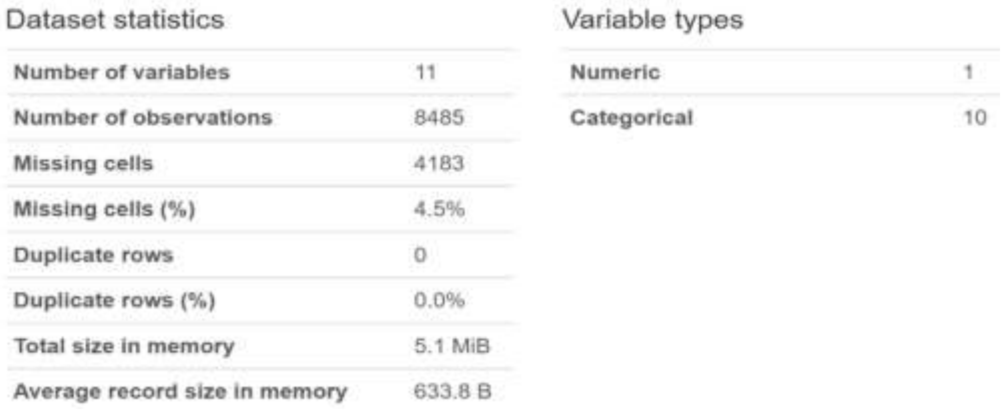
CENSUS PROJECT REPORT

Abstract

This project report shows the census data cleaning and exploratory data analysis on the demographics of a modestly sized town, centrally located between two large cities.
Further explorations using pandas profiling and data visualizations were carried out with the aim to gain insights to proffer recommendations to the Local Government to decide what future investments to consider and what facility should be built on the unoccupied plot of land.

Data Features and Overview

The dataset statistics reported a total of 11 variables and 8485 observations, showing 1 numeric and 10 categorical variables. 4.5% of these entries were missing (a total of 4183 missing units). The ".info()" and ".isnull()" methods were called to show the missing cells. The features revealed 2065 and 2118 missing cells (entries) in 'Marital Status' and 'Religion' variables respectively.

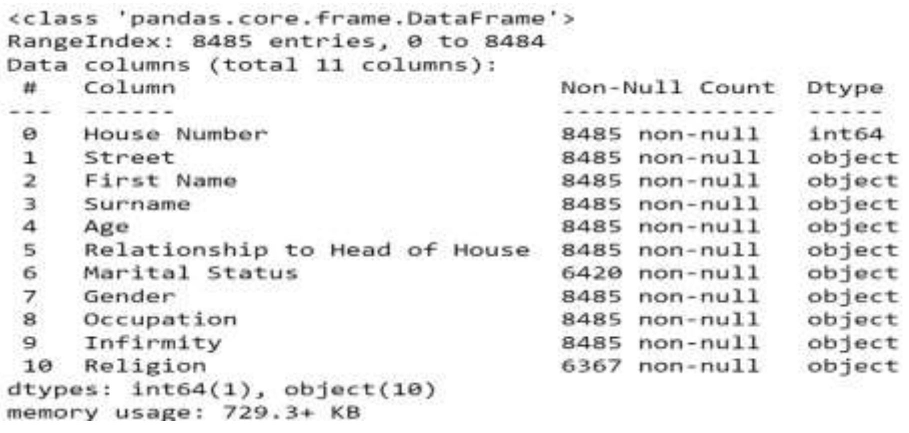


Dataset statistics

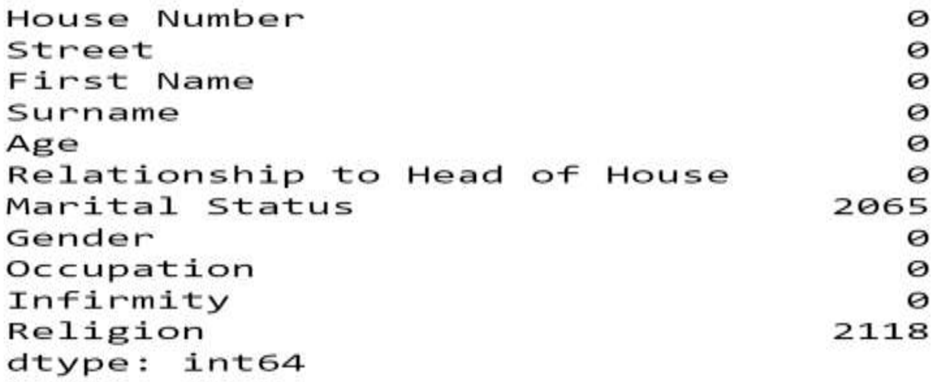
Number of variables	11
Number of observations	8485
Missing cells	4183
Missing cells (%)	4.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	5.1 MiB
Average record size in memory	633.8 B

Variable types

Numeric	1
Categorical	10



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8485 entries, 0 to 8484
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   House Number                  8485 non-null   int64 
 1   Street                        8485 non-null   object
 2   First Name                    8485 non-null   object
 3   Surname                       8485 non-null   object
 4   Age                           8485 non-null   object
 5   Relationship to Head of House 8485 non-null   object
 6   Marital Status                6420 non-null   object
 7   Gender                        8485 non-null   object
 8   Occupation                    8485 non-null   object
 9   Infirmity                     8485 non-null   object
 10  Religion                      6367 non-null   object
dtypes: int64(1), object(10)
memory usage: 729.3+ KB
```



```
House Number                        0
Street                              0
First Name                          0
Surname                             0
Age                                 0
Relationship to Head of House       0
Marital Status                   2065
Gender                              0
Occupation                          0
Infirmity                           0
Religion                         2118
dtype: int64
```

Figure 1.0: Showing missing entries in Marital Status and Religion

Hence, we will be looking at,

1. Data Cleaning and Preprocessing
2. Exploratory Data Analysis
3. Data Visualizations
4. Recommendations

Data Cleaning and Preprocessing

During this process, the errors were identified by calling the unique method (unique ()) on all variables. The dataset was conditioned and filtered to view the entries where these errors were present. The identified errors were resolved to make the data fit for further exploration and analysis. These errors and how they were handled are stated below. The codes are provided in the Jupyter Notebook submitted with this report.

