Fundamentals of Data Science Census Project Report

Abstract

This report presents the findings from the thorough data cleaning and analysis of a town's mock census data. The objective was to inform the local government on the best use of an unoccupied plot of land and the allocation of resources and investment based on insights gleaned from the analysis. Key findings included a median age of 34, an increasing demand for high-density housing driven by a 27.4% marriage rate and 11.4% immigration rate, a youthful population, and an employment rate of 53.19% percent. Additionally, investment in schooling and general infrastructure was recommended to accommodate the town's growing population and ensure sustainable development.

Introduction

1.1 Background to the Study

This project aims to analyze a small town's census data and provide insights to help the local government determine the optimal use of an unoccupied plot of land and prioritize investments in community services. The analysis began with a comprehensive data cleaning process, which involved identifying and correcting errors in the dataset, all conducted using Jupyter Notebook, an Integrated Development Environment. The analysis results are then presented, offering key insights to support informed decision-making. Finally, based on these findings, recommendations on investment and resource allocation are provided.

Methodology

This section contains the thorough data cleaning approach taken and done.

2.1 Data Cleaning

The initial dataset contained incomplete records, missing values, and various errors, which were meticulously addressed using a combination of data-cleaning techniques. The census dataset comprised eleven columns with information on 11,082 town residents. The cleaning process began with the following strategies:

1. Identifying missing values in the dataset
2. Exploring the dataset for duplicate values
3. Thorough data cleaning of the entire dataset per column
4. Converted empty strings into NaN values for better cleaning.
5. Data Imputation for three columns: 'Surname', 'Religion' and 'Relationship to Head of House'.
6. Correcting general errors in the dataset.

The following key steps were implemented to address the errors in each column:

1. Examine the datatype of a column
2. Examine the unique values of a column
3. Examine ‘Not a Number’/ Null or Missing values in a column
4. Identifying general errors within a column, such as detecting and correcting misspellings or outliers.
5. Corrected false responses from the population. For example, marital status and age.

Columns with missing Values and General Errors

Surname: The missing values in this column were cleaned using data imputation. The missing surnames were inferred using information from other columns such as ‘Relationship to Head of House’, ‘Street’, and ‘House Number’.

Relationship to Head of House: Rows where the age was less than 18 were reviewed and filled based on information from other columns. For rows where the occupation was listed as "University Student" and the relationship to the head of the house was missing, the value was set to "Lodger," reflecting the likelihood that university students schooling in the nearby cities would be lodgers. Additionally, a custom function was implemented to assign or infer values for the "Relationship to Head of House" column. The dataset was grouped by household, defined by "House Number" and "Street," and for each group, the function identified the head of the house using their age, gender, and surname as references. Missing relationships were inferred based on criteria such as surname matching and relative age. For instance, individuals at least 18 years younger than the head and sharing the surname were categorized as "Son" or "Daughter." Those with different surnames were labeled as "Lodger" or "Visitor" based on random sampling, depending on their age relative to the head. If no head was identified, missing relationships were defaulted to "Visitor." Finally, any remaining missing values were replaced with "unknown."

Marital Status: Upon exploration, all the 2685 missing values in this column were for individuals aged less than 18. It is important to note that 18 is the legal age of marriage in the United Kingdom (UK Government, 2023). Hence, ‘Underage’ was assigned to all the nan values.

Religion: All unrealistic or fake religions were corrected and replaced with ‘Other’. Only religions listed in the 2021 UK Census Data were adhered to. One misspelling was corrected – ‘Buddist’. For the missing values, religion was assigned based on the relationship of the religion to the head of the house for those aged less than 18. For the remaining missing values, ‘Other’ was assigned.

Infirmity: Terms such as ‘Unknown Infection’, and ‘Healthy’ were renamed to ‘No Infirmity’. Appropriately, missing values meant the absence of infirmity and so were corrected as such. Terms such as Blind, Deaf, Physical Disability, Disabled, and Mental Disability meant the presence of life-altering conditions that require long-term support/care.

Four new columns: Employment Status, Commuter, Age Category, and House Identifier were added to the dataframe for further analysis.

Results

This section presents the key findings from the analysis of the census data, highlighting important trends and patterns. The insights gained from these results will serve as the foundation for our discussion and guide the recommendations to be proposed.

3.0 Population Demography

The nomenclature of this town's population comprises of 11,082 residents, including 5,355 males (48.32%) and 5,727 females (51.68%). As depicted in Figure 1 below, 5.83% of the population are aged 0-5, 9.92% are aged 6-12, 10.75% are aged 13-19, 24.33% are aged 20-35, 23.49% are aged 36-50, 16.18% are aged 51-65, 7.31% are aged 66-80, 2.10% are aged 81-100, and finally, 0.10% are over 100 years old. The median age of the population is 34. The town has a youthful population, from Toddlers to Middle-aged adults, compared to Seniors and Centenarians.

