Assessing childhood obesity through geospatial placement of fast food restaurants

A Coursera Applied Data Science Capstone project

Introduction: Business Problem

The World Health Organization terms childhood obesity as "one of the most serious public health challenges of the 21st century", with prevalence increasing globally and affecting low and middle-income countries, as well as low and middle-income areas in more developed countries. In 2016, there were about 41 million children under the age of five that were categorized as overweight [1].

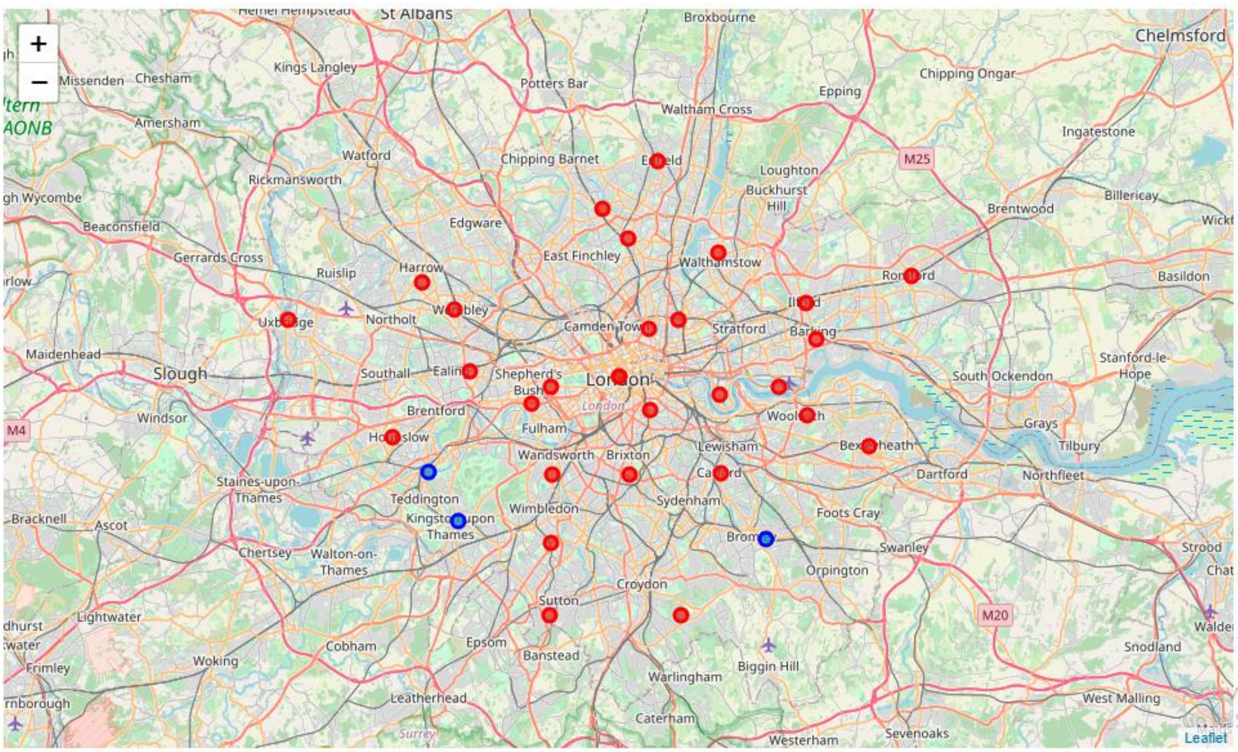
Although there are different ways to measure obesity within different age brackets (i.e. it is measured in a different way for 0-5 years, 5-19 years, and 19+), one thing is for certain: obesity has detrimental effects on the health of the individual in addition to putting strain on the healthcare system of the community.

There are several contributing factors to obesity. This report will analyse one aspect only: the possible impact of energy-dense foods that are high in fat and sugars and low in healthy vitamins and minerals. One particular source of such foods is fast food restaurants that offer foods such as fried chicken, burgers, pizza, etc. While there are many ways that a possible association between such restaurants and childhood obesity can be analysed, this report will investigate only whether there is a correlation between childhood obesity and the proximity of fast food restaurants to schools.

