MSc Artificial Intelligence and Data Science

Module 771766 - Fundamentals of Data Science

Census Project Report

By

Student ID – 202403820 | Samuel Datubo Jaja

Abstract

This report looks at various data science processes completed for the United Kingdom census project of 2021 to be specific. It includes aspects such as cleaning data, analyzing them and making recommendations to local government concerning land development as well as investment decision. The analysis looked at age distribution patterns, unemployment trends, religious groups, married individuals and divorcees, house occupancy rates, immigration and emigration rates, birth rates, death rates, number of university students, and commuters. Through careful analysis, the report suggests/ranks constructing a train station for the significant 50.53% of the population and investing in employment and training programs based on an unemployment rate above 8%. In making these recommendations, statistical analysis and hypothesis testing were conducted where necessary, to suit various needs of the town.

1.0 Introduction

The United Kingdom conducts a census every ten years to examine the population's status. This census helps them to make better-informed, data-backed decisions and policies like resource allocation and planning. (Office for National Statistics, n.d). The most recent census was conducted in 2021 (GOV.UK, 2022). As a member of the local government team, I used mock census data in this project to decide how to utilize an unoccupied piece of land best and determine the most beneficial investments for the new government.

2.0 Methodology (Data Cleaning & Analysis)

2.1 Data Cleaning

Data cleaning is always done on a copy of the original data (Skiena, 2017). The fuel of Data Science, Artificial Intelligence and Machine Learning is data and to achieve the most accurate insight from any data, data must be of high quality (Datascientest, 2023).

In columns of the census dataset the following operations were carried out:

1. Check for duplicate rows, empty strings, unique values and null values.
2. Check for appropriate data type.
3. Computation of percentage of missing value.
4. Data integrity check and standardization.
5. Saving cleaned and analyzed data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10457 entries, 0 to 10456
Data columns (total 12 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Unnamed: 0                      10457 non-null  int64
 1   House Number                    10405 non-null  float64
 2   Street                          10405 non-null  object
 3   First Name                      10405 non-null  object
 4   Surname                         10405 non-null  object
 5   Age                             10405 non-null  object
 6   Relationship to Head of House   9616 non-null   object
 7   Marital Status                  7879 non-null   object
 8   Gender                          10405 non-null  object
 9   Occupation                      10405 non-null  object
 10  Infirmity                       92 non-null     object
 11  Religion                        4302 non-null   object
dtypes: float64(1), int64(1), object(10)
memory usage: 980.5+ KB
```

Figure 1: Census Dataset Summary Statistics

The census dataset contained 10457 rows and 12 columns as seen in Figure 1, where each column was cleaned rigorously as detailed in the Jupyter Notebook associated with this report.

1. **Unnamed: 0**: I checked whether the column was a sequential index or useful for tracking rows before deciding to rename it to **'Serial Number**,' as noted in the Jupyter Notebook.
2. **Street**: The order of NaNs (Not a Number) in 'Street' and 'House Number' is based on the hierarchical nature of census data. A house number depends on the street. NaNs in the Street Column can be filled by leveraging the high likelihood that individuals with the same house number and surname live on the same street. Similar methods were applied to clean **House Number** and **Surname**.
3. **First Name**: As seen in the Jupyter Notebook was filled with Unknown to maintain data integrity.
4. **Age:** I converted the 'Age' column to numeric, handled errors by setting them to NaN, and filled missing values with the median due to the skewed distribution. I ensured ages were within 0 to 122 years and used bin widths of 5 for visualization.
5. **Relationship to Head of House**: I assigned "Head" to individuals aged 18 or above, but only if no other head exists in the household, while those under 18 are labeled as "Minor." Any remaining cases are marked as "Unknown" to avoid ambiguity.
6. **Marital Status:** To fill missing 'Marital Status' values, I labelled individuals under 18 as 'Minor' and those 18 or older as 'Unknown'.
7. **Gender**: I filled missing gender values by first mapping the gender based on known relationships using a dictionary (e.g., mapping 'Son' to 'Male'). Any remaining missing values are then filled using the mode, but only if less than 10% are missing to avoid skewing the data; otherwise, they are marked as 'Unknown'.
8. **Occupation:** I created a function to fill missing 'Occupation' values based on age, labeling those within the student age range as 'Student' and others with the most frequent occupation.
9. **Infirmity:** To clean the 'Infirmity' column, I filled the NaN values with 'Not Specified' to retain the existing data and avoid losing information.

10. **Religion**: To clean the 'Religion' column, I first converted it to string format and replaced NaN values with ‘Unknown’. I then calculated the frequency of each religion, setting a threshold of 5 occurrences to identify low-frequency religions. These low-frequency entries were replaced with the most common religion (mode) bearing in mind this can skew my data if there are so many missing values. For children under 18 with "Unknown" religion, they inherited the religion of the head of their household, ensuring consistency within households. Also, I visualized frequency of religions and low-frequency religions.