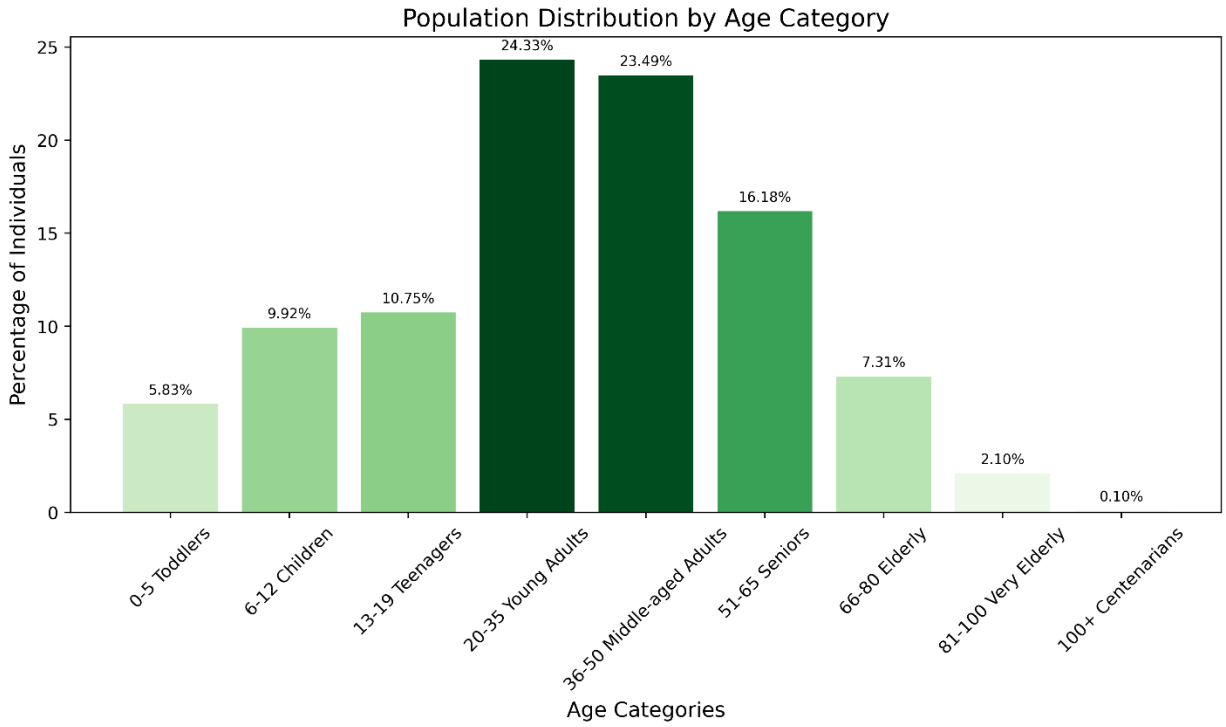

Figure 1: Distribution by Age Category

In Figure 2 below, an age pyramid shows the population count by gender.

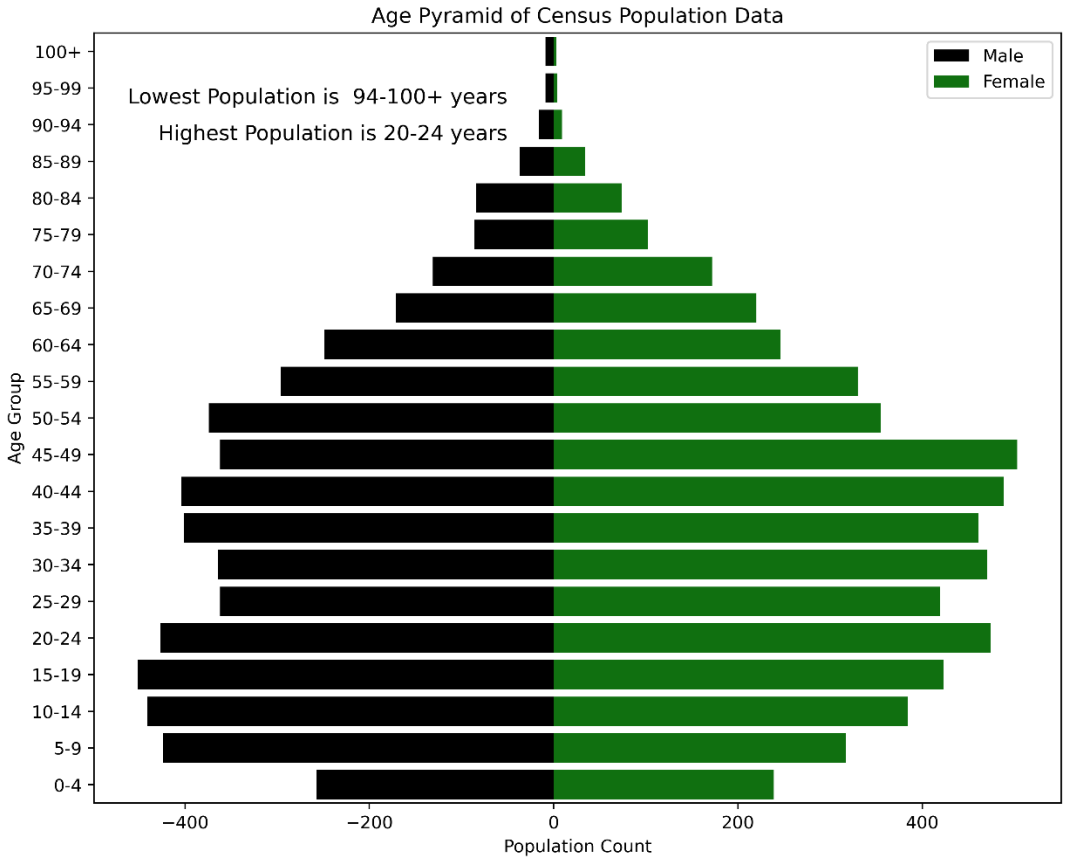

Figure 2: Age Pyramid Distribution by Gender

As illustrated in Figure 2, there are more males aged 15-19 and more females aged 45-49 in the population.