Data Errors and Fixing

Data Type:
The entries of age variable were cast into their right data type, ‘Integer’ after its entries in floats were rounded off to their nearest whole numbers and the age, ‘Nine’ in word was replaced to figure 9.

Misspelling:
The misspelt entry “Neice” in the ‘Relationship to Head of House’ was replaced with “Niece”.

Blanks and Missing Cells:

Blanks and missing cells (also known as null or NaN) were rightly imputed by either inferring information from the individual's household or from obvious records about the individuals themselves.

In a case where the First Name of the individual is blank, it is difficult to fill that entry. In row '3052', a single female university student didn't indicate her First Name. Such error will be so difficult to input.

The blank entry in 'Age' was filled by considering the age of other male individuals' having the same occupation. The average age of these individuals was then used to fill the blank age.

The blank entries in "Relationship to Head of House" were decided and inputted based on the status of those in the household. For one to be 'Head' of a household, it is expected that he or she is 18years and above (Gov.uk). I observed two persons who declared to be "Head" of a household but under the age of "18". These persons' households were checked. It was observed that one of the respondents misstated her claim to be 'Head' of house seeing that she is below 18, while her husband is 19. Her husband's record was updated to 'Head' of the household.

The second respondent, an unemployed Single mother aged 16 possibly misunderstood the criteria for being a 'Head, and such errors is hard to fix, but in this case, it might be advised to drop this entry since it won't influence the data (NSPCC, 2022). A married female respondent's 'Relationship to head of house' blank entry was inputted as 'Wife' having identified the husband as 'Head' of house.

"Marital Status" null entries were considered and updated based on the legal age to be married, with or without the consent of parents. With parental consent, one can be allowed to marry from the age 16 but without the consent of the parents, the expected legal age is from 18 years and above (MyLawyer, 2022). The age of individuals with null entries is between 0 - 17. Individuals below 16 years of age were regarded as "Minors" and those above 16 years were casted as "Single".

Religion status had “Nope” and “NaN” entries. Most individuals whose religion status were null values, were mostly Minors below 18 years of age.

It is obvious that some persons did not declare their religion, and they indicated ‘Private’. Although such entry can be considered as undeclared with the possibility that individual has a religion but prefers not to disclose it. However, the entry was inferred from their household so as not to have too many unique religion entries.

Some entries in the ‘Infirmity’ variable were blank and they were all filled with None.

Inconsistencies:

There were inconsistencies in the unique entries of Marital Status and Gender; ‘W’, ‘S’, ‘D’, ‘M’, were replaced with ‘Widowed’, ‘Single’, ‘Divorced’ and ‘Married’ respectively in Marital Status while ‘M’, ‘m’, ‘f’, ‘F’, were updated with “Male” and “Female” in Gender.

In the Occupation entries, some individuals filled occupation titles such as ‘Sub’, ‘Copy’, ‘Land’, ‘Make’, ‘Best Boy’. It is either they didn’t complete the occupation titles or did not rightly name their occupation. I conditioned my dataset to show these individuals, considered their age and found out they were of employable age. Hence, I decided to reclassify them as being ‘Employed’. Also, there was an empty ‘Occupation’ response of a 24-year-old single male. This empty cell was filled with Unemployed given the fact that he is of employable age.

Furthermore, I filtered the occupation to verify if there were persons from age 65 and above who were unemployed. These persons I also considered as “Retired” persons with some assumptions that they might have started receiving their pension and they are no longer working (Gov.uk).

EXPLORATORY DATA ANALYSIS

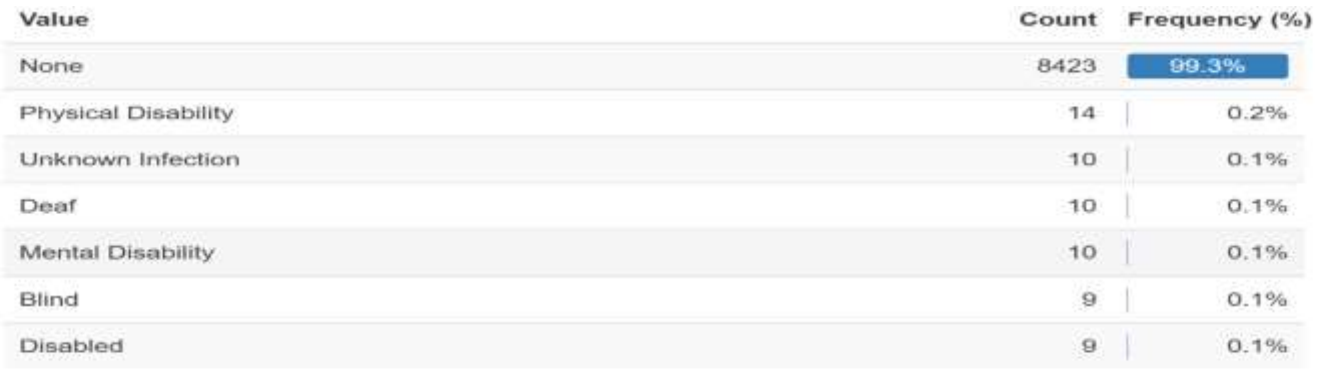
After the data cleaning process, further exploration was carried out to understand the relationship that exist between these variables and if there were other hidden insights not revealed during the data cleaning process.

To achieve this, the following were done.

