APPLIED DATA SCIENCE CAPSTONE

Obesity, deprivation and venues in Buckinghamshire

1. Introduction

Buckinghamshire is a county in England north-west of Greater London. The local authority for the county is Buckinghamshire Council. The Council has a public health department, responsible for improving the overall health of its population.

The department has funding to build and run a fitness and nutrition centre to encourage exercise and healthier diets, which in turn will prevent poor health. The department wishes

to locate the most appropriate location for this fitness and nutrition centre. The criteria for finding the most appropriate location should take into account rates of obesity, poverty and the presence of services that encourage healthy or less-healthy behaviours.

1.1 Data

The first dataset I used can be found using Public Health England's Fingertips API and covers the Index of Multiple Deprivation (IMD) score for each ward in England, which is a measure of poverty. The higher the IMD score, the higher prevalence of poverty there is in a ward.

The second dataset I used was the rate of obesity among Year 6 age children by ward, which is also provided by the Fingertips API. Currently there is no publicly-available measure for obesity among adults at ward level, so this was the next most relevant proxy for measuring obesity.

The third dataset I used was provided by the Foursquare API. I used this API to identify the venues/services in each ward listed on Foursquare. I looked for the presence of services that encourage healthy behaviours, such as gyms or leisure centres, and those that encourage less-healthy behaviours, such as fast-food restaurants.

2. Methodology

2.1 Visualising the geographical data

First of all, I did a little exploration with the data I had extracted. This started with a visualisation of the wards across Buckinghamshire. This involved combining data taken from the ONS Geography Portal to produce the following data frame shown in Figure 1.

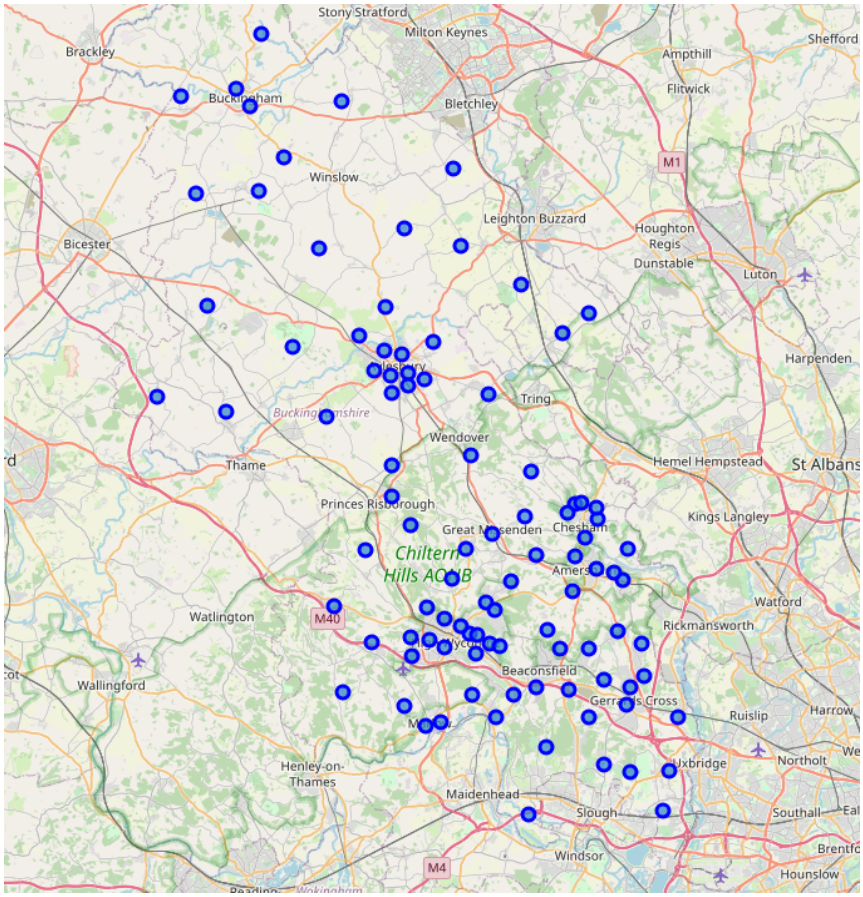
Then, using the Folium library, I produced a simple map of Buckinghamshire showing points for each ward in the county (see Figure 2).

