FUNDAMENTALS OF DATA SCIENCE PROJECT REPORT

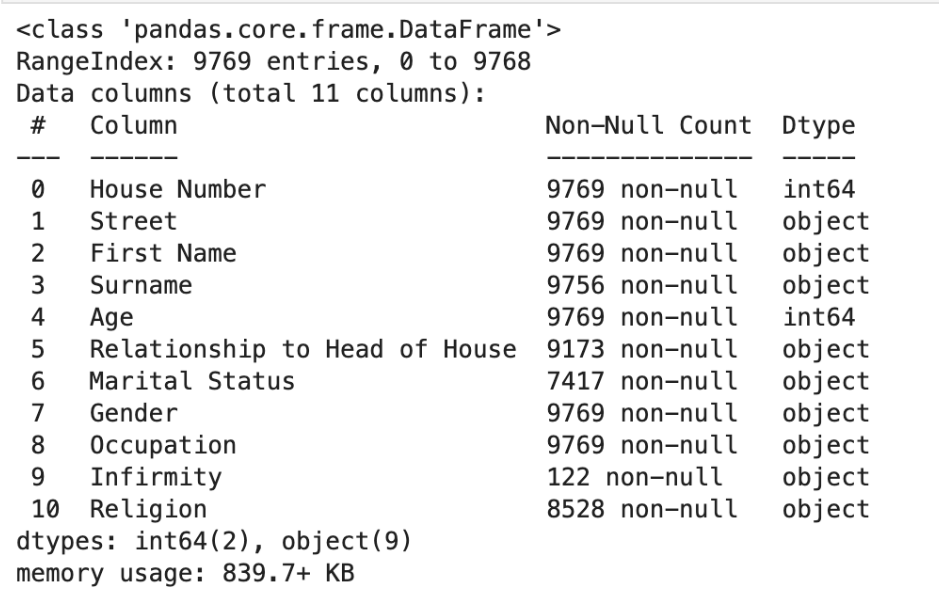
1.0 INTRODUCTION

The aim of this project is to analyze the mock census data and provide insights and recommendations for the local government on how best to utilize an unoccupied plot of land and where to allocate funds in. The analysis underwent thorough data cleaning processes, which involved identifying missing values, suspicious or inconsistent entries, errors and replacing with more suitable values using the Jupyter notebook executed in a Python environment. The results and suitable recommendations of the analysis has been presented.

2.0 DATA CLEANING

The mock census data presented contained missing values, errors and suspicious entries which were thoroughly cleaned using a combination of different cleaning techniques. The dataset contained 11 columns which provided information of 9769 individuals living in the town. The data cleaning process started by:

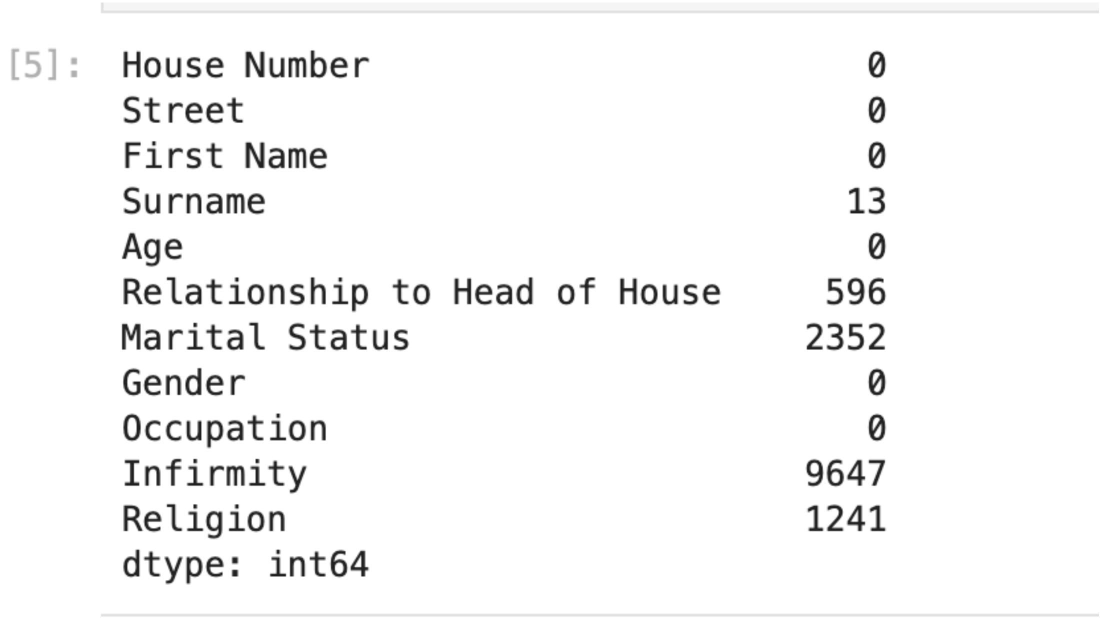
1. Examining the summary of the dataset
2. Identifying the columns with missing values
3. Checking for empty strings and converting to NaN for quicker cleaning
4. Identifying suspicious / false entries and replacing with appropriate values
5. Identifying general errors



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9769 entries, 0 to 9768
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House Number                  9769 non-null   int64
 1   Street                        9769 non-null   object
 2   First Name                    9769 non-null   object
 3   Surname                       9756 non-null   object
 4   Age                           9769 non-null   int64
 5   Relationship to Head of House  9173 non-null   object
 6   Marital Status                7417 non-null   object
 7   Gender                        9769 non-null   object
 8   Occupation                    9769 non-null   object
 9   Infirmity                     122 non-null    object
 10  Religion                      8528 non-null   object
dtypes: int64(2), object(9)
memory usage: 839.7+ KB
```

Figure 1: Summary of the dataset

Figure 1 above shows the details of the dataset, number of columns and entries



```
[5]:  House Number                      0
      Street                            0
      First Name                        0
      Surname                          13
      Age                               0
      Relationship to Head of House   596
      Marital Status                 2352
      Gender                            0
      Occupation                        0
      Infirmity                      9647
      Religion                       1241
      dtype: int64
```

Figure 2: Columns with NaN/missing values

Figure 2 above shows columns with NaN/missing values

The data cleaning techniques for each column are as follows.