1. Exploratory Data Analysis (pandas profiling)
2. Data Visualization

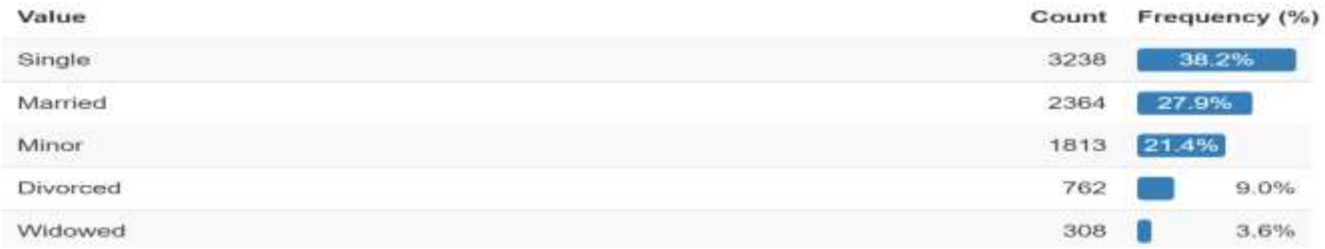
Descriptive Statistics

Infirmity



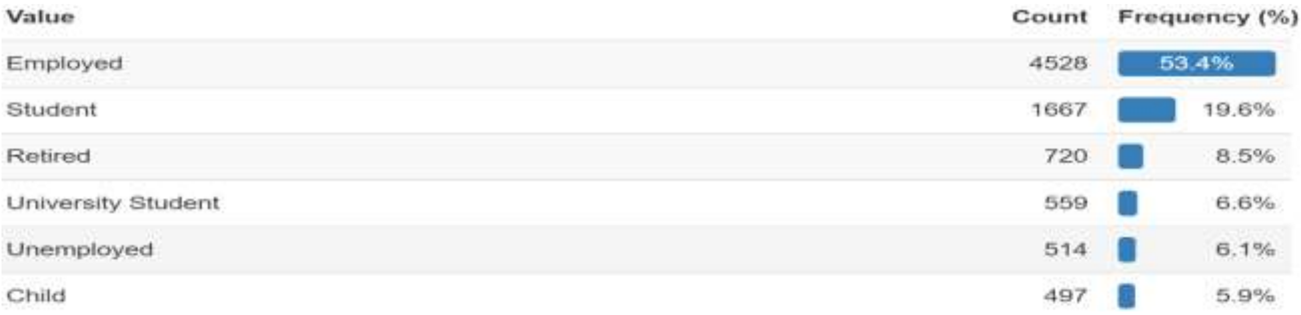
Value	Count	Frequency (%)
None	8423	99.3%
Physical Disability	14	0.2%
Unknown Infection	10	0.1%
Deaf	10	0.1%
Mental Disability	10	0.1%
Blind	9	0.1%
Disabled	9	0.1%

Marital Status



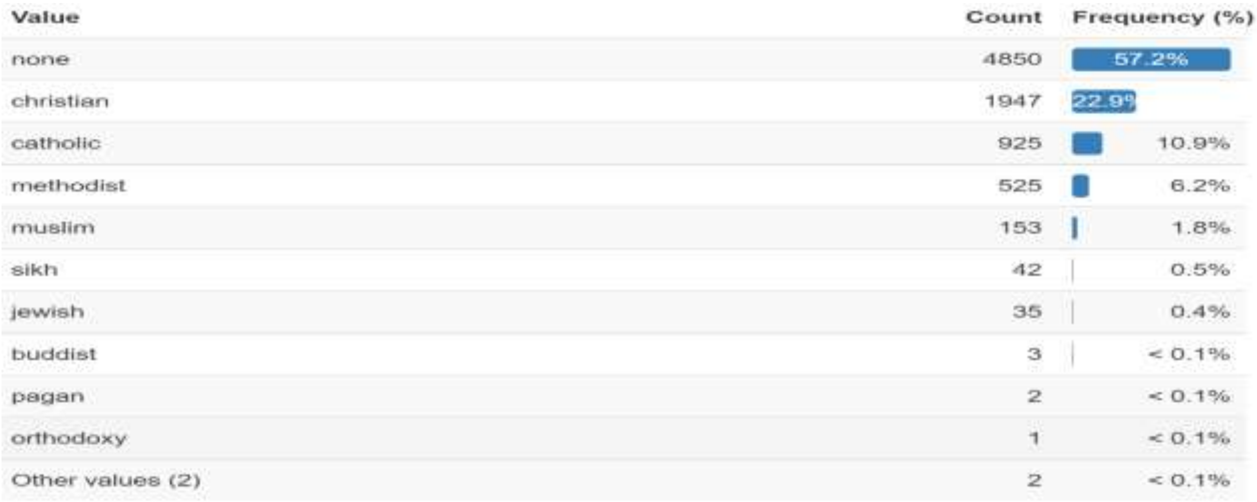
Value	Count	Frequency (%)
Single	3238	38.2%
Married	2364	27.9%
Minor	1813	21.4%
Divorced	762	9.0%
Widowed	308	3.6%

Occupations



Value	Count	Frequency (%)
Employed	4528	53.4%
Student	1667	19.6%
Retired	720	8.5%
University Student	559	6.6%
Unemployed	514	6.1%
Child	497	5.9%

Religion



Value	Count	Frequency (%)
none	4850	57.2%
christian	1947	22.9%
catholic	925	10.9%
methodist	525	6.2%
muslim	153	1.8%
sikh	42	0.5%
jewish	35	0.4%
buddist	3	< 0.1%
pagan	2	< 0.1%
orthodoxy	1	< 0.1%
Other values (2)	2	< 0.1%

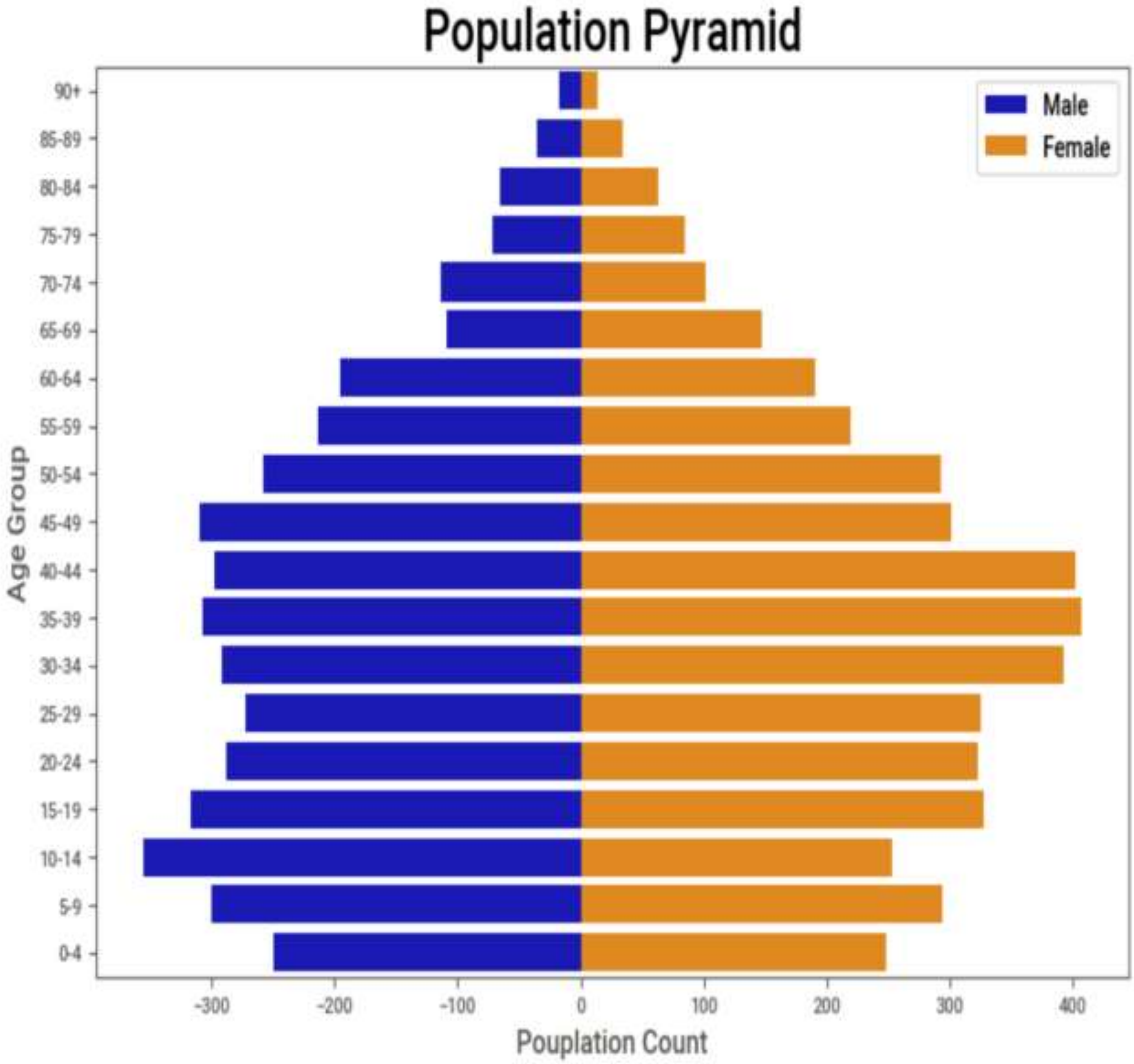


Figure 1.1: An Age Population Pyramid showing population Distribution

52.1% and 47.9% of the population are females and males respectively. This population pyramid shows a slightly low population of minors (0-4, 5-9) compared to young adults (15-39) and middle-aged persons (40-65).

