Data Cleaning and Analysis: Census Project Report

Table of Contents

1. Overview

This report explores the census data of a relatively small town located between two considerably bigger cities which it is connected to by motorways, and makes recommendations related to:

- What to build on an unoccupied plot of land that the local government wishes to develop, and;
- What social and welfare services program the government should invest in.

To make these recommendations, the raw census dataset was cleaned – removing errors and making imputations where necessary, after which the data was analysed, and insights developed. These insights were ultimately used to guide the decision-making process relating to the decision of the local government.

2. Data Cleaning

The first step for the project involved cleaning the dataset to remove and correct any data errors. The raw data contained 9,387 rows and the data cleaning actions can be found in the corresponding Jupyter notebook. The dataset contained errors such as letters in place of words, blanks values, data in the wrong datatype and nan values.

Figure 1: Dataframe information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9387 entries, 0 to 9386
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House Number                  9387 non-null   int64
 1   Street                        9387 non-null   object
 2   First Name                    9387 non-null   object
 3   Surname                       9387 non-null   object
 4   Age                           9387 non-null   object
 5   Relationship to Head of House 9387 non-null   object
 6   Marital Status                7165 non-null   object
 7   Gender                        9387 non-null   object
 8   Occupation                    9387 non-null   object
 9   Infirmity                     9387 non-null   object
 10  Religion                      7112 non-null   object
dtypes: int64(1), object(10)
memory usage: 806.8+ KB
```