Figure 1

	Ward ID	Ward Name	longitude	latitude
0	E05002636	Ballinger, South Heath and Chartridge	-0.67319	51.71306
1	E05010335	Buckingham South	-0.97332	51.99219
2	E05010336	Central & Walton	-0.80040	51.81117
3	E05010337	Coldharbour	-0.83723	51.81315
4	E05002637	Central	-0.54285	51.60405
...	...	...	...	...
93	E05010575	Gerrards Cross	-0.56235	51.58500
94	E05010576	Iver Heath	-0.51430	51.53985
95	E05010577	Iver Village & Richings Park	-0.52161	51.51195
96	E05010578	Stoke Poges	-0.58652	51.54336
97	E05010579	Wexham & Fulmer	-0.55764	51.53903

98 rows × 4 columns

Figure 2



One thing to notice here is that there are far more wards in the southern side of the county, where it gets closer to Greater London and more populous. There is also a cluster of wards around Aylesbury. The wards are much larger in the northern part of Buckinghamshire. Normally, we would expect to see higher rates of poverty in more populous areas, which will be located in these dense ward clusters.

2.2 Obesity and poverty effects

Using the PHE Fingertips API, I was able to extract the deprivation and childhood overweight and obesity rate data for each of these wards. I consolidated this in one dataframe, as presented in Figure 3.

Figure 3

	Ward ID	Ward Name	longitude	latitude	IMD score	Excess weight rate
0	E05002636	Ballinger, South Heath and Chartridge	-0.67319	51.71306	7.687	15.4
1	E05010335	Buckingham South	-0.97332	51.99219	5.163	26.3
2	E05010336	Central & Walton	-0.80040	51.81117	12.502	33.3
3	E05010337	Coldharbour	-0.83723	51.81315	9.723	35.4
4	E05002637	Central	-0.54285	51.60405	4.978	26.9
...	...	...	...	...	...	...
93	E05010575	Gerrards Cross	-0.56235	51.58500	5.948	28.1
94	E05010576	Iver Heath	-0.51430	51.53985	10.579	40.0
95	E05010577	Iver Village & Richings Park	-0.52161	51.51195	11.884	42.9
96	E05010578	Stoke Poges	-0.58652	51.54336	8.689	28.6
97	E05010579	Wexham & Fulmer	-0.55764	51.53903	17.537	33.3

98 rows × 6 columns

One of the things I initially wanted to know is whether obesity and poverty are related. This was important in figuring out whether I was on the right track on focusing on these two factors, or whether they were losing focus by moving different directions. I performed a simple linear regression analysis to test this, using the Seaborn library and the obesity and poverty data for each ward to produce the chart shown in Figure 4. Each dot represents a ward.