Surname: The imputation technique used to fill in missing value in this column was achieved by first grouping by households by leveraging House Number and Street and then assigning the head's surname to household members with familial relationships (e.g., Husband, Wife, Son, Daughter) and then assigning 'Unknown' to unidentified occupants.

Age: A check was carried out to find any suspicious entries, some individuals below 18 had their marital status as "Married," "Divorced "and "Widowed", these are minors (GOV.UK, 2023) so the entries was replaced with N/A(Minors).

Relationship to Head of House: For single individuals, random sampling was adopted assigns relationships. A check for suspicious entries was also carried out i.e. individuals that have been assigned head and less than 18. In the UK only an adult (above 18) can own head of a house, and yan individual must be 18 to be considered an adult (Office for National Statistics, 2019). These entries were replaced with values more appropriate.

Marital Status: A check was carried out to see if the missing values in the column are minors, this returned true. In the UK the legal age to get married is 18 (**GOV.UK, 2023)** the marital status for individuals below 18 were filled with 'N/A(Minors)'

Infirmity: According to the 1881 census style only "deaf and dumb, blind, imbecile or idiot, or lunatic" were recognized as Infirmity (Office for National Statistics, 2022), Broken finger, while can be very painful is not a long-term illness (Infirmity). Unknown infection is not an Infirmity, these were replaced with No Infirmity. Infirmity are disabilities and it will not be appropriate to assign any to those who have not been recorded to have one, all missing values were replaced with Unknown. "Ethical considerations are paramount in health data science, particularly when dealing with sensitive patient information (MacPherson and Pham, 2020)."