The results of such an analysis and others within this topic are very important because such results can make a strong case for policy intervention in ensuring that areas close to schools are kept free of such eateries where children can have quick and easy access to unhealthy foods either with their parents after school or (if they are old enough) by themselves on the way back after school. While there are bound to be other factors that impact the obesity rate, the geospatial placement of fast food restaurants within proximity of schools is certainly one that should be analysed to help policymakers in reducing childhood obesity in areas of high incidence.

This report will analyse the 32 boroughs in London, UK. Each borough in London is a local authority district that makes up Greater London (thereby excluding the City of London), and is governed by a London borough council. The selection of London for this report is because it is one of the most culturally-diverse cities in the world whilst being one of the most important economic and financial hubs globally. Additionally, it is also the city of residence of this author.

The prevalence of childhood obesity in London, as last measured in 2015/16, was 23.2% compared to a global average of around 18% [2]. The following map shows the different boroughs of London (with each borough centred around its respective borough council address) in which all boroughs with childhood obesity prevalence higher than the global average are marked with red and those with lower are marked in blue. This indicates the severity of the problem within London and is representative of how high rates of childhood obesity are prevalent in urban centres in developed countries.



The null hypotheses for this report are :

H01: There is no correlation between the prevalence of childhood obesity in a borough and the average distance between a fast food restaurant and a school in the borough.
H02: There is no correlation between the prevalence of childhood obesity in a borough and the average number of fast food restaurants within a 1 km radius of a school in the borough.

Data

To evaluate the possible correlation, the variables of interest will be:

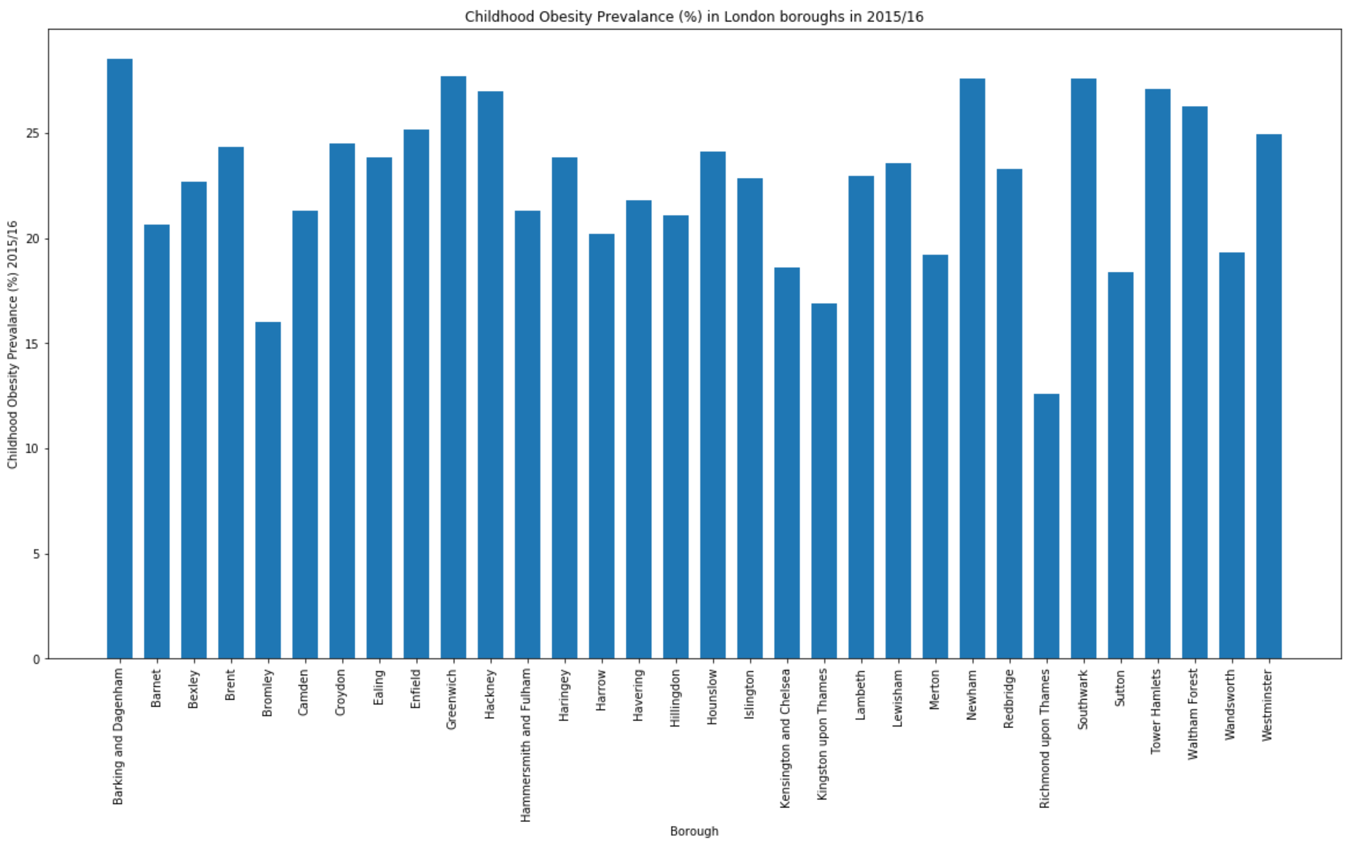
- Prevalence of childhood obesity in each London borough (%)
- Average distance of a fast food restaurant from a school in the borough (km)
- Average number of fast food restaurants within a 1 km radius of the school