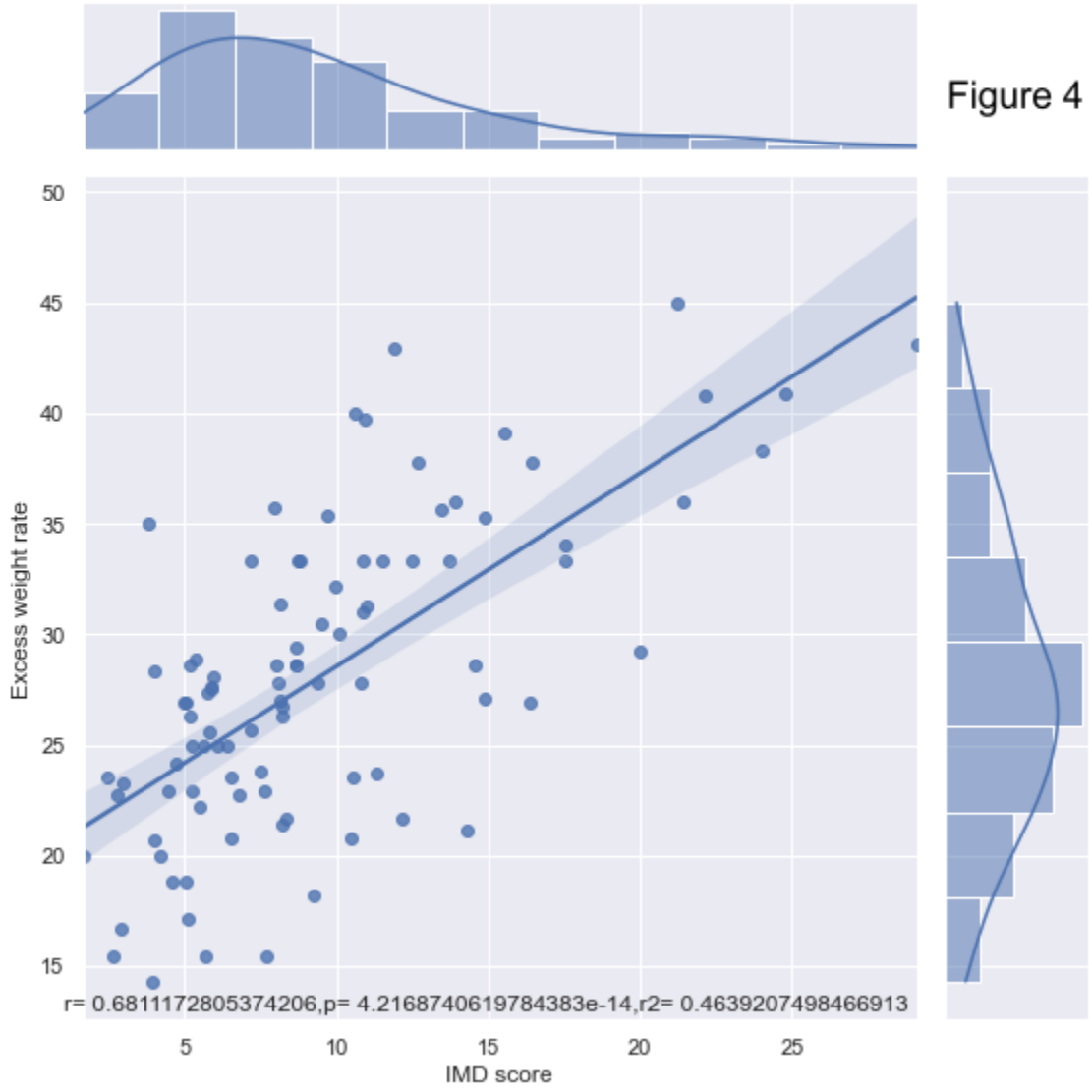


Figure 4

Using the Scipy library, I also calculated the Pearson correlation coefficient (r), the p-value (p) and the coefficient of determination (r2), included in the bottom of the chart. They suggest that there is a positive relationship between poverty and obesity in Buckinghamshire: the higher IMD score, the higher the rate of excess weight among Year 6 children. Though the effect is mild, it is highly significant. So the two factors are indeed related and I had evidence to believe I was on the right track.

2.3 Looking at the venues in Buckinghamshire

Using the Foursquare API, I was able to pull out the top 100 venues within a 2.5km radius of the coordinates for each ward. I had to make the radius this large as some of the wards are on a far larger scale. Converting my results into a grouped one hot table produced the data frame below (Figure 5), showing the number of venues in each venue category by ward.

Figure 5

	Neighborhood	African Restaurant	Airport	Alternative Healer	American Restaurant	Antique Shop	Arts & Crafts Store
0	Abbey	0	0	0	1	0	0
1	Amersham Common	0	0	0	0	0	0
2	Amersham Town	0	0	0	0	0	0
3	Amersham-on-the-Hill	0	0	0	0	0	0
4	Asheridge Vale and Lowndes	0	0	0	0	0	0
...	...	...	...	...	...	...	...
93	Wendover & Halton	0	0	0	0	0	0
94	Wexham & Fulmer	0	0	0	0	0	0
95	Wing	0	0	0	0	0	0
96	Wingrave	0	0	0	0	0	0
97	Winslow	0	0	0	0	1	0

98 rows × 159 columns

I used this information to identify the prevalence of healthy and unhealthy services. There was no method here for choosing what was 'healthy' and what was 'unhealthy', I just used the arbitrary method of common sense.

So, under **healthy** services/venues/locations, I had the following categories: Athletics & Sports, Cave, Forest, Fruit & Vegetable Store, Garden, Golf Course, Gym, Gym/Fitness Center, Gym Pool, Hill, Indoor Play Area, Lake, Martial Arts School, Other Great Outdoors,

Outdoors & Recreation, Park, Playground, Pool, Recreation Center, River, Scenic Lookout, Soccer Field, Sports Club, Squash Court, Stables, Tennis Court, Trail, Water Park.

Under **unhealthy**, I had the following categories: American Restaurant, BBQ Joint, Burger Joint, Burrito Place, Chinese Restaurant, Dessert Shop, Diner, Donut Shop, English Restaurant, Fast Food Restaurant, Fish & Chips Shop, Food Truck, Indian Restaurant, Turkish Restaurant, Noodle House, Pizza Place, Pub, Steakhouse.