Religion: All Nan values were filled with "Not Answered", some suspicious entries like 'Agnostic'. 'The Chantry', 'Asia', 'Atheist', 'The Templars' were changed to "Other Religion" as well as this aren't recognized as a religion in the UK (Citizens Advice, 2024). The "Catholic" religion was condensed to Christian in line with the 2021 census (Office for National Statistics, 2022).

3.0 POPULATION DISTRIBUTION

The town has a population of 9,769, with females making up 52.55% and males 47.45%. The age distribution is balanced: 6.22% are toddlers (0-5), 9.78% Children (6-12), Teenagers (13-19) make up 10.15%, Young Adults (20-35)– make up the highest of 23.92%, Middle-aged (36-50)– 22.84%, Seniors(51-65)– 16.16%, Elderly(66-80)– 8.24% , Very Old(81-100) -2.56% , and 100 + make up the least 0.12%. These statistics shows that it is a very youthful population.

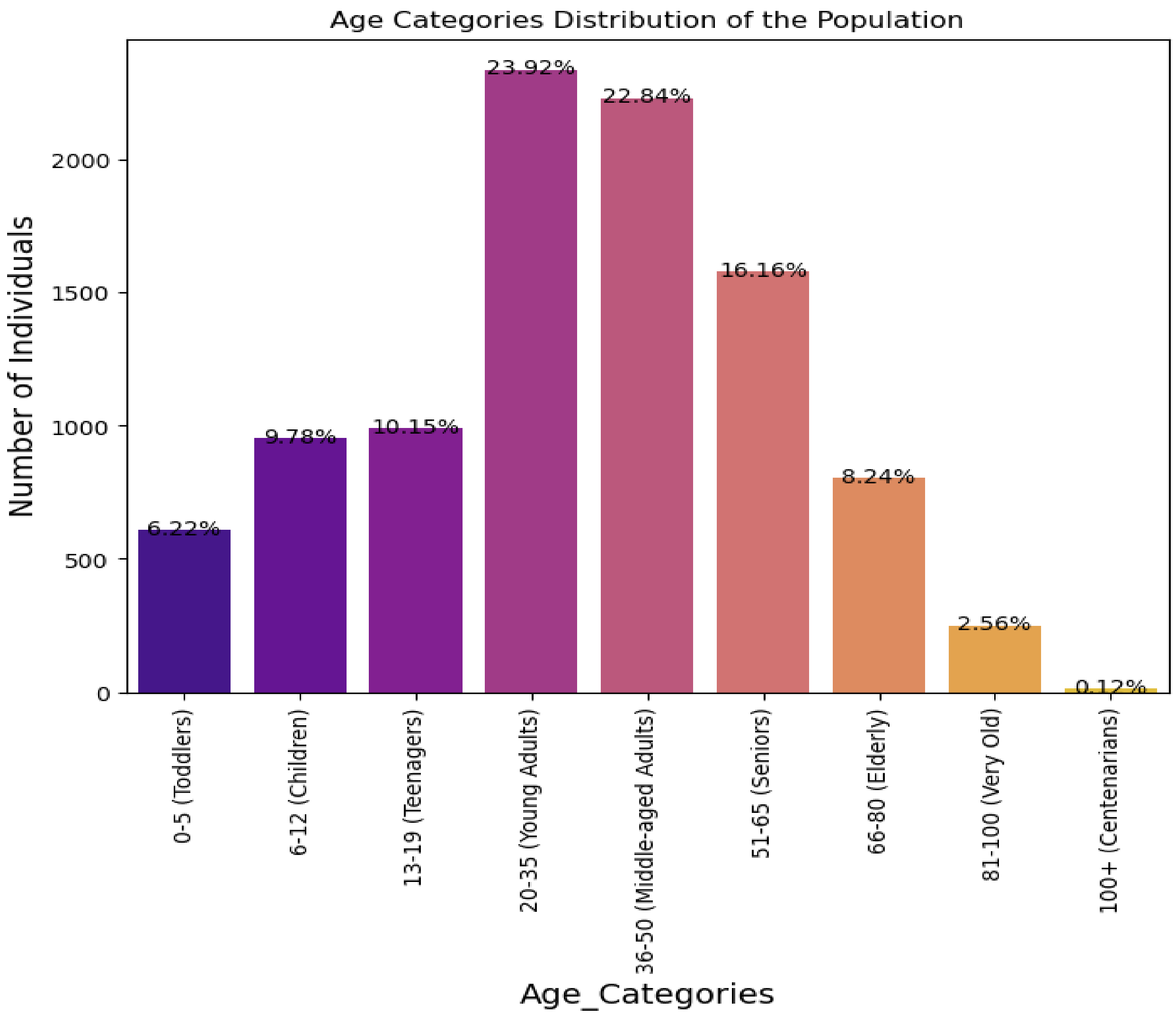


Figure 3: Age Distribution of Population

Figure 3 above shows the age distribution of different age groups and percentage

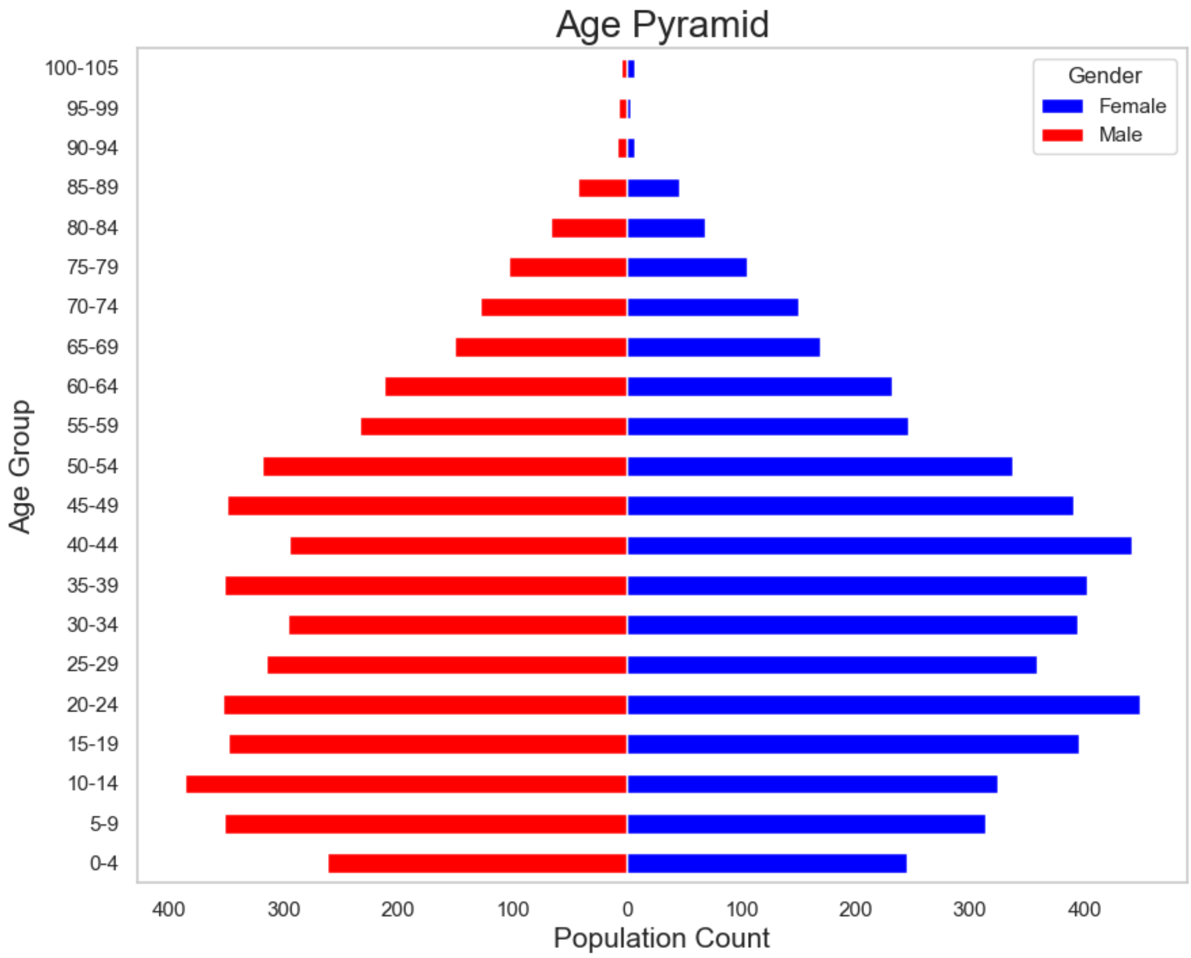


Figure 4: Population Age Pyramid showing Gender

In Figure 4 above we can observe that there are more males aged 10-14 and females aged 40-44. It also depicts a growing and expanding population, there will be more people of retirement age in the future, the strong presence of the 20-39 age group signifies potential childbearing.

3.1 UNEMPLOYMENT TRENDS

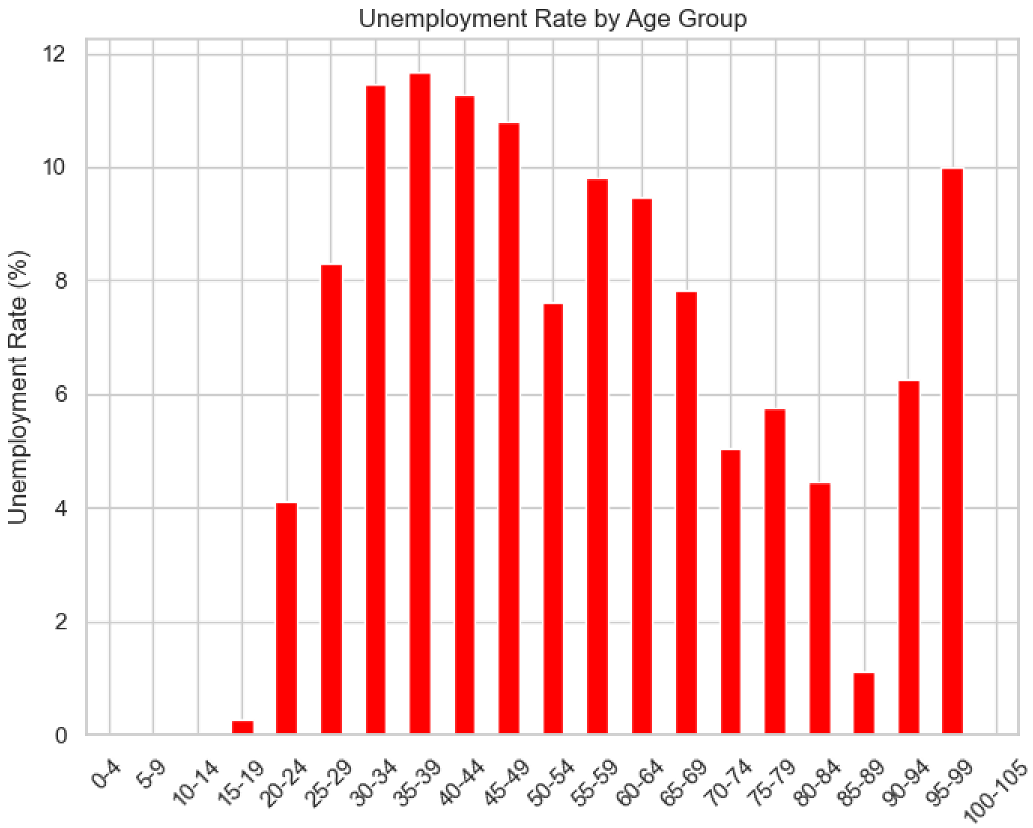


Figure 5: Unemployment rate by Age group

Figure 5 shows that middle aged adults age range 30-39 have the highest unemployment rate amongst the age groups which could be due to mid-career transitions or career shifts.