The main data set that will be used is the London Borough Profiles and the London Borough Atlas from the London Datastore (www.data.london.gov.uk). These two files are available in Excel format and they will provide data on prevalence of childhood obesity in each of the boroughs. The other two variables, i.e. the average distance of a fast food restaurant from a school in the borough, and the average number of fast food restaurants within a 1 km radius of the school, are measured using the Foursquare API and are explained in more detail in the Methodology section. Further, the following assumptions are taken for the purpose of analyzing the data:

- The coordinates representing each borough are those of the respective borough council office (or town hall).
- Only 10 schools within a 2 km radius of each borough council are taken as a representative sample for this analysis.
- Only those fast food restaurants are considered that are within a 1 km radius of each sample school.

Methodology

To begin, the London Borough Profile datasheet was downloaded from (https://data.london.gov.uk/dataset/london-borough-profiles) and saved into a dataframe. After a quick review of the dataframe, only the data pertaining to the 32 London borough councils was kept. There were 85 columns in the sheet, with variables ranging from population estimates to turnout at local elections. While some of these variables may help in further analysis of the problem using other statistical and machine learning tools, this report is only considering one aspect and the variable of interest in this case is 'Childhood Obesity Prevalance (%) 2015/16'. The bar chart below shows the different levels of childhood obesity prevalence within the 32 London boroughs in 2015/16, clearly showing the disparity that exists between different boroughs.



Next, the council postcode for each borough is found manually from Google Maps and the coordinates for each council (that will eventually represent the borough itself) is then extracted using *pgeocode.* The first few rows of the ensuing dataframe is shown below.

	Borough	Childhood Obesity Prevalance (%) 2015/16	Post Code	Latitude	Longitude
0	Barking and Dagenham	28.542	IG11 7LU	51.5333	0.0833
1	Barnet	20.6579	N11 1NP	51.6197	-0.1441
2	Bexley	22.7093	DA6 7LB	51.4629	0.1394
3	Brent	24.3444	HA9 0FJ	51.5531	-0.3023
4	Bromley	15.9988	BR1 3UH	51.4013	0.0305

Since the process of finding a sample school and then finding fast food restaurants near each school has to be repeated for the 32 boroughs, it is better to develop a function that can be looped through the dataframe.

The function is defined as *mainfunction* that will take an argument from the for loop. The function itself includes two embedded for loops (the first for each school and the second for each restaurant within the school's proximity).

The function also includes declaring the author's Foursquare API credentials and setting the search radius of schools from the borough centre to be 2 km and the search radius of fast food restaurants from each sample school to be 1 km. The search query used the Foursquare API category ID of 'schools' (4bf58dd8d48988d13b941735) and the category ID of 'fast food restaurants' (4bf58dd8d48988d16e941735) exactly as mentioned in the Foursquare Developers guidance documents.

A counter is also embedded into the for loop to keep count of how many restaurants and how many schools are being returned with each search query.

The 'mean distance' for each borough is calculated by averaging the mean distance of each restaurant from a particular school, and then averaging that mean distance for all the sample schools within the borough. The distance is calculated using Vincenty's formula, which provides a more accurate estimation of geospatial distance between two coordinates than the Haversine formula.