3.1 Birth, Death, and Fertility Rate

In the United Kingdom, the Office for National Statistics (ONS) produces the statistics for birth, fertility, and death rates. Annually, the Crude Birth Rate (CBR) is calculated by dividing the total number of live births in a year by the mid-year population and then multiplying the result by 1,000. In this census project, the number of live births was 107, and the total population was 11,082. This resulted in an estimated birth rate of 9.66 per 1,000 individuals. In comparison, the crude birth rate in the United Kingdom for 2021 was 10.4 live births per 1,000 population, as reported by the Office for National Statistics (ONS) in their birth summary tables for England and Wales.

Similarly, according to the Office for National Statistics, the average life expectancy in the United Kingdom at birth is 78.6 for males and 82.6 for females. The weighted average is 80.67 years. Thus, the death rate was estimated based on those aged 81-100+. Hence, the estimated death rate per 1000 individuals was 1%. This suggests a net increase in population.

General Fertility Rate (GFR) is calculated based on the number of live births per 1,000 childbearing women aged 15-44. Office for National Statistics considers this age range as childbearing age. In this project, the estimated fertility rate was 40.23 births per 1,000 women, suggesting a potential for population expansion. In comparison, the General Fertility Rate for England and Wales in 2021 plummeted to 54.1 births per 1,000 women, as reported by the Office for National Statistics (ONS, 2021).

3.2 Immigration and Emigration

Immigration was calculated by analyzing the percentage of lodgers and visitors in the dataset. Similarly, emigration was calculated by examining the percentage of divorcees in the population who are expected to leave the town based on assumption. The immigration rate was estimated to be 11.41% while the emigration rate was estimated to be 9.21%.

3.3 House Density/Occupancy Rate

Analysis revealed that there are 3701 houses in this town with an average occupancy of 2.99, approximately 3 people per household. Further analysis was conducted to determine if the houses were underused or overused. According to Statista (2024), the average number of people per household in the United Kingdom was 2.36 as of 2022. This was employed to set the threshold for normal occupancy, underusage, and overusage. Figure 3 illustrates the Occupancy rate with 74.55% showing normal occupancy, 22.59% reflecting underused houses, and 2.86% representing overused houses.

