Determining the Best Location for a New High School in North Yorkshire, UK

Table of Contents:

1. Introduction

1.1. Background

The population of the UK is growing steadily, as is the population of the North Yorkshire county region (Fig. 1). The growth trend in the region is likely to continue and possibly accelerate due to the movement from urban to rural living stimulated by Covid-19 and the increased acceptance of home working. The prohibitively high cost of living in many southern areas is also causing many people to consider a move further north.

Fig. 1 – UK map with North Yorkshire county highlighted in red.[1]

Although the population of the UK (and North Yorkshire) is ageing, the overall rate of population growth is significantly higher, and therefore the number of children in the region is increasing. This results in an increased requirement for education and high schools. As most high schools in the region are already near capacity, it is unlikely that current education provision will be sufficient. It is therefore necessary to determine where a new school should be located.

As much of the North Yorkshire county is very rural, many children are already required to travel significant distance to reach their school. This may have a negative impact on a child's education due to increased tiredness and reduced learning effectiveness. This is also a disadvantage for the local authority as they are required to fund transport for all children to their nearest school.

1.2. The Problem

North Yorkshire county council needs to consider the increased requirement for high school places in the next few years, and determine the best location to site a new high school.

Data that is likely to influence the decision includes population distribution within the region, the locations of existing high schools, the current capacity situation of existing high schools, and the average distance children must travel to their nearest high school. This project aims to use these data to propose the location of a new high school in North Yorkshire.

1.3. Interest

North Yorkshire county council would be the key interested party in this analysis, as they would be able to use the analysis to locate a new school in the area with the most benefit to the local population, and reduced requirements to fund transport.

Other interested parties may include families looking to move to the North Yorkshire region, and considering the proximity to local high schools as part of their decision.

2. Data Acquisition

Geographic data for the North Yorkshire region was taken from Doogal[2], which compiles a UK wide postcode directory and hierarchical assignments (e.g. assignments of postcodes to constituency, local authority city, county) using data from Ordnance Survey, Royal Mail and the Office for National Statistics.

Relevant data on schools including name, location (latitude/longitude), postcode and school type (primary school, high school etc.) can be accessed using Foursquare API .

Data on population distribution was sourced from the North Yorkshire County Council Data Hub[3]. This contains the most accurate and localised data on local population distribution as it is compiled from UK census information. The lowest level data available at the time of analysis is by 'Lower Super Output Area' (LSOA). The UK is divided into LSOAs, each of which is a geographic region with a minimum population of 1000 and a mean of 1500[4].

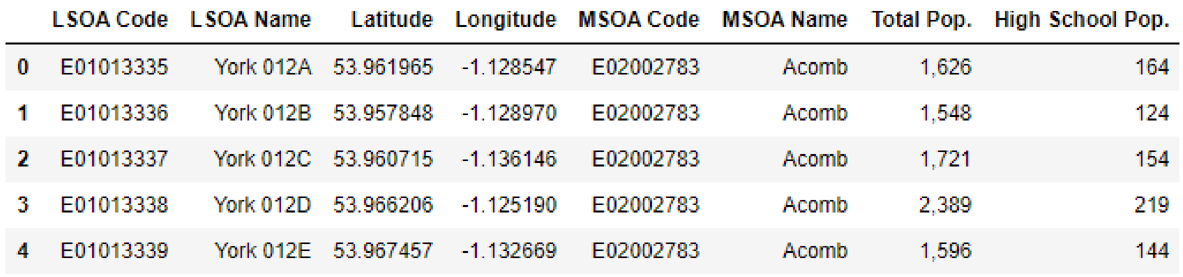
In order to perform a more detailed analysis, supplementary data on each school in the region of interest was sourced from the British Government Department of Education[5], including the classification of each school as rural/urban, and the current school capacity information.

3. Methodology

3.1. Data Cleaning

Data on population distribution was first cleaned to contain only the North Yorkshire region, and all unneeded columns dropped. The data is supplied at postcode level, but as the population distribution data is only available by LSOA, the data was aggregated to LSOA, and the mean latitude and longitude calculated for each LSOA.

The population data was reduced to contain only the total population for each LSOA, and the high school age population calculated for each by summing the data for ages 11-18 (Fig. 2).

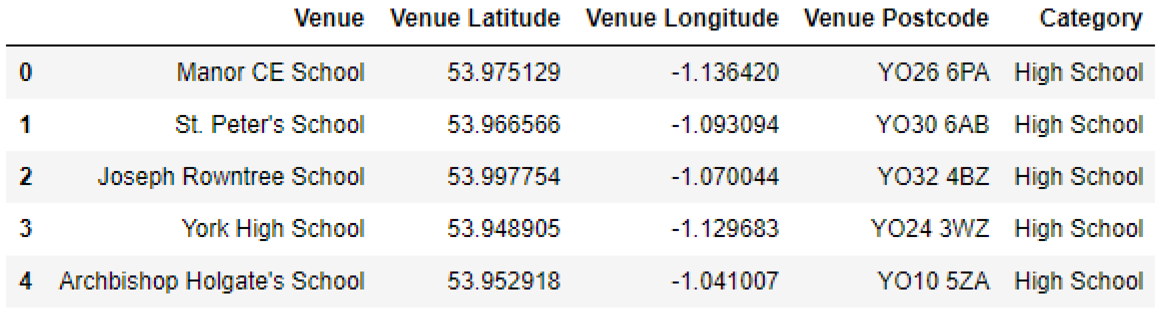


	LSOA Code	LSOA Name	Latitude	Longitude	MSOA Code	MSOA Name	Total Pop.	High School Pop.
0	E01013335	York 012A	53.961965	-1.128547	E02002783	Acomb	1,626	164
1	E01013336	York 012B	53.957848	-1.128970	E02002783	Acomb	1,548	124
2	E01013337	York 012C	53.960715	-1.136146	E02002783	Acomb	1,721	154
3	E01013338	York 012D	53.966206	-1.125190	E02002783	Acomb	2,389	219
4	E01013339	York 012E	53.967457	-1.132669	E02002783	Acomb	1,596	144

Fig. 2 – Sample of dataframe containing the geographical coordinates, total population and estimated high school population for each LSOA in the North Yorkshire.