	Missing Values	Percentage
Unnamed: 0	0	0.000000
House Number	52	0.497275
Street	52	0.497275
First Name	52	0.497275
Surname	52	0.497275
Age	53	0.506838
Relationship to Head of House	843	8.061586
Marital Status	2578	24.653342
Gender	52	0.497275
Occupation	54	0.516400
Infirmity	10377	99.234962
Religion	6160	58.907909

Figure 2: Census Data before cleaning

	Missing Values	Percentage
Serial Number	0	0.0
House Number	0	0.0
Street	0	0.0
First Name	0	0.0
Surname	0	0.0
Age	0	0.0
Relationship to Head of House	0	0.0
Marital Status	0	0.0
Gender	0	0.0
Occupation	0	0.0
Infirmity	0	0.0
Religion	0	0.0

Figure 3: Census Data after cleaning

Figure 2 and 3 displays the dataset columns percentage missing values before and after data cleaning.

2.2 Data Analysis

2.2.1 Age Distribution and Population Demographics

After thorough data cleaning, an age group column was added to aid analysis. An age pyramid by gender was plotted as shown in Figure 4 which shows the population distribution of 10457 people by age and gender, with 47.7% (4985) males in green on the left and 52.3% (5472) females in red on the right. The age group with the highest population is 40-44, while the 110-114 age group has the fewest people. The base of the pyramid, particularly among the youngest age group (0-4 years), is narrowing, indicating a declining birth rate. Additionally, the large number of people aged 45-64 suggests an aging population, with an overall media of 34 years, 27.9% of the population are between 0-19 years, 30.1% between 20-39 of years, 32.7% between 40-64years and 9.2% are 65years and above.

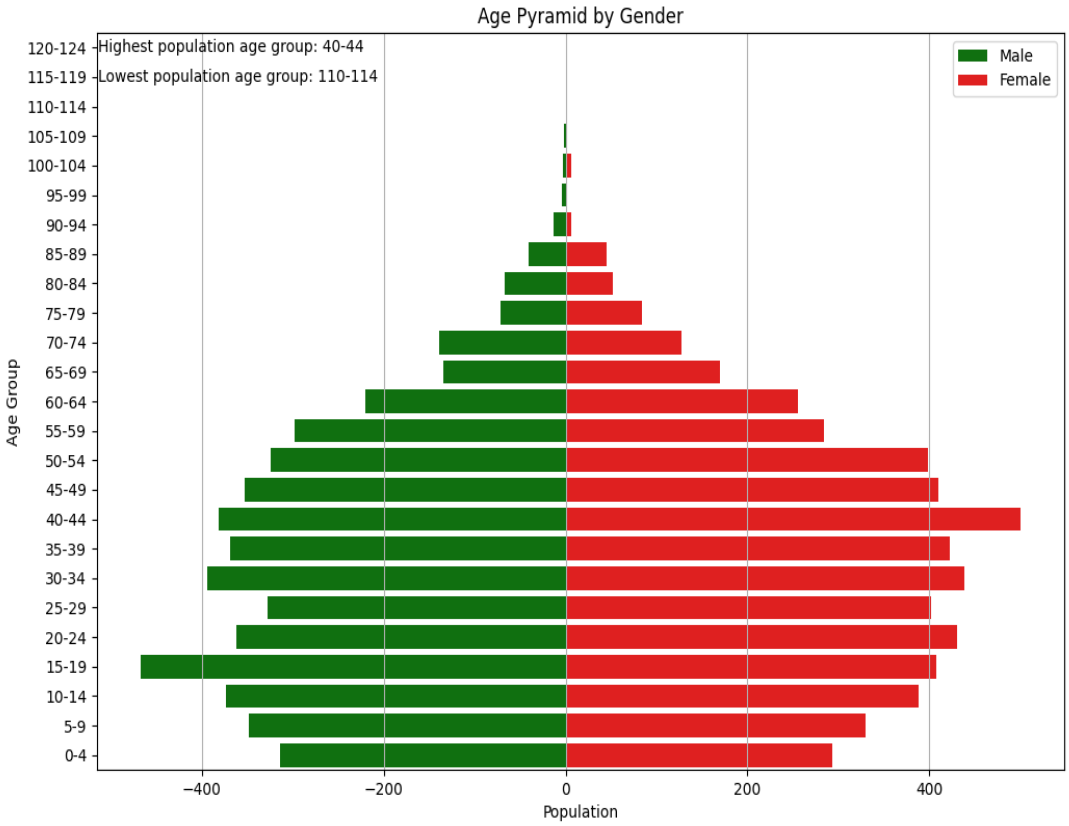

Figure 4: Age Pyramid by gender showing the age distribution of males and females

2.2.2 Birth Rate, Death Rate, and Migration & Emigration

The analysis indicates a shrinking population with a birth rate of 27047 per 100,000 and a death rate of 93,733 per 100,000. Birth rates were calculated using the number of children aged 0 and women aged 25-29, while death rates were summed across adjacent age groups. For migration, half of the lodgers and visitors were considered immigrants and half of the divorcees as emigrants, resulting in an immigration rate of 2,400 per 100,000 and an emigration rate of 4,528 per 100,000. The net growth rate, combining these factors, was "-68,813" per 100,000, confirming a population decline.

