Census Project Report

The objective of this project is to examine a dataset containing information about a specific population. This dataset contains information about individuals' demographics, occupation, religion, and health status. The goal is to use this data to gather insights about various aspects of this population such as their overall health status, age groups, unemployment and any correlations between demographic factors and their various outcomes.

To achieve this objective, I undertook a rigorous process of data cleaning, which included handling missing data. After cleaning the dataset, I conducted an exploratory data analysis to gain insights into the data and identify any patterns or trends that could assist me in my analysis. This report will present my discoveries and provide insights into the sample population. I will start by presenting descriptive statistics of the variables in our dataset, followed by an exploratory data analysis to detect any relationships or trends. Finally, I will discuss my conclusions and any recommendations for future research.

Data cleaning

To handle the missing and NaN values, I used various techniques, including replacing them with the most common value in the column, and in some cases, imputing missing values using statistical methods such as median imputation.

To address the issue of incorrect data types, I converted the columns to their appropriate data types. For example, the age column was converted from an object to an integer data type.

Overall, the data cleaning process ensured that the dataset was accurate, consistent, and ready for further analysis and modeling.

Age: The column was found to contain values in string format instead of integer or float. To identify the locations of these string values, error exception handling constructs were used which pointed to the indices of these values. Further explorations were carried out on the two discovered empty strings. This is to ensure that we do not use statistical methods like median or mode obtained from the whole dataset as that may result to a son been assigned with age that is greater than those of his parents. For the first empty string, I used the median age of all the husbands living in Andrews Mill Street and married to replace the empty string. For the second string, I obtained the median age of all the university students that are related to the

head of the house as son and used it to replace the empty string.

```
count    8329
mean       35
std        21
min         0
25%        18
50%        35
75%        50
max       106
Name: Age, dtype: int32
```

Table 1.0 is the descriptive properties of the Age column with numerical outputs.

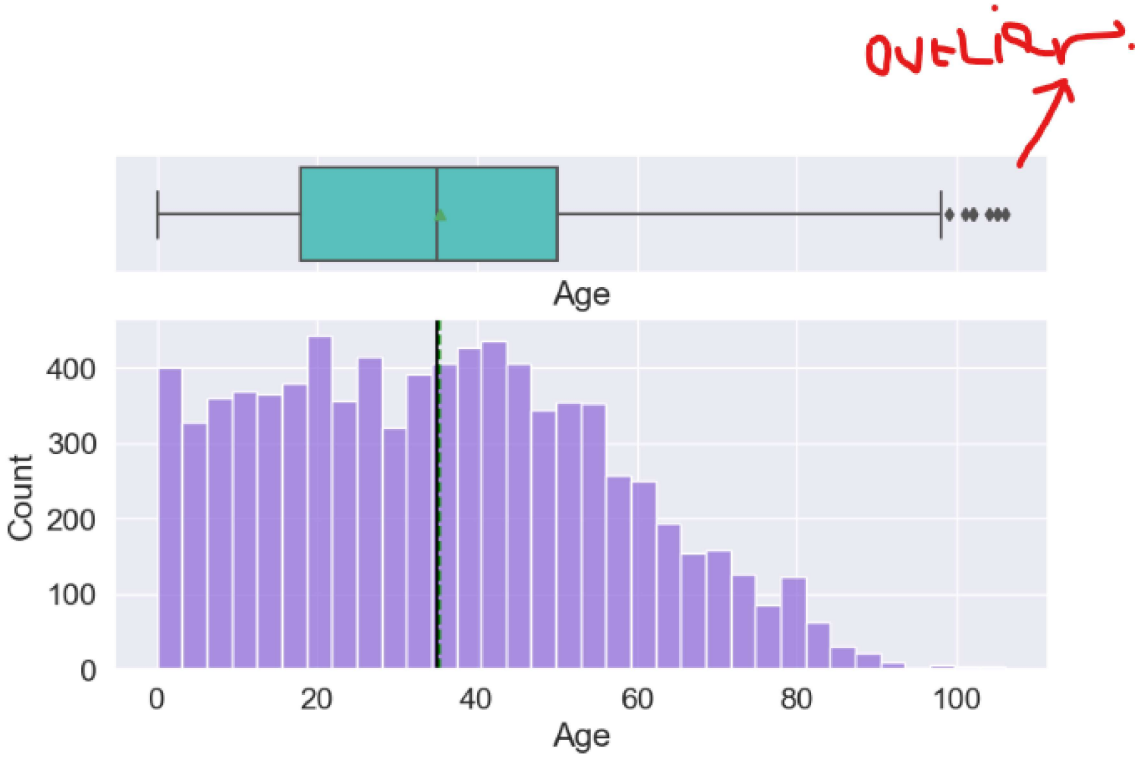

Fig. 1.0 Univariate and Box plot showing the outliers.

Looking at the Table 1.0 and Fig 1.0, you will see that I have an outlier in my Age column. It is preferable to use median imputation when the distribution of a dataset is skewed since median is less affected by outliers than mean (Kumar, 2023). This helped me to make the final decision to use the median value to replace the empty strings and subsequently converted the Age column to an integer.