The young adults (30-44) tend to have the highest population, mostly, females. This also shows high population of females in childbearing age from (ages 19 to 44). However, the number of the females in this childbearing age is not a major influence in the population with regards to the number of the recent births i.e., people aged between 0 – 4 years. It was observed that majority of these group of people are employed. It is probable that they are more committed to their careers (Forbes, 2022).

Birth and Death Rate

$$Crude\ Birth\ Rate = \frac{Number\ of\ births\ in\ 1\ year}{Total\ population} per\ thousand$$

$$Crude\ Death\ Rate = \frac{Number\ of\ deaths\ in\ 1\ year}{Total\ population} per\ thousand$$

As calculated in the Jupyter Notebook, the crude birth and death rates in a year is approximately 10 births and 3 deaths per thousand population. We can estimate that this town will experience 80 births and 24 deaths per 8,000 population in a year respectively. This shows a slight growth in the population.

Correlation Matrix

The correlation diagram below shows a strong relationship between “Age” and “Relationship to Head of House”, “Marital Status” and a slightly correlated with “Religion”. This means the “Marital Status” and “Relationship to Head of House” changes with increase in age, while Religion status changes fairly with increase in age.

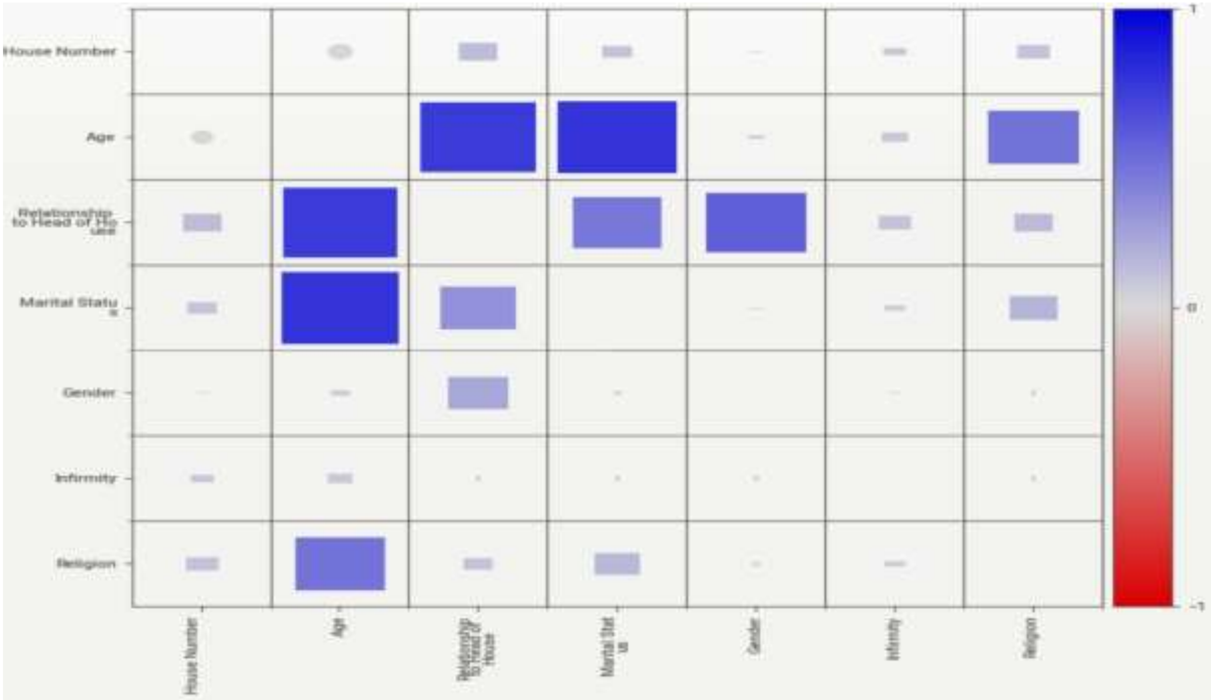


Figure 1.2: Correlation Matrix showing the Relationship between Variables.