2.2.3 Unemployment Trends

Figure 5 shows various high-level occupation categories, (a column added to aid analysis) within the population. The "Student" category has the highest frequency, with 2,156 individuals, followed by the "Other" category with 1,551 individuals, and "Tech" with 1,435 individuals. The "Clerical" and "Media & Entertainment" categories have the lowest frequencies, with 94 and 83 individuals, respectively. This distribution highlights the predominant roles within the population, with a significant number of students and individuals classified under "Other," while fewer people are engaged in clerical and media & entertainment roles.

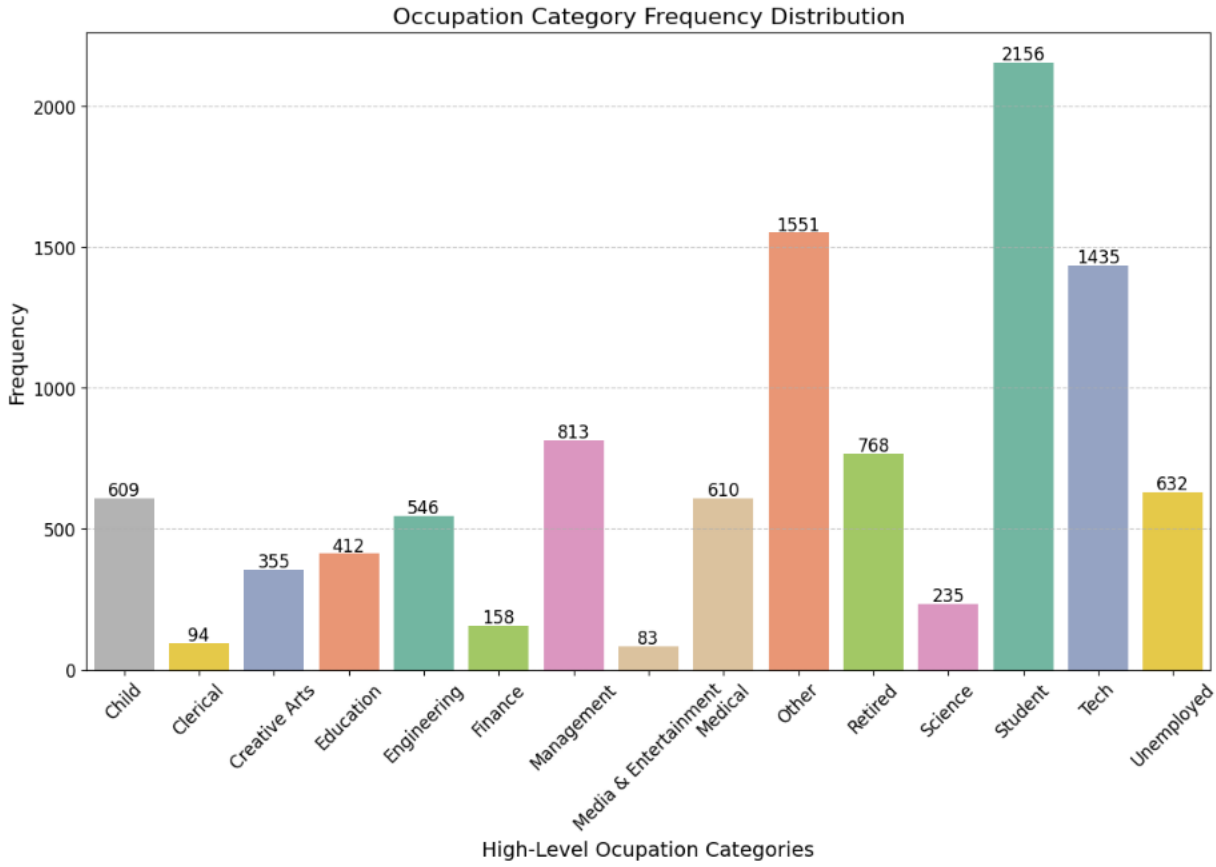

Figure 5: Occupation Category Frequency Distribution

Figure 6 shows the frequency distribution of unemployed individuals (with 8.45% of the labor force unemployed) across various age groups. The highest frequency of unemployment is observed in the 40-44 age group, with 96 individuals, followed closely by the 30-34 and 50-54 age groups, with 88 and 79 individuals respectively. Figure 6 shows a significant concentration of unemployed individuals in the middle age ranges (30-54 years), with

relatively fewer unemployed individuals in the younger and older age groups (above 60 years).