Marital Status: This column was discovered to have a 2056 nan value and 3 empty strings after carrying out sanity check. Further exploration of the dataset showed that people with this NaN values are younger than 18 years. The UK Marriage and Civil Partnership (Minimum Age) Act 2022, states that people that are not up to 18 years are not allowed to be married or have civil partners (Ministry of Justice, 2023). This points out that they are not eligible for marriage hence my reason for replacing the NaN values of these 2056 with Not_Eligible. The empty strings in the remaining three people were explored further and it was discovered that Lyon Hilton living in House Number 53 has a common surname with the other 5 people that are living in the house. This helped me to replace the empty string with 'Married.' For the remaining two people, further exploration of the dataset did not lead to any tangible information hence, I replaced it with 'Unknown.'

Occupation: A total of 656 individuals reported being retired from their previous occupation. The column contained both common occupations (e.g., student, unemployed, child) and specific, less common ones (e.g., Retired immunologist, control and instrumentation, hydrogeologist). Due to the wide range of occupations and potential difficulty in cleaning and standardizing the data, one approach is to group all individuals reporting "Retired" as their occupation into a single category or occupation called 'Retired'. This is because the information provided by these individuals may not offer enough meaningful data to draw accurate conclusions and treating them as a separate occupation could introduce noise.

I also did a check on my dataset to ensure that I don't have people that are not up to working age among the employed group as the UK law stipulates that children can only start full work when they have gotten to 16 years of age which is the minimum year for school leaving age (UK Government, 2012). This showed that all the people less than 16 years in my dataset do not have any occupation.

Religion: A check on this Colum showed that there are 2105 NaN values and 5 empty strings. I used conditional statements in relationship with Head of the house, Surname, House number and Religion to obtain the correct information that helped to reduce the NaN values from 2105 to 388. Further exploration of the dataset did not provide me with enough information on what is required to replace the remaining NaN values and empty strings, hence the mode was used to clean up the column.

```
None          3864
Christian     2264
Catholic      1223
Methodist      699
Muslim         143
Sikh            83
Jewish          50
Quaker           2
Buddist          1
Name: Religion, dtype: int64
```

Table 2.0 Descriptive details of the Religion column

The number of people with None values stood at 3864, followed by the Christian with the population of 2264.

Infirmity: The Infirmity column appears to indicate if a person had a physical or mental disability or infection. There are 8261 "None" entries and only a few other values, making it less useful for analysis due to its high skewness. Also, Infirmity is not something you can assign to someone using mode or median as this has to do with health. I also noticed that there is a feature called Unknown Infection. What that means to me is that there is Infirmity for that group of people, but that infection is unknow. The discovered 8 empty strings in this column were assigned Unknown_Infirmity to them.

```
None                   8261
Physical Disability      17
Mental Disability        13
Blind                    10
Unknown Infection         8
Unknown_Infirmity         8
Deaf                      8
Disabled                  4
Name: Infirmity, dtype: int64
```

Table 3.0 Descriptive details of the Religion column

First Name: My exploration of this row showed that there is only one empty string. Since First name is unique, I used ‘Unknown’ to replace the empty string.

Surname: One empty string was found, prompting further investigation to clean and prepare the data for analysis. The house number (5) and street name (Gill Mews) were extracted, and this information was used to extract details of all the individuals living at that house with a common surname. A common surname (ADAMS) was identified among the residents of that

house number and was used to replace the missing value.

Final CSV File: The final step in my data cleaning was running the sanity check on my dataset and saving it with a new name 'census06_cleaned.csv'. My final data produced the following features:

```
House_Number                    0
Street                          0
First_Name                      0
Surname                         0
Age                             0
Relationship_to_Head_of_House   0
Marital_Status                  0
Gender                          0
Occupation                      0
Infirmity                       0
Religion                        0
dtype: int64
```