School data was obtained from foursquare using a Venue Search using search term 'school'. This was chosen in place of a category search as schools may be assigned to many categories within Foursquare (e.g. School, High School, Community College). Data was requested for the geographical coordinates of each LSOA, within a radius of 10km and using a limit of 300. The high limit and radius were used to maximize the probability of returning all schools within the region.

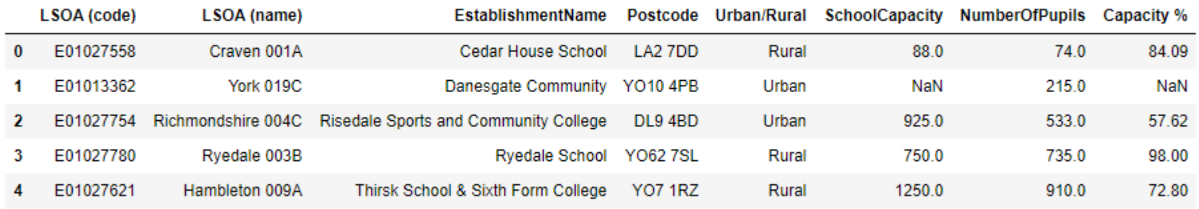
The data was cleaned by removing duplicates and dataframe slicing based on category to remove anomalies (e.g. driving school) and restrict the data set to only high schools, resulting in a list of 130 high schools with their geographical coordinates, and postcodes where available (Fig. 3).



	Venue	Venue Latitude	Venue Longitude	Venue Postcode	Category
0	Manor CE School	53.975129	-1.136420	YO26 6PA	High School
1	St. Peter's School	53.966566	-1.093094	YO30 6AB	High School
2	Joseph Rowntree School	53.997754	-1.070044	YO32 4BZ	High School
3	York High School	53.948905	-1.129683	YO24 3WZ	High School
4	Archbishop Holgate's School	53.952918	-1.041007	YO10 5ZA	High School

Fig. 3 – Sample of High School data obtained from Foursquare.

The Department of Education data on school capacity was cleaned to retain only those high schools within the LSOAs in the North Yorkshire region, and the Urban/Rural status and Capacity % obtained for each (Fig. 4).



	LSOA (code)	LSOA (name)	EstablishmentName	Postcode	Urban/Rural	SchoolCapacity	NumberOfPupils	Capacity %
0	E01027558	Craven 001A	Cedar House School	LA2 7DD	Rural	88.0	74.0	84.09
1	E01013362	York 019C	Danesgate Community	YO10 4PB	Urban	NaN	215.0	NaN
2	E01027754	Richmondshire 004C	Risedale Sports and Community College	DL9 4BD	Urban	925.0	533.0	57.62
3	E01027780	Ryedale 003B	Ryedale School	YO62 7SL	Rural	750.0	735.0	98.00
4	E01027621	Hambleton 009A	Thirsk School & Sixth Form College	YO7 1RZ	Rural	1250.0	910.0	72.80

Fig. 4 – Sample of dataframe containing school capacity and rural/urban classification for schools in North Yorkshire.

It was noted that there some capacity data was missing and this was populated with mean capacity %. This does result in a small decrease in confidence in the results but should not artificially skew the results themselves. As this represents only one part of the analysis, this was deemed to be more appropriate than dropping those data points altogether as that.

The Foursquare school location data was joined on the capacity data from the Department of Education, using School name, and cross-checked using postcode. This identified remaining anomalies where schools did not meet analysis criteria (incorrect age range, school facility (e.g. gym) returned as result). These were dropped from the dataset.

Finally, the list of schools was plotted by location on a map, overlaid with a GeoJson file[6] displaying the UK administrative boundaries (Fig. 5). This identified 2 schools with location outside of the analysis area (York and North Yorkshire). These were dropped from the dataset, leaving a list of 38 high schools within the region.