3.2 RELIGIOUS AFFILIATIONS

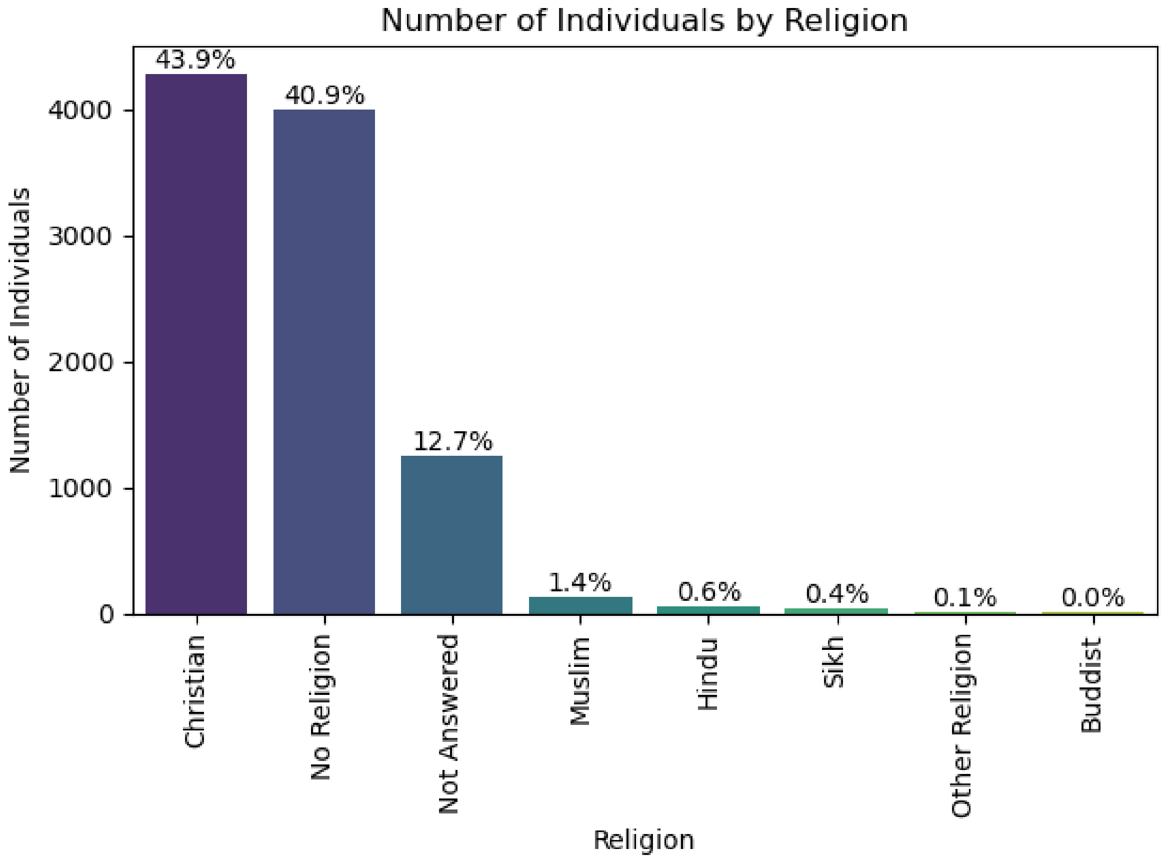


Figure 6: Religion Distribution Plot

The data in Figure 6 shows that while Christian remains the largest religion with 43.9% of the population, there is a noticeable increase in individuals identifying as No Religion (40.9%). This trend aligns with 2021 census reports which reveals that the UK is witnessing a rise in the non-religious demographic, reflecting a broader shift away from "Christians" (Humanists UK ,2023).