For every ward, I produced a total for locations with healthy and unhealthy categories. I then converted these two totals into a percentage of the total venues in each ward, to try to control for population and venue density. I then added this data to our obesity and poverty data, to produce the following dataframe (Figure 6).

	Ward Name	IMD score	Excess weight rate	Total healthy	Total unhealthy
0	Ballinger, South Heath and Chartridge	7.687	15.4	40.000000	0.000000
1	Buckingham South	5.163	26.3	7.142857	14.285714
2	Central & Walton	12.502	33.3	6.849315	20.547945
3	Coldharbour	9.723	35.4	5.479452	28.767123
4	Central	4.978	26.9	12.500000	12.500000
...	...	...	...	...	...
93	Gerrards Cross	5.948	28.1	8.695652	8.695652
94	Iver Heath	10.579	40.0	10.344828	27.586207
95	Iver Village & Richings Park	11.884	42.9	37.500000	25.000000
96	Stoke Poges	8.689	28.6	22.222222	16.666667
97	Wexham & Fulmer	17.537	33.3	21.428571	7.142857

94 rows × 5 columns

Figure 6

Just like with the poverty data, I wanted to find out how useful the rate of healthy or unhealthy locations was in understanding obesity. Also just like with the poverty data, I decided to do this again by two simple linear regressions, using the same method. I compared the rate of unhealthy locations with the rate of excess weight among Year 6 children (Figure 7), as well as the rate of unhealthy locations with the same obesity measure (Figure 8).