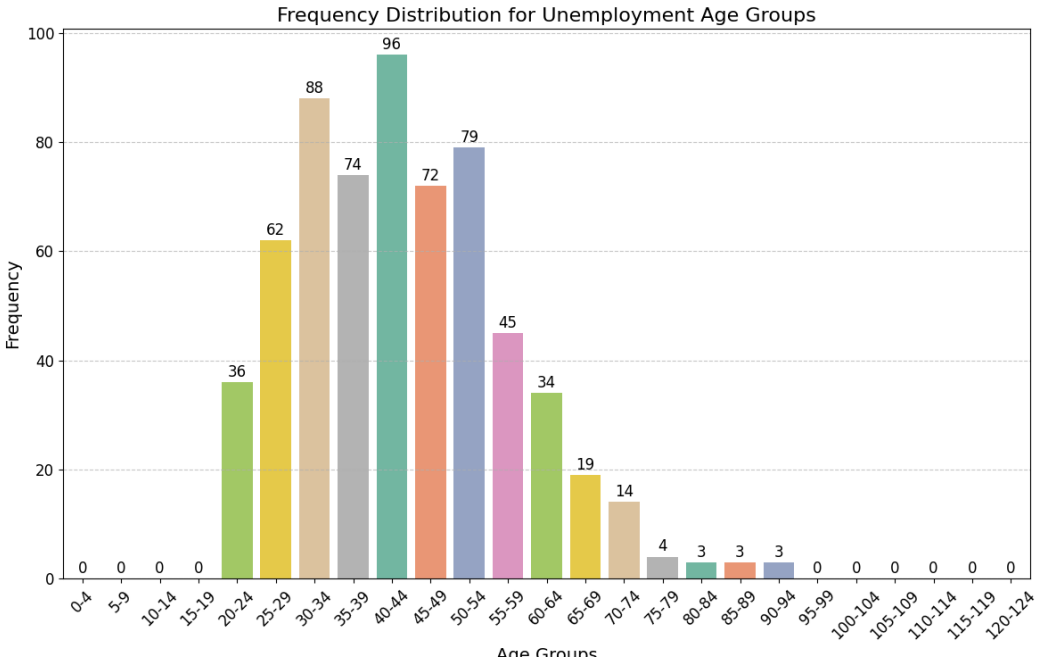

Figure 6: Frequency Distribution for Unemployment Age Groups

2.2.4 Religious Affiliations

Figure 7 and Figure 8 shows frequency of religions and frequency of low-frequency religions respectively, whereas Figure 9 displays the age distribution across different religious affiliations. The Unknown religion has a wide age range with a median around the late 40s, indicating a diverse age group. Sikh and Muslim populations are younger, with medians in the early 30s and late 20s, respectively, suggesting growing communities. Catholic and Methodist groups have median ages around the early 40s, reflecting stable, multi-generational communities. Christian and Jewish populations have older medians, particularly Jewish, which suggests an older demographic.

