

## Census Project Report

The report aims to conduct a thorough examination of the census data for a moderate size town and offer a well-informed suggestion on what should build on an unoccupied plot of land.

To make this recommendation, it is necessary to first clean and analyzed the census data. Therefore, the following sections will concentrate on (1) insights gained from examining the town's population demographics, employment patterns, and other significant factors, including a detailed assessment of the religious and population growth, (2) correlations between the census information and migration patterns within the town, and (3) our final recommendation.

### Data cleaning

The census data underwent a comprehensive cleaning process to rectify any data errors, which is documented in detail in the corresponding Jupyter Notebook. The census data contains about 11 features from 8877 residents and the tabular description of the census data is shown in Fig.1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8878 entries, 0 to 8877
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   House_Number                         8878 non-null   int64
1   Street                               8878 non-null   object
2   First_Name                           8878 non-null   object
3   Surname                              8878 non-null   object
4   Age                                  8878 non-null   float64
5   Relationship_to_Head_of_House        8878 non-null   object
6   Marital_Status                       6911 non-null   object
7   Gender                               8878 non-null   object
8   Occupation                           8878 non-null   object
9   Infirmary                            8878 non-null   object
10  Religion                             6857 non-null   object
dtypes: float64(1), int64(1), object(9)
memory usage: 763.1+ KB
```

Fig.1 Dirty census data

The Fig.1 depicted above shows the presence of missing values in both the marital status and religion columns. Furthermore, it was observed that the age column was in a float format instead of an integer format, and it was corrected accordingly. The data set for the house number, street number, and first name column did not contain any missing values, empty strings, and inconsistencies. The cleaning process for the remaining columns is outlined below

**Religion:** All the missing values were found to correspond to minors (under 18). Therefore, these missing values were filled as undecided. Moreover, some individuals reported their religion as "nope", "Sith", and "private" which are not recognized religions in the UK. To rectify these errors, information was gathered from the participant's immediate family or close relatives, using their street, house number, and surname as filters. Additionally, some participants reported specific Christian denominations such as Methodist or Catholic instead of their actual religion. These were classified as Christian for analysis purposes.

**Marital Status:** The primary entry error observed was the recording of NaN (Not a Number) values for minors (under 18). According to the Marriage Act of 1949 (s3), In the UK, the minimum age for marriage in the UK is 18. Thus, the NaN values were modified to indicate "NA(minor)". The only other missing value was updated to indicate "husband" after verifying the relevant relative information.

**Relationship to the head of the house:** Three single mothers, all under the age of 17, claimed to be the head of their respective households. However, this is impossible, as one must be at least 18 years old to assume that role. It was concluded that they had lied about their age. Consequently, their age was adjusted to 26, which is the median age of women between 18 and 34 (fertile age group).

**Others:** The gender, surname, infirmity, and occupation columns in the data contained empty strings, and some individuals had provided inaccurate occupation information. To rectify the error, we used data from an individual's record or from others in their household (for gender) to make inferences. When referential information was unavailable, missing records were replaced with 'Unknown' (in the case of surname). The 9 missing values in the infirmity column were all filled with 'None', which is the most frequently occurring data point, and the occupation column was also corrected accordingly.

Individuals who were recorded as unemployed and over 65 years of age had their employment status changed to "Retired" Fig.2. Although it's possible for someone can be unemployed at retirement age, it's common practice to exclude such cases from future analyses of unemployment, as individuals over 65 are typically not considered eligible for work as per UK government guidelines (Gov. UK).

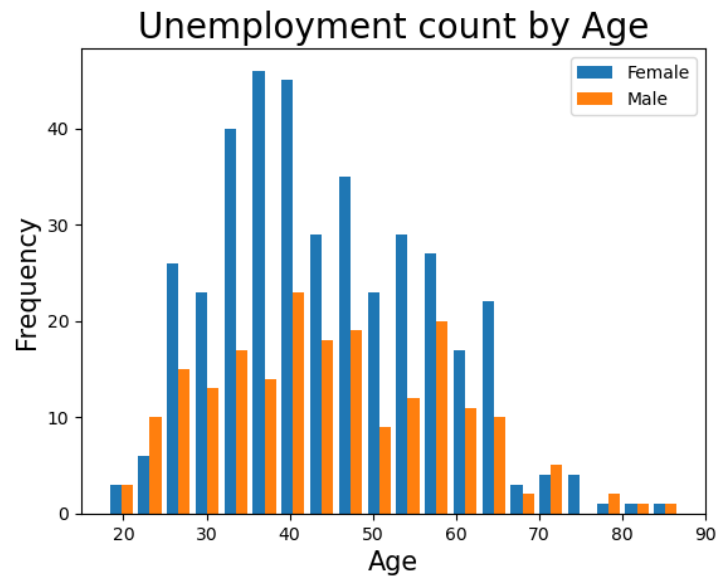


Fig.2: Plot of Unemployment against age

## Population Demographics

After completing the data cleaning process, the finalized census data will possess the following features Fig. 3.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8878 entries, 0 to 8877
Data columns (total 12 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   House_Number                         8878 non-null   int64
 1   Street                               8878 non-null   object
 2   First_Name                           8878 non-null   object
 3   Surname                              8878 non-null   object
 4   Age                                  8878 non-null   int32
 5   Relationship_to_Head_of_House        8878 non-null   object
 6   Marital_Status                       8878 non-null   object
 7   Gender                               8878 non-null   object
 8   Occupation                           8878 non-null   object
 9   Infirmary                            8878 non-null   object
10  Religion                             8878 non-null   object
11  updated_occupation                   8878 non-null   object
dtypes: int32(1), int64(1), object(10)
memory usage: 797.8+ KB
```