A summary of the errors and data cleaning steps taken for each of the column in the dataset is described in the table below. No data was deleted in the dataset.

Table 1: Summary of data cleaning steps

Columns in the dataset	Example errors	Actions taken
House Number	Error free	No actions needed
Street	Error free	No actions needed
First Name	Blank values	Handled using the most common first name.

Surname	Error free	No actions needed
Age	Words in place of letters, floats in place of integers – wrong datatype	Changing words to numbers and converting floats to integers
Relationship to head of house	Blank and Nan values	Imputations based on household dynamics
Marital status	Letters in place of words, blank and nan values. Also, deciding on what the law says	Changing letters to words and imputations based on household dynamics (and the law).
Gender	Letters in place of words and blank values	Changing letters to words and imputations based on household dynamics
Occupation	Blank values	Imputations based on common denominators
Infirmity	Blank values	Imputations based on common denominators
Religion	Wrong, blank and nan values	Handling obvious errors and imputations based on common denominators, such as nan values handled with common surnames or the religion of the head of house.

3. Data Analysis

The second step of the project, after data cleaning, involved exploring the dataset for trends and insights which will guide the process of making recommendations. To help with this process, additional columns were created, and they are:

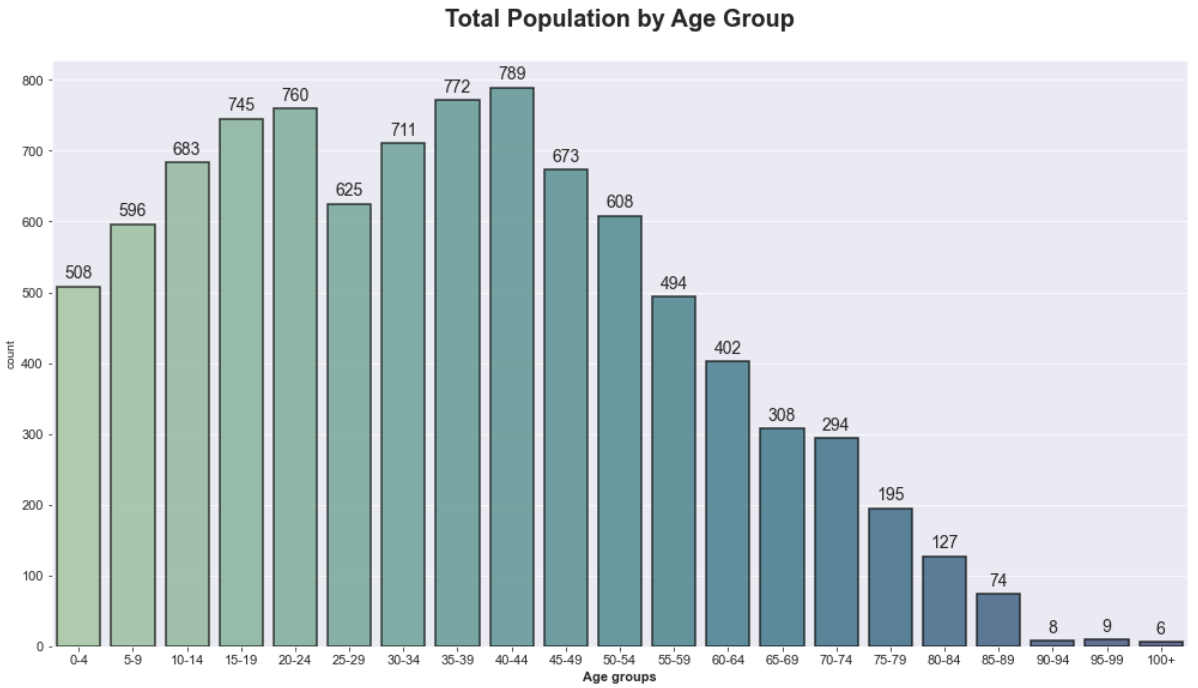
- **Age range**: 5-year age groups
- **Labour options**: used to categorise the population into employed, unemployed, students etc.

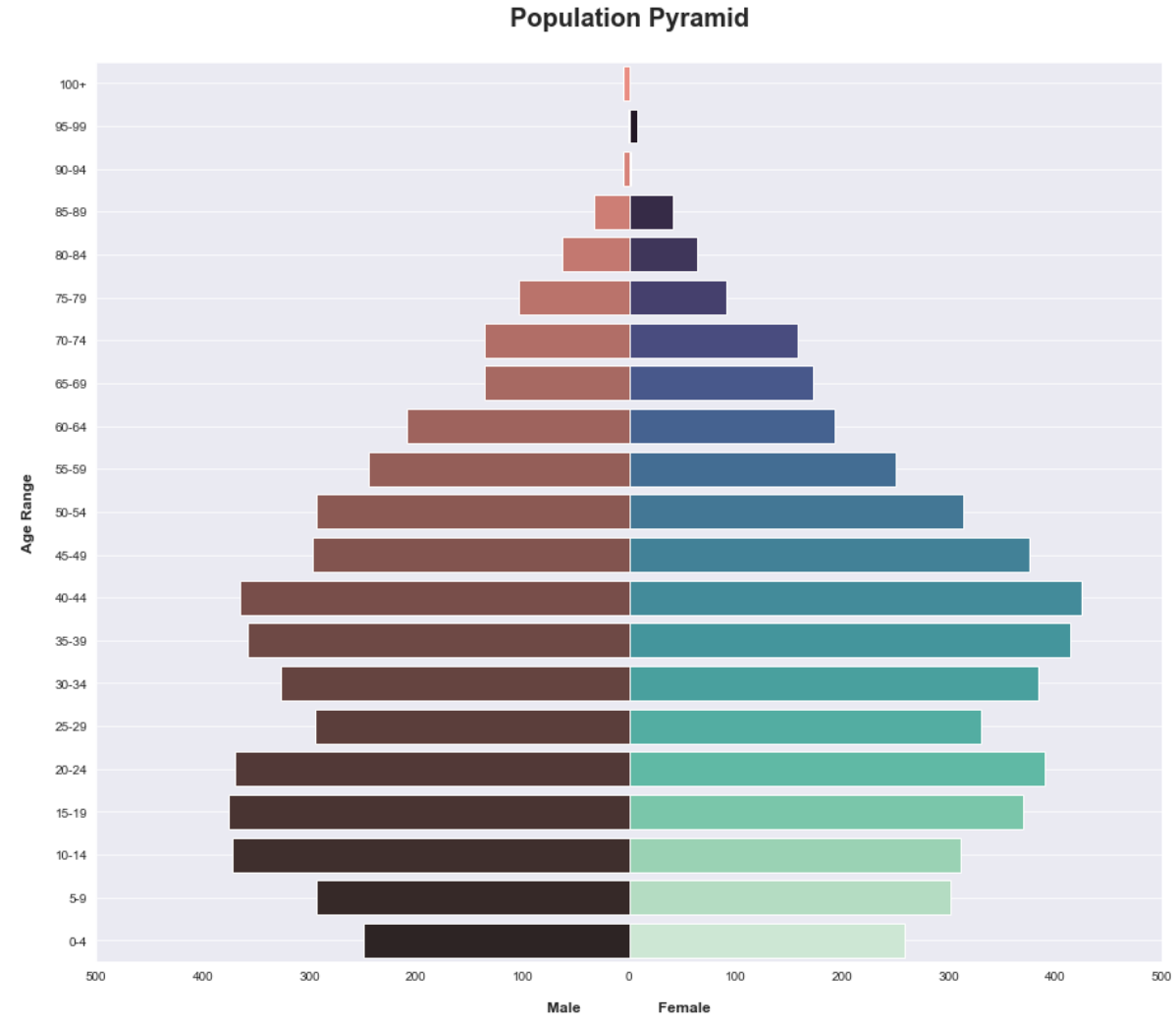
3.1. Population Dynamics

The town's population dynamics are noted below:

- A total of 9,387 people live in the town, of which 4,853 people or 51.7% of the population are female and 4,534 people or 48.3% of the population are female.