Similarly, the average number of restaurants within a 1 km radius of each sample school in each borough is calculated by summing the total number of restaurants found for each school within each borough and averaging them.

Finally, a Pearson's correlation analysis is carried out with the three variables (prevalence of childhood obesity in the borough, mean distance of restaurants from the school, and average number of restaurants within a 1 km radius of a sample school.

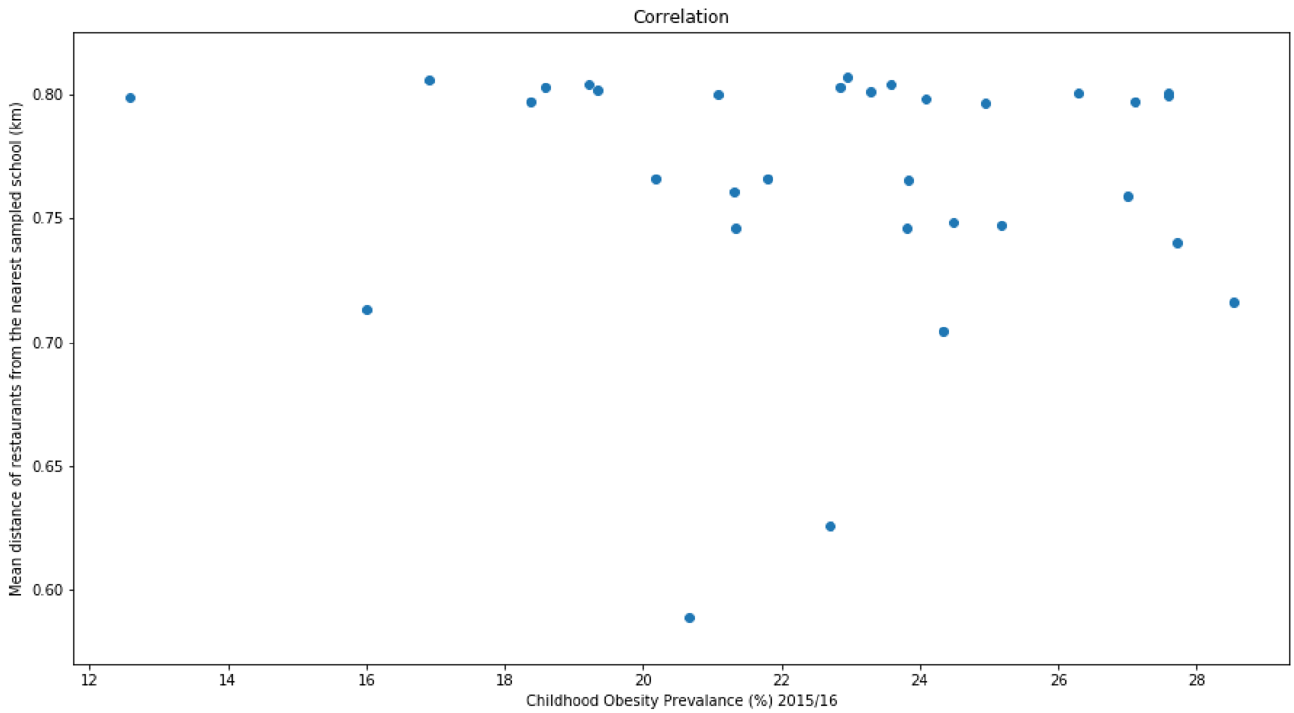
Results

The final results of the running the *for* loop through *mainfunction* are shown below.