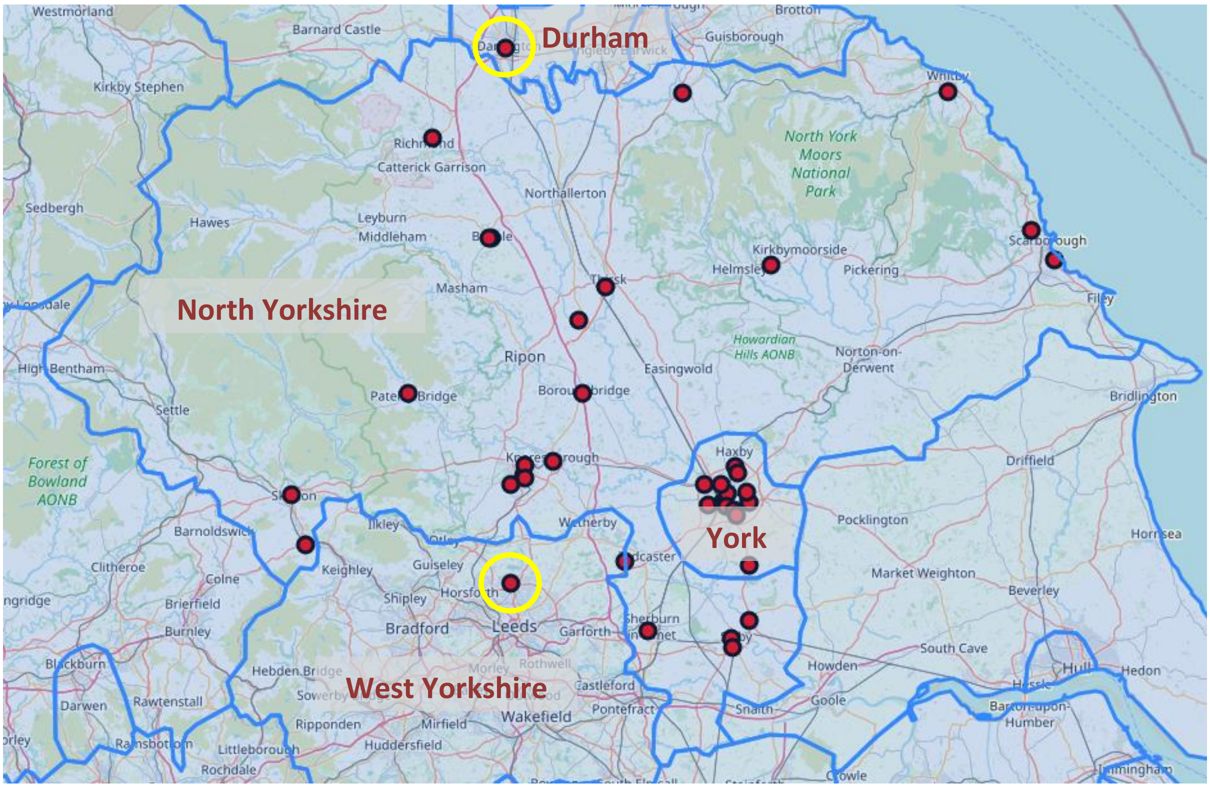


Fig. 5 – Folium map displaying school locations, overlaid with GeoJson of UK administrative boundaries (two schools outside the York and North Yorkshire region are circled in yellow).

3.2. Feature Selection

The features considered for analysis are listed below:

- Existing school location
- Current school capacity %
- Urban/Rural school categorization
- Population Distribution
- Average distance travelled to nearest high school

Most features were determined to be critical to the analysis, but it was unknown whether the urban/rural categorization of a school would have any impact. To determine whether this could be a useful predictor, the School Capacity % was plotted against the urban/rural status (Fig. 6). This shows that urban schools are significantly more likely to have higher capacity utilization. The feature was therefore determined to be a valid predictor and was retained in the analysis.

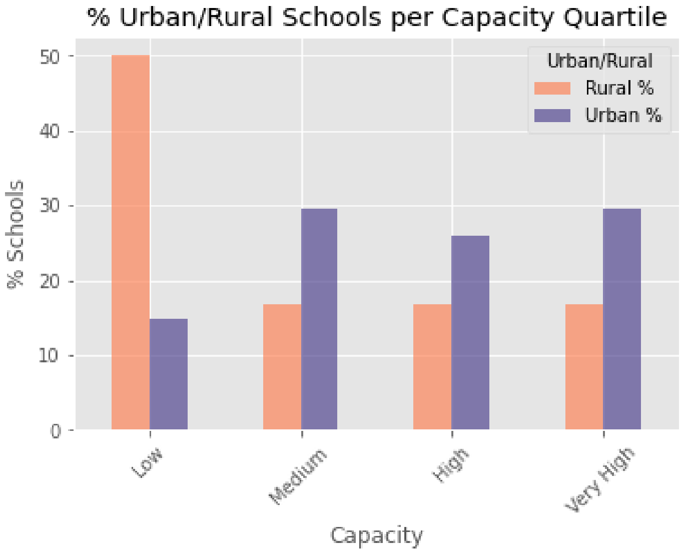


Fig. 6 – Bar chart displaying the % of Urban and Rural Schools in each Capacity Quartile

3.3. Data Analysis

In the data analysis stage, some data visualisation will be carried out first to identify any obvious trends. This will include visualising population distribution by heat map, and also looking at whether there is any obvious school shortage from plotting school capacity information by location.

Once this is complete, further, more detailed analysis will be carried out. The total number high school age students associated with each school (by assuming that each student attends the nearest school) will be analysed, and the average distance pupils must travel to each school will be calculated. By comparing the results of these 2 analyses we will propose the most impactful locations to site a new school. This should then be a starting point for a more detailed analysis of these areas.

3.3.1. Relationship between Capacity and Location

It was hypothesized that there could be a correlation between location and current capacity status of schools. A histogram was plotted of capacity distribution (Fig. 7), which was then used to split the schools into 4 groups based on capacity status (Low – <60%, Medium – 60-85%, High – 80-95%, Very High - >95%).

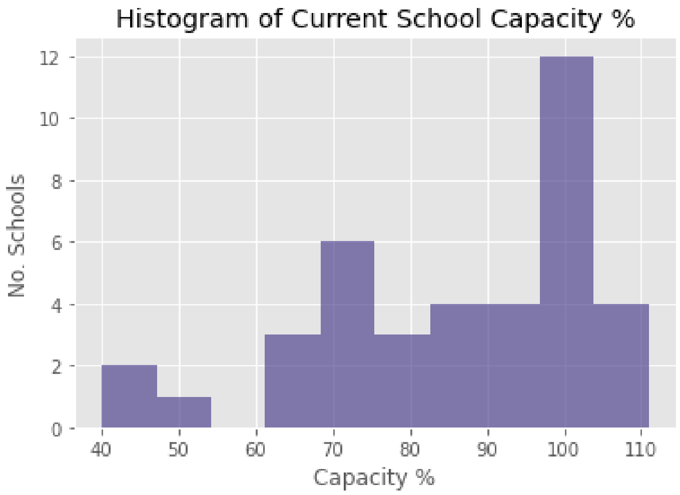


Fig. 7 – Histogram of School Capacity % Distribution

The schools were then plotted by location, coloured according to their capacity group and with size determine by urban/rural categorization (Fig. 8). It can be seen that there is no clear geographical pattern as to where schools with the most capacity strain are located. It is clear that a greater proportion of urban schools have capacity utilization > 85%, but there are also several rural schools in the highest capacity group. It is clear therefore that location and capacity alone are not enough to identify the location most in need of a new school.