- The median age in the town is 35 years, and looking at the age pyramid, the town has a lower number of young people compared to the middle-aged population.
- Of the total population, 328 people are either lodgers or visitors.

Figure 2: Population by age group and age pyramid





3.2. Birth Rate and Death Rate

- The town's crude birth rate is an estimated 10 births per thousand. This estimate assumes that live births are the number of children aged 0 (the total number of children aged 0 are 91).
- The crude death rate is an estimated 16 deaths per thousand. Here, the ONS 1915 data on the total number of deaths in England and Wales (the earliest data available) - an estimated 1.59% of total mid-year population - was applied to the town's population data (an estimated 149 deaths).
- Based on the crude birth and death rate data above, the town's rate of natural increase is an estimated -6 person per thousand.

3.3. Migration: Immigration and Emigration

Estimates relating to migration were calculated by taking the difference in population between the different age groups, with the age-group 60 and above excluded from the calculations; when it increases this is assumed as immigration and when it reduces this is assumed as emigration.

While the above is a simplistic approach, and although it would have been easy to identify a component of the immigration numbers using university students whose relationship with head of house is none (suggesting they are not related to people in the town) and lodgers, it would have been a lot trickier identifying other components of immigration relating to employment and emigration out of the town.

Using the approach noted earlier the migration statistics were:

- An estimated 416 people are immigrants.
- An estimated 430 people have emigrated from the town.

3.4. Population Rate of Change

Based on the town's birth and death rates, and the migration data, the estimated population growth rate is -0.77%.