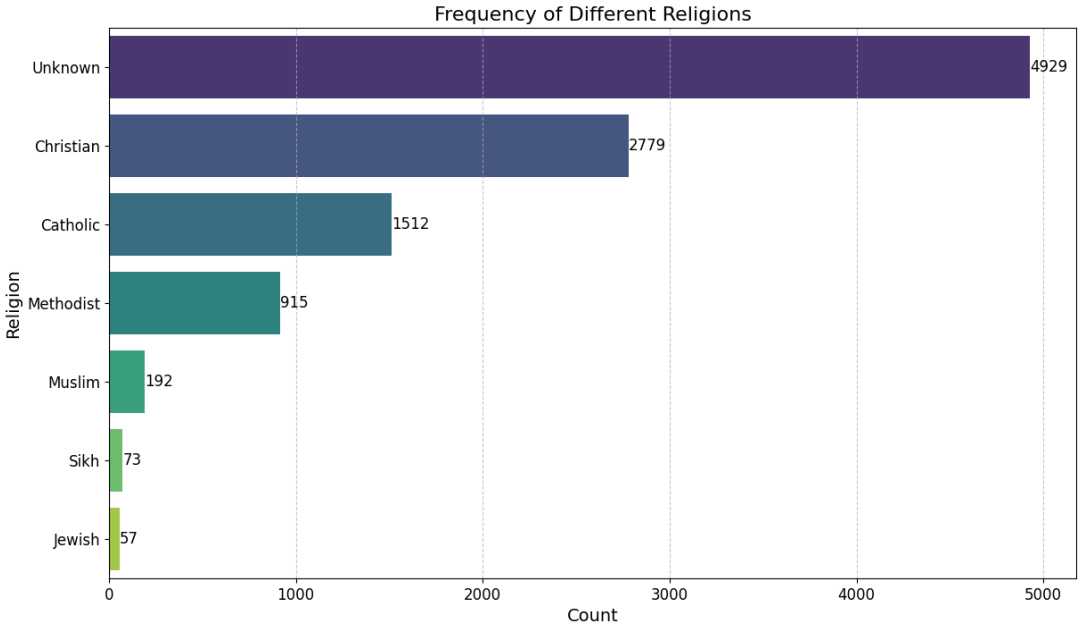

Figure 7: Frequency of different religions

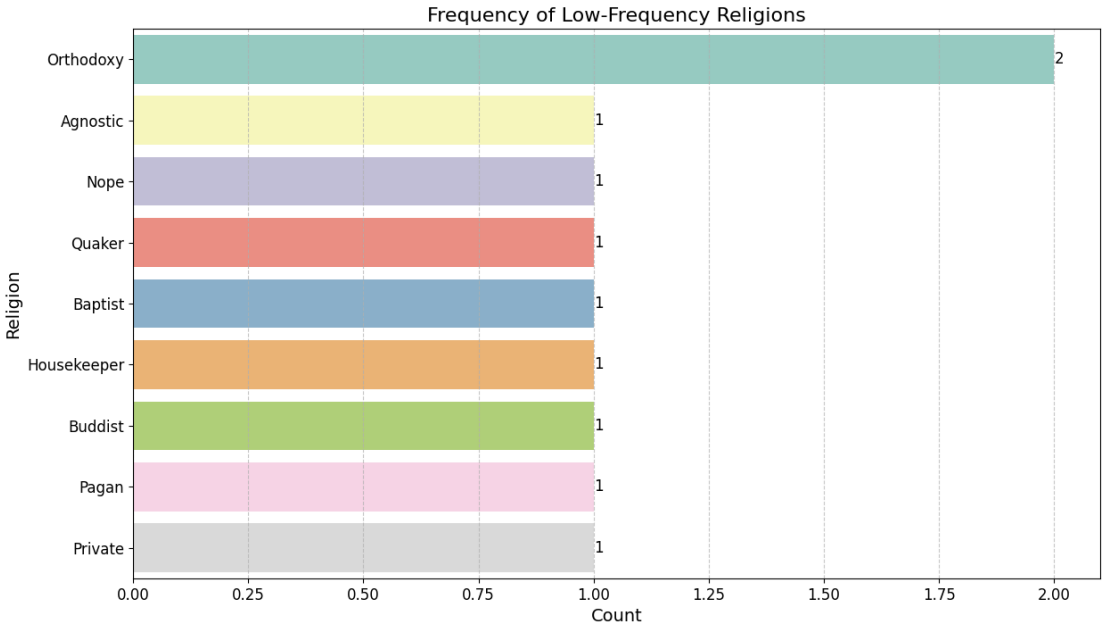

Figure 8: Frequency of Low-Frequency Religions

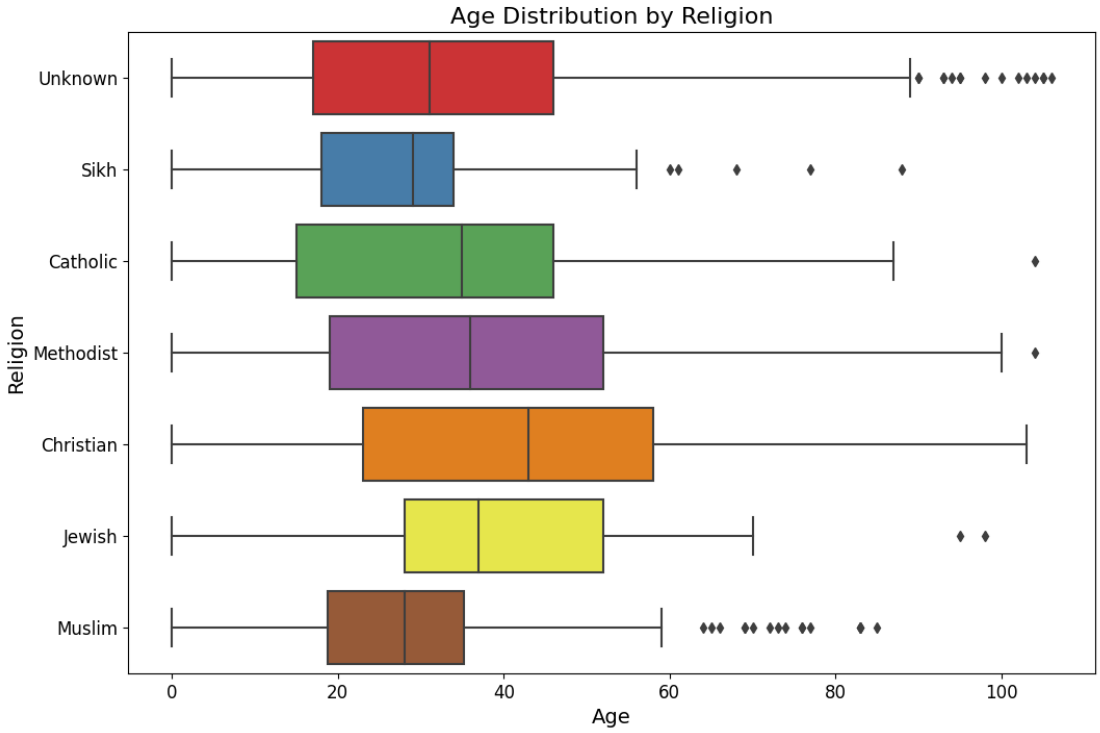

Figure 9: Age Distribution by Religion using box plot

Figure 10 shows the projected population by religion in 10 years, based on inheritance rates and current distribution. The likelihood that children will inherit their parents' religion is: Unknown (47.61%), Christian (25.53%), Catholic (14.85%), Methodist (8.72%), Muslim (2.27%), Sikh (0.62%), and Jewish (0.36%). The largest projected population is for the "Unknown" category, followed by Christian and Catholic. These projections provide insight into the potential future religious composition of the population.