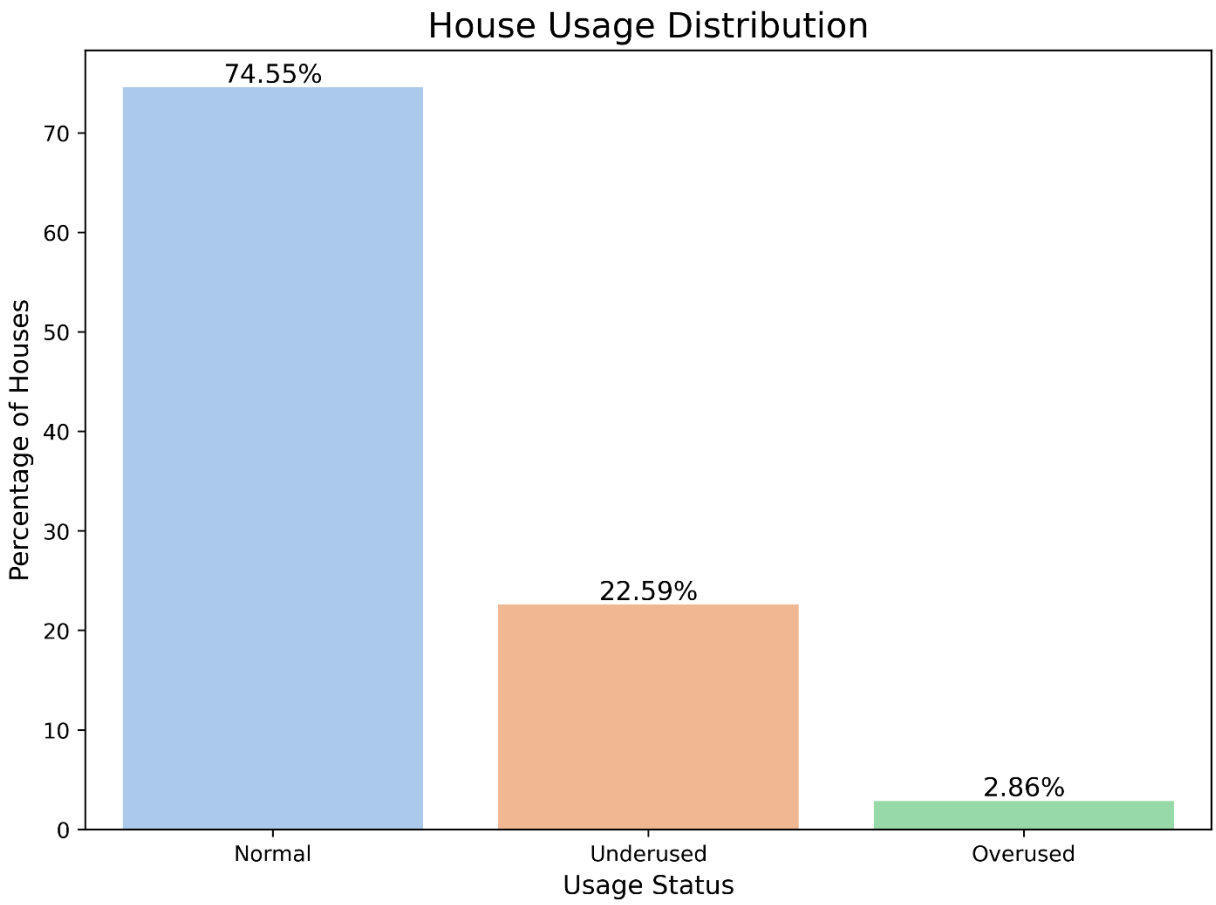

Figure 3: House Occupancy Distribution

3.4 Employment/Unemployment Trends

The 'Employment Status' column reveals that 53.19% of the town's population is employed, slightly below the England and Wales average (ONS, 2023). Students account for 20.31%, reflecting a young, education-focused population, while retirees (7.23%), university students (6.65%), unemployed individuals (6.70%, nearly double the national rate of 3.4%), and children (5.91%) represent smaller groups. This indicates a balanced demographic mix of workers, dependents, and those in transition stages like education or retirement. The unemployment rate, though not dominant, warrants attention to address economic challenges. With nearly 27% of the population comprising students and university attendees, there is a clear focus on education and future workforce development. Targeted programs for students, unemployed individuals, and retirees could boost well-being and productivity.

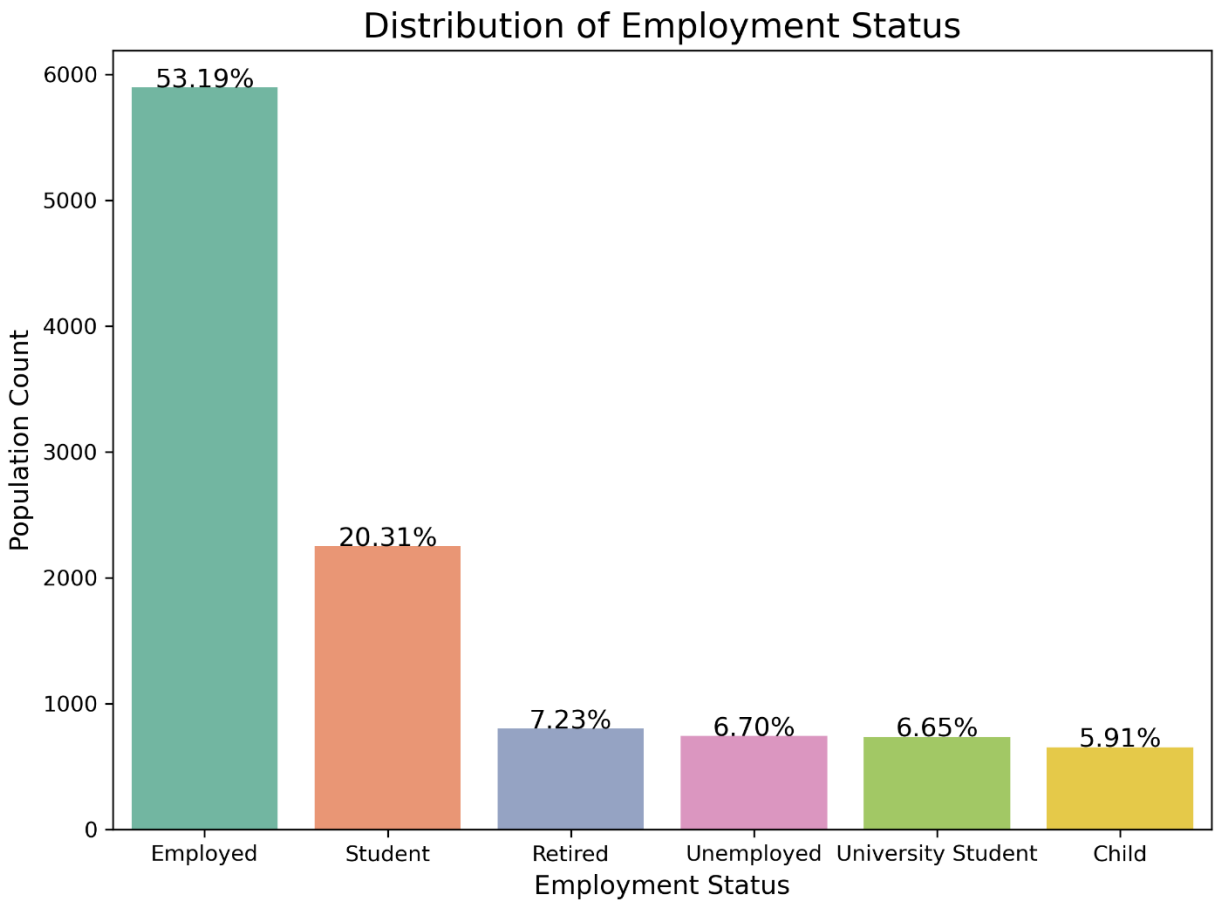

Figure 4: Employment Status Distribution

3.5 Commuters

The analysis reveals that the total number of commuters is 2,152, representing approximately 19.42% of the town's population. This group primarily consists of 737 university students and individuals with white-collar jobs. The town is relatively small and situated between two large cities, leading to the assumption that individuals with white-collar jobs likely commute to work using motorways. Similarly, university students are expected to commute, as the town does not have a university of its own.