Table 3.0 Main file showing absence of NaN values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8329 entries, 0 to 8328
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House_Number                  8329 non-null   int64
 1   Street                        8329 non-null   object
 2   First_Name                    8329 non-null   object
 3   Surname                       8329 non-null   object
 4   Age                           8329 non-null   int32
 5   Relationship_to_Head_of_House 8329 non-null   object
 6   Marital_Status                8329 non-null   object
 7   Gender                        8329 non-null   object
 8   Occupation                    8329 non-null   object
 9   Infirmity                     8321 non-null   object
 10  Religion                      8329 non-null   object
dtypes: int32(1), int64(1), object(9)
memory usage: 683.4+ KB
```

Table 4.0 Main file showing data types of all the columns.

I was able to retain all my columns and rows after the cleaning of the dataset (8329, 11). This cleaned dataset will now be used for exploratory data analysis to answer research questions.

Detailed Analysis

Age

The pyramid analysis revealed a relatively low number of young people aged 0-4 and many middle-aged and older individuals, particularly those aged 65 and above. This may indicate a lower birth rate and a population that tends to live longer.

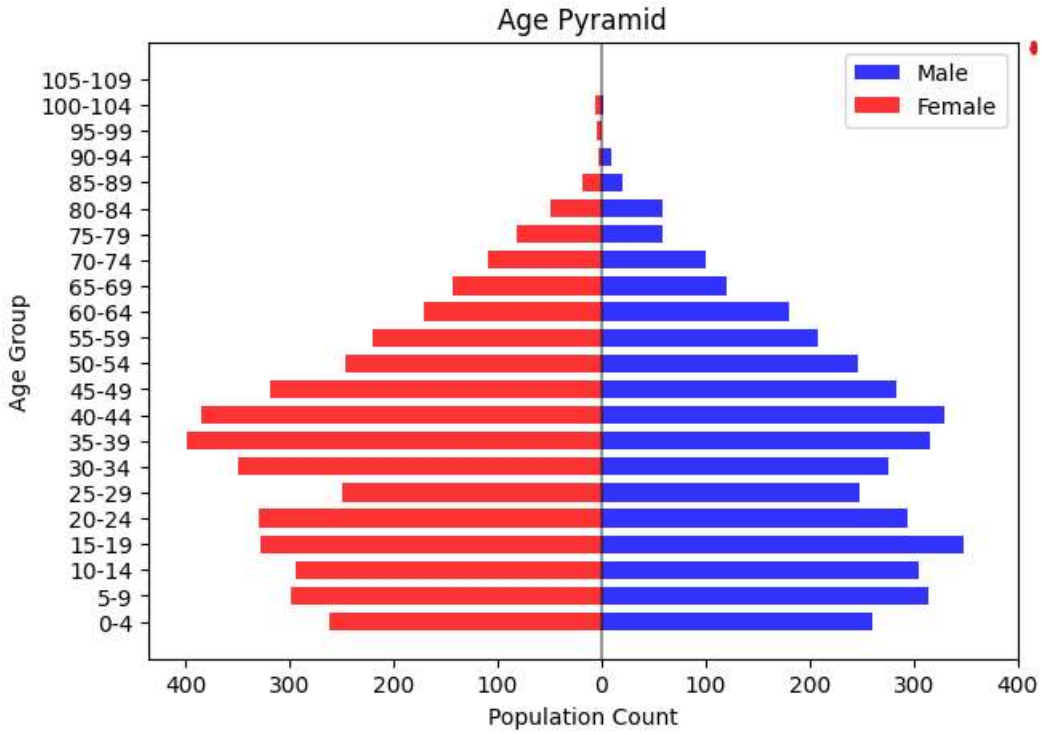

Figure 2.0 Description of the Age Pyramid

Furthermore, the age distribution shows that there is a consistent number of people across the age groups from 10 to 59 years old. This group of people (10- 59) are the most active and mobile people who can be classified as students (university students included) or working-class people. They also made up the larger number of people and this suggests that they will be commuting in and out of the town. The consistency across this age group could suggest that the population has a relatively stable demographic structure in terms of people of working age and students. However, there is a decline in the number of people in the older age groups, particularly those aged 70 and above.