Fig.3 Cleaned census data

To aid analysis, the following has been added:

- Age grouping: Ages placed into 5-year age bands for the population pyramid.
- Occupation category:
  1. All the retired occupations were grouped as retired.

2. An "Updated\_occupation" column was added to the database. It includes types such as Retired, Employed, Unemployed, Child (Age group: 0-4 years), and Student (Age group: 5-30 years).

The age distribution plot Fig 4 indicates show a Gaussian distribution, with a mean and median age of 37 & 36 respectively. Moreover, more than 75% of the population is 53 years old or younger, and the oldest person is 105. overall, it indicates a predominately young community.

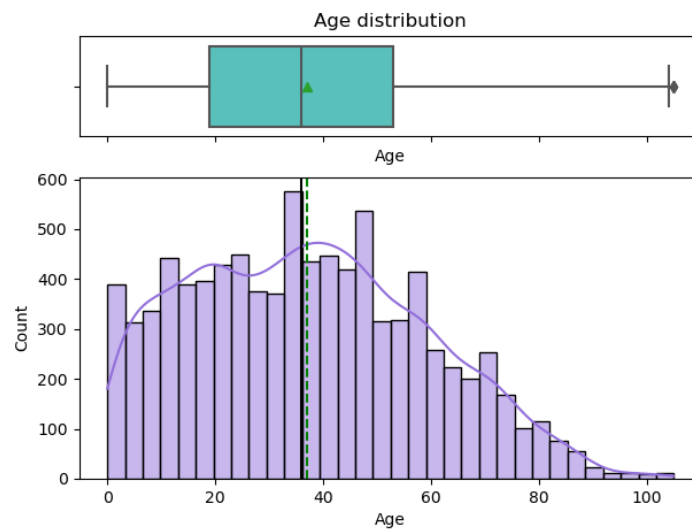


Fig.4 Plot of age distribution

To gain a better understanding of the demographic characteristic of the population, an age pyramid was plotted as shown in Fig.5. The pyramid shows that there are more middle-aged individuals than younger people. There is also a significant drop in the early age group (0-4 years), indicating a low birth rate. One notable deviation is a decrease in population between ages 24 to 34. This fluctuation could be explained by university students and early career seekers leaving the town.