3.6 Religious Affiliation

Figure 5 shows the distribution of religious affiliations in the census data. Christianity is the most prevalent at 50.01%, followed by those identifying with No Religion at 46.11%. Other religions such as Muslim (1.48%), Hinduism (0.85%), Sikhism (0.43%), Buddhism (0.01%), and minority affiliations like Humanism (0.02%) and Agnosticism (0.05%) have significantly smaller representation. The large proportion of individuals identifying with "No Religion" suggests a paradigm shift from traditional religious practices, potentially indicating growth in secular or non-religious beliefs. Meanwhile, the smaller representation of newer or less common religions

like Humanism or Agnosticism may indicate emerging trends, albeit with minimal impact. The data does not directly indicate growth or decline trends for specific religions but suggests an overall diversification of religious affiliations in the population.

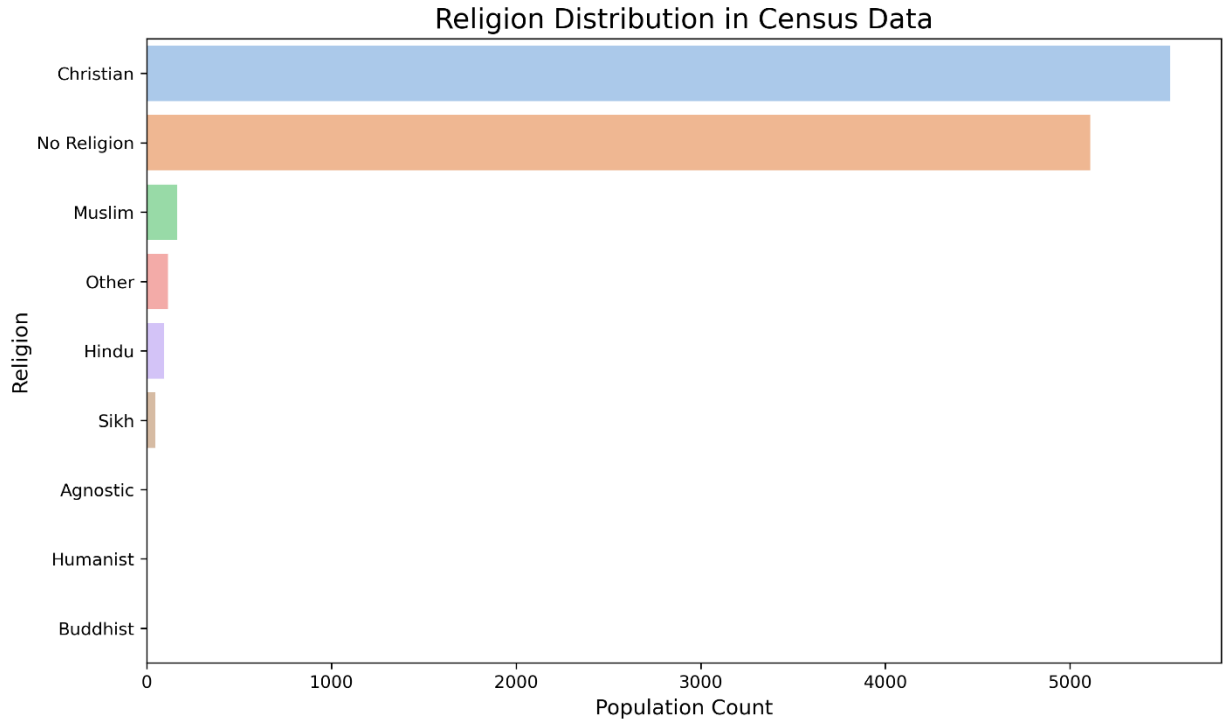

Figure 5: Religious Affiliation Distribution

In Figure 6, The average ages reveal a relatively balanced distribution among some smaller religious groups, such as **Agnostic (38)**, **Christian (38)**, and **Sikh (37)**, suggesting these groups have consistent representation. However, the dominant groups, like **Christianity** and **No Religion**, still stand out in overall numbers despite their similar averages to smaller groups due to their larger absolute population.

Humanist (18), **Other (24)**, and **Buddhist (28)** show lower averages, which may indicate smaller but stable populations. Meanwhile, **Muslim (30)** and **Hindu (31)** averages suggest modest representation, potentially growing in specific demographic pockets.