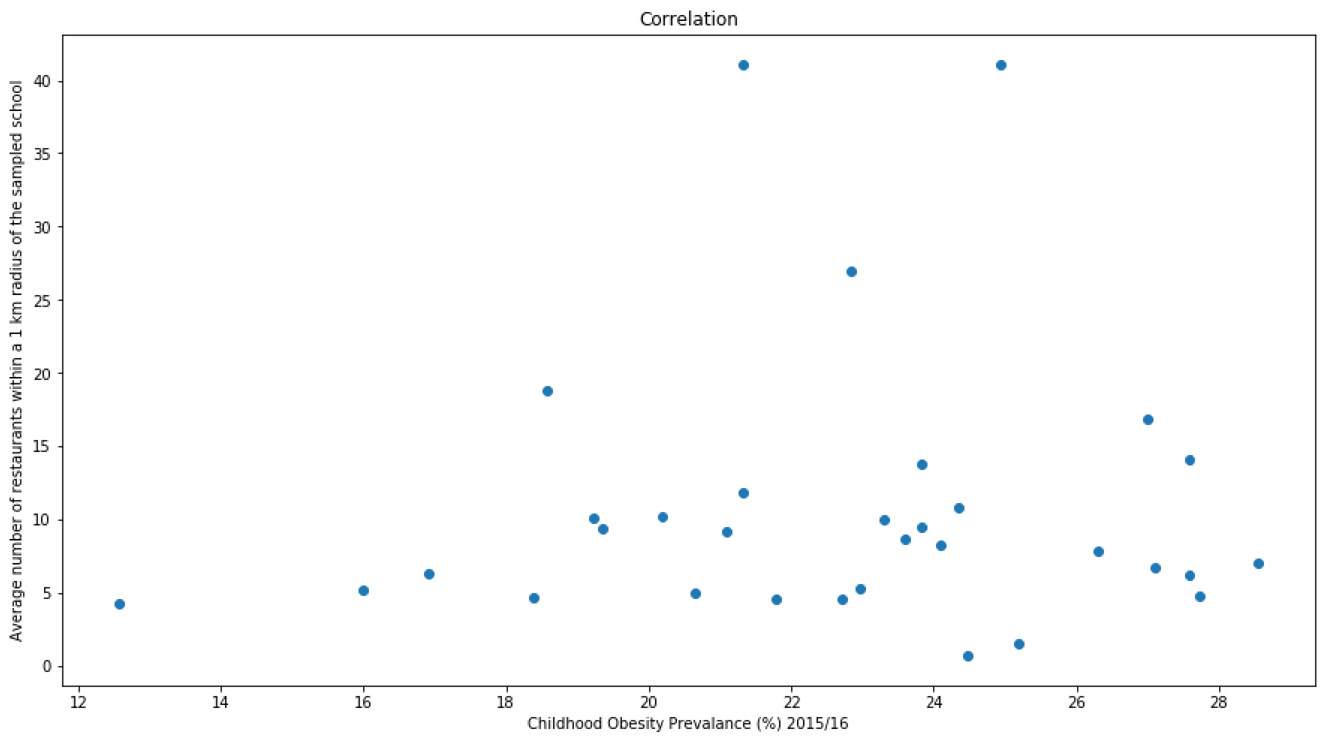
	Borough	Childhood Obesity Prevalance (%) 2015/16	Post Code	Latitude	Longitude	Mean distance	Ave Number of restaurants	Mean distance	Ave Number of restaurants
0	Barking and Dagenham	28.542	IG11 7LU	51.5333	0.083300	0.716213	7.000000	0.716213	7.000000
1	Barnet	20.6579	N11 1NP	51.6197	-0.144100	0.588785	5.000000	0.588785	5.000000
2	Bexley	22.7093	DA6 7LB	51.4629	0.139400	0.625980	4.625000	0.625980	4.625000
3	Brent	24.3444	HA9 0FJ	51.5531	-0.302300	0.704619	10.857143	0.704619	10.857143
4	Bromley	15.9988	BR1 3UH	51.4013	0.030500	0.712979	5.250000	0.712979	5.250000
5	Camden	21.3267	WC1H 9JE	51.5085	-0.125700	0.746144	41.100000	0.746144	41.100000
6	Croydon	24.4814	CR0 1EA	51.3500	-0.060975	0.748286	0.714286	0.748286	0.714286
7	Ealing	23.8228	W5 2HL	51.5122	-0.285189	0.746089	9.500000	0.746089	9.500000
8	Enfield	25.1879	EN1 3XA	51.6515	-0.085000	0.747375	1.500000	0.747375	1.500000
9	Greenwich	27.7255	SE18 6PW	51.4833	0.074067	0.740115	4.800000	0.740115	4.800000
10	Hackney	27.0025	E8 1EA	51.5460	-0.062900	0.759120	16.875000	0.759120	16.875000
11	Hammersmith and Fulham	21.3181	W6 9JU	51.4911	-0.220400	0.760615	11.888889	0.760615	11.888889
12	Haringey	23.824	N22 8LE	51.6000	-0.116700	0.765465	13.750000	0.765465	13.750000
13	Harrow	20.1817	HA1 2XY	51.5714	-0.336150	0.765888	10.250000	0.765888	10.250000