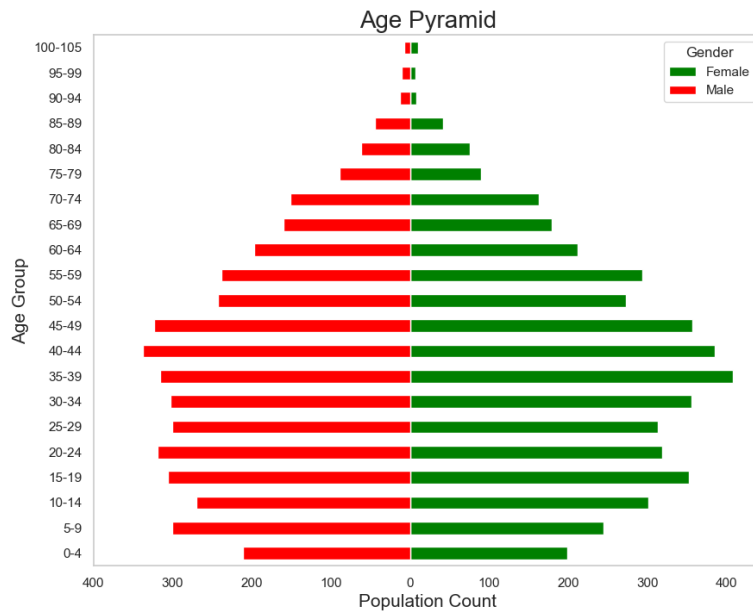


Fig.5 Age pyramid

Further analysis revealed that around 80% of the population comprises either students or workers, with workers making up roughly 53% of the total population. This suggests that the unemployment rate is relatively low (~ 6.3%) as shown in Fig. 6. The male-female population ratio is almost equal, with only a slight difference of 410 more females than males. Most adults in the town are either married or single Fig. 7, and the incidence of infirmity in the town is relatively low compared to variables, with less than 1% of the overall population reporting any type of infirmity. This indicates a generally healthy society.

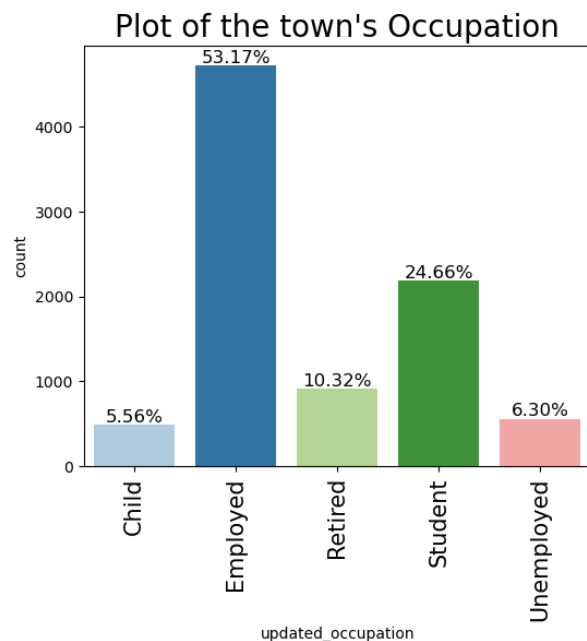


Fig.6 Plot of the town's occupation

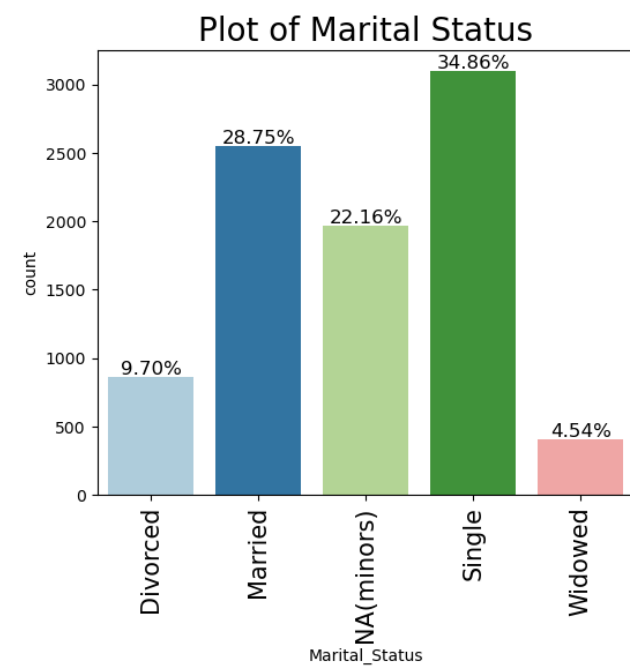


Fig.7 Plot of Marital Status

Tabel.1: Descriptive statistics

#### Gender

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0

Female	4644
Male	4234

#### Infirmity

Categorical

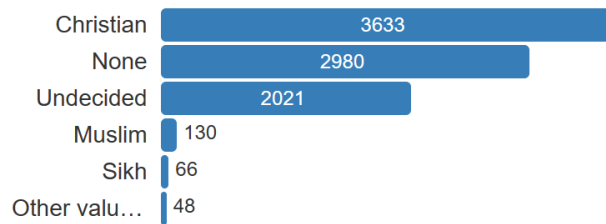
Distinct	7
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%

None	8813
Physical D...	18
Disabled	11
Mental Dis...	10
Deaf	9
Other valu...	17

## Religion

Categorical

Distinct	9
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%



### Data Insights and Analysis:

Given that infirmity constitutes such a small proportion of the population, it is not considered significant enough to warrant extensive analysis. Therefore, no recommendations will be provided based on infirmity.