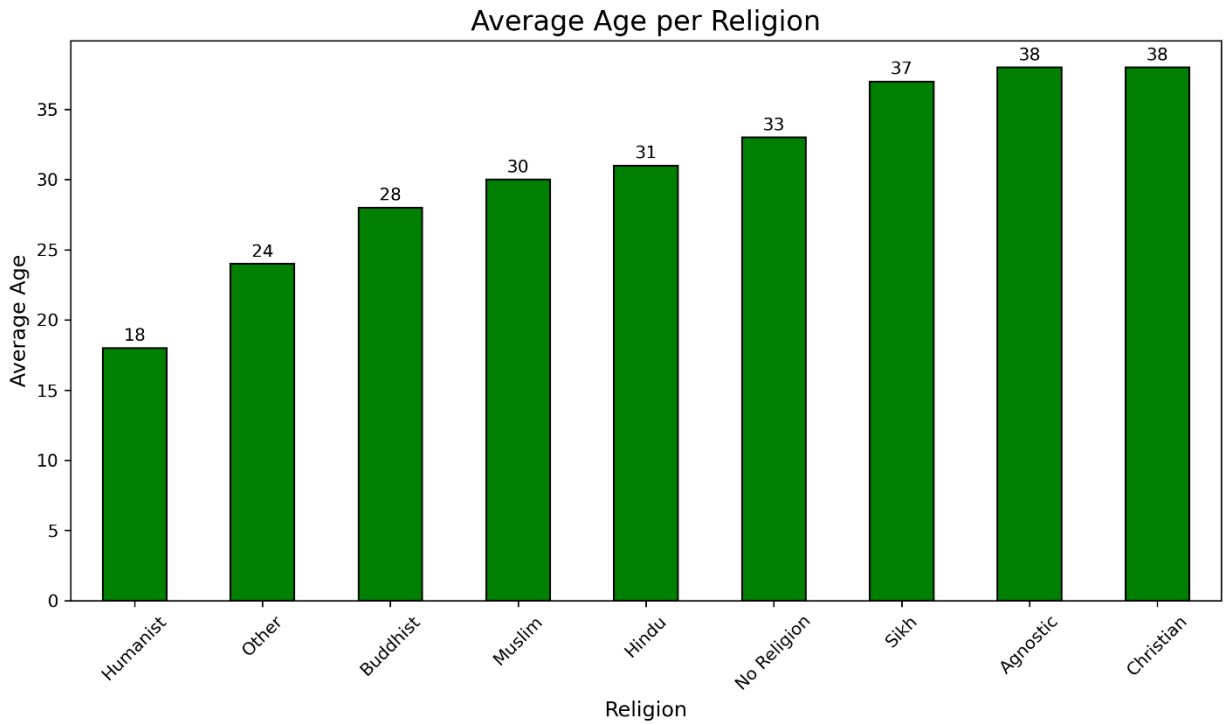

Figure 6: Average Age per Religion

3.7 Marriage and Divorce Rate

The town's population includes an estimated 27.24% who are married, totaling 3,019 individuals, and 9.21% who are divorced, accounting for 1,021 individuals. This indicates that the marriage rate is substantially higher than the divorce rate as shown in Figure 7.

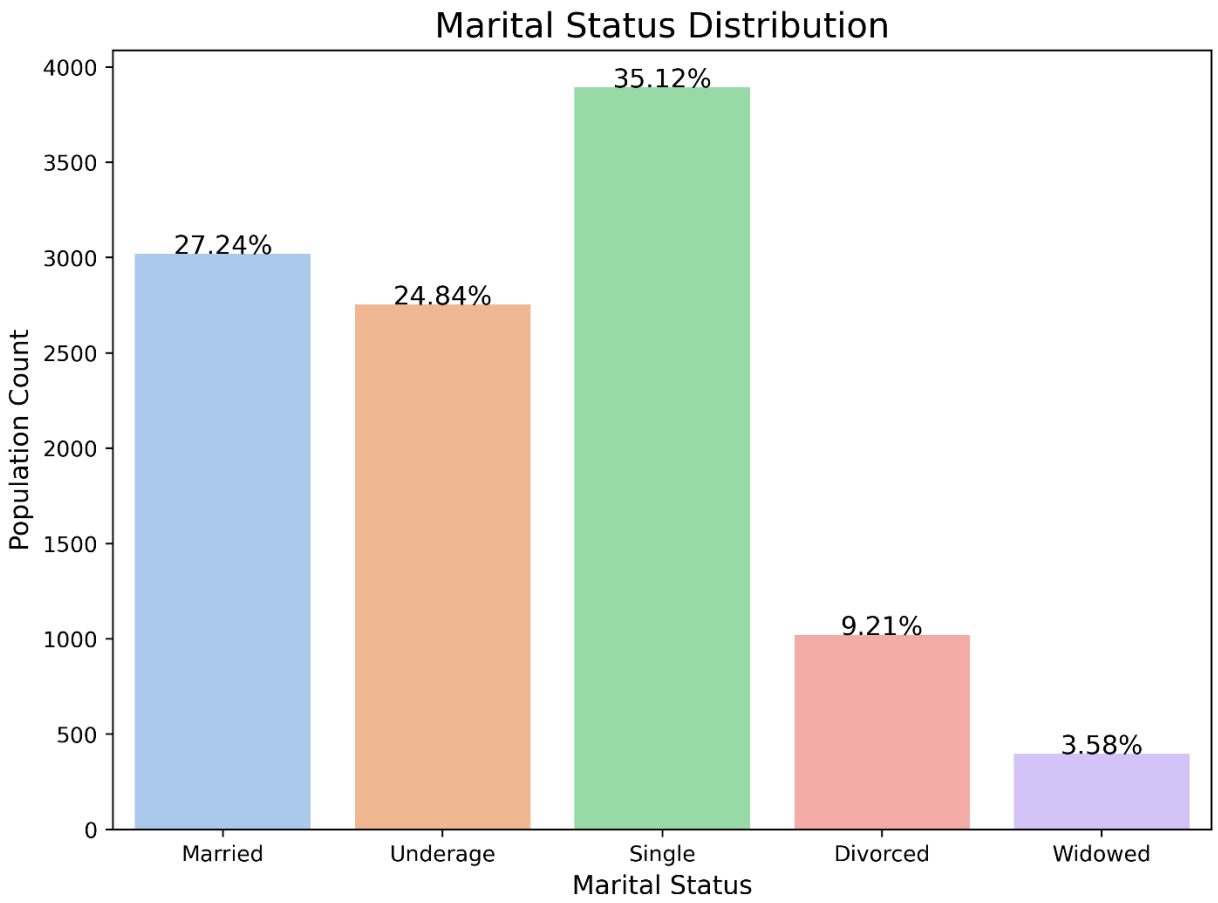

Figure 7: Marital Status Distribution

3.8 Infirmity Status

The population appears to be generally healthy, with a significant majority (11,007 individuals) reporting no infirmity. The number of individuals with disabilities is relatively low in comparison, with 17 blind individuals, 16 mentally disabled, 15 deaf, 15 with physical disabilities, and 12 with other forms of disability as illustrated in Figure 8. This suggests that the majority of the population experiences no major health issues or impairments. However, while the numbers are small, it is important to consider providing appropriate support and resources for those with disabilities to ensure their inclusion and well-being.