3.5. Marriage and Divorce Rates

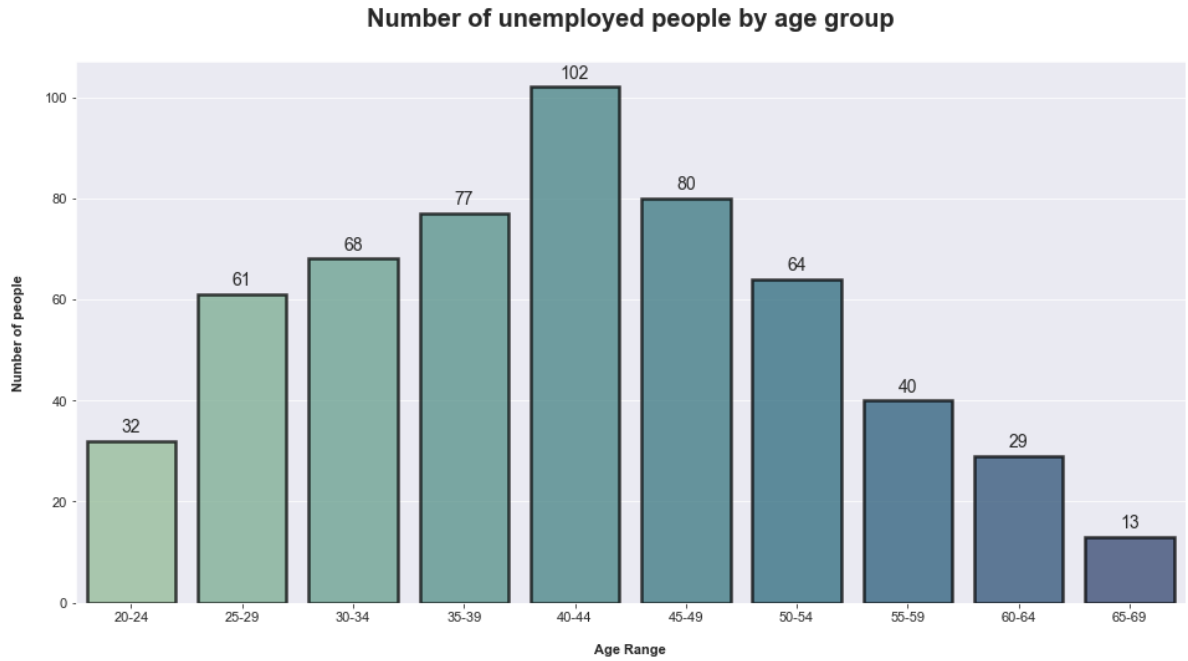
- There are 2,660 married people in the town and a total of 1,330 marriages
- The general marriage rate is 179 marriages per 1,000 (of the population age 16 and above)
- A total of 845 divorced people live in the town of which 502 of them are female
- The general divorce rate is 67 divorces per 1,000 (of the population age 16 and above)

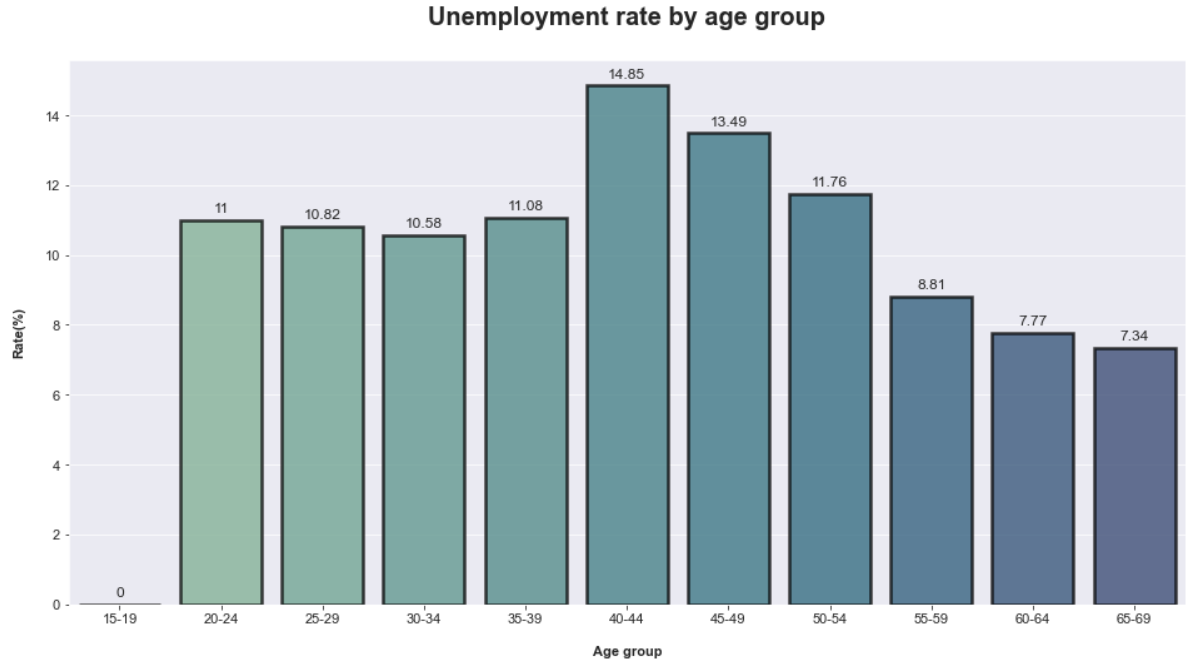
3.6. Employment and Unemployment

Based on the town's dynamics (as it relates to school enrolment, employment, and retirement), an estimated 6,217 fall under the working age population of age 18 to 67 and the labour force (working age population minus university students who are assumed are not looking for work) comprises of 5,605 people. Accordingly, of the 5,605 people in the labour force:

- 5,039 people are employed either in industry or as PhD students (89.9% of the labour force)
- 566 people are unemployed (an unemployment rate of 10.1%)
- Unemployment is highest amongst the age-group 40-44 at 14.85% which is higher than the town's average of 10.1%
- Unemployment is lowest amongst the age-group 65-69 at 7.34% which is understandable considering the retirement age is 68 years old.

Figure 3: Number of unemployed people by age group and the age range unemployment rate



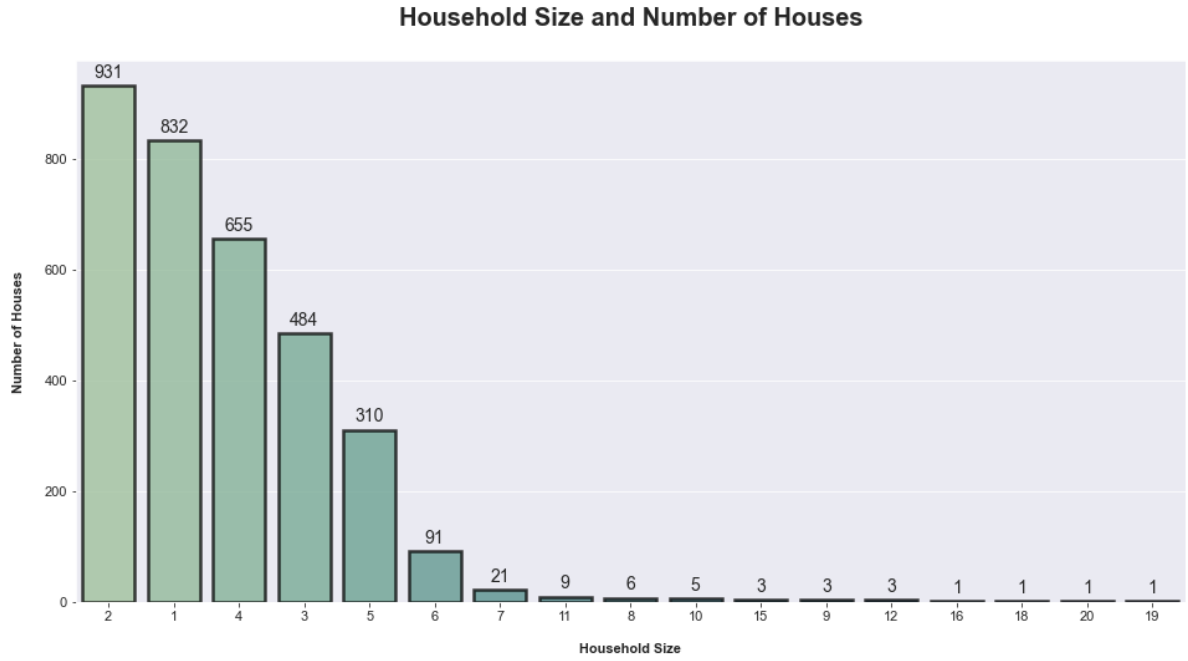


3.7. Household Occupancy

- There are 3,357 houses (households) in the town (count of house numbers and street).