### Religion

The impact of replacing the religion of young children with "undecided" is clearly visible in the figure below. It becomes challenging to analyze the religious transmission effectively from parents to children since most children now have an undecided religion as shown in Fig.8. However, it is still possible to make a rough estimation using the major religions (Christianity, Islam, Sikhism, and Judaism). Out of 2799 parents with a declared religion, only 353 children had a declared religion, resulting in a low transmission rate of 13%.

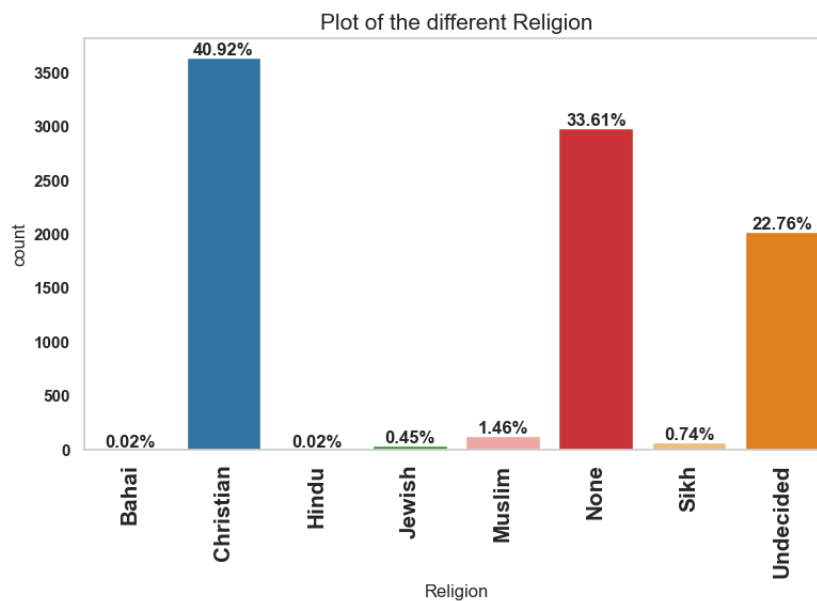


Fig.8 Plot of the different religions

Moreover, the portion of residents who identify as having "none" as their religion suggests that younger individuals in the town are less likely to affiliate themselves with a particular religion as shown in Fig. 9. This trend may be the reason why older religions in the UK such as Christianity and Judaism have fewer young adherents as shown in Fig.10.