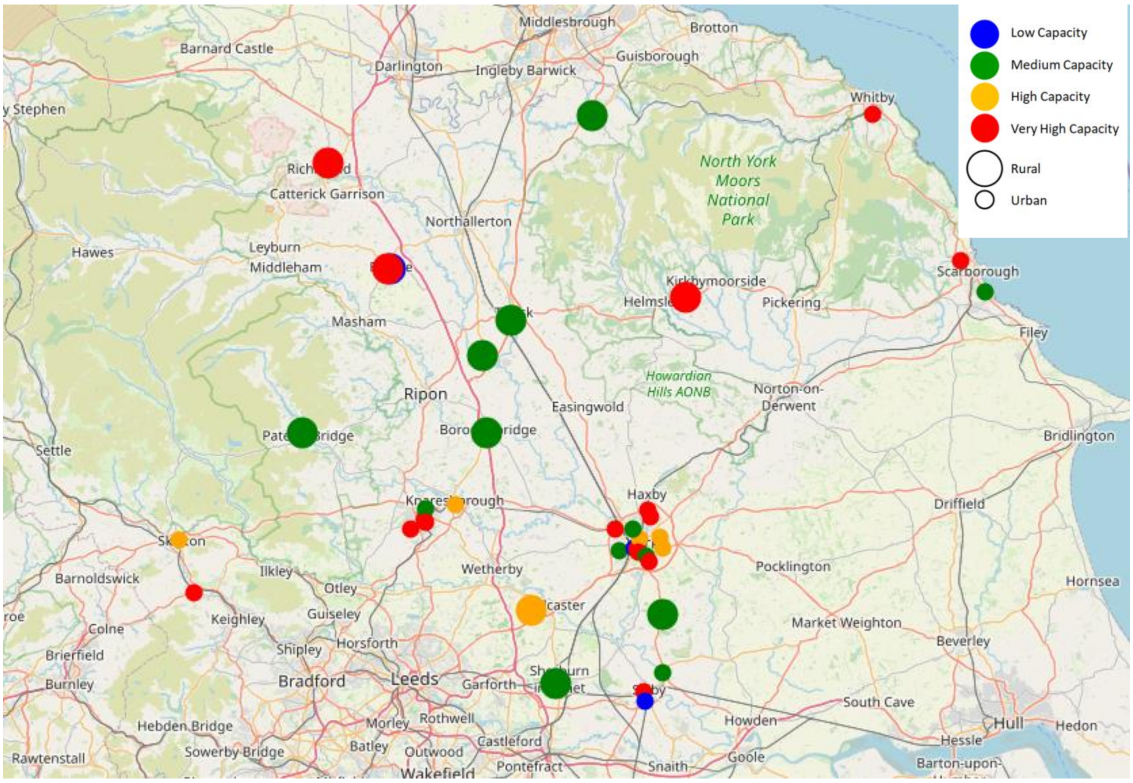


Fig. 8 – Folium map displaying school locations, grouped by capacity status & rural/urban categorization.

3.3.2. Visualisation of Population Distribution and Comparison with Capacity

When compared with the map of school locations and their associated capacity status (Fig. 8), the population distribution generally aligns well, as would be expected. It can be seen that areas with higher population density, also generally have a higher density of schools. Although there are no obvious gaps, there are some areas that could warrant further investigation. The areas of high population density around Harrogate and Scarborough, visually look to have slightly lower school provision relative to the population. This will be kept in mind as the analysis continues.

