879.614173	1248	70	529	1847	4	461.75	33246.58
3	Lambeth	10.36	314242	2796.569792	1590	362	447	2399	7	342.71	30332.24
3	Tower Hamlets	7.63	272890	2399.982381	1335	0	469	1804	5	360.80	35765.40
3	Westminster	8.29	226841	2501.629847	2388	222	1329	3939	12	328.25	27363.21

This cluster are some of the most densely populated areas of London. As well as this, outside of Enfield they have the highest number of cycle routes, including several new planned ones. There are a lot of cycle routes compared to cycle shops in these boroughs too. These three factors should help in having shops that would have a relatively high turnover of quick sharp commuter bike sales

Cluster 5 - Lewisham and Waltham Forest

Cluster Labels	Borough	Area	Population	Radius	Open Routes	In Progress Routes	Planned Routes	Total Routes	Shop Count	Route/Shop Ratio	Pop Density
5	Lewisham	13.57	286180	3200.631744	409	728	137	1274	3	424.67	21089.17
5	Waltham Forest	14.99	265797	3363.927296	1614	0	0	1614	3	538.00	17731.62

This cluster I believe sits within the middle of the other 2. There are relatively few cycle shops for the amount of routes in these boroughs but the population density in these areas is also average.

Conclusion

Due to the current pandemic and the race to get fitter, the interest in cycling has boomed recently. The city of London is working hard to help make this a reality for it's 9 million residents.

This is an ideal time for anyone interested in opening up a cycling business to help serve this expanding market.

Within this report, not only have we identified suitable boroughs within which to open a business but we have been able to identify the type of sales and market each location would attract.

Any investment comes with risk and this must be thoroughly assessed beforehand.

This report has not identified any particular sites within any borough. Accessibility and leasing costs for any sites will be a large factor in any decision.

It is always recommended that further detailed study continues to take place, especially locally in relation to any potential sites that are selected for investment.