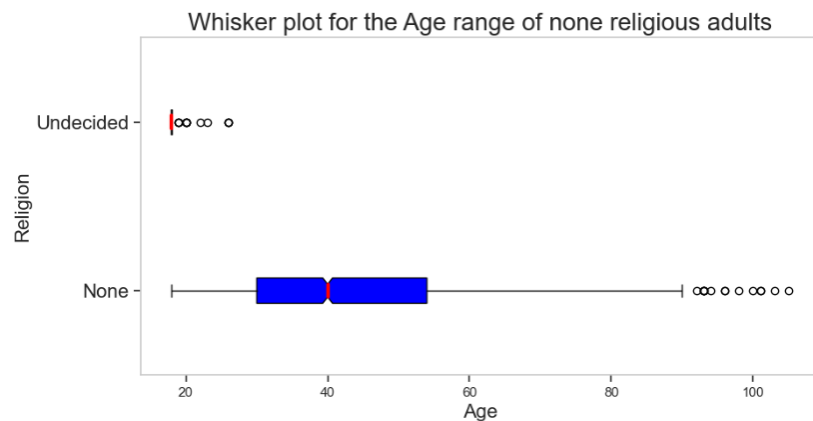


Fig. 9 Plot of none religious adults

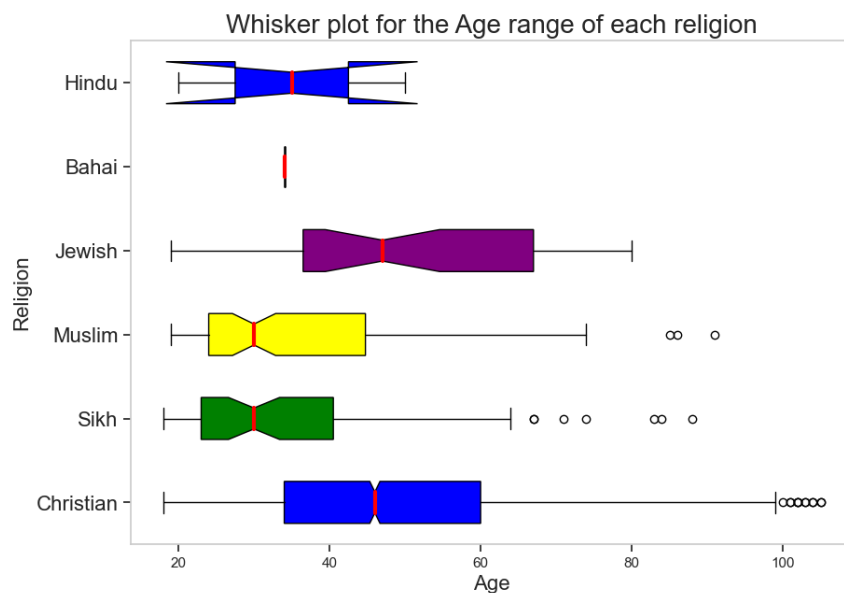


Fig.10 Plot of the age range of each religion



Despite the potential for growth in some emerging religions, such as Islam and Sikh, as evidenced by their slightly lower median age of followers Fig.10, they still represent a relatively small percentage of the population and are not significant enough to justify building a new church.

### Divorce and Marriage

As shown in Fig. 11 below, it was found that divorce occurs across all age groups and there are more female divorcees than males, which suggests that male divorcees may leave the town.

To calculate the crude divorce rate, marriage rate, and divorce-to-marriage ratio, we need to use the number of divorced women as the baseline. The marriage rate can be calculated by counting the number of 'married' individuals and dividing it by two. The resulting divorce-to-marriage ratio is 8.84% divorces per 1276 marriages. The number of female divorcees will be used for calculating the net migration in the subsequent section.

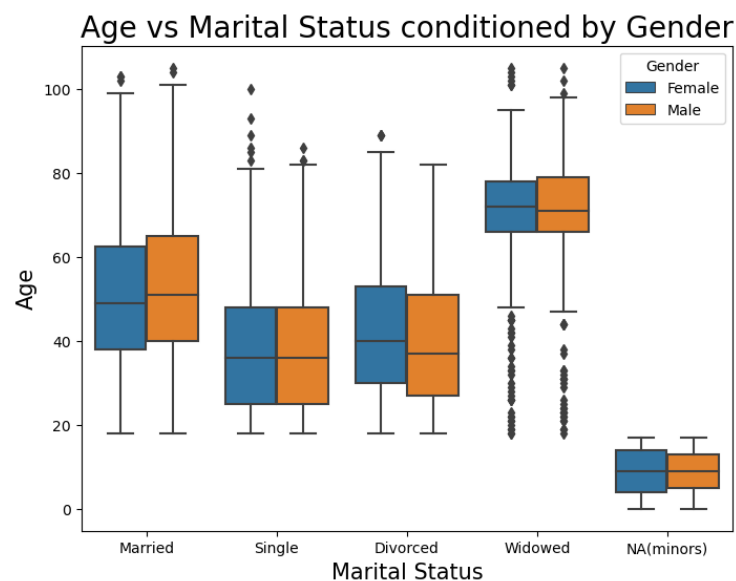


Fig.11 Plot of age vs marital status conditioned by gender

### Occupation:

The available data on occupation is limited, however, it is possible to make some general classifications based on available information. The occupations can be classified into three categories: outbound (those who work outside of the town), inbound (those who work within the town), and unknown (those whose occupations cannot be categorized due to insufficient information about the town).

Based on the available information, certain jobs such as community worker, elementary, secondary, and continued education officer, police officer, lawyer, and more can be classified as inbound occupations. This suggests that there are individuals who work in the town and commute within it. Approximately 89 jobs were considered outbound jobs, indicating that there are individuals who work outside of the town and commute in. This means that a total of 650 people, which is roughly 14 % of the working population, are verified commuters. However, the actual number of commuters could be higher.

The presence of English as a second language teachers and other similar jobs is an indication of possible foreign immigration into the town. Furthermore, the town has two established industries, namely the movie industry and the mining industry.