14	Havering	21.7873	RM1 3BD	51.5751	0.185800	0.766113	4.555556	0.766113
15	Hillingdon	21.0931	UB8 1UW	51.5462	-0.479133	0.799808	9.222222	0.799808
16	Hounslow	24.0913	TW3 4DN	51.4684	-0.368200	0.798041	8.250000	0.798041
17	Islington	22.8464	N1 2UD	51.5407	-0.094400	0.802832	27.000000	0.802832
18	Kensington and Chelsea	18.5841	W8 7NX	51.5016	-0.198500	0.802615	18.750000	0.802615
19	Kingston upon Thames	16.9056	KT1 1EU	51.4126	-0.297400	0.805891	6.285714	0.805891
20	Lambeth	22.9528	SW2 1RW	51.4437	-0.115800	0.806706	5.300000	0.806706
21	Lewisham	23.5846	SE6 4RU	51.4449	-0.018500	0.803966	8.714286	0.803966
22	Merton	19.2151	SM4 5DX	51.3982	-0.199200	0.804158	10.111111	0.804158
23	Newham	27.5862	E16 2QU	51.5022	0.044200	0.800713	6.250000	0.800713
24	Redbridge	23.2937	IG1 1DD	51.5576	0.072800	0.801283	10.000000	0.801283
25	Richmond upon Thames	12.5733	TW1 3BZ	51.4454	-0.329700	0.798517	4.285714	0.798517
26	Southwark	27.5827	SE1 2QH	51.4869	-0.093038	0.799362	14.125000	0.799362
27	Sutton	18.3804	SM1 1EA	51.3500	-0.200000	0.797177	4.666667	0.797177
28	Tower Hamlets	27.1038	E14 2BG	51.4969	-0.018765	0.797025	6.777778	0.797025
29	Waltham Forest	26.2931	E17 4JF	51.5907	-0.020480	0.800739	7.875000	0.800739
30	Wandsworth	19.342	SW18 2PU	51.4439	-0.197100	0.801587	9.444444	0.801587
31	Westminster	24.9423	SW1E 6QP	51.5085	-0.125700	0.796319	41.100000	0.796319

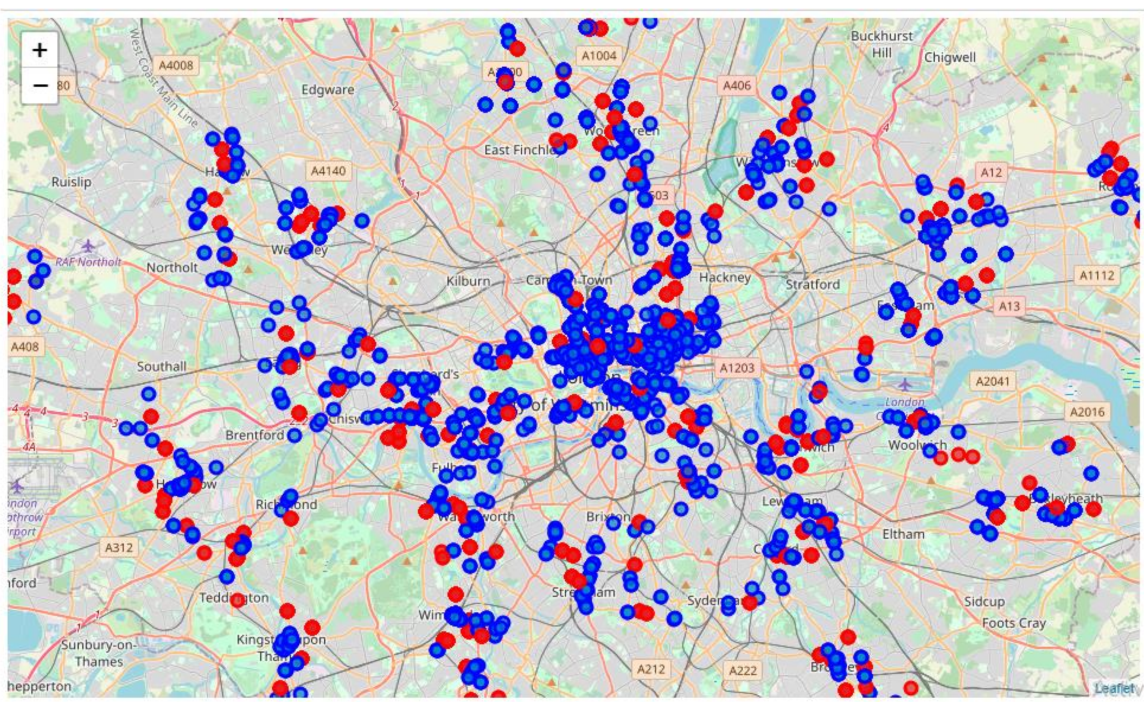
The Pearson's correlation coefficient between prevalence of childhood obesity and the mean distance of restaurants from the nearest sampled school is -0.05 with a p-value of .8 and the scatter plot is shown below. Since the p-value is greater than .05, we fail to reject the null hypothesis H01; this implies that there is insufficient evidence to prove that there is at least a moderate correlation between these two variables.