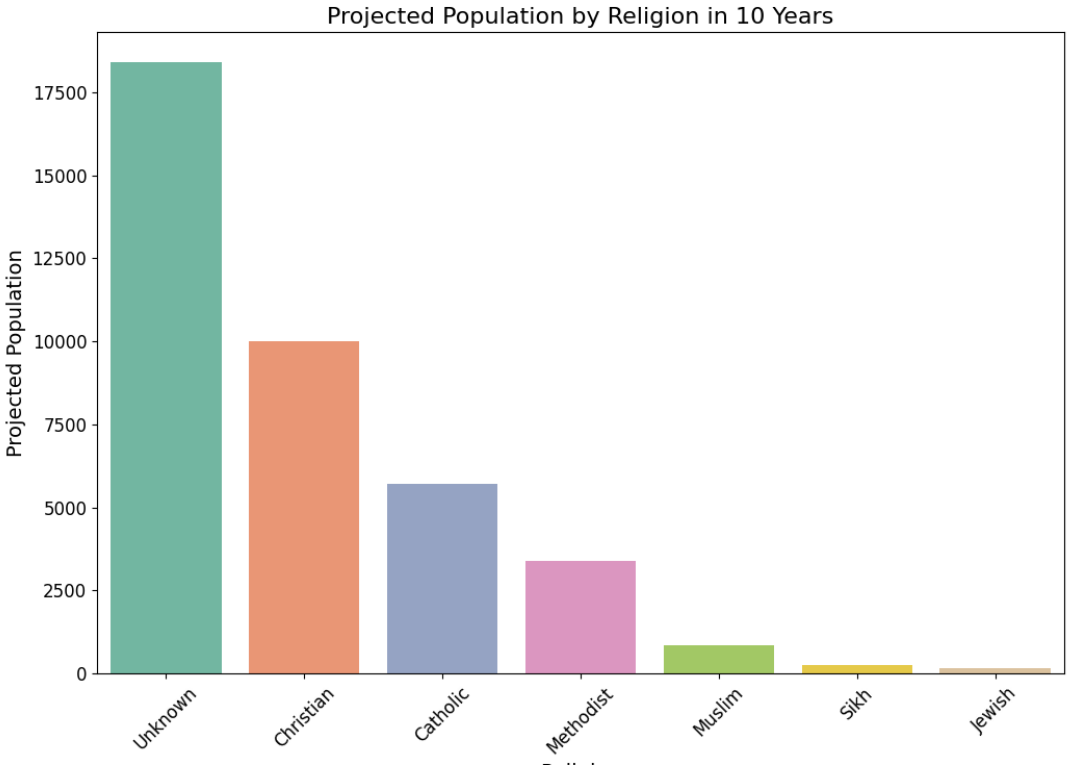

Figure 10: Projected Population by Religion in 10 Years

2.2.5 Divorce and Marriage Rates

Figure 11 highlights the transition of marital status as people age, the "Minor" status dominates the youngest age groups (0-14), while "Single" is most common in the 15-29 age range. "Married" status peaks around middle age (30-64) and gradually decreases with age, while "Widowed" status becomes more prevalent in the older age groups (65+), with noticeable shifts from "Single" to "Married" and then to "Widowed" in later years.

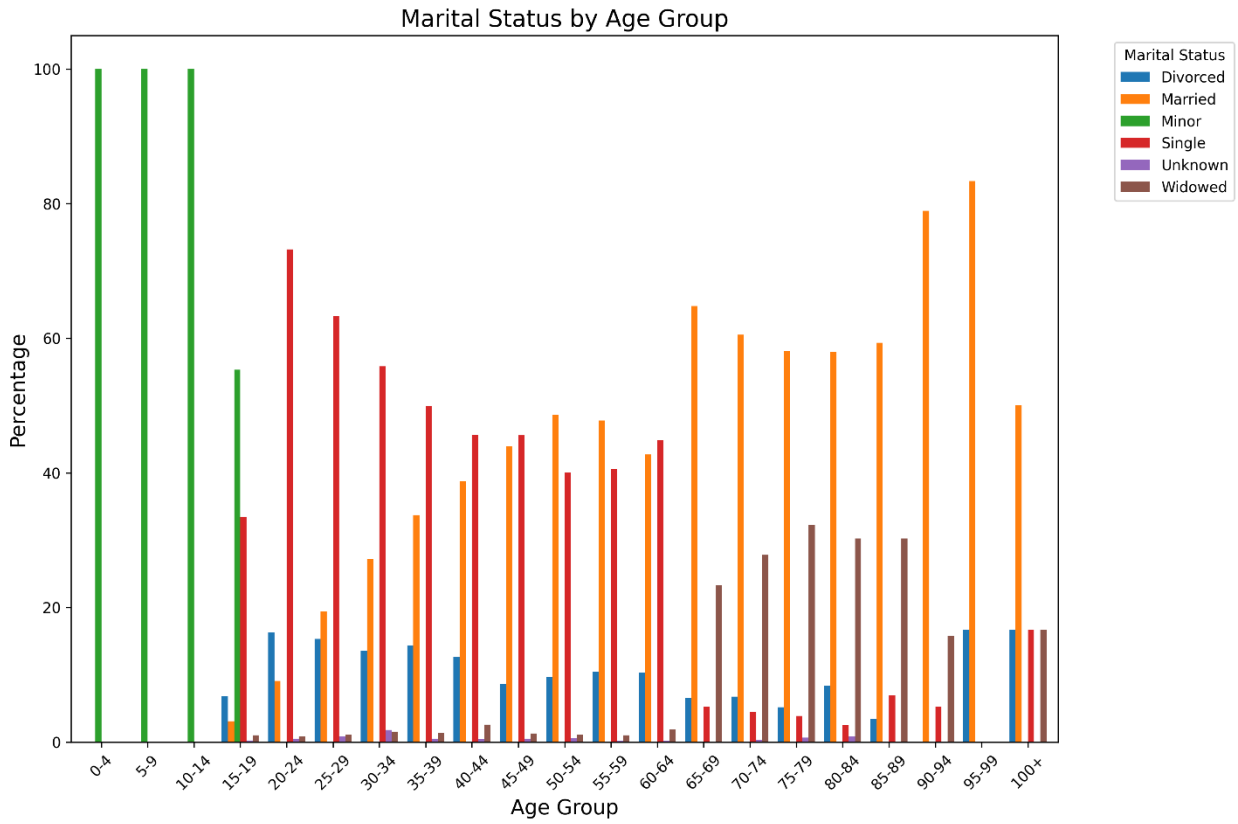

Figure 11: Marital Status by Age Group

The marriage rate in the town stands at 36.70%, while the divorce rate is 12.25%. These statistics have significant implications for housing policies and planning. A higher marriage rate typically indicates a demand for family-sized homes, as married couples often look for larger living spaces to accommodate their families. Conversely, a notable divorce rate may increase the need for smaller, single-occupancy homes or apartments, as individuals transitioning out of marriages may prefer more compact living arrangements according to (GOV.UK, n.d). The Hypothesis test also showed that marital status is dependent on age group.