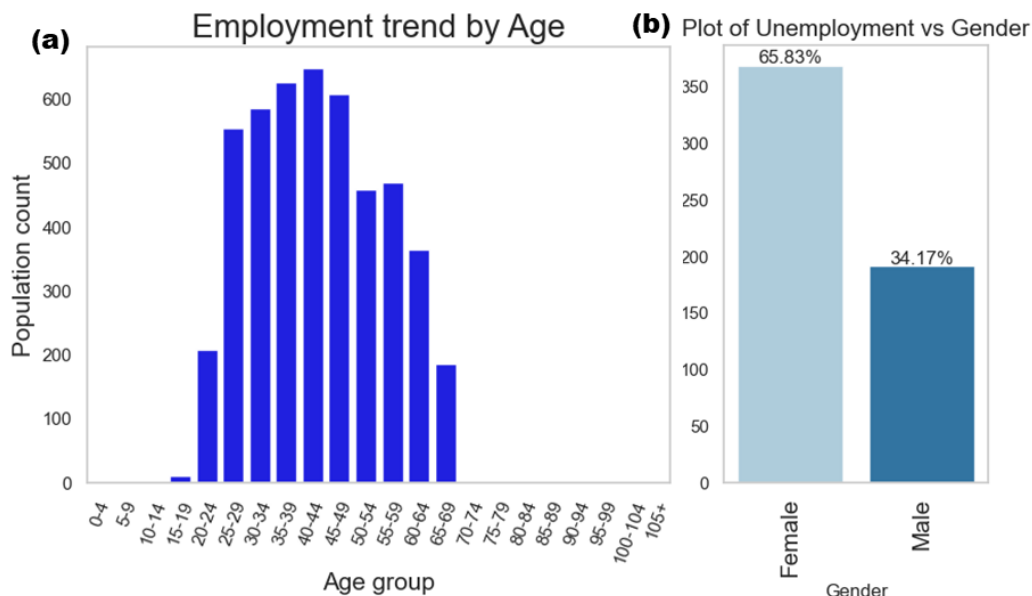


Fig.12 Employment trend

As illustrated in Fig.12(a), Individuals aged 20 to 69 years old are typically considered the most effective for employment, but there is still a small number of individuals aged 15 to 19 who are employed. Furthermore, the minimum age for unemployment is 18 years old and the retirement age is 65. Female residents experience higher unemployment rates than males, even though there are more females than males within most age groups as shown in Fig. 12(b).

Finally, the town's overall unemployment rate is 6.3% and since there is a sizable portion of the population made up of students, there will be a need for job creation in the coming years.

### **Birth and Death Rate:**

When calculating the birth rate, two factors were taken into account. Firstly, the total number of women in the population was used instead of the total population. Secondly, women age 25 to 34 at (five-year intervals) were considered.

$$\left( \frac{\# \text{ births in 1 year}}{\# \text{ women in 5 years}} \right) * 1000 = \text{Births per } \#1000$$

Currently, the birth rate is approximately 270 births per 1000 individuals. However, five years ago, the birth rate was estimated to be 290 births per thousand. As a result, the birth rate has decreased by 20 children per 1000 individuals over the past five years.

To estimate the birth rate over 20 years, one additional consideration was made. (1) the maximum age band for birth rate estimation was set to 50-54 to avoid bias introduced because the female population decreases drastically after this age group.

$$\left( \frac{\# \text{ births in 1 year}}{\# \text{ women in 5 years}} \right) * \# \text{ total women in town} = \text{Births per } \# \text{ women}$$

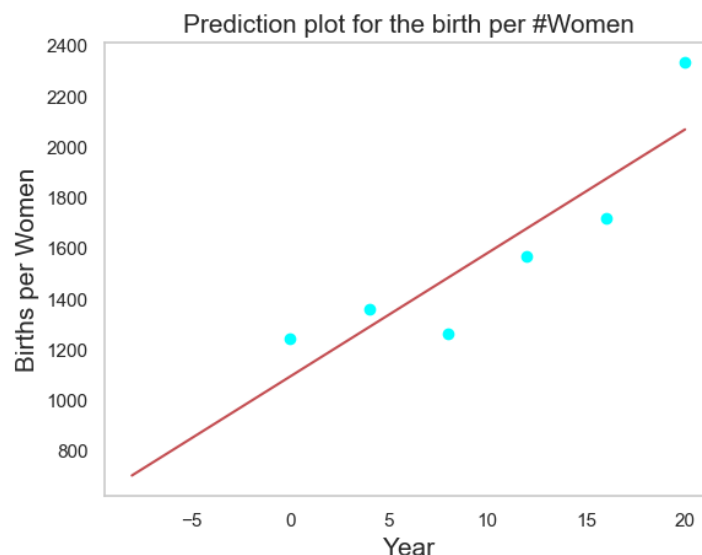


Fig. 13 Predicted plot of birth per women

The plot in Fig 13 indicates a general decrease in the birth rate over 20 years. Based on the overall trend, it is projected that there will be an estimated 897 and 702 births per total number of women over the next 4 and 8 years, respectively.

$$\left( \frac{\# \text{ deaths in 5 year}}{\# \text{ people in previous 5 years} * 5} \right) * \# \text{ total population in town} \\ = \text{deaths for \#women} * 2 = \text{deaths for \#population}$$

The death rate is determined by using the provided equation to estimate the number of deaths based on the difference in age groups. Negative values in the death rate in Fig. 14 indicate an increase in the population instead of a reduction. Typically, the death rate rises with age, except for the 95-year-old age group, which stands out as an outlier as shown in Fig. 14. The projected death per #population for the next 5 and 10 years is 934 and 900 death, respectively. Compared with the projected birth rate analysis, the population of the town is expected to shrink. However, to gain a more comprehensive understanding of the town's population, we will now examine the influence of immigration and emigration.