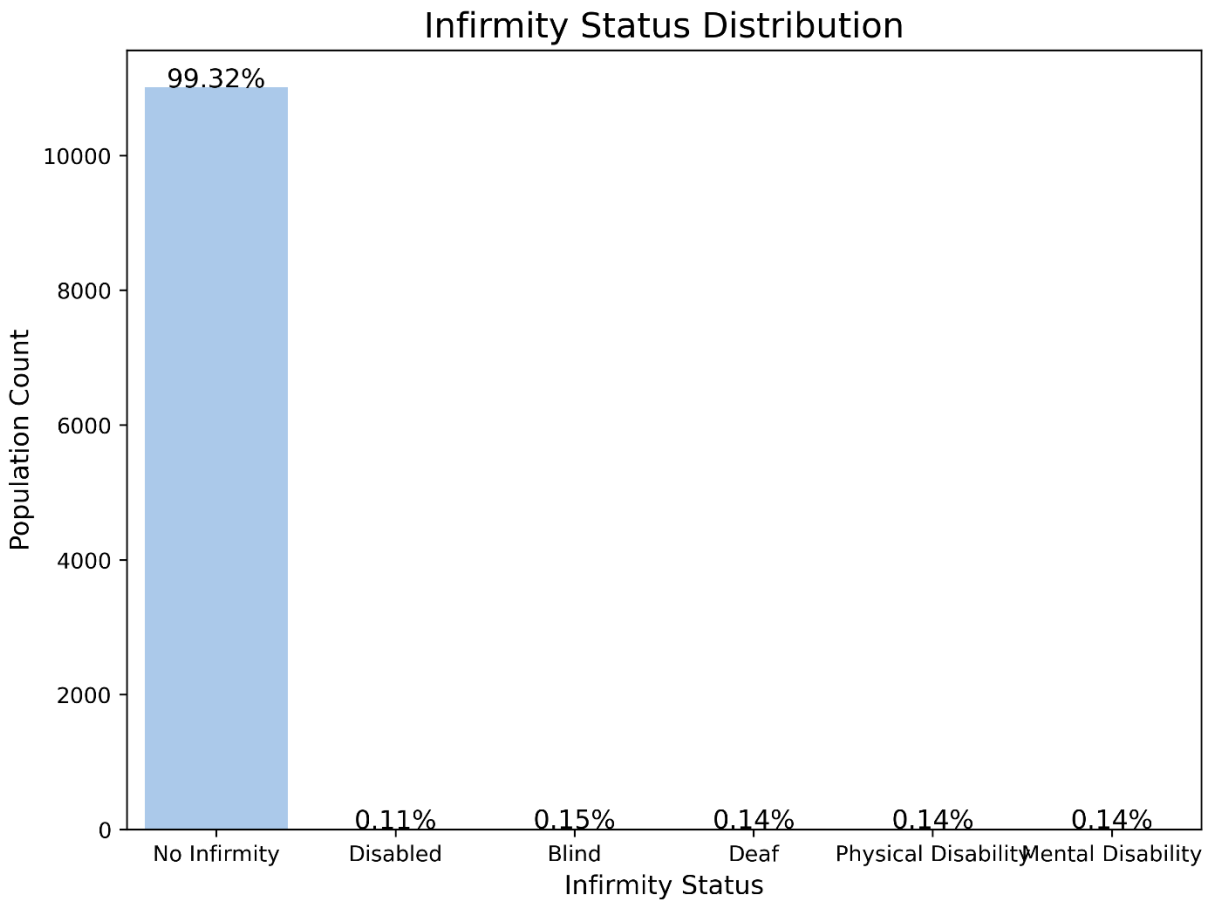

Figure 8: Infirmity Status Distribution

Discussion

This section includes discussions that will inform the recommendations regarding the use of the unoccupied plot of land and the allocation of resources to social services. All the alternatives presented in the project brief will be assessed, and the decision-making process will follow a criteria-based selection approach to answer the presented questions, gradually narrowing down the options and justifying the preferred choice over the others until the final recommendation is made.

Question A: What should be built on an unoccupied plot of land that the local government wishes to develop?

The consideration of high-density housing development as the first choice is strongly supported by the demographic characteristics of the population, which indicate a significant potential for growth. The relatively young age profile of the population aged 20-35 making up 24.33% of the town's population suggests an inclination toward expansion, particularly as young adults are more likely to form new households. This trend is further substantiated by a marriage rate of

27.4%, which often correlates with an increased demand for housing. Additionally, the birth rate of 9.66% per 1,000 individuals, although moderate, combined with a notable immigration rate of 11.41%, points toward sustained population growth over the coming years.