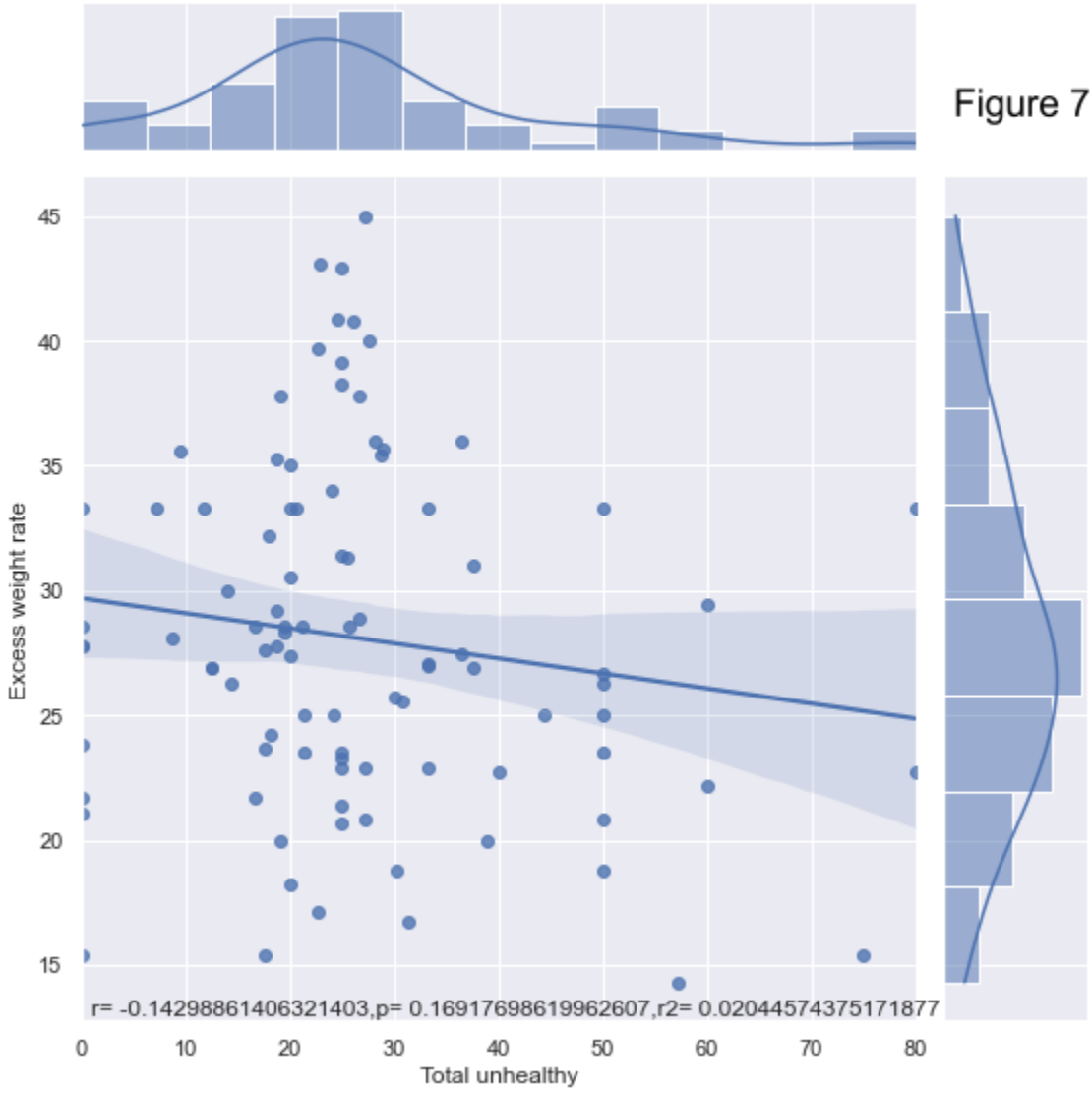


Figure 7

Neither of these analyses suggest that the rate of healthy/unhealthy locations are useful factors when considering obesity. Neither showed a positive or negative relationship and showed high variance, suggesting no influential effect. I still used these metrics in the clustering analysis as it was part of the initial use case, but these are important considerations for the future if taking this project further.