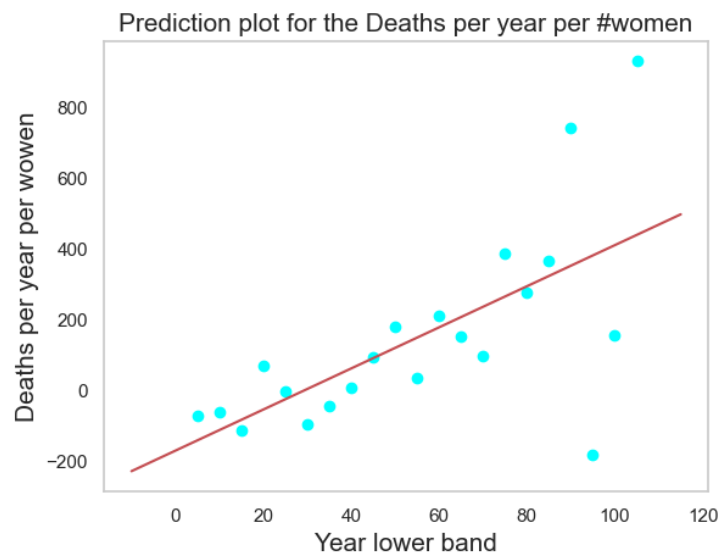


Fig.14 Prediction plot for death per year per women

### Immigration and Emigration:

The majority of emigrants in the town are university students and divorced couples, while interestingly, university and Ph.D. students make up the most significant groups of immigrants. As a result, migration by students is not considered in the major decision population estimation.

The focus of immigration analysis is on lodgers, visitors, and individuals with unknown values, while the focus of emigration analysis is on divorced couples leaving the town. Based on this assumption, it is expected that 721 immigrants will remain in the town, while 212 are projected to leave after their stay. Additionally, 497 divorced couples chose to stay, while 364 left. Women accounted for 377 of the immigrants and 190 of the emigrants. Overall, the net

migration is positive (523) and will significantly contribute to the overall towns of about 146 per 1000 people.

### Occupancy Rates:

The town occupancy rate varies from street to street and with a range of 1 to 5 occupants per house. This range of values was utilized to measure the density of each street by computing the difference between the occupancy rate and the actual number of houses. The resulting range of values for each street is displayed in Fig.15.

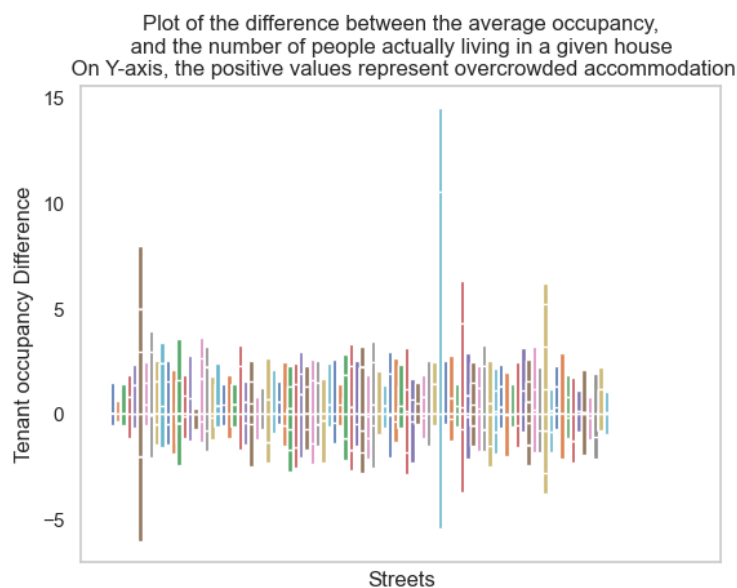


Fig.15 Plot showing the difference between average occupancy and the actual number of people living in a given house