3.3 MARRIED AND DIVORCED RATES

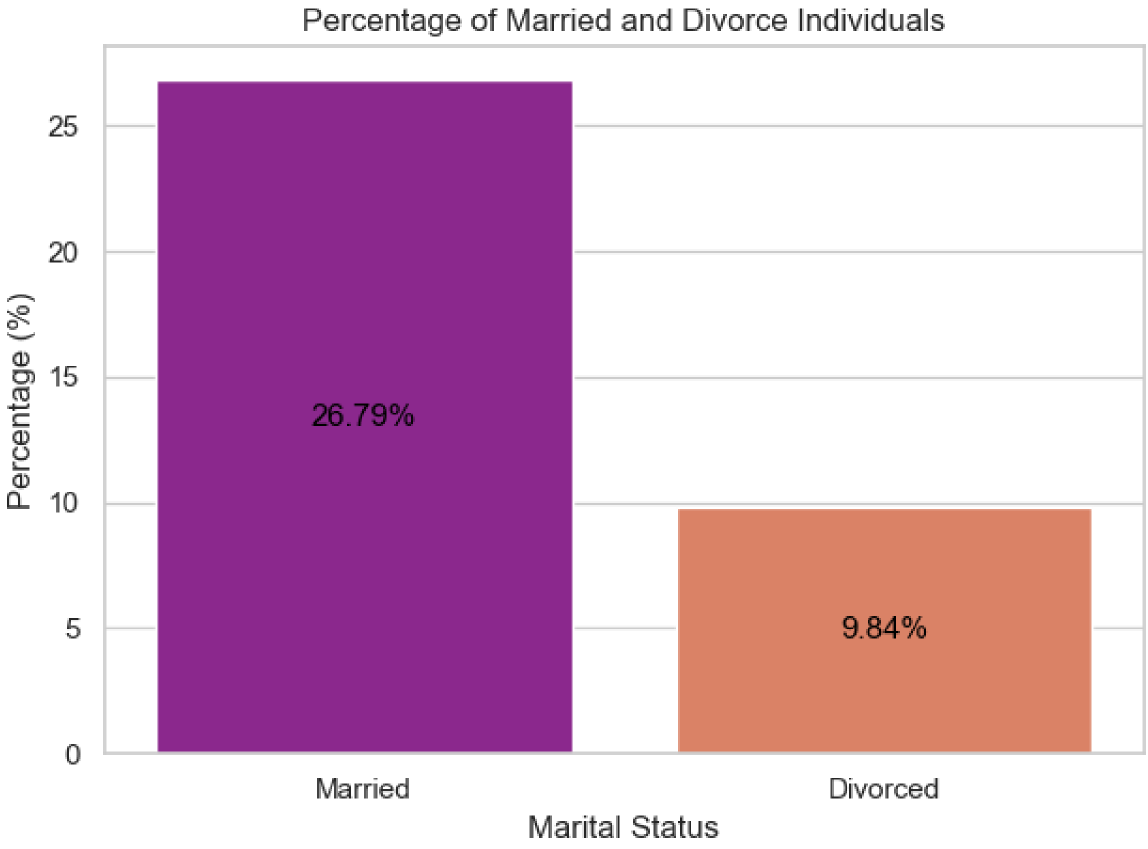


Figure 7: Married and Divorced Rates

Figure 7 shows we have more married people making up 26.79% than divorced (9.84%) in our population, this can be a factor to consider for low-density housing.

3.4 OCCUPANCY RATES

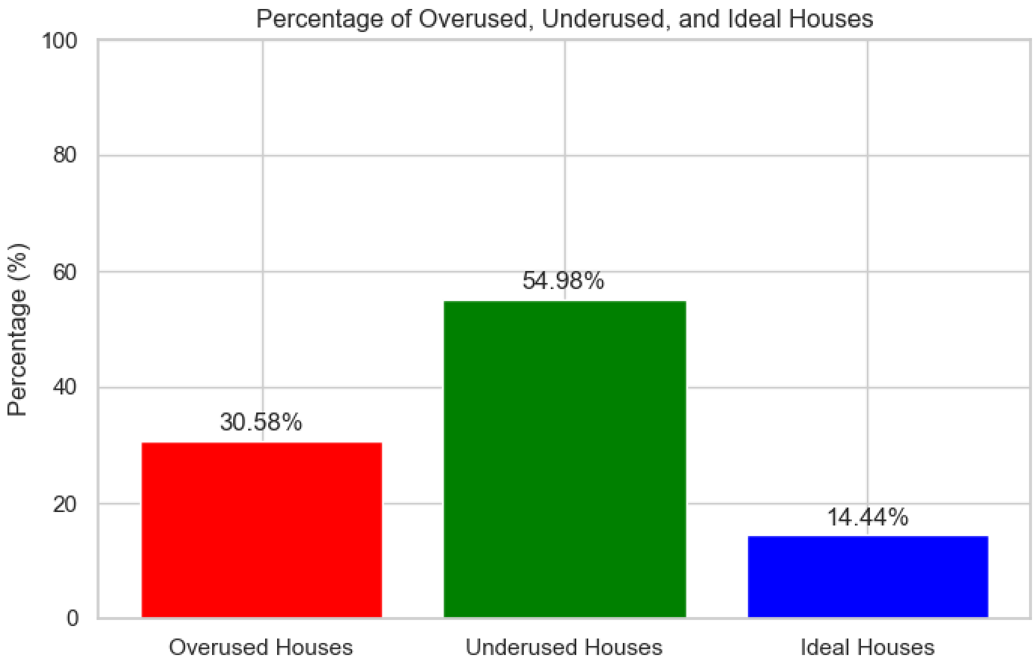


Figure 8: House Usage

From the result in Figure 8 we can observe that most houses are underused with a percentage of (54.98%) followed by over used houses (30.58%) and then ideal houses which are houses with the average number of people in a household (3). The existing houses are therefore underused.

3.5 UNEMPLOYMENT TRENDS

The grouping of the occupation was achieved by filtering through the occupation column and grouping all other Occupations except Child, Students, Retired, Unemployed, University Student into a group, this one done to get all the working citizens into a single group for easy analysis.