Population Demography:

Occupation Groups:

The Occupations were grouped into six namely, ‘Student’, ‘University Student’, ‘Employed’, ‘Unemployed’, ‘Child’ and ‘Retired’. The University Student comprises of both University and PhD students. Retired group are all Retirees and those over 65 who specified Unemployed (ONS, 2016).

Age Class: Grouped age into 5-year class intervals for population pyramid.

Household: Conditioned based on ‘House Number’, and ‘Street Name’ and in some cases, included those who are married and Heads to show number of families.

Family Size: Conditioned on ‘House Number’, ‘Street Name’ and ‘Surname’

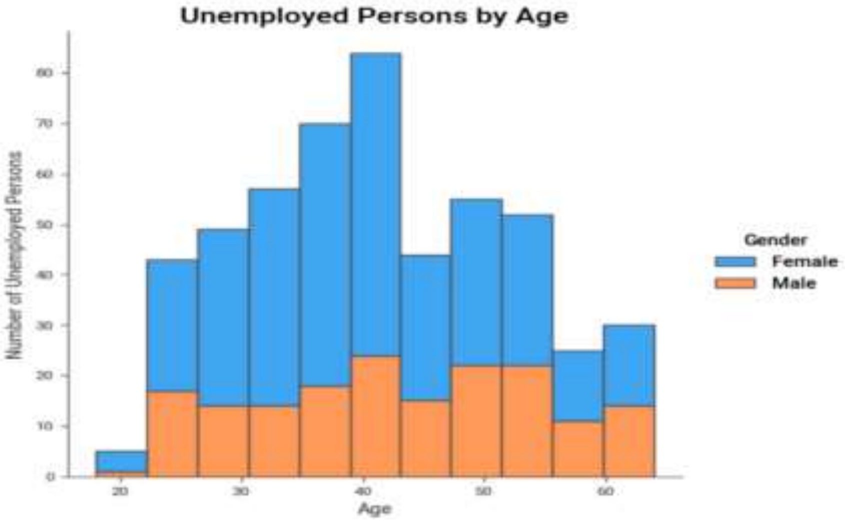


Figure 1.3: 6.1% of the population are Unemployed

Fig 1.3 above shows that there are greater percentage of unemployed females compared to males. These unemployed persons fall mostly between the age (30 - 45) which is an individual’s most active age.

Religious Affiliations and Infirmity

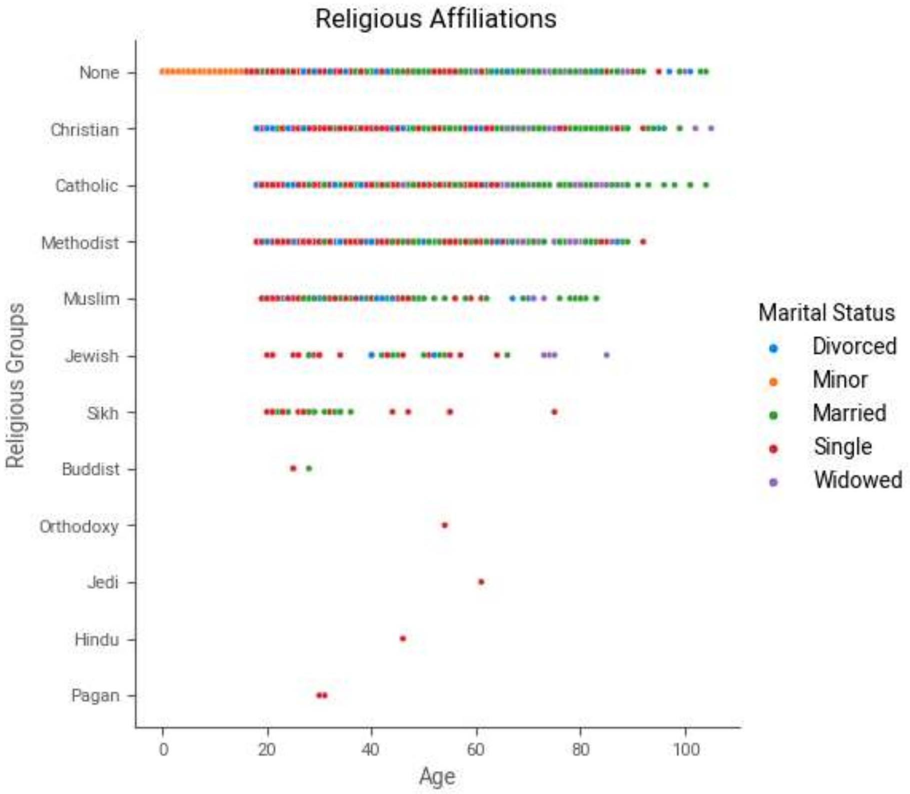


Figure 1.4: The Religious Affiliations showing Age distribution to their Marital Status

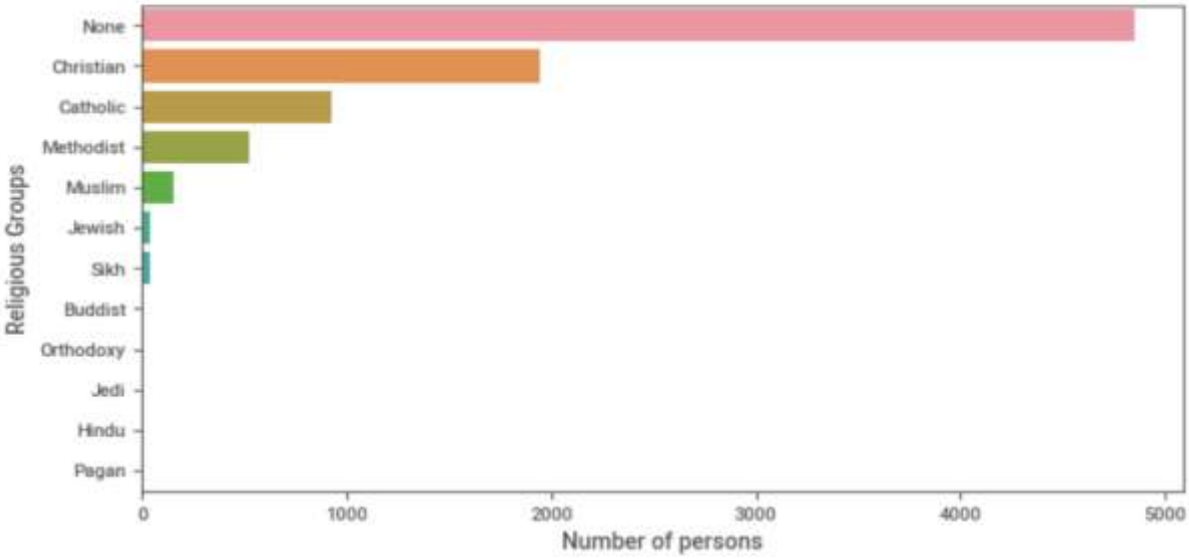
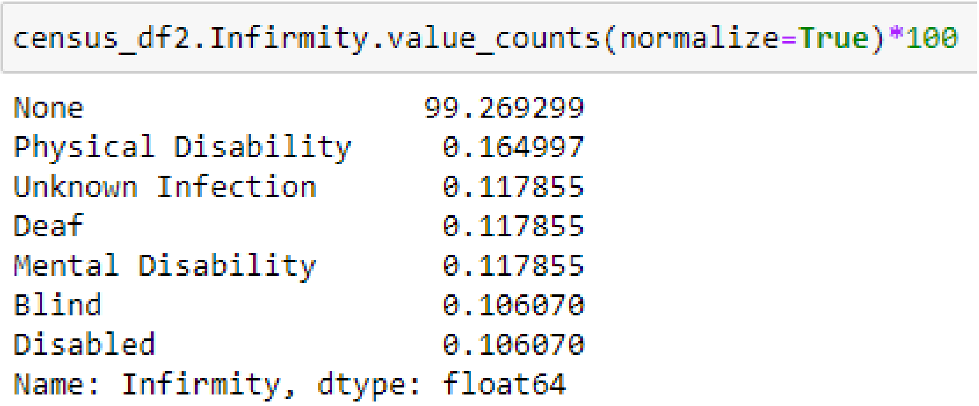


Figure 1.5: The population of Different Religious Groups

In figure 1.4 and 1.5, 43.8% of the population are religiously affiliated (which is less than half of the total percentage).The population with none or undeclared religious affiliations are about 57.2% of the total population.

The more affiliated religious groups, are highly correlated with young adults to old age.



```
census_df2.Infirmity.value_counts(normalize=True)*100
```

```
None                     99.269299
Physical Disability       0.164997
Unknown Infection         0.117855
Deaf                      0.117855
Mental Disability         0.117855
Blind                     0.106070
Disabled                  0.106070
Name: Infirmity, dtype: float64
```