2.2.6 Analysis of Occupancy Level

To assess housing needs, average occupancy per house was calculated by street. The analysis revealed that 1,342 overcrowded and 2,036 underutilized houses based on comparisons with average occupancy levels per street.

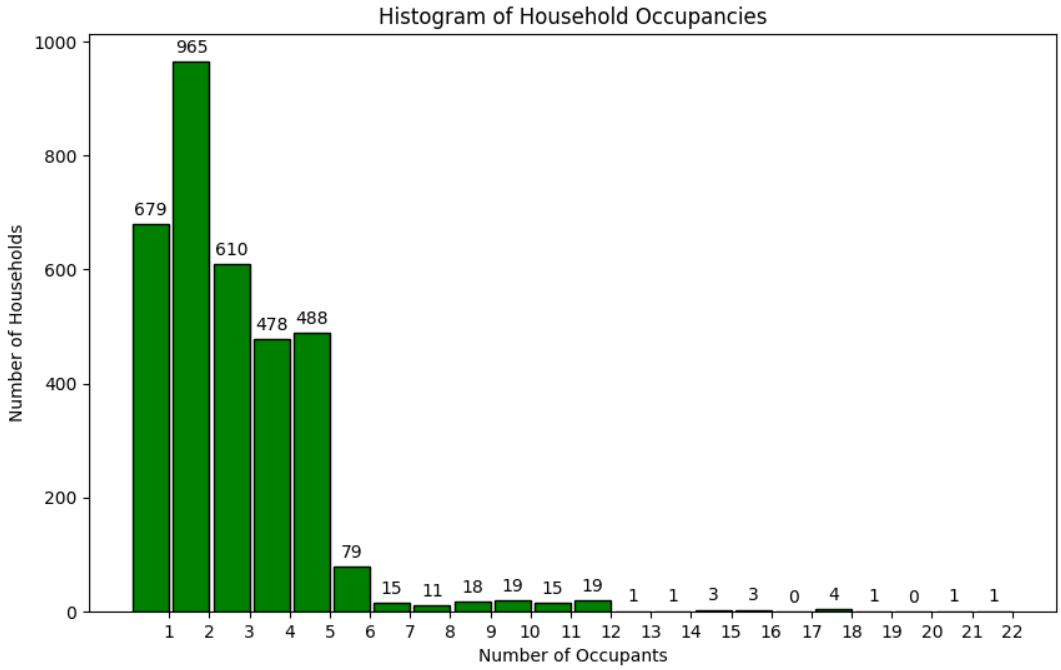

Figure 12: Histogram of Household Occupancies

Figure 12 shows the number of occupants per household, with most households having one to five occupants and two-occupant households being the most common. Households with more than six occupants are rare, indicating that larger households are uncommon.

2.2.7 University Students and Potential Commuters

The town has 669 university students, all commuters due to the lack of local universities. 4615 residents work in potential commuting-intensive professions like IT consultants, sales executives, lecturers, teachers, engineers, and doctors. This amounted to 50.53% of potential commuters, excluding retirees, the unemployed and children. This significant number of commuters highlights the need for efficient transport infrastructure.

2.2. 8 Population Health Status

"The hypothesis test revealed that the proportion of healthy individuals in the town is significantly higher than the expected 98%, indicating that the town's population is in good health (Public Health England, 2022)."

3.0 Results and Discussion

This section discusses the results and gives recommendations to the government. Table 1 below gives a brief of the land and investment development plan by the government. From which a data driven decision will be recommended.