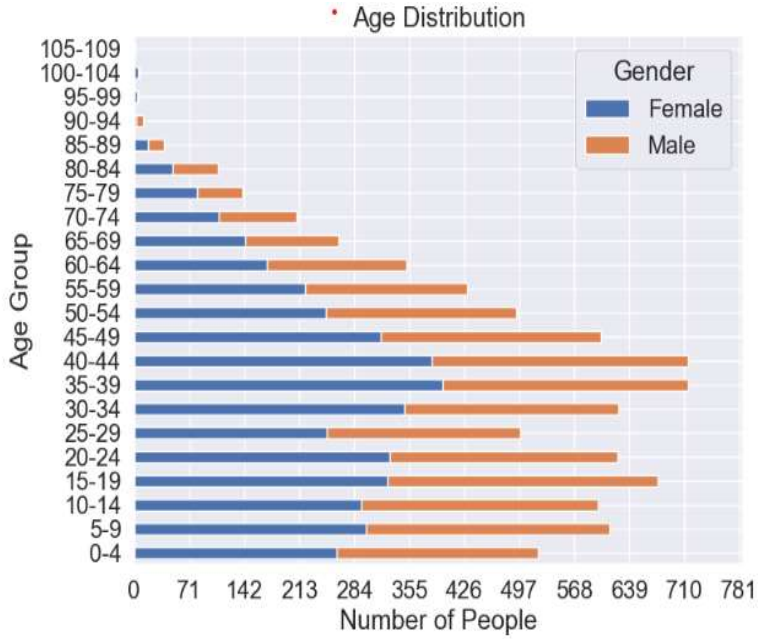

Figure 3.0 Age distribution

In terms of gender, the number of males and females is relatively similar across most age groups. However, there is a slightly higher number of females in the older age groups, particularly those aged 75 and above. This could suggest that females tend to live longer than males in the population.

In summary, the provided dataset suggests that the population has a stable demographic structure in terms of people of working age and students, but a relatively low birth rate and a tendency to live well into old age. There is also a slight gender difference in terms of life expectancy, with females tending to live longer than males.

Age and unemployment trends

The provided data reveals that the unemployment rate varies significantly across different age groups. Individuals in the age group of 0-19 have an unemployment rate of zero, likely due to not yet being in the labor force. The rate increases steadily from age 20 to 69, peaking at 12.84% for the age group 30-34. There is a slight decline in the rate after age 44 but remains high for older age groups. It is worth noting that the unemployment rate in this community is higher than the UK unemployment rate of 3.7% as of March 14, 2023(Office for National Statistics, 2023).

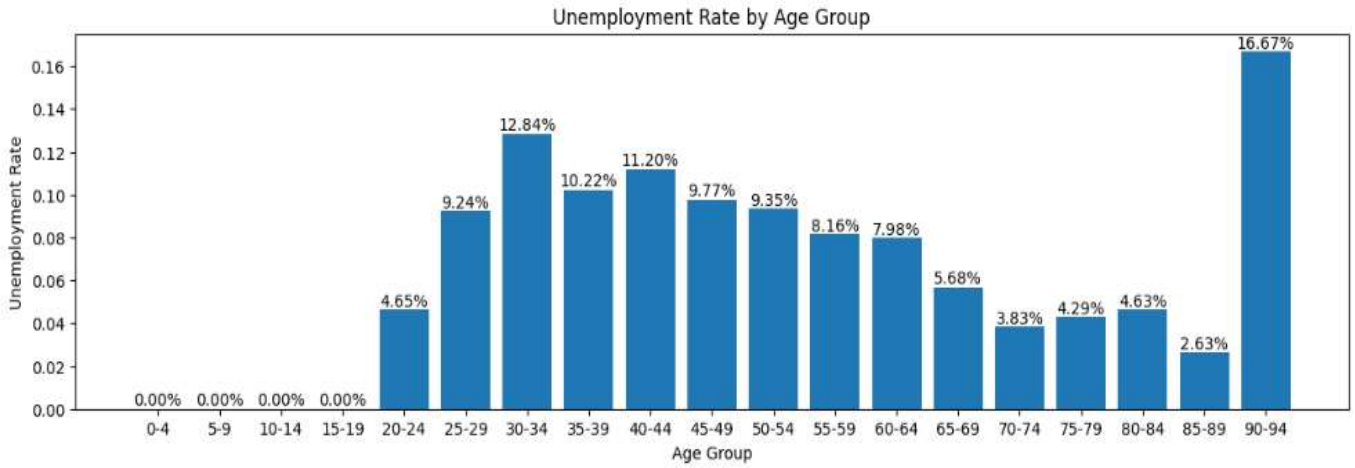

Figure 4.0 Evolution of unemployment rate with the age

The unemployment rate was also above national level highest in the age group 80-84, reaching 4.63%, due to retirement. Surprisingly, it drops to 2.63% for ages 85-89 and spikes to 16.67% for ages ninety and above, indicating an outlier.

Overall, age is a significant factor in determining unemployment rates, with younger and near-retirement individuals having lower rates, and those in their Middle Ages having higher rates.

Religion distribution