Figure 1.6: The Infirmities as it affects the Town

During further exploration, the descriptive statistics revealed less than 1% of the entire population having infirmities and this mostly comprised of young and middle-aged population, and 99.3% of the population without infirmities across the age groups.

This shows a healthy population and clearly indicates that there may not be a need for a large medical facility in the nearest future.

Commuters

University Students (PhD students included) are considered as commuters together with all those who are employed and have occupations that requires them to travel to their place of work outside the city.

Here, it will be difficult to determine all the occupations that requires commuting due to the various occupations specified. But it is assumed that all those who are employed, will commute for one reason or another. From the calculation in the jupyter notebook, 75% out of this population was chosen to represent employed commuters.

With these considerations, 6.6%(a total of 559) of the entire population are University students and 40.2%(a total of 3,396 out of 4528 employed persons) are employed commuters such as Researchers, Journalists, Tourists, Armed forces and lots of others, which is more than half of the employed population consisting of the entire population. Therefore, 46% of the total population was estimated to commute (refer jupyter notebook for calculation).

$$Total\ No\ of\ Commuters = 75\%(employed\ persons) \quad + \quad University\ Students$$

$$\%Commuters = \frac{Total\ No\ of\ Commuters}{Total\ Population} \quad * \quad 100$$

Occupancy Rates

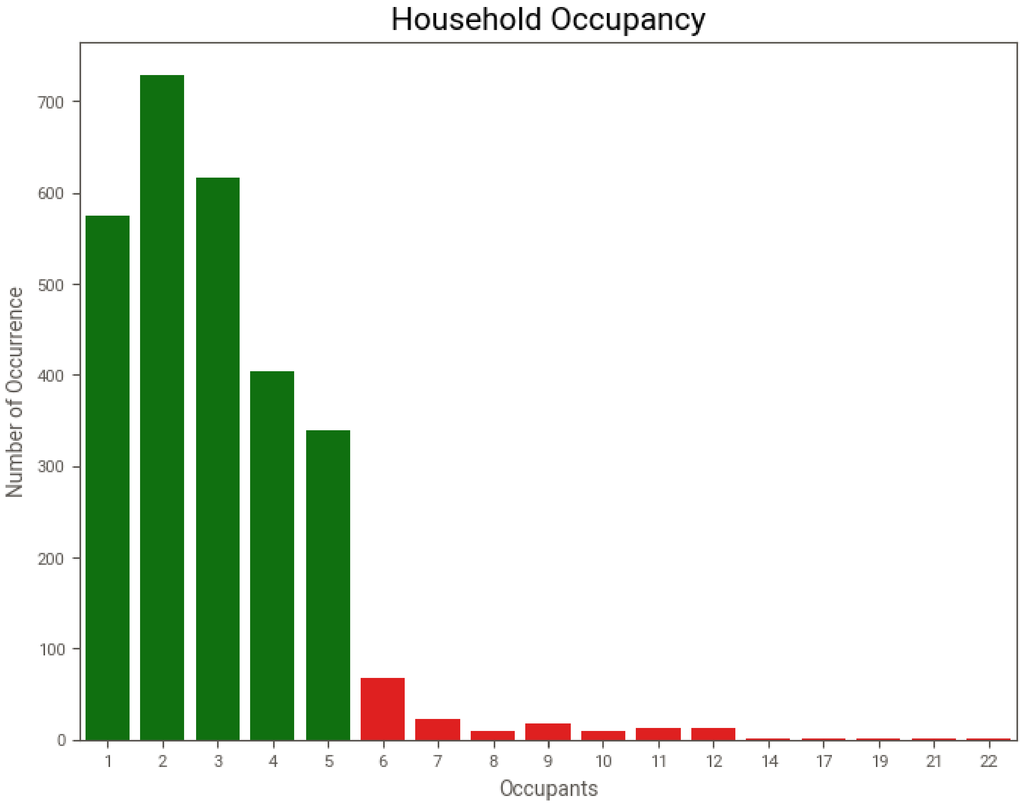


Figure 1.7: Showing Unique Occupancy per Household

About 66% of the houses have a maximum of three occupants, while 26% have 4 or 5 occupants showed in figure 1.7 above. Hence, just about 8% of the houses in the town are accommodated by more than five persons.