Table 1: Decision -Making Criteria for land and Investment Development

Category	Considerations	Decision
Land Development	If the town's population is rapidly increasing, allocate land for high density housing projects.	High-Density Housing
Land Development	If the town is affluent with a demand for large family homes, allocate land for low-density housing projects.	Low-Density Housing
Land Development	Evaluate the necessity for an additional place of worship if there's already one for Catholics and demand exists for another denomination.	Religious Building
Land Development	Construct a minor injuries center if there are frequent injuries or expected future pregnancies.	Emergency Medical Building
Land Development	If there are many commuters, consider building a train station to alleviate road traffic.	Train Station
Investment	Implement retraining programs if there's significant unemployment, to equip residents with new skills.	Employment and Training
Investment	Prepare for increased funding for end-of-life care due to a growing	Old age care
Investment	Ensure adequate investment in essential services if the town is not expanding.	General Infrastructure
Investment	Boost educational spending if there is a rising number of school-aged children, either from new births or families moving into the town.	Increased spending for schooling

Source: Module 771766 Project brief

(a) What type of development should be constructed on an empty plot of land that the local government intends to develop?

Decision on High-Density Housing

The population is not experiencing significant growth, with a birth rate of 28.85% (27,047 per 100,000), and a much higher death rate of 93.73% (93,733 per 100,000). This results in a negative natural population change rate of -66.68% (−66,686 per 100,000), indicating that the population is decreasing naturally. Therefore, the growth is likely due to immigration, and significant population expansion is not expected in the short term. **high-density housing due to significant population expansion will be ruled out**.

Decision on Low-Density Housing

Analysis shows that 8.22% of the population holds high-paying jobs, indicating some affluence. However, 80.09% of households have up to four occupants, housing 60.73% of the population. This indicates no substantial demand for larger family homes. **Therefore, the option of developing large family housing is not justified and will be ruled out.**

Decision on Train Station:

Based on the analysis of commuter professions within the census data, it is found that 50.53% of the town's population are potential commuters. This percentage indicates that half of the population rely on regular road travel for their activities, which justifies the need to build a train station in the town.

Decision on Religious Building:

The Christian population makes up 26.57% of the town’s population, a significant portion. In contrast, the Catholic population, at 14.59%, already has an established place of worship. Given this disparity, and to cater to a large portion of the town’s religious needs, it is recommended and justifiable to construct a place of worship for Christians.

Decision on Emergency Medical Building:

Given the negative natural population growth rate of -66,686 per 100,000 and the very low disability rate of 0.77%, constructing an emergency medical building is not advised at this time. The population isn't growing significantly, with only 2835 expected births, indicating that existing medical facilities should meet the town's needs for now. **So, this option is ruled out.**

While a place of worship for 26.57% of Christians could strengthen community ties, building a train station would benefit 50.53% of potential daily commuters, addressing broader transportation needs. Thus, the train station should be prioritized. High-density and low-density housing are not required due to low population growth and minimal demand for larger homes, current medical facilities are adequate given the low birth and disability rates.

(b) Which one of the following investment options should be invested in?

Decision on Employment and Training

Based on the analysis of the town's demographic data, the unemployment rate is 8.04%. Given this high rate, it is essential to focus on employment and training initiatives to enhance job opportunities and reduce unemployment according to (Office for National Statistics, 2024).

Decision on Old age care

Approximately 9.25% of the population is aged sixty-five or older, 6.34% are seventy or older, and only 3.80% are seventy-five or older. Although providing care for the elderly is important, the data shows a relatively small segment of the population reaching these older ages. Consequently, the option for old age care will be dismissed. The age pyramid can also visualize this as seen in Figure 4.

Decision on General Infrastructure

The population is not significantly expanding, as indicated by the negative net growth rate. This suggests there is no indication of stress on the current systems and infrastructure in place. **Therefore, the option for general infrastructure investment will be eliminated.**

Decision on Increased spending for schooling

Approximately 20.34% of the population are students, indicating that a significant portion of children in the town are enrolled in school. Given that the population is shrinking, with a natural rate of increase of -66,686 per 100,000, there is no indication of significant pressure on the current school system. **Therefore, the option to increase spending for schooling will be eliminated.**

The only option kept based on data is retraining people for new skills due to the high unemployment rate.

4.0 Conclusion

The recommended options are building a train station for the significant 50.53% of population and investing in employment and training programs based on unemployment rate above 8%. Other development options were of lower priority based on the current data.

5. References

Datascientest (2023) *Data cleaning: Definition, methods and relevance in Data Science* Available online: https://datascientest.com/en/data-cleaning-definition-methods-and-relevance-in-data-science [Accessed 03/08/2024].

GOV.UK (2022). *Census 2021 first results, England and Wales*. Available Online: Census 2021 first results, England and Wales - GOV.UK (www.gov.uk) [Accessed 01/082024].

GOV.UK (n.d). *Births, deaths, marriages and care* Available online: https://www.gov.uk/browse/births-deaths-marriages [Accessed 03/08/2024]

Office for National Statistics (n.d). *About the Census*. Available online: https://www.ons.gov.uk/censushttps://www.ons.gov.uk/census [Accessed 02/08/ 2024].

Office for National Statistics (2024). *Employment in the UK: July 2024*. Available Online https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/employmentintheuk/july2024#unemployment [Accessed 03/08/2024]

Public Health England (2022). *National Health Survey: General Health of the Population*. Available Online https://www.gov.uk/government/statistics [Accessed 05 /08/ 2024].

Skiena, S.S. (2017). *The Data Science Design Manual*. Springer International Publishing.