The Pearson's correlation coefficient between prevalence of childhood obesity and the average number of restaurants within a 1 km radius of the sampled school is 0.06 with a p-value of .7 and the scatter plot is shown below. Since the p-value is greater than .05, we fail to reject the null hypothesis H02; this implies that there is insufficient evidence to prove that there is at least a moderate correlation between these two variables.



All the search results of schools and fast food restaurants for each borough are shown below on a Folium-based map of London. The blue markers represent fast food restaurants while the red markers represent the sample schools.



Discussion

The analysis shows that there is no significant correlation between the prevalence of childhood obesity in the London boroughs and the mean distance of fast food restaurants from schools within the boroughs as well as the average number of fast food restaurants within a 1 km radius of the schools.

There can be a number of reasons why there may still be an association between these variables but the analysis has failed to prove it. Firstly, we are evaluating a sample size of only 32 and it is possible that the strength of the correlation becomes more dominant once the sample size increases further. Secondly, we have taken a limited number of schools within the proximity of the borough council address; in reality, there will be many more schools that fall within the borough jurisdiction but may not have been included in this analysis. The effect of this may be more pronounced if, for example, the borough council itself is located in an area where there is a lower density of schools and restaurants which may affect the overall analysis. Thirdly, the data for childhood obesity prevalence was measured in 2015/16, whereas the geospatial data of schools and restaurants is current; it is possible that many fast food restaurants may have closed during this period which also has led to a very weak correlation between the variables. Finally, it is also important to note that consumer behavior also plays an important role in this, especially in the context of consumption habits. Perhaps the proximity of the fast food restaurants to schools is less of an influencing factor in the cause of childhood obesity compared to the proximity of such restaurants in densely populated or low-income residential areas.

There are many other factors and tools that must be used and considered to enable a more holistic analysis to take place. Other variables that can be considered for future analysis are population density, average income, household size, dietary habits, availability of parks and recreational spaces, etc. Tools such as multivariate regression can be used to predict the relation, if any, between the prevalence of childhood obesity and these variables.

Conclusion

This exploratory analysis aimed to assess whether there is a correlation between the prevalence of childhood obesity in London boroughs and a) the geospatial proximity of fast food restaurants from schools within the borough and b) the average number of restaurants within a 1 km radius of the sampled schools. The data on the boroughs was taken from the London Datastore while a Foursquare API account was used to query and collect geospatial data on schools and fast food restaurants near the schools for each of the boroughs. Pearson's correlation found that there was no strong or significant correlation between the variables in question; however there are several reasons for this as discussed in the Discussion section. Further research should include other variables, such as population density and household income, as well as consider other statistical tools such as multivariate regression. The results of this report highlight the importance of considering other factors in order to help reduce the prevalence of childhood obesity, which has now become a severe health issue that threatens to endanger the lives of millions of people globally.

References:

1. https://www.who.int/dietphysicalactivity/childhood_why/en/
2. https://www.worldobesity.org/about/about-obesity/prevalence-of-obesity