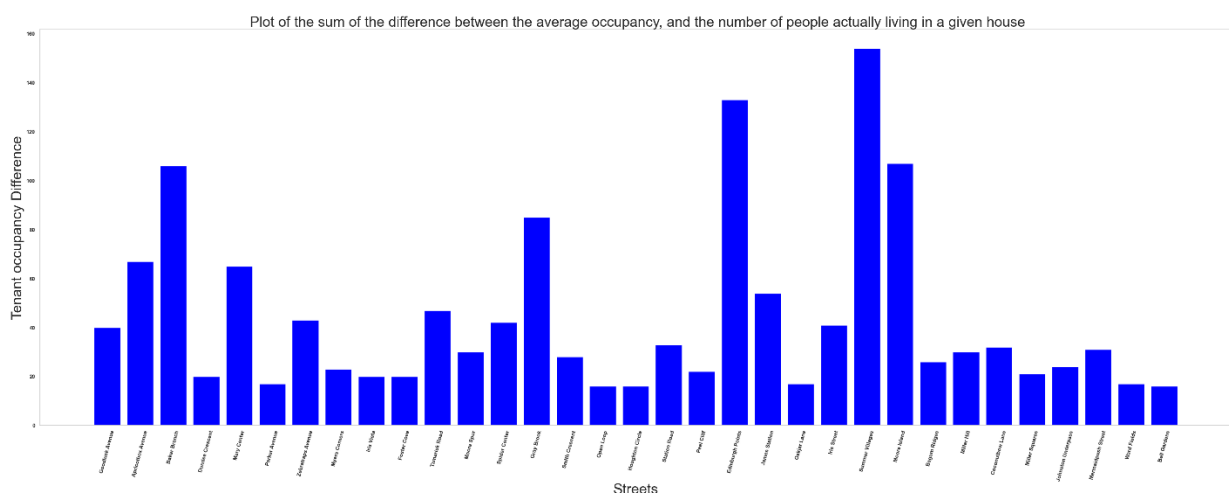


Fig.16 Plot showing the most densely occupied streets

By analyzing the plot of the sum of differences per street, it was noticed that most streets have a dense population, as shown in the positive bar chart. Only streets where the sum difference was greater than 10 were included in the plot to improve clarity. These streets were grouped into three categories: less dense, dense, and highly dense. The high population density may be attributed to various factors such as families with high incomes living together, a large group of students sharing a residence, lodgers staying with families that have children, and divorcees renting rooms to supplement their income, among others. Overall, it's evident that the town faces a shortage of housing.

### **Recommendations and Conclusions:**

The town is relatively healthy, with a low incidence of illness, making the construction of a new health facility unnecessary. Additionally, the town is well-connected to two larger neighboring towns, and the existing transportation system is sufficient, as evidenced by the high number of commuters. Moreover, the town's major religions are seeing a decline in their younger members, making the construction of a new religious center unnecessary.

However, although the town's population is projected to decline in the next decade, there is still a significant overcrowding issue in the current housing situation, and the town's high life expectancy means that people will continue to live longer. Therefore, the construction of new, low-income housing is strongly recommended to address the problem of congestion and to meet the need of large families.

Furthermore, unemployment levels in the town are high, and most graduates either leave or remain to compete for scarce employment opportunities. Therefore, investment in employment and training, or retraining people for new skills, is essential.

Investment in old age care is also a possible option, given the expected increase in the number of retirees due to the high life expectancy in the town. Additionally, with the declining birth rate, the town will increasingly rely on immigration to sustain its workforce in the coming years.

## **Bibliography:**

Census 2021 results - Census 2021 (no date) Census 2021 Results - Census 2021. Available online: <https://census.gov.uk/census-2021-results> [Accessed 28/04/2023].

Data Visualization with Python (2022). Available online: <https://www.youtube.com/watch?v=icOcecahe5Y> [Accessed 28/04/2023].

Geeks for Geeks (2023) Box Plot in Python using matplotlib.  
Available online: <https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib>  
[Accessed 5/4/2023]

Lesson 8: The Cohort Component Population Projection Method — MEASURE Evaluation (no date). Available online: <https://measureevaluation.org/resources/training/online-courses-and-resources/non-certificate-courses-and-mini-tutorials/population-analysis-for-planners/lesson-8.html> [Accessed 28/04/2023].

Marriage Act (1949) Section 3.  
Available online: <https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3>  
[Accessed 26/4/2023]

Office for National Statistics (2023) Labour market in the regions of the UK: April 2020. Available online: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/regionallabourmarket/july2023>  
[Accessed 16/4/2023]

Wikipedia (2023) Religion in the United Kingdom.  
Available online: [https://en.wikipedia.org/wiki/Religion\\_in\\_the\\_United\\_Kingdom](https://en.wikipedia.org/wiki/Religion_in_the_United_Kingdom)  
[Accessed 16/04/2023]