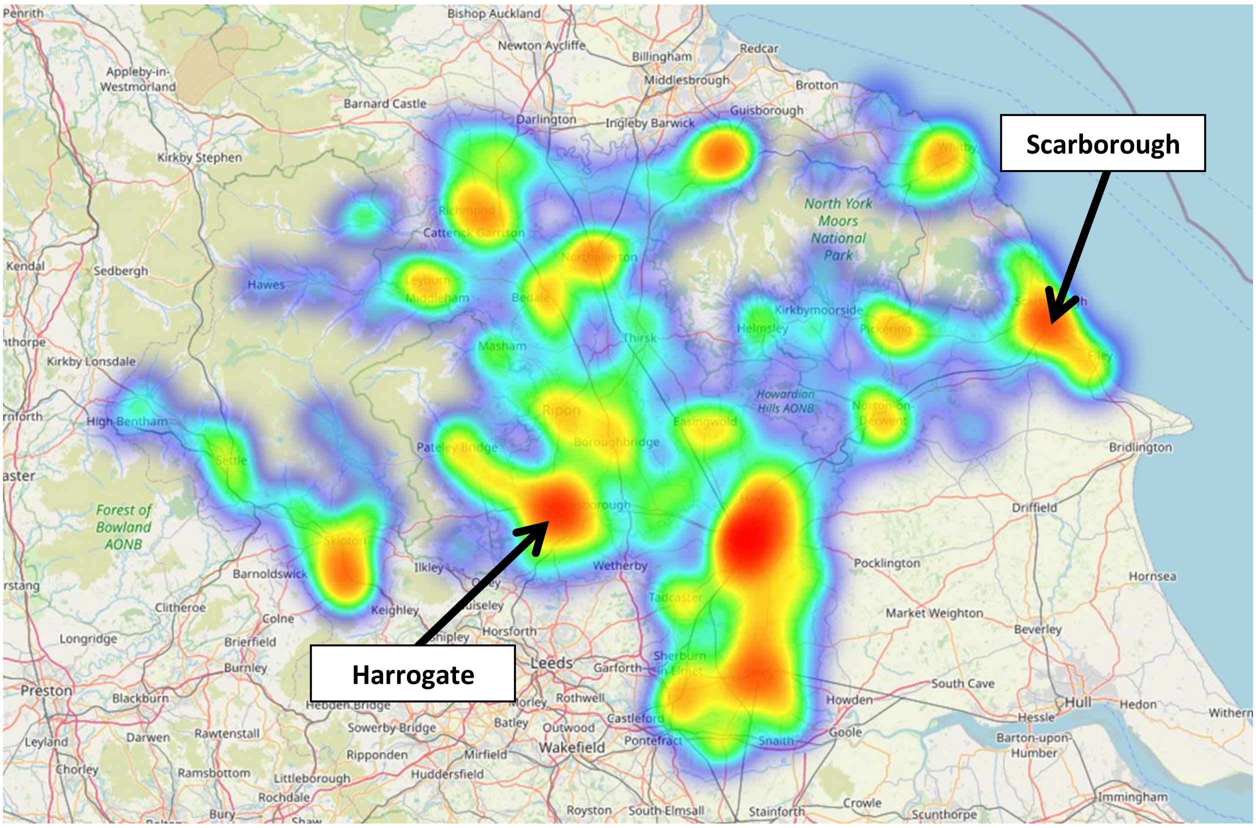


Fig. 9 – Folium heat map of high school age population distribution

3.3.3. Analysis of Total High School Age Population Associated with Each School

The number of students of high school age (11-18) who will attend each school, is considered to be a good indicator of where a new school may be most required. This can be considered alongside the current capacity status of the schools.

To carry out this analysis it is necessary to make an assumption about which school a student will attend based on their home address. It is assumed that the each student will attend the nearest school (calculated based on ‘point-to-point’ distance between home location and school location). Therefore the number of students allocated to each school can be calculated by identifying the nearest school to each LSOA (Fig. 10).

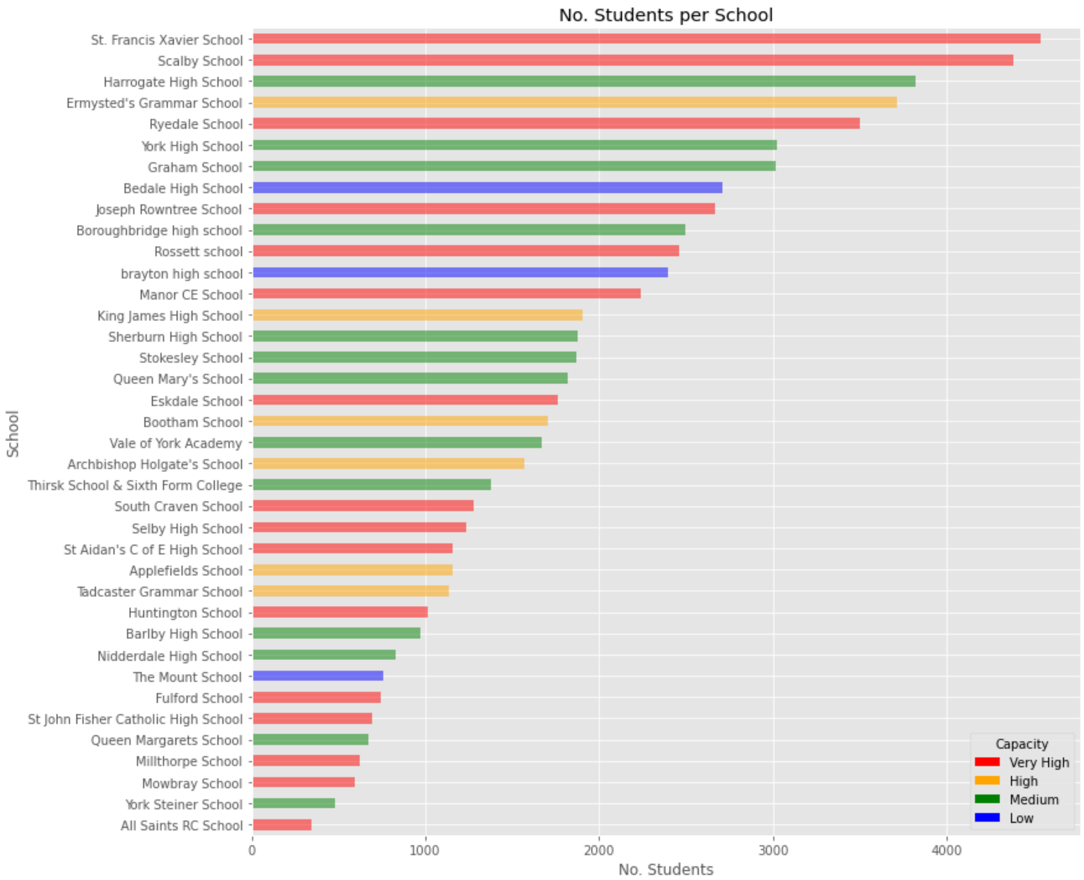


Fig. 10 – Number of students allocated to each school based on shortest distance between home location (LSOA lat/long) & school location (school lat/long). Bar colour represents the current capacity utilization of the school.