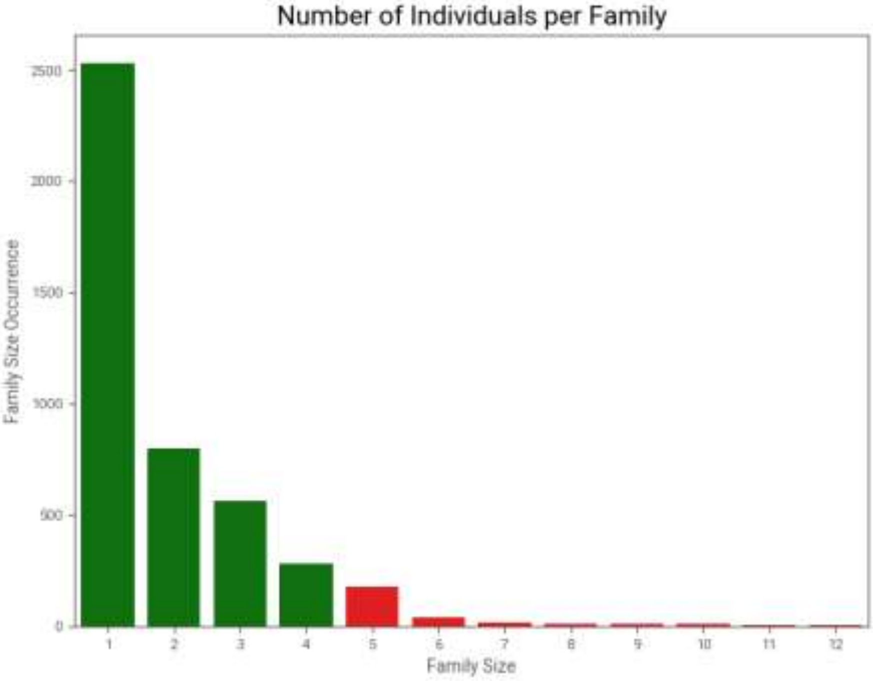


Figure 1.8: Showing Size of Family per Household

Additionally, the chart (figure 1.8) above revealed that 86% of the households(families) are made up of a maximum of three individuals. This suggests there is no need for large housing.

Migration

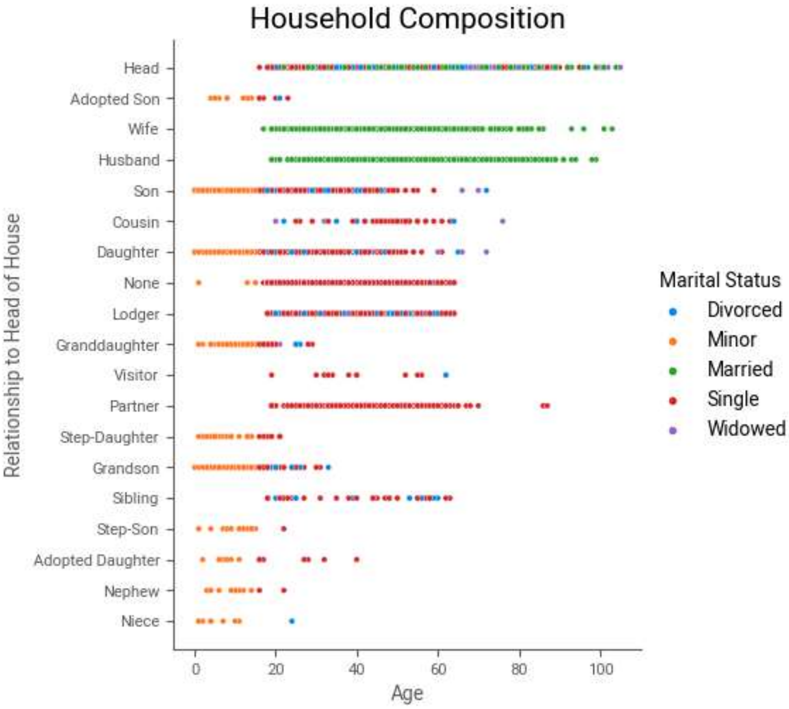


Figure 2.0: Household Composition in relation to Relationship to Head of House

The migrants in the town are the University Students, Lodgers and Visitors. Lodgers and Visitors as shown in the figure above, majorly constitute persons who are Single and these persons are probably eligible to commute, though there are a few of the Lodgers who are Divorced. There are 186 emigrants per 1,000 population in this town (refer Jupyter Notebook).