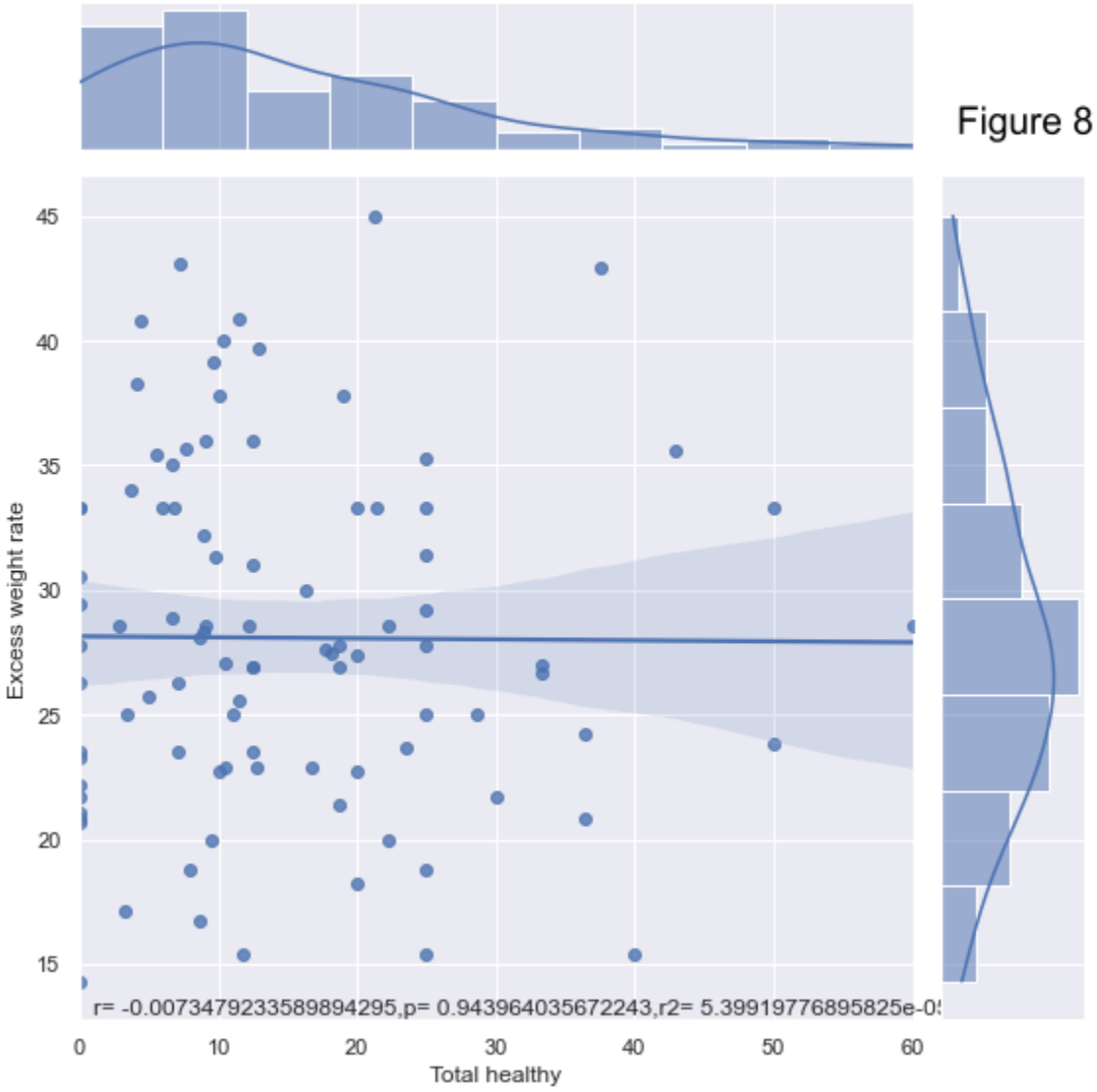


Figure 8

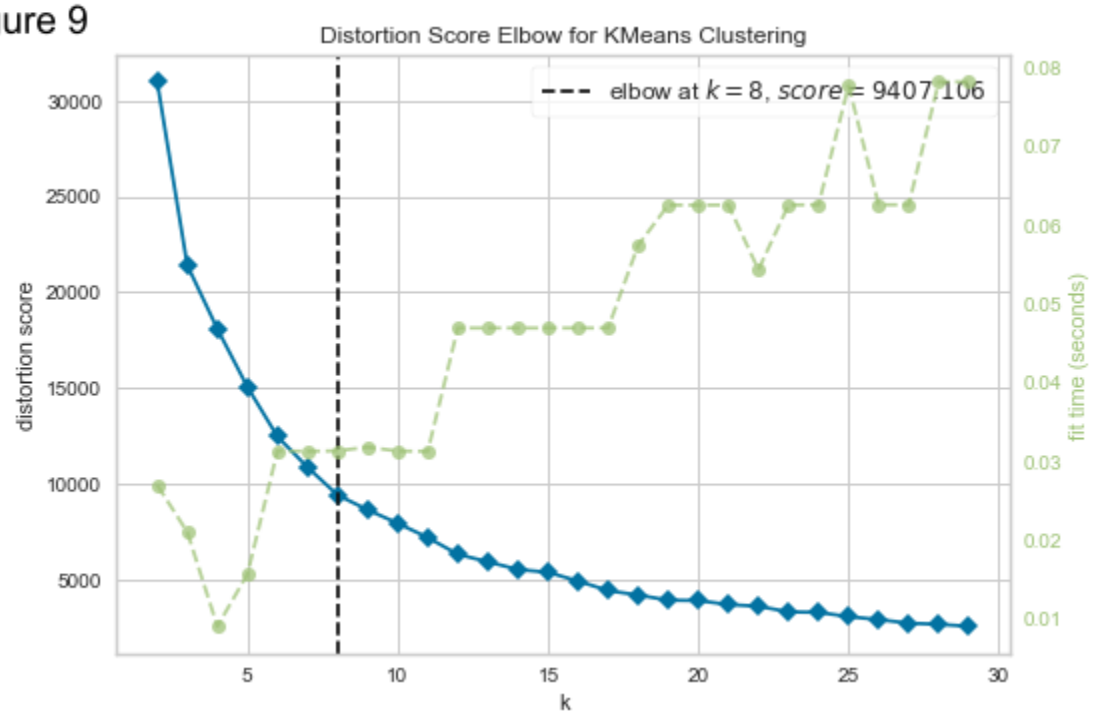
2.4 The cluster analysis

With data on obesity, poverty and unhealthy/healthy locations all in one place, I was able to start clustering it all. I chose k-means clustering as the method, an unsupervised algorithm, that groups data to k number of clusters who have the closest means.

This requires identifying an optimum value for k, the number of clusters chosen for the model. I did this by using the elbow method, which goes through various iterations of clustering with different values for k, testing for distortion for each iteration. The optimal value for k via this method is at the “elbow”, i.e. the point after which the distortion/inertia