There is large disparity regarding the number of students allocated to each school, ranging from less than 500 to more than 4000. Those with a very high number of students allocated, would be worthy of further investigation as to their requirement for a new school. Although, there is no clear correlation between current capacity status and the number of students for which the schools is the nearest, but we can see that 4 of the 5 top schools are in the 'high' or 'very high' capacity groups.

3.3.4. Analysis Average Distance from Nearest School

The average distance students need to travel to their nearest school is another good indicator of where there may be greatest need for a new school. Students who travel must travel further often have longer school days resulting in increased tiredness and lower learning effectiveness.

For each LSOA, the nearest school has already been identified in section 3.3.3. Using this the distance between each LSOA and its nearest schools has been calculated and the average distance travelled for all schools plotted (Fig. 11).

The first thing to note from this analysis is that the average distance travelled for all of the schools is not particularly high (the maximum is still less than 14 km). This suggests that the distribution of the current schools is reasonably good. However, it is still worth focusing on the schools where the average distance travelled is highest, as these are the more rural schools, which implies greater variation in the distance travelled, and therefore it is likely that a number of students must still travel a significant distance to reach their nearest high school.

Similarly to the previous analysis, there is no clear correlation between current capacity status and average distance travelled, but 4 of the 5 schools requiring students to travel the furthest are also in the 2 top capacity groups.

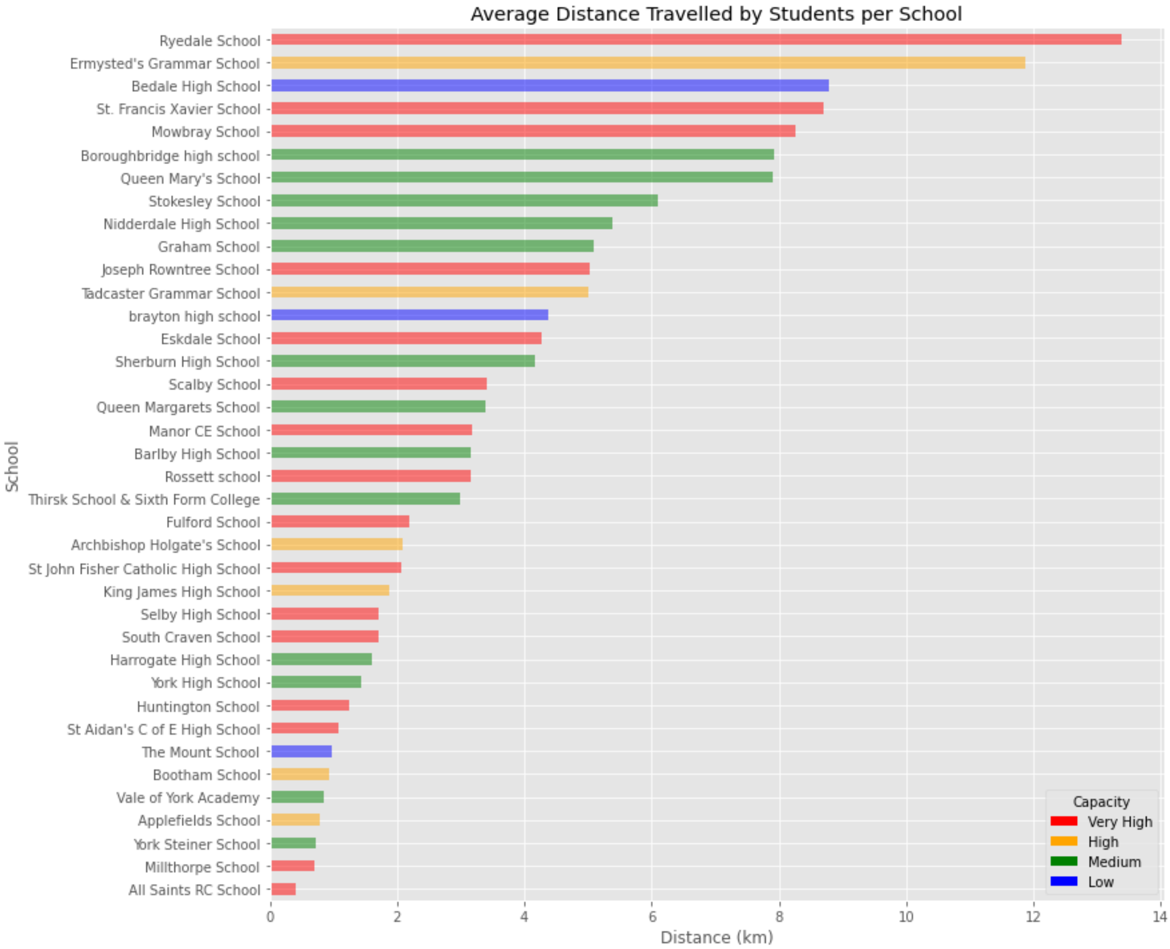


Fig. 11 – Average distance travelled by students allocated to each school. Bar colour represents the current capacity utilization of the school.

3.3.5. Comparison Between Average Distance Travelled to School & Number of Students Allocated to Each School

In the final stage of the analysis, the results from the 2 previous sections were compared.

Each school was ranked (best to worst) according to the number of students allocated to it, and the average distance students must travel to reach it.

The 2 rankings were plotted against one another to identify any correlation (Fig. 12). It is proposed that schools achieving poor rankings in both areas would be an appropriate starting point for further analysis into further, more detailed analysis as to where a new school should be sited.