Divorced and Marriage Rate

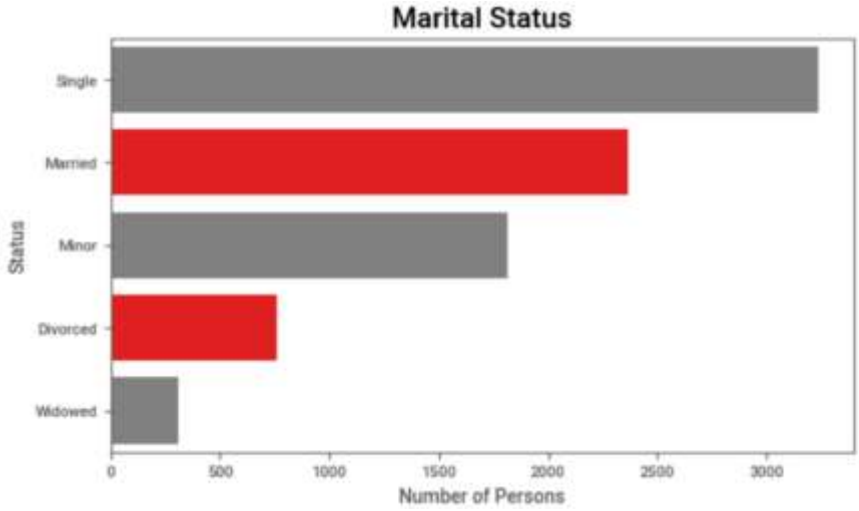


Figure 2.1: Showing the Marital Status with focus on Married and Divorced Persons

A total of 3238 persons, representing 38.2% of the population are singles, 27.9% and 9.0% are Married and Divorced respectively.

It was observed that the highest divorces were females, and it is most likely that male divorcees relocated after being divorced. The 'Age by Marital Status' plot below shows that divorce rate is more prevalent among persons of the middle age group.

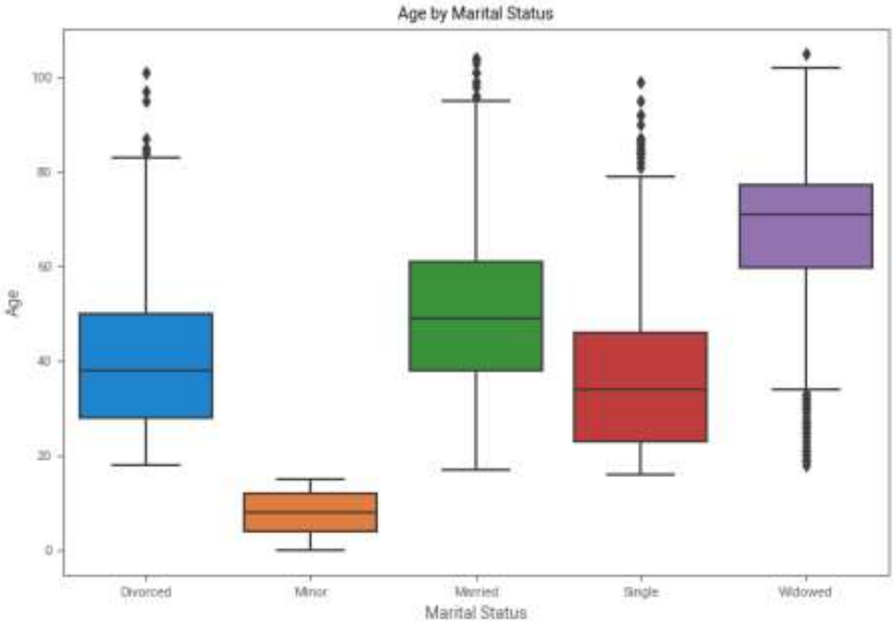


Figure2.2: Boxplot of Marital Status by Age

The Divorced and Marriage rate per thousand population was calculated in the Jupyter notebook (Eurostat, 2017).

$$\mathit{Marriage\ Crude\ Rate} = \left(\left(\frac{\mathit{Number\ of\ Married\ Persons}}{2}\right) \div \mathit{Total\ Population}\right) * 1000$$

$$\mathit{Divorced\ Crude\ Rate} = \frac{\mathit{Number\ of\ Female\ Divorcee}}{\mathit{Total\ Population}} \ {*1000}$$

The divorce to marriage ratio is 55 divorces per 139 marriages per thousand population.

Recommendations

Having established that a greater percentage of the population are commuters, there is a need for a train station in order to ease pressure on the roads. This will further lead to additional revenue for the local government as well as aid easy commuting for workers, lodgers, visitors and university students.

Employment training will benefit the unemployed persons most of whom are within the active age (30 – 45 years)

Given the large population of the ‘Retired’ and ‘Employed’ individuals who are either receiving their pension or earning a living, we could probably say that the population is more of middle class. In addition, the population is not expanding significantly, single units, single-family homes or households with small number of occupants constitutes a larger part of the population. Hence, there is no need for investment in housing.

Less than 1% of the entire population recorded any form of infirmity. This with no expected increase in birth means that there is no pressing need for a medical facility.

The local government should consider these future investments: general infrastructure as that may arise from the construction of the train station. The station is expected to also bring about increase in the number of people in the community. A care home should be considered as the population as the middle-age class transitions to old age.

References

Eurostat (2017). *Marriage and Divorced Statistics.*
Available Online: Marriage and divorce statistics - Statistics Explained (europa.eu) [Accessed 06/12/2022]

Forbes (2022). Why are Young Women More Career Driven than Young Men?
Available Online: Why Are Young Women More Career Driven Than Young Men? (forbes.com) [Accessed: 08/12/2022].

Gov.uk, *Universal Credit Eligibility*.
Available Online: Universal Credit: Eligibility - GOV.UK (www.gov.uk) [Accessed 06/12/2022]

My Lawyer (2022). *The Law on getting Married*.
Available Online: The law on getting married | MyLawyer [Accessed 06/12/2022]

NSPCC (2022). *Moving Out*.
Available Online: Moving out | NSPCC [Accessed 06/12/2022]

Office for National Statistics (2016). *Standard Occupation Classification*.
Available Online: SOC 2010 - Office for National Statistics (ons.gov.uk) [Accessed 06/12/2022]

Wikipedia, The free Encyclopaedia (2022). *Population Pyramid.*
Available Online: Population pyramid - Wikipedia [Accessed 06/12/2022]