- The median household occupancy (number of people in a household) is 2.

- 24.78% of the households in the town are single person households, 27.73%, 14.42%, 19.51%, 16.51% are households with 2, 3, 4 and 5 people respectively.
- 89% of the population occupy 96% of houses in the town;
    - ❖ 25% of the houses have one person living in them; with 9% of the population occupying them.
    - ❖ 28% of the houses have two people living in them; with 20% of the total population occupying them.
    - ❖ 14% of the houses have three people living in them; with 15% of the total population occupying them.
    - ❖ 20% of the houses have four people living in them; with 28% of the total population occupying them.
    - ❖ 9% of the houses have five people living in them; with 17% of the total population occupying them.

Figure 4: Household occupancy dynamics in the town



3.8. Commuters

The number of commuters in the town were estimated based on the following approaches:

- As there is no university in the town, university and PhD students are categorised as commuters.
- Lodgers and visitors are assumed to remain constant all year round and are commuters.
- Of the people categorized as employed, a sample 10% of the roles with the most people were shortlisted and a manual categorization of whether they would require commuting was carried out.

Based on the steps noted above, the estimated total number of commuters are 3,852 people; 41% of total population or 64.4% of total university & PhD students, lodgers, and visitors.

3.9. Infirmity

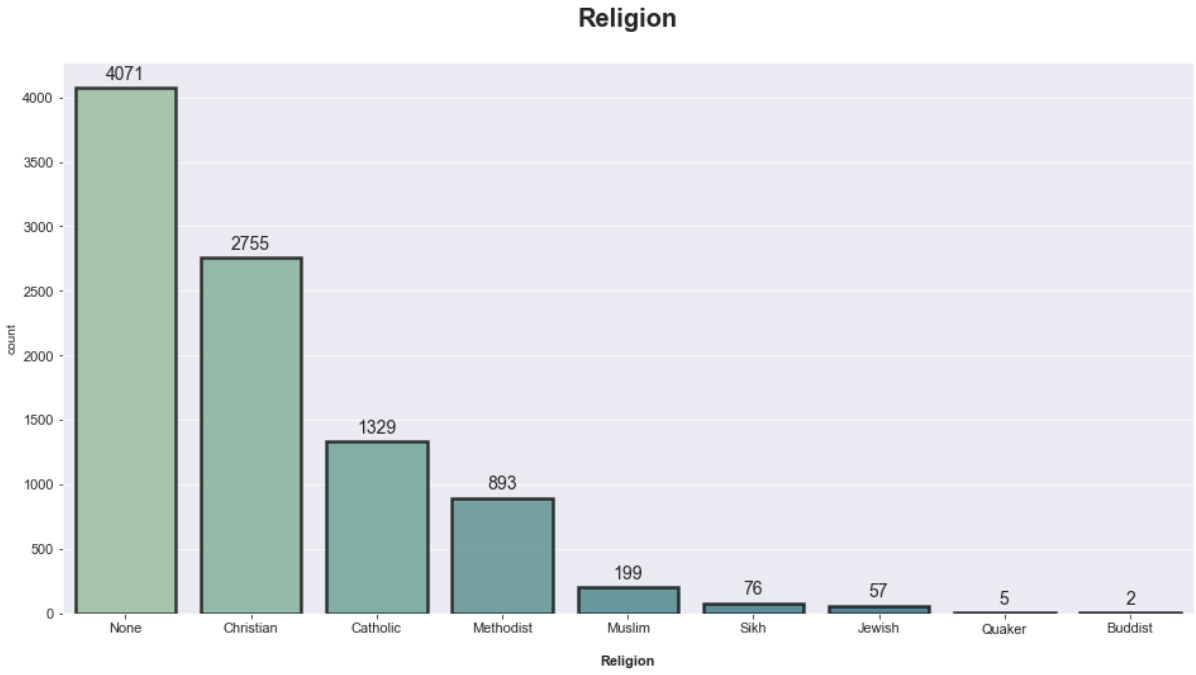
There are only a handful of people with infirmity in the town, about 61 people and roughly 0.65% of the town's population. Due to the small percentage of the population with any infirmity, any recommendation reached will not be influenced by infirmity, therefore, an extensive analysis was not conducted.