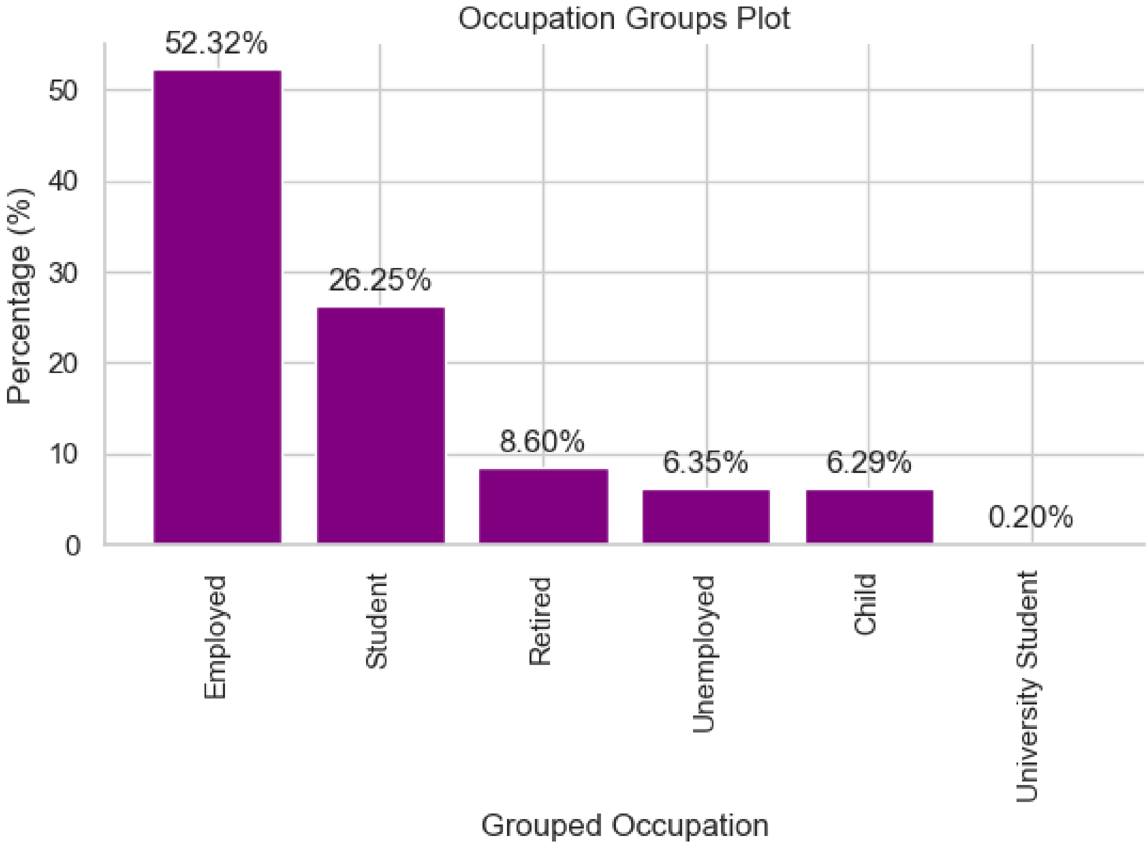


Figure 9: Grouped Occupation Plot

The data in figure 9 we can observe that we have more Employed individuals (52.32%) in the town's population followed by the students (26.25%), Retired (8.60%), then Unemployed (6.35%), Child (6.29%) and University Student (0.20%).

3.6 COMMUTERS VS NON-COMMUTERS

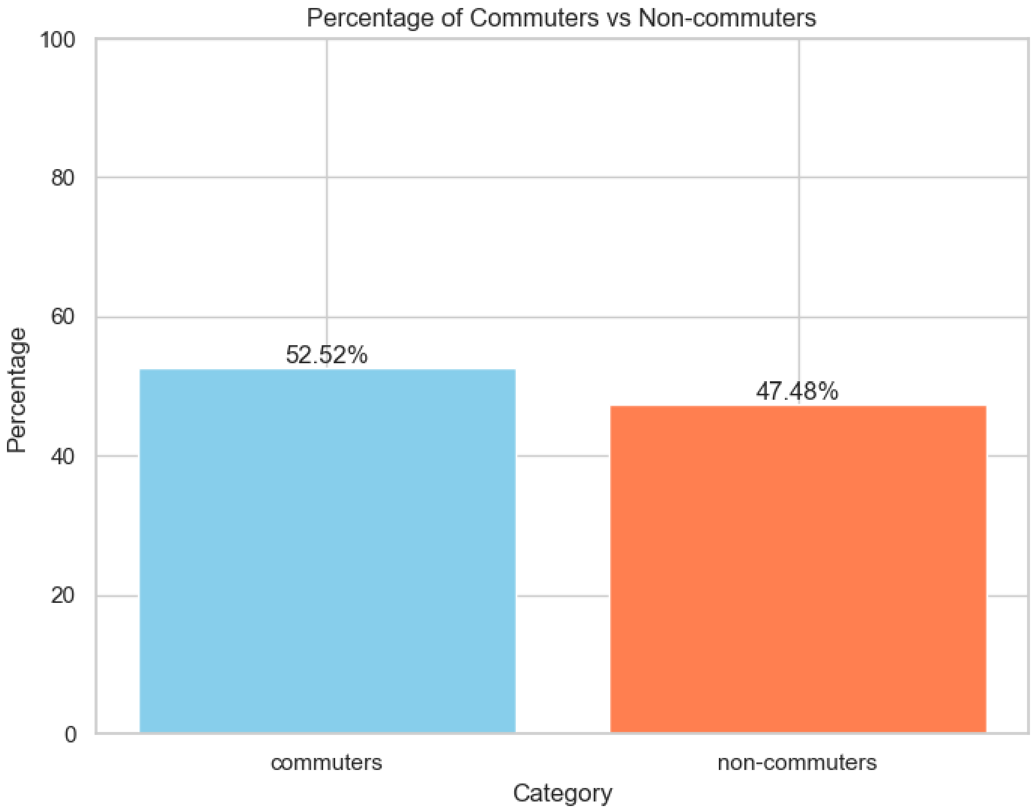


Figure 10: Commuters vs non-commuters

The data in Figure 10 shows there are more commuters (52.52%) in the populations, this category contains university students who need to commute to school, and other professions who needs to commute to work while the non-commuters (47.48) include students, children and retired citizens.

3.7 BIRTH AND DEATH RATE

These estimates were derived based on the Office for National Statistics standard population measurement unit, which calculates rates per 1,000 individuals (Office for National Statistics, 2024). The dataset was filtered to count the number of children with an age of 0 (101), representing newborns/live births. The birth rate was then calculated by dividing live births by the total population and multiplied the result by 1,000. The multiplication by 1,000 standardizes the birth rate, showing it as the number of live births per 1,000 individuals. For the death rate, it was estimated based on those aged 55-110+ instead. Another factor that impacts population growth is the fertility rate, which can be estimated rate was estimated by calculating the number of fertile women between 18 and 45 years old (Office of National Statistics, 2023) in the population divided by number of live births and then multiplied by 1000.