The largest group of individuals reported having no religious affiliation were 3864 in total. The Christian was 2264. Catholic and Methodist were the third and fourth largest groups, respectively, with 1223 and 699 individuals. Muslim, Sikh, Jewish, Quaker, and Buddhist populations were small, with 143, 83, 50, 2, and 1, respectively. Without additional information or data over time, it is difficult to determine if any religions are growing or shrinking or if newer religions are increasing in numbers.

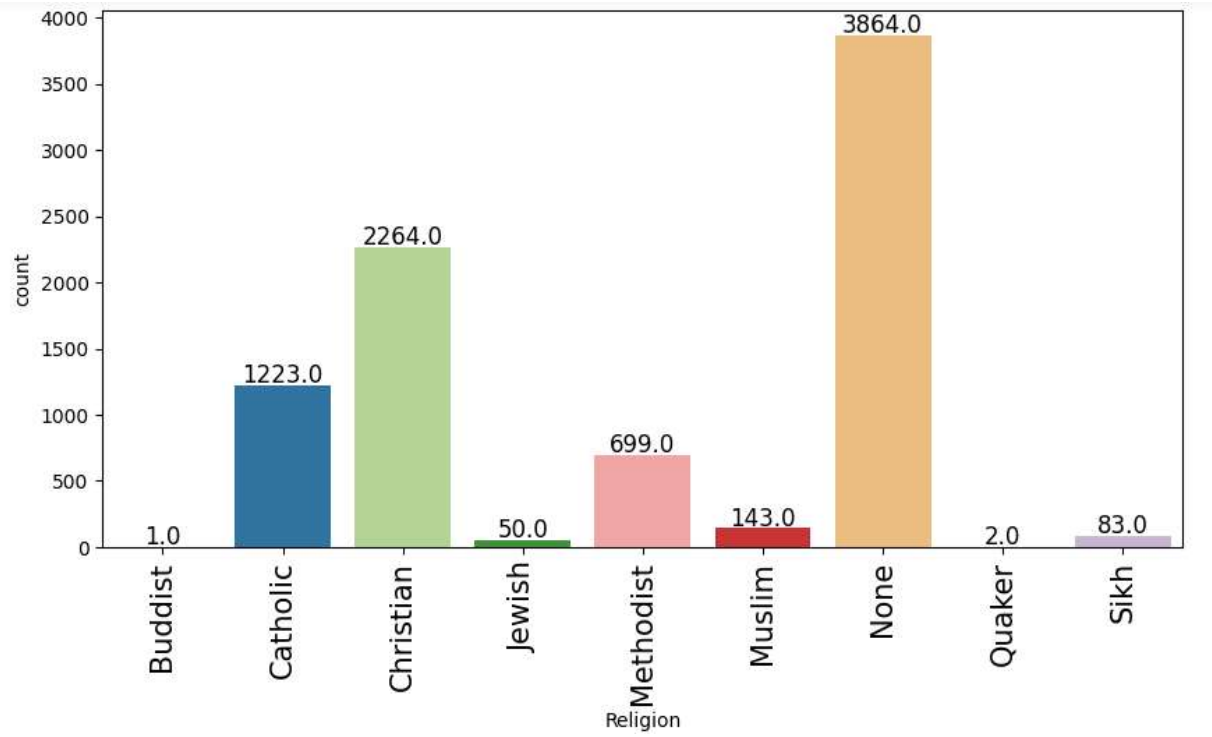

Figure 5.0 Religion distribution

Marital Status

The marital status showed that greater number of people (2941) are single, the married are 2226, the not eligibles are 2056, the Divorced are 744 in number while widows are just 360. The unknowns are two in all.

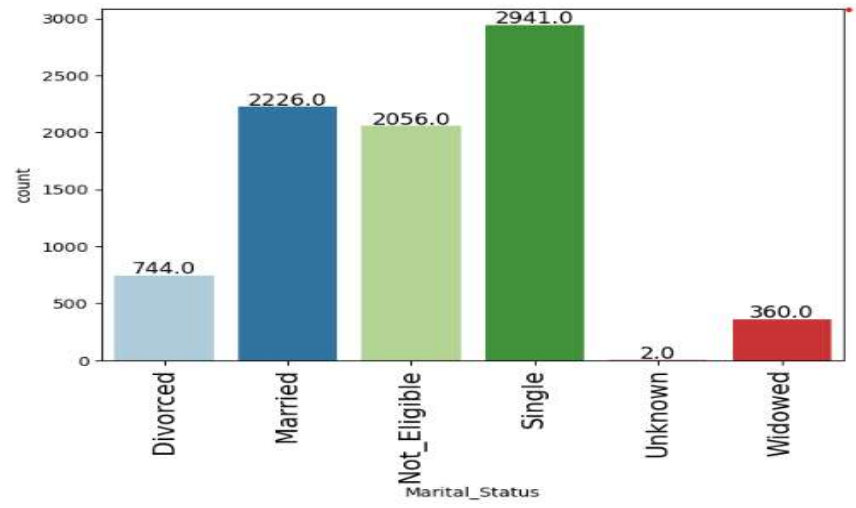

Figure 6.0 Religion distribution

Marriage and divorce rates

The marriage rate increases with age until around age 65-69, where it begins to decline slightly. The highest marriage rates are among those aged 30-94. The plot showed that marriage rate is higher than the divorce rate.

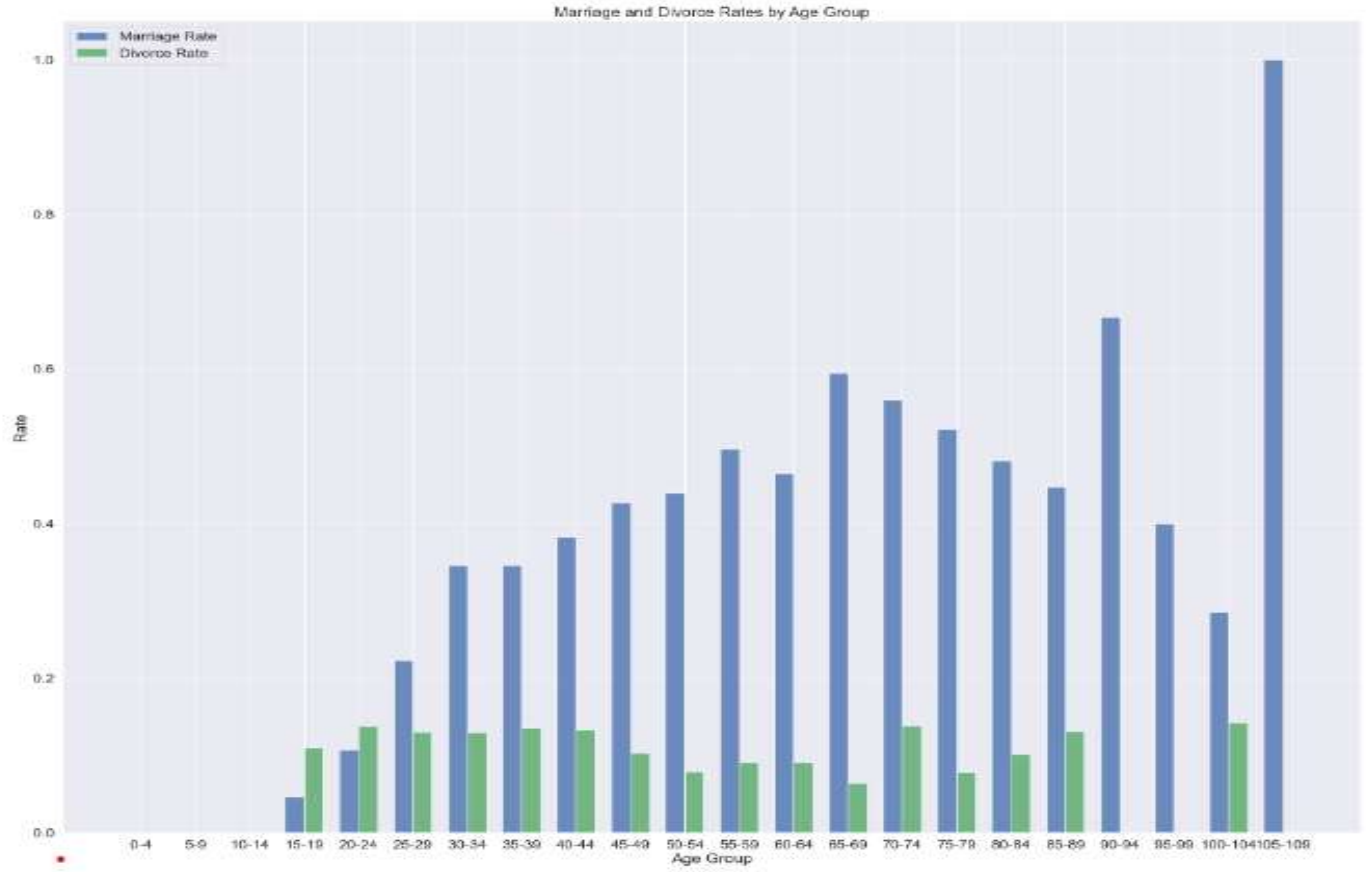

Figure 7.0 Divorce and marriage rates.

These trends may impact how we think about housing. For example, if we expect people to get married and form households, we may need more family-sized housing units. Conversely, if divorce rates are high, we may need smaller housing units for single adults or single parents. Additionally, if we expect people to marry later or not at all, we may need more housing units for single adults.

Occupation

The dataset showed a sizable number of students and university students (2228) which is above 25% of the population, but it is notable that the town has no universities. This suggests that many students are likely to commute to nearby towns or cities for their education. Other

groups like Retired with the value of 659, unemployed with 513 and those record as child with 500. Table 5.0 below shows the occupation with five hundred and above.