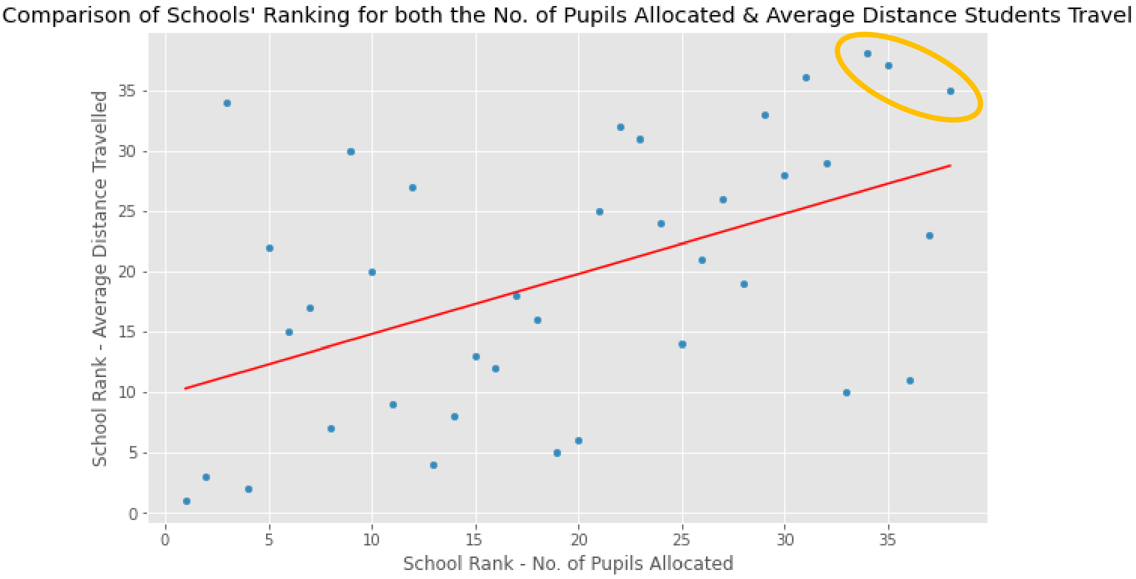


Fig. 12 – Comparison between School Ranking based on number of pupils allocated to it and the average distance students must travel to reach it. The 3 locations with the worst combined ranking are circled in orange.

Although the data is well spread, there is clear positive correlation, showing that schools with a greater number of pupils allocated are also more likely to be those where the average distance students must travel is greater.

In order to narrow down the proposed locations for a new school, the 3 locations with the worst combined ranking should be focused on. Each of these ranks in the bottom 5 of both analyses. Their location can be seen as the **red** markers in Fig. 13, and their associated locations (nearest towns) are Richmond, Helmsley and Skipton.

It is noted that these 3 schools do not have others in the vicinity, and therefore it is unlikely that students have many alternative options of schools to attend. Additionally, they all have current capacity utilisation higher than 85% (in the top 2 groups), and 2 of the 3 have capacity utilisation higher than 95%. This further supports the suggestion that these areas would be good candidates for a new school.