High-density housing is a strategic response to accommodate this anticipated growth efficiently, offering a sustainable solution to urban planning challenges. According to United Nations data, urban areas worldwide are experiencing increased densification due to demographic pressures and migration trends (United Nations, 2018). Therefore, prioritizing high-density housing aligns with global best practices in managing population growth in urban environments while maximizing the use of available land resources.

The second option of low-density housing is not viable in this context, as the population demographic does not indicate a predominance of affluence or a strong demand for large family homes. Furthermore, given the anticipated population growth, this approach would inefficiently utilize land resources, failing to meet the housing needs of a growing and dynamic community.

The third option of building a train station could alleviate road congestion and benefit the 19.42% of the population identified as commuters. However, this option may not be cost-effective given that it serves a relatively small proportion of the population, making it less impactful than alternatives that address broader community needs.

The fourth option of building an additional religious building may not be the most effective use of resources given the current distribution of religious affiliations. Christianity, while the most prevalent at 50.01%, already has a place of worship in the town, and the substantial portion of the population identifying with "No Religion" (46.11%) suggests limited demand for additional religious spaces. Smaller religious groups, such as Muslims (1.48%) and Hindus (0.85%), might benefit from dedicated spaces, but their representation is relatively low, making the overall demand insufficient to justify prioritizing this option over others addressing broader community needs.

The fifth option of an emergency medical building, such as a minor injuries center, may not be a high priority based on the population's current health indicators. While the fertility rate is 40.23% and the birth rate is 9.66%, suggesting a moderate potential for pregnancies, the infirmity data reveals very low numbers of severe disabilities or chronic conditions requiring urgent care. Given these figures, the existing medical infrastructure may be sufficient, and resources could be better allocated to address other pressing community needs.

Based on the analysis, it is recommended that the unoccupied plot of land be developed into high-density housing. This option aligns closely with the town's demographic profile, which features a significant proportion of young adults aged 20-35 (24.33%) and a marriage rate of 27.4%, both of which indicate a growing demand for housing. High-density housing would efficiently utilize the available land while accommodating the projected population growth driven by factors such as the birth rate of 9.66% and an immigration rate of 11.41%.

This development would not only address immediate housing needs but also provide a sustainable solution to urban planning challenges. In contrast, options like low-density housing

or additional religious buildings would fail to meet the broader needs of the community, while a train station or emergency medical facility, though beneficial for specific groups, lack the wide-reaching impact necessary to prioritize them over high-density housing.

Question B: Which one of the following options should be invested **in?**

The first option, which is employment training, should not be prioritized since 53.19%—a majority of the town's population—is gainfully employed, while 6.70% are unemployed, which is somewhat higher compared to the UK 2021 Census data on unemployment, recorded at 4.3% (Office for National Statistics, n.d.).

The analysis reveals that individuals aged 51 and older make up 8.91% of the town's population, with retirees accounting for 7.23%. Given these relatively low numbers, prioritizing end-of-life care at this time may not be necessary, as the existing measures are likely sufficient to meet current needs.

The third option, increasing spending on schooling, should be prioritized due to the substantial presence of school-aged children and evidence of new births. Children currently account for 6.62% of the population, toddlers make up 5.83%, and teenagers represent 10.75%, totaling 23.2%. This significant proportion indicates a high demand for education that is likely to grow in the future. Consequently, increased funding for schooling is essential to accommodate this rising need and ensure adequate resources for the town's younger population.

Given the rapidly growing and youthful population, there is a pressing need to allocate more resources toward general infrastructure. Additionally, the immigration rate of 11.41% coupled with the 35.12% rate of single individuals suggests further expansion, likely resulting in more marriages in the near future. Therefore, substantial and increased investments are necessary in key infrastructure areas such as road maintenance, water and waste management, railway stations, security, healthcare, and education.

As such, investment in general infrastructure and increased school spending should be strongly considered out of the four options considered.

Conclusion

In conclusion, comprehensive data cleaning and analysis of the mock census for a small town situated between two cities were conducted to guide the town council on how to utilize an unoccupied plot of land and determine priority investments based on population dynamics and growth trends.

References

Office for National Statistics (ONS), 2021. *Birth summary tables, England and Wales: 2021*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2021 [Accessed 1 January 2025]

Office for National Statistics (ONS), 2021. *Employment in local authorities, England and Wales: Census 2021*. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/employmentinlocalauthoritiesenglandandwales/census2021 [Accessed 5 January 2025].

Office for National Statistics (ONS), 2021. *Birth summary tables, England and Wales: 2021*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2021 [Accessed 3 January 2025].

Office for National Statistics (n.d.) *Unemployment*. Available at: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment (Accessed: 2 January 2025).

Statista. (2022). *Average household size in the UK 2022*. Available at: https://www.statista.com/statistics/295551/average-household-size-in-the-uk/#:~:text=In%202022%2C%20the%20average%20number%20of%20people%20per,Kingdom%20was%202.36%20compared%20with%202.37%20in%202020 [Accessed 26 Dec. 2024].

UK Government. (2023). *Legal age of marriage in England and Wales rises to 18*. Available at: https://www.gov.uk/government/news/legal-age-of-marriage-in-england-and-wales-rises-to-18 [Accessed 23 Dec. 2024].

United Nations, Department of Economic and Social Affairs, Population Division. (2018). *World Urbanization Prospects: The 2018 Revision.* Retrieved from https://population.un.org/wup/