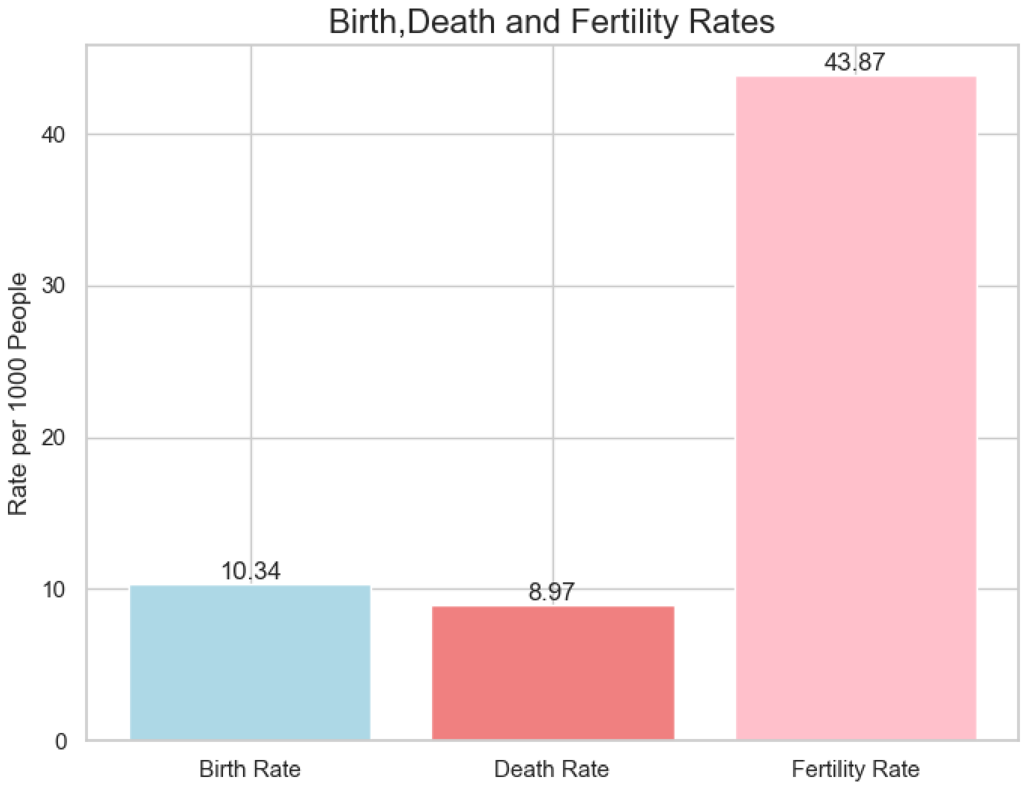


Figure 11: Birth, Death and Fertility Rates

Figure 11 above shows that the rate of birth 10.34 in the population is higher than the death rate 8.97 signifying a growing population. The high fertility rate 42.87 per 1000 women in the population further shows a potential for population growth.

3.8 INFIRMITY RATES

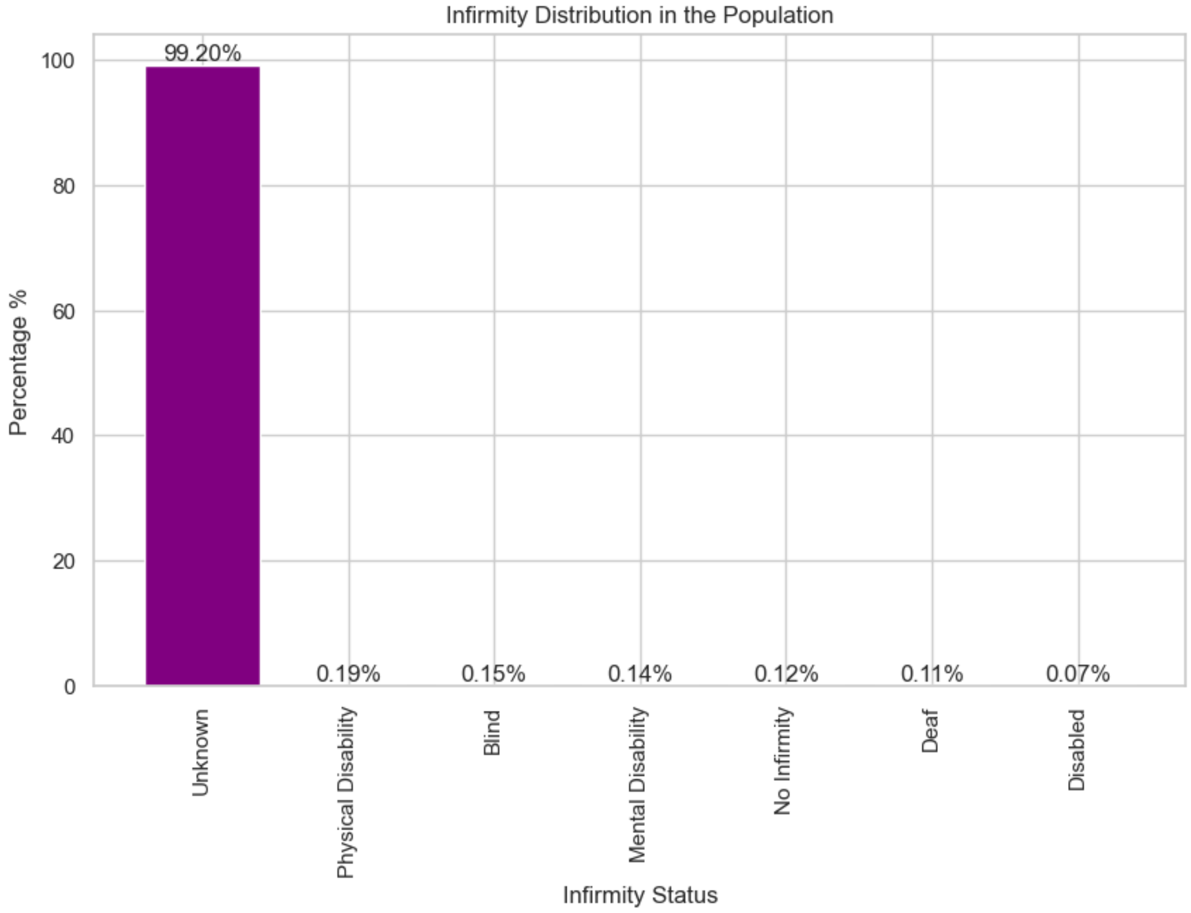


Figure 12: Infirmity Distribution in the Population

In Figure 12 we can observe that the infirmity results show a large amount of the population (99.20%) falls under the "Unknown" category. Among the reported disabilities Physical disability (0.19%) is the most recorded, followed by Blind (0.15%) and mental disability (0.14%). No Infirmity (0.12%), Deaf (0.11%), and "Disabled" (0.07%) categories make up smaller proportions of the recorded infirmities.