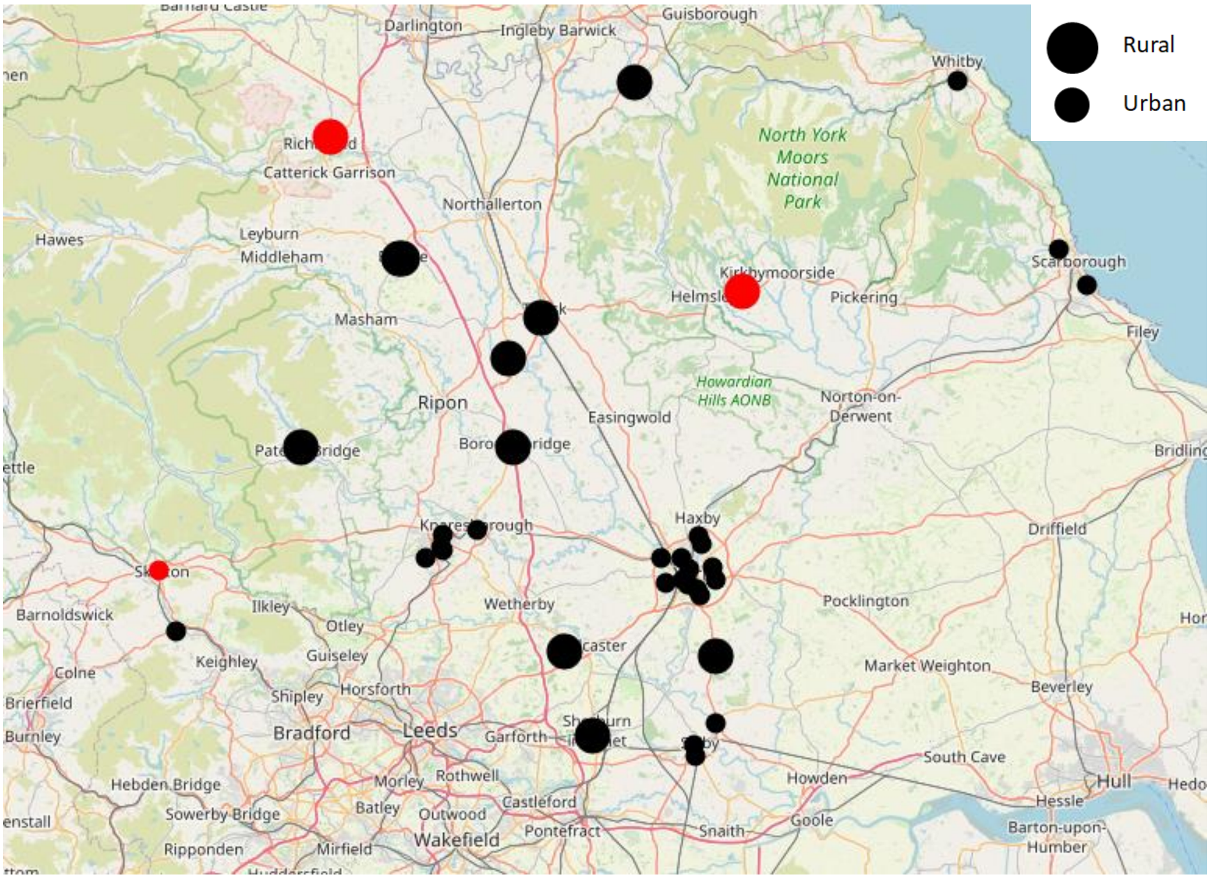


Fig. 13 – Folium map showing the location of schools, with the 3 schools identified as both requiring students to travel higher distances, and having a high number of students allocated in red.

4. Results

Schools were split according to their current capacity utilization and plotted geographically, but there was found to be no clear pattern between the location of a school and its capacity status, therefore on its own, this is not a valid predictor. Schools were also found to align well with population distribution (with more schools in areas of higher population density), although the 2 areas of Scarborough and Harrogate were identified as possible areas with high population density, and not the same density of schools.

Analysis of the number of students allocated to each school allowed schools with a very high student catchment to be identified. Several of these were also shown to have a high current capacity utilization. Similarly, the average distance students travel to each school, also allowed schools supporting a very large geographical area to be identified.

There was clear correlation between schools with a very high student catchment and those supporting a large geographical area. Three schools were focused on as ranking in the bottom 5 for both analyses. This analysis suggests that these are schools situated in areas which are struggling from a lack of school provision.

The local areas for the 3 schools previously identified (Richmond, Helmsley and Skipton) meet several of the criteria for requiring a new school. Analysis of existing schools in these areas show that they already have high capacity utilization, so cannot accept an increase in student numbers, represent the nearest school for a very high number of high school age students, and the average distance travelled by students to reach the existing schools is well above average. These areas are therefore proposed as the most worthy of further investigation as to which would most benefit from a new school.

5. Discussion

The results of this analysis give some indication of locations which would benefit most from a new school. There were 3 areas identified which meet several criteria for requiring a new school – namely the existing schools in the area have high capacity utilization, the allocated number of students is high, and the average distance students must travel is high.

These factors indicate that the existing school provision in these three areas may be insufficient, and therefore benefit could be derived from opening a new school. However, subsequent analysis would be needed to determine exactly where a new school should be situated. The analysis in this project has focused on the existing schools (status and location), but a new school should clearly not be sited in the exact location of an existing school. Therefore, a follow up analysis should focus on where (within the areas identified) a new school would result in the maximum positive impact on number of students allocated per school and on average distance travelled by students.

Although this analysis has provided a basic understanding of schools in the area, and allowed potential locations for a new school to be identified, there are also some points which have not been considered, which could impact these results. A significant point of interest is that, particularly within the York area, there are schools with limited or no spare capacity very near to schools with reasonable spare capacity. This casts some level of doubt on the assumption that each student attends the nearest school, and it is hypothesized that some of this variation may be due to perceived standard of the school (e.g. OFSTED rating). In a more follow-up analysis this may need to be included as a data feature.

5.1. Future Direction

The analysis completed in this project is relatively simplistic, and to build on this several other features could be included.

The population growth rate has not been considered within this analysis, and therefore is assumed to be consistent across all areas of the region. This is unlikely to be the case, as it growth rate is typically higher in urban areas (although this is likely to be affected by Covid-19 as there has been an increase in relocation to more rural areas).

A subsequent analysis could also include the average travelling time to the nearest school as well as average travelling distance, as this would likely to provide additional insights. Although travelling distances are typically much further in rural areas, the difference in time is likely reduced due to higher traffic levels in urban areas.

Inclusion of school type (State/Independent) is also recommended in a future analysis as complete coverage of State schools is require regardless of nearby Independent school locations.

6. Conclusion

The purpose of this project was to identify the locations within the North Yorkshire county region most in need of a new high school. By visualizing the population distribution of the region and the capacity status of existing high schools in the region, it was clear that the existing school locations provide good coverage of the region and align well with the population distribution. By further analysing existing schools on 2 factors – the number of students allocated to them, and the average distance students travel to attend the school, three schools were identified as performing poorly in both areas, suggesting that existing school provision in these areas is not sufficient. The areas surrounding these schools are proposed as starting points for further investigation as to where would be most appropriate to site a new high school

A final decision on optimum location for a new school would require significant further analysis by North Yorkshire county council, considering additional factors such as education standard of existing schools, transport links and localized rates of population growth.

7. References

1. https://www.twinkl.co.za/illustration/united-kingdom-map-with-north-yorkshire, accessed 23 January 2021.
2. https://www.doogal.co.uk/PostcodeDownloads.php, accessed 23 January 2021
3. https://hub.datanorthyorkshire.org/dataset/population-estimates/resource/d95b8528-c4fc-444f-a691-b5d8b90c365d, accessed 23 January 2021.
4. https://datadictionary.nhs.uk/nhs_business_definitions/lower_layer_super_output_area.html, accessed 24 January 2021.
5. https://www.get-information-schools.service.gov.uk/Downloads, accessed 23 January 2021.
6. https://www.ukpostcode.net/shapefile-of-uk-administrative-counties-wiki-16.html, accessed 30 January 2021.