3.10. Religion

In the town, 4,071 do not practice any type of religion while 5,316 people (56.6% of the total population) practice some sort of religion. There are 8 religions practiced in the town and the major ones are:

- Christians: practiced by 29.4% of the population (51.8% of the people who practice a religion)
- Catholics: practiced by 14.2% of the population (25% of the people who practice a religion)
- Methodists: practiced by 9.5% of the population (16.8% of the people who practice a religion)
- Muslims: practiced by 2.1% of the population (3.7% of the people who practice a religion)

Figure 5: Religion dynamics in the town



4. Recommendations

4.1. Assessing the investment options

The table below highlights the options being considered by the town and the corresponding criteria for the decision-making process.

Table 2: Decision-making criteria

Question	Answer	Details	Corresponding Decision
Is the population significantly expanding?	**No**	The population is not expanding. The growth rate is negative at -0.77%.	Build low density housing
Is the population affluent?	**Yes**	An indicator for this is the housing occupancy situation in the town, where roughly 44% of the population live in 67% of the houses in the town.	Build low density housing
Are there lot of commuters?	**Yes**	There are an estimated 3,852 people classified as regular users of the road network; 41% of total population or 64.4% of total university & PhD students, lodgers, and visitors.	Build train station
Is there demand for a second church?	**Yes**	The only place of worship in town is used by Catholics even though Christians are the dominant group in the town accounting for 29.4% of the town's population; 51.8% of the people who practice a religion in the town.	Build a church
Are there many injuries in the town or likelihood of future pregnancies?	**No**	There aren't many injuries in the town, going by the infirmity numbers. Also, while the number of marriages in the town are relatively high, the median age of 35 and low number of childbirths suggests it is possible that future pregnancies will most likely not experience significant growth and remain flat.	No need for an emergency medical building
Is there evidence of a lot of unemployment?	**Yes**	An estimated 10.1% of the town's labour force is unemployed. Also, across the age groups from 20-24 to 50-54, unemployment rates are relatively high, ranging between 10.58% to 14.85%.	Invest in employment training programs
Is there evidence of increasing number of retired and old people?	**No**	The share of old people as a % of total population is 8.9%, while the old-age dependency ratio is 13.1%. Also, from the age of 45 till 69, there is a rapid decline in the population difference which could be construed to be emigration out of the town in these age groups.	Old age care while important is not a priority in the context of the town
Is there evidence of demand for schools?	**No**	Age group 0-4 is less than 5-9 which is less than 10-14. Since there is no indicator that the school system is currently stretched (there is 100% school enrolment) and considering the town's relatively low birth rate (when compared with death rate) there isn't a strong argument to increase school spending (in the absence of other data).	No need to increase spending for schooling
Is the town expanding?	**No**	The population is not expanding. The growth rate is negative at -0.77%. In addition, net migration (immigration at 416 vs emigration at 430) is relatively low at -14 which means more emigrants vs immigrants. Also, there are more deaths (149) vs births in the town (91).	New public services infrastructure is not priority as there are no indicators of an expanding town.