Student	1669
Retired	656
University Student	559
Unemployed	513
Child	500

Table 5.0 occupation with 500 people and above.

Infirmity

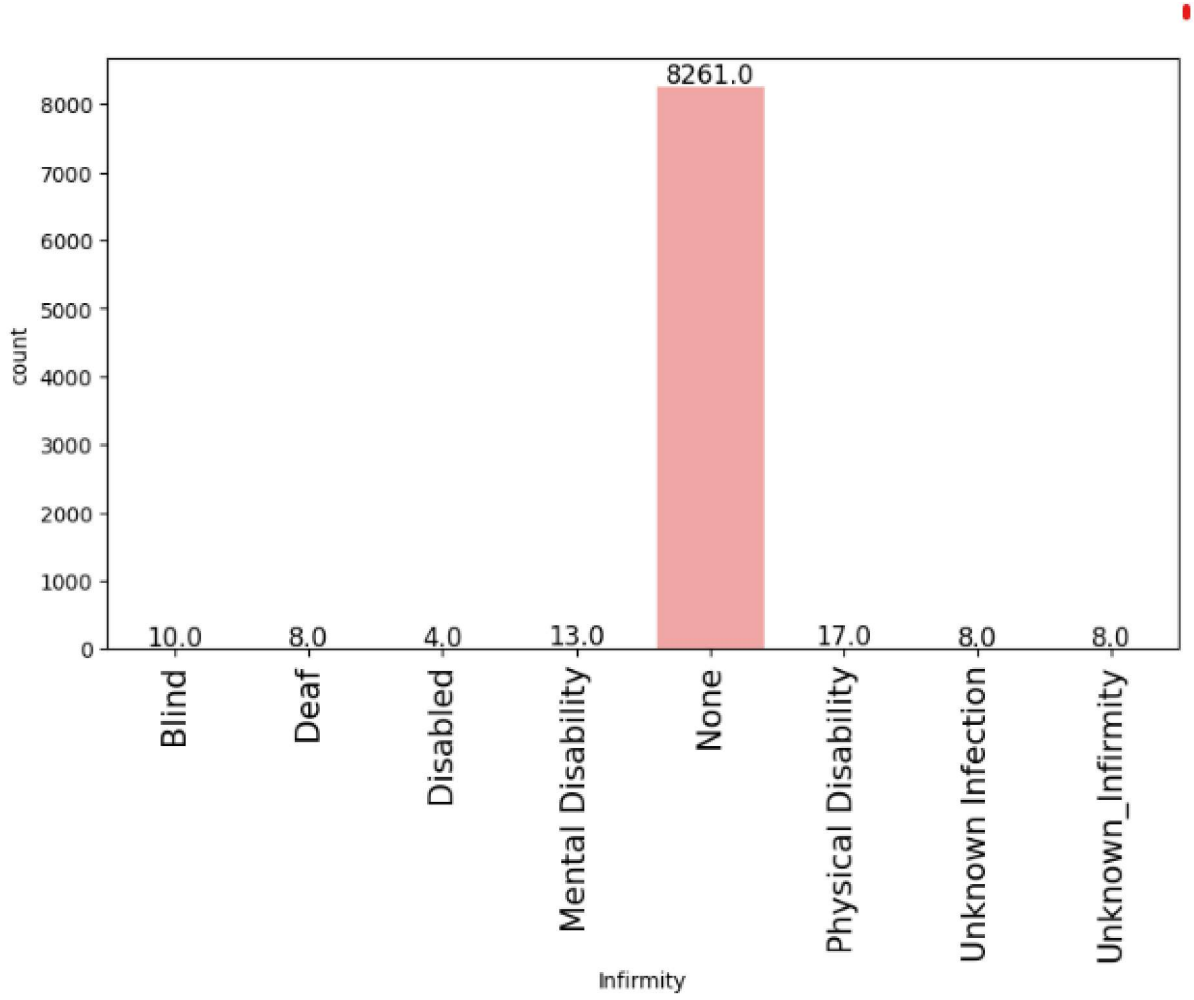

Figure 8.0 Infirmity distribution

From the dataset, most of the populations (8261 instances) reported no infirmities. Of the remaining instances, the most common infirmities reported were physical disability and mental disability, with 17 and 13 instances, respectively.

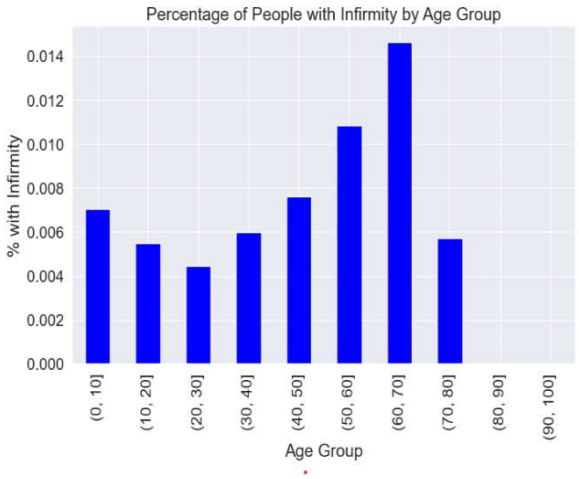

Figure 9.0 Infirmity distribution

Infirmity	None	Yes	Yes_percentage
Age			
(0, 10]	1126	8	0.007055
(10, 20]	1266	7	0.005499
(20, 30]	1117	5	0.004456
(30, 40]	1331	8	0.005975
(40, 50]	1306	10	0.007599
(50, 60]	912	10	0.010846
(60, 70]	606	9	0.014634
(70, 80]	347	2	0.005731
(80, 90]	146	0	0.000000
(90, 100]	17	0	0.000000

Table 6.0 Infirmity distribution

To further analyze the relationship between infirmity and age, I binned the ages into 10-year intervals and calculated the percentage of "yes" responses for each bin. The results show that the percentage of people reporting infirmities increases with age, with the highest percentages in the 60-70 age bin and the lowest percentages in the 20-30 age bins. However, it is worth noting that the number of instances in the older age bins is much smaller, which could skew the percentages.

Overall, this analysis suggests that while infirmities are rare in this population (0.728% of instances reported infirmities), they are more common in the age group 60-70.

Relationship to the Head of the House

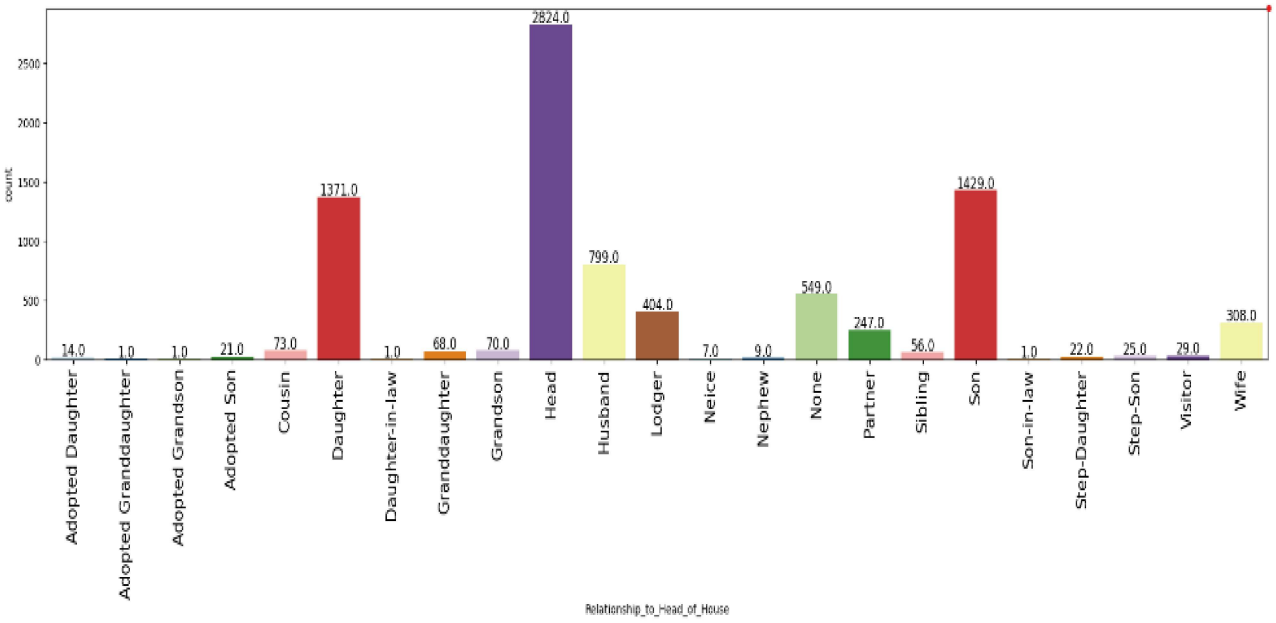

Figure 10.0 Relationship to the head of the house

The dataset contains information on the relationships between individuals and the head of the household. There are 2824 individuals identified as heads of household, 1371 daughters, 1429 sons, 799 husbands, and 308 wives related to them. In addition, there are 404 lodgers and 29 visitors related to the heads of households. This information can be useful in planning additional transportation and housing needs for individuals coming into the city.