4.0 ANALYSIS

In this section, all options outlined in the project brief will be assessed. The decision-making process will follow an elimination-by-aspects approach, systematically narrowing down the choices. Each option will be evaluated and justified based on its relevance and advantages, culminating in a final recommendation that prioritizes the most suitable choice over the alternatives.

4.1 What should be built on an un-occupied plot of land that the government wishes to develop?

a) **High-density housing (If the population is significantly expanding)**

The population growth has a birth rate of 10.34 and a death rate of 8.97 which, further analysis shows that there is a Total Fertility Rate (TFR) of 42.87 per 1000 women and 0.04 per Woman this is relatively low compared to the fertility rate recorded in 2021(ONS, 2022) this shows no population expansion.

There is no pressing need for high-density housing in the town. The data supports focusing on other priorities. This option will be **eliminated.**

b) **Low-density housing (if the population is "affluent" and there is demand for large family housing)**

Some indicators of Affluence include Income, Housing, Employment (Berry, 2013). From our plot in Figure 8 we can observe that most houses are underused with a percentage of 54.98% followed by over used houses (30.58%) and then ideal houses which are houses with the average number of people of household (3). Figure 9 also shows 52.35% of the population are employed. For housing analysis as well, the data shows that there are more married people making up 26.79% than divorced (9.84%) in Figure 7. Housing does not seem to be an issue in the town which suggests needs for low-density housing however, the existing houses are still underused.

With a high percentage of underused housing, low-density housing (which typically involves more space per dwelling and fewer dwellings per area) is also not a pressing need. This option will be **eliminated**

c) **Train station. (If there are a lot of commuters in the town)**

More than half of the population 52.52% are commuters which means that there are more people who commute by road frequently, building a train station could take pressure off the roads. We will **keep this option**

d) Religious building. Is there demand for a religious building?

From the analysis we had just 1 Catholic entry in the population result which was condensed to Christian in line with the 2021 census and 5127 Christians about 43.59% (Figure 6) of the population, and there is only a Catholic Church, this means there will be demand for another religious building for the Christian denomination. However, the Catholic place of worship can be converted to be used for Christians. We will keep this Option for now.

e) Emergency medical building (if there are many injuries or future pregnancies likely in the population)

The analysis (Figure 12) shows a relatively low rate of disabilities accounting for only 0.66% in total of the entire population, the TFR of 0.04 birth per Woman is not quite low too, this option will be **eliminated.**

4.2 Which one of the following options should be invested in?

a) Employment and training. (If there is evidence for a significant amount of unemployment)

As shown in Figure 9 most individuals are employed with a high employment rate of 52.32% and low unemployment rate of 6.35%. There is no significant amount of employment, there will no need to invest in employment and training. This option will be **eliminated.**

b) Old age care (if there will be an increased number of old people in the future)

The analysis shows that the town has 27.08% total of seniors to centenaries while this is not a large portion of the population, many young people who are in the Toddlers to Middle-aged adults age categories, accounting for about 72.91% of the town's population. This suggests that there may be a significant number of old people in the future and old age care is essential. We will **keep this option**.

c) Increase spending for schooling (If there is evidence of a growing population of school-aged children)

In the town's population 6.22% are toddlers, 9.78% are Children (6-12), and Teenagers make up 10.15%, accounting for a total of 26.15%, in Figure 9

26.25% accounts for students, which is not so high, however another factor that can determine this is the capacity of the current schools. We will <u>keep this option</u> for now.

d) General Infrastructure (If the town is expanding)

From the analysis, the town's population generally is not expanding, investing in general infrastructure is not really a priority. This option will be **eliminated.**

5.0 RECOMMENDATIONS

In conclusion of all the five options to be considered, after implementing the elimination-by-aspects approach the option of building a religious building or train station were the only options left. In order of priority, the train station emerges as a critical need for the town. A train station is recommended to be built on the unoccupied land as this would address accessibility gaps, reduce congestions on the roads, and align with sustainable development goals.

The analysis highlights that of all the four options, the option of investing in schools and old age care were left after the elimination-by-aspects approach. Old-age care should be invested as it is the most critical need due elderly individuals in the population and future. Prioritizing old-age care by Investing in facilities and services tailored to their needs, such as nursing homes, assisted living, and accessible healthcare infrastructure. ensures the well-being and dignity of the elderly population (Ghebreyesus, 2017).

BIBLOGRAPHY

Berry, M. (2013). *The Affluent Society Revisited*. Oxford: Oxford University Press. Online edition published January 23, 2014. Available at: https://doi.org/10.1093/acprof:oso/9780199686506.001.0001 (12/12/ 2024).

Ghebreyesus, T. A. (2017). *Health is a fundamental human right*. Human Rights Day 2017. World Health Organization, 10 December.

Government of the United Kingdom. (no date). *Your rights to housing if you're under 18*. Available at: https://www.gov.uk/your-rights-to-housing-if-youre-under-18 (Accessed: 9/12/2024).

Humanists UK. (2023). *2021 Census: More non-religious than Christians among those under 67*. Available at: https://humanists.uk/2023/01/30/2021-census-more-non-religious-than-christians-among-those-under-67/#:~:text=In%20Wales%2C%20the%20results%20were,46%25%20ticking%20'Christian' (Accessed: 11/12/ 2024).

MacPherson, Y. and Pham, K. (2020). Ethics in Health Data Science. In: Celi, L., Majumder, M., Ordóñez, P., Osorio, J., Paik, K. and Somai, M. (eds) *Leveraging Data Science for Global Health*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-47994-7_22

Office for National Statistics (ONS). (2023). *Conception statistics: 2021*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/bulletins/conceptionstatistics/2021 (Accessed: 12/11/2024).

Office for National Statistics (ONS). (2024). *Childbearing for women born in different years, England and Wales: 2021 and 2022*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/bulletins/childbearingforwomenbornindifferentyearsenglandandwales/2021and2022 (Accessed: 10/122024)

Office for National Statistics (ONS). (2022). *Religion, England and Wales: Census 2021*. Available at:

https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/religion/bulletins/religionenglandandwales/census2021 (Accessed: 10/ 11/ 2024).

Oldest in Britain (n.d.). *Oldest in Britain*. Available at: https://oldestinbritain.nfshost.com/ (Accessed: [10/12/2024]).