PROJECT REPORT: CENSUS

INTRODUCTION

This report is an analysis of a census of a moderately sized town of about eight thousand and it presents recommendations to be considered by city council on investment decisions and development plans for an unused plot of land. To aid the proposal, the data analysis process was undertaken; from data cleaning to exploratory data analysis (EDA), however, no model was built.

It also features further analyses used to buttress the yielded exhortation. This encompasses the town's population/age demography, population growth, percentage of commuters in the census, occupancy rates and occupation demography which are discussed below.

DATA CLEANING

Data collected from the census was cleaned by making corrections to the errors observed. These are demonstrated in the accompanying Jupyter Notebook file.

The data contained null values in the Marital Status and Religion columns as well as blank entries across the dataset in the following columns:

1. Age
2. Relationship to Head of House
3. Marital Status
4. Gender
5. Occupation
6. Infirmity
7. Religion

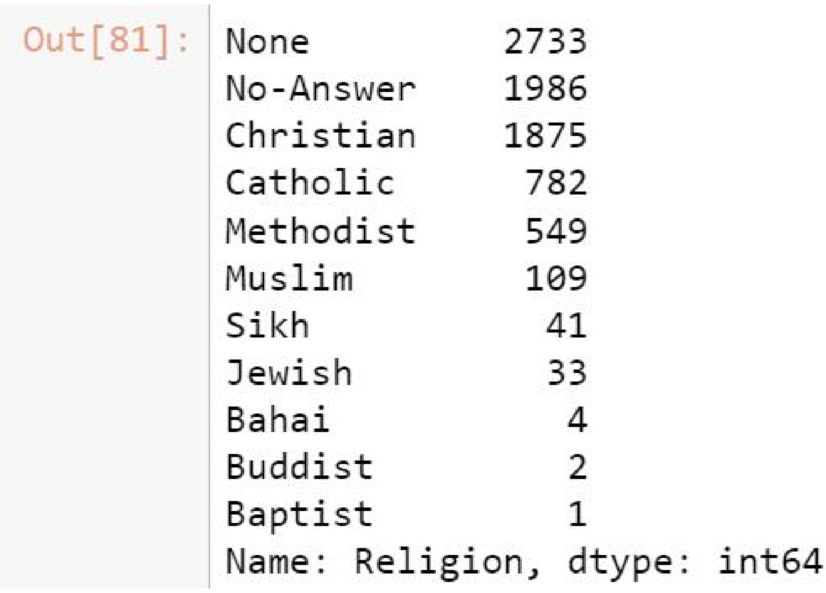
These blank spaces and null values were changed to "No-Answer" to preserve the integrity of the data such that data is recorded exactly as intended in order to prevent data misconceptions (Wikipedia, 2022: https://en.wikipedia.org/wiki/Data_integrity). For Religion, those who answered "Nope" was replaced with "None" as it is inferred they meant they do not belong to a religious group. However, the Age column with such blank and null entries were replaced with the most occurring age, which is "39" in this dataset, to infer with proper and near accurate calculations relating to age.

Some information was not properly entered, thereby not conforming to the uniformity of the dataset. For example, an Age entry was in words and some entries in Marital Status and Gender were in short forms such as "D", "F" and so on. This was properly handled and carefully replaced with their appropriate full forms.

Underaged "Head" of houses were discovered and there were two (2) of them. To proffer a solution, a new column was created called "Is-Head" and the status of their *'headship'* was set to *'No'*. Minors (Under 18) were considered with respect to their Marital Status and was it was set to "NA" as they are not adults and cannot have a Marital Status (Gov.uk). The same was done for children under 16 with religious affiliation. An exception has been made for those who are under 18 and have a Marital Status.

One household contained a ‘Husband’ over 18, a ‘Head’ whose ‘Female’ aged 15 and a baby of age 0 (Zero). This is an illegal household according to the Marriage and Civil Partnership (Minimum Age) Act 2022 (legislation.gov.uk, 2022) and hence, was dropped as it was not deemed significant enough to affect the overall analysis of the dataset in a negative way.

Outliers were kept in the dataset as they were too few to influence the data. This was noticed in one record where the census has only one (1) Baptist in the population. This is either untrue or the individual is an immigrant. Further investigation showed that this individual has no relation to the head of the house and the rest of household. This indicates that every member of the house is a tenant. This is shown in the figures below:



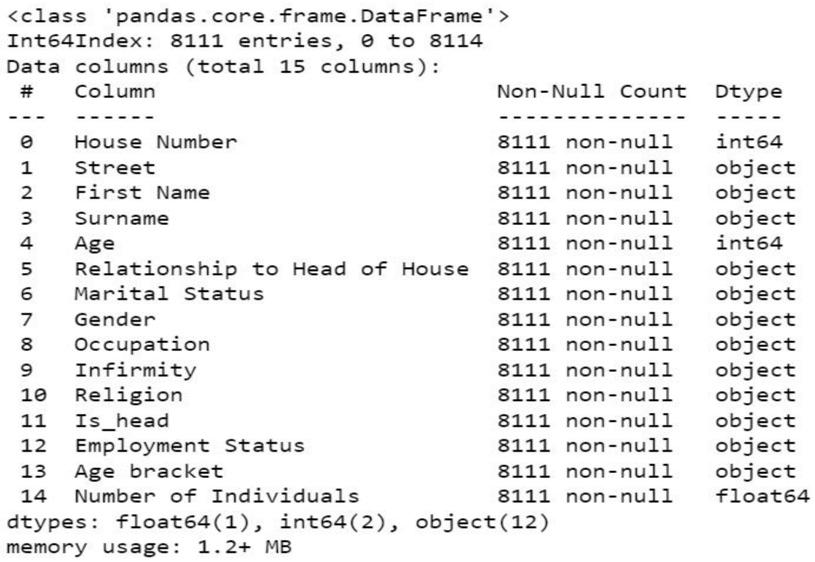
```
Out[81]: None          2733
         No-Answer     1986
         Christian     1875
         Catholic       782
         Methodist      549
         Muslim         109
         Sikh            41
         Jewish          33
         Bahai            4
         Buddist          2
         Baptist          1
         Name: Religion, dtype: int64
```