Commuters:

The likely commuters in the town are the students (university students included), Lodgers, Visitors, unemployed and some employees that may be frequenting the nearby towns for work.

Recomndations:

I recommend constructing a train station on the unoccupied plot of land. The data shows that the largest population in the town falls between the ages of 10-59, indicating that there are a high number of active and mobile individuals who are likely to commute within and outside

the city for work or education. The university students and students which are above 25% of the population need the train to commute within and outside the town. Furthermore, the relationship instances for the head of the house where we have Lodgers and Visitors suggest that people do come into the community, which could increase the demand for transportation. A train station would provide a safe and convenient means of commuting for both the town's residents and visitors.

The population is stable but not expanding as can be seen from the age pyramid where the population of people between the age range of 0 -14 are lower than those of the middle age group. This is an indication that High density housing are not required.

While Low-density housing could be considered due to the stable marriage rate in the population, the data shows a slightly low birth rate, indicating that the demand for larger family housing may not be significant.

Additionally, the data does not suggest a significant demand for an emergency medical building as only 0.728% of the population reported infirmity.

Also, to note is that out of 8329 people in the dataset, 3864 reported no religion affiliation. The data does not show any demand for a new religious building.

Based on the data provided, a train station is the most practical and beneficial option for the town.

I also recommend employment and training investments as the best option for the town. The high unemployment rate, particularly in the age group of 20-34, indicates a need for upskilling and reskilling programs to facilitate employment.

The low-density housing also implies potential demand for larger family housing, which could be met through job creation and business attraction.

While the number of retired people is expected to increase, the town already has a high proportion of older individuals, and thus, investing in old age care may not be an immediate priority.

Similarly, although the population of school-aged children is growing, the data does not suggest an urgent need to increase spending on schooling.

Overall, expanding towns require investment in general infrastructure, but it is crucial to prioritize investments based on immediate needs. Given the high unemployment rate, investing in employment and training is the most appropriate option to address.

Conclusion:

Based on the listed challenges currently facing the town, I advise that they embark on the construction of a train station and invest in employment and training.

References:

Kumar, A. (2023) *Python - Replace Missing Values with Mean, Median & Mode*. Data Analytics. Available online: https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/?utm_content=cmp-true [Accessed 31 Mar. 2023].

Ministry of Justice (2023) *Legal age of marriage in England and Wales rises to 18*. GOV.UK. Available online: https://www.gov.uk/government/news/legal-age-of-marriage-in-england-and-wales-rises-to-18.

Office for National Statistics (2023) *Unemployment - Office for National Statistics*. ons.gov.uk. Available online: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment.

UK Government (2012) *Child employment*. GOV.UK. Available online: https://www.gov.uk/child-employment.

UniHull Canvas (n.d.) *Sign in to your account*. login.microsoftonline.com. Available online: https://canvas.hull.ac.uk/courses/66260/assignments/207481 [Accessed 2 Apr. 2023].