4.2. Evaluating the shortlisted choices

After analysing the data and based on the table above, the choices available to the town and the argument for or against them will be evaluated below. The evaluations will be based on a best-interest and investment returns case.

Table 3: Evaluation of choices

Shortlisted Options	Decision	Evaluation
Building low density housing	**No**	Going by the typical size of houses in the UK (2 and 3+ rooms per houses) and seeing that the town has a lot of spare rooms with 29% of population living in 53% of houses in the town (single- and two-bedroom households), more houses is potentially not a priority (also given the dynamics in other households) in a best-interest and investment return scenario.
Building a church	**No**	There clearly is a demand for a new place of worship for Christians. However, in a best-interest and investment return scenario, building a new church will not necessarily improve quality of life in the town, and will not provide significant returns for the council.
Building a train station	**Yes**	**The recommendation will be for the town to build a new train station.** Having explored the other options above and based on the data regarding the people who frequently use the motorways, and considering a best-interest and investment scenario, this seems to be the best option available. The reasons are: ▪ There are potentially jobs in the cities and the town could position itself as a hub for people looking to work in those cities. One criterion for jobseekers is traffic situation and how quickly they can get to work. With the high number of commuters, building a train station will have the effect of reducing traffic on the roads. ▪ Also, building a train station could open opportunities for logistic and other industries to move into the town, thereby in the long-term alleviating the high unemployment situation in the town. ▪ In addition, the benefits that comes with having new industries in the town means tax income which will then be used to invest in the options noted above.
Investing in employment training programs	**Yes**	**The recommendation will be for the town to invest in employment and training.** There is a high unemployment rate in the town. It is highly probable that this is because there are not a lot of industries in the town to accommodate graduating students and migrants moving into the town. In this case, investing in training programs will mean training people to go and provide their services in a different city (with the town positioning itself as a hub for talents seeking for the best place for their families) or start new businesses in the town which will potentially create jobs for other unemployed people. Also, investments in training would mean that potential businesses attracted to the town would have a ready pool of talents to choose from.

4.3. Recommendation after assessing options

After assessing all the options available to the local government based on the data analysis and insights gleaned, the recommendations are that a train station should be built, and the government should invest in employment and training programs.

5. Bibliography

Office for National Statistics (2017) Causes of death over 100 years. Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/causesofdeathover100years/2017-09-18 [Accessed 04/12/2022]

National Society for the Prevention of Cruelty to Children (2022) When is it legal for a child to leave home? Available online: https://www.nspcc.org.uk/keeping-children-safe/in-the-home/moving-out/ [Accessed 03/12/2022]

UK Government (2022) Implementation of the marriage and civil partnership (minimum age) act 2022. Available online: https://www.gov.uk/government/news/implementation-of-the-marriage-and-civil-partnership-minimum-age-act-2022#:~:text=The%20Marriage%20and%20Civil%20Partnership%20(Minimum%20Age)%20Act%202022%20received,on%20Monday%2027%20February%202023.&text=The%20Act%20will%20raise%20the,the%20scourge%20of%20forced%20marriage. [Accessed 03/12/2022]

World Bank (2022) Metadata Glossary. Available online: https://databank.worldbank.org/metadataglossary/gender-statistics/series/SP.DYN.CDRT.IN#:~:text=Subtracting%20the%20crude%20death%20rate,in%20the%20absence%20of%20migration. [Accessed 04/12/2022]

Appalachian State University (2002) Population Growth - An Introduction. Available online: http://www.appstate.edu/~neufeldhs/bio1102/lectures/lecture18.htm [Accessed 05/12/2022]