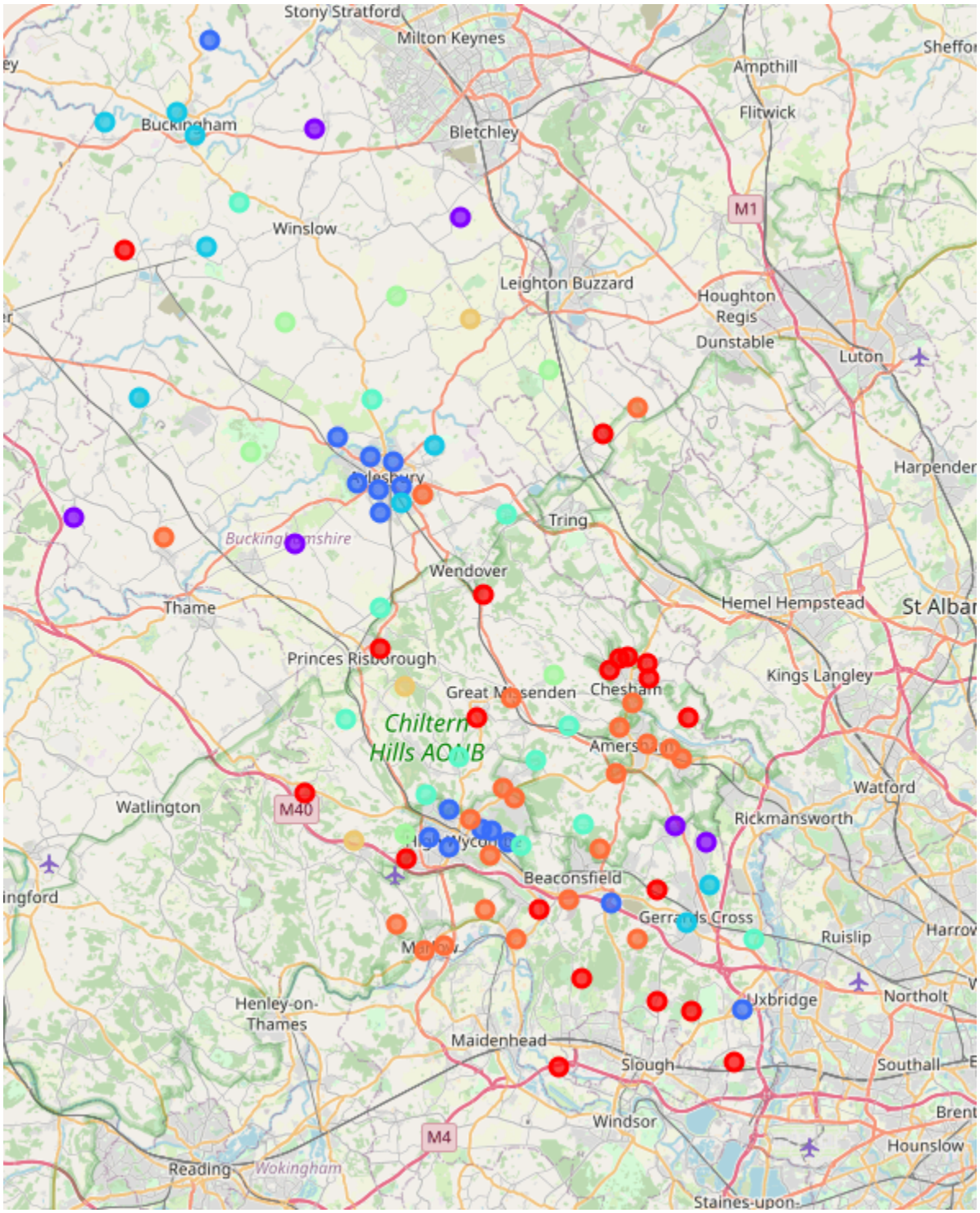
starts decreasing in a linear fashion. To identify the elbow, I used the Yellowbrick library to plot the following chart shown in Figure 9.

Figure 9



This identified 8 as my optimal value for k. This meant I set up my k-means model to place every ward in Buckinghamshire into one of 8 groups, assigning a label for each ward. I produced the model using the scikit-learn library, then plotted the results (all wards with a colour for each cluster) onto the Folium map shown in Figure 10.

Figure 10


Stony Stratford
Milton Keynes
Ampthill
Flitwick
Buckingham
Bletchley
M1
Winslow
Leighton Buzzard
Houghton Regis
Dunstable
Luton
Harpenden
Tring
Wendover
Thame
Hemel Hempstead
St Alban
Princes Risborough
Kings Langley
Chesham
Chiltern Hills AONB
Watford
Watlington
M40
Rickmansworth
Beaconsfield
Ruislip
Harrow
Henley-on-Thames
Uxbridge
Northolt
Maidenhead
Slough
Southall
Windsor
M4
Hounslow
Reading
Wokingham

3. Results

The mean value for each variable for every ward cluster is displayed in the data frame shown in Figure 11 below.