Fig 1: Religion Statistics

Data visualization of the “Occupation” of the census was difficult due to the high number of occupation types. The solution to this approach was to categorise the occupations of the populace. These classifications are ‘Employed’, ‘Unemployed’, ‘Student(child)’, ‘University Student’, ‘Child’ and ‘Retired’ and were stored in the new column; “Employment Status”. This way a proper demography can be visualized.

ANALYSIS

After the data cleaning, the census data had these features in the figure 3 below:



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8111 entries, 0 to 8114
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House Number                  8111 non-null   int64
 1   Street                        8111 non-null   object
 2   First Name                    8111 non-null   object
 3   Surname                       8111 non-null   object
 4   Age                           8111 non-null   int64
 5   Relationship to Head of House  8111 non-null   object
 6   Marital Status                8111 non-null   object
 7   Gender                        8111 non-null   object
 8   Occupation                    8111 non-null   object
 9   Infirmity                     8111 non-null   object
 10  Religion                      8111 non-null   object
 11  Is_head                       8111 non-null   object
 12  Employment Status             8111 non-null   object
 13  Age bracket                   8111 non-null   object
 14  Number of Individuals         8111 non-null   float64
dtypes: float64(1), int64(2), object(12)
memory usage: 1.2+ MB
```

Fig 2: Dataframe Attributes Info

New columns were added in order to aid the visualization of the data like in the example of the "Occupation". The new columns are:

- **Is_head:** Fields where the head of the household is underaged is set to 'No' and where it is true that the individual is an adult, 'Yes', whereas other relationships are set to 'Other'.
- **Employment Status:** Occupations in a trimmed category set of which the values are easier to plot.
- **Age bracket:** Ages placed in 10-year groups classified as a generation, used in plotting the population pyramid.

Further discussion on the Analysis in given in the pages to follow.

The Population

Survey of the census using the population pyramid shows the basic statistics of the population, revealing what suggests a low birth rate and higher number of young males. The females in their middle-ages are more in number than that of the young female population. The population seems to be doing well into their eighties (80's) for both male and female.

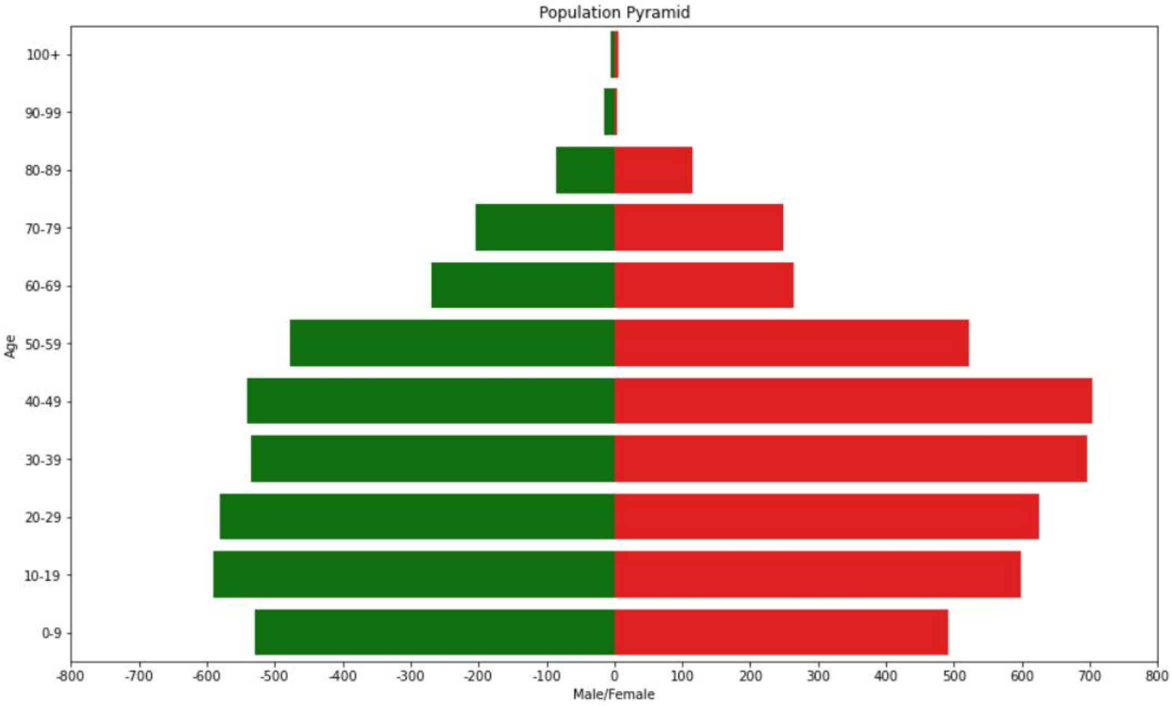