Figure 11

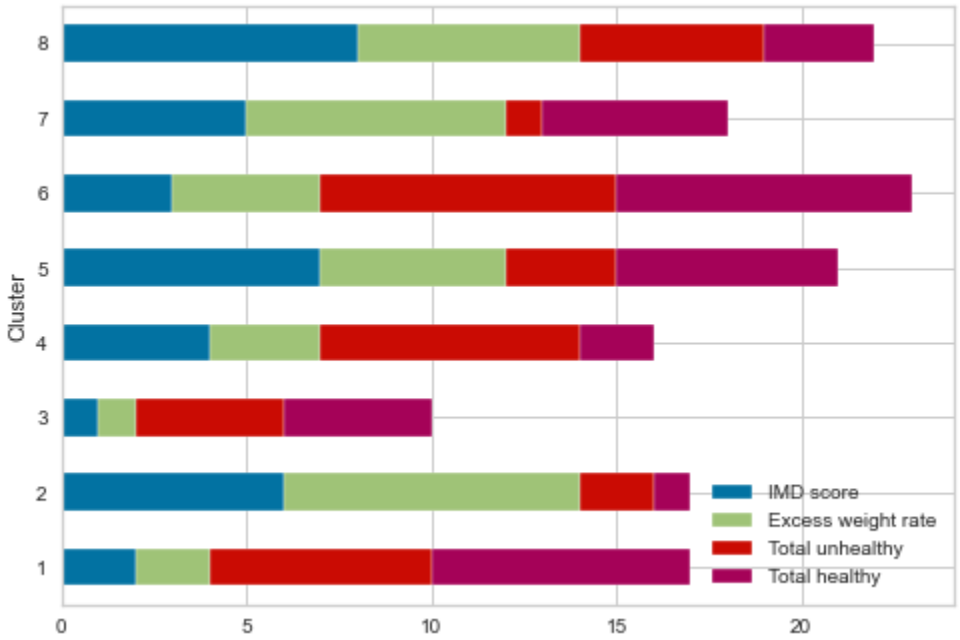
	IMD score	Excess weight rate	Total healthy	Total unhealthy
1	10.492850	28.630000	23.760133	18.136893
2	7.212333	22.750000	0.000000	54.523810
3	16.966125	38.025000	8.538129	26.733262
4	8.900667	27.544444	4.797343	11.313687
5	6.088750	25.566667	20.404040	41.780303
6	9.304667	26.400000	45.476190	1.587302
7	7.997667	23.800000	15.000000	78.333333
8	5.990045	24.431818	7.747641	24.809221

Cluster 3 (dark blue dots in Figure 10) presents the highest average rate of obesity (as measured by the rate of excess weight among year 6 age children), the highest average rate of poverty (as measured by the IMD score), and has a higher number of unhealthy services than those that are healthy.

Cluster 1 also stands out as having a relatively high rate of obesity and poverty, but it also has a relatively high rate of healthy services and a low rate of unhealthy services. Cluster 2 is notable for having the highest rate of unhealthy services, however it has a relatively low obesity and poverty rate compared to other clusters.

The below chart in Figure 12 shows the mean value for the four variables for each cluster, ranked out of 8 (the lower the number, the higher the cluster scored for a variable). The rank for the 4 variables is summed for each cluster. With the lowest total rank, Cluster 3 again appears to be the most notable cluster.

Figure 12



4. Discussion

My k-means model produced 8 clusters, as shown in Figure 10. A closer look at each cluster shows that Cluster 3 looks to be the most appropriate to act as a short list of locations for the fitness and nutrition centre, with the highest average rate of obesity and poverty, while the rate of unhealthy services outsizes the rate of healthy services.

This means that Aylesbury and High Wycombe are optimal targets for the fitness and nutrition centre, because they contain wards in Cluster 3. Without any wards in Cluster 3, the towns of Chesham, Amersham, Buckingham and Gerrard Cross are not not optimal targets.

My regression analyses of the rate of healthy/unhealthy services, however, suggested that they may not be relevant metrics to use when looking for appropriate locations for the centre. If I were to take this project further, I would consider either different ways of treating the venues data (e.g. using counts instead of rates, or including different venues in the healthy/unhealthy totals), or use metrics for entirely different variables to use for the k-means model.

5. Conclusion

Going back to the Public Health department with my cluster analysis results, I would suggest that Aylesbury and High Wycombe as the key targets for a new fitness and nutrition centre.

However, moving forward, I would suggest that we test various other metrics and factors to use in clustering Buckinghamshire wards, according to how impactful they are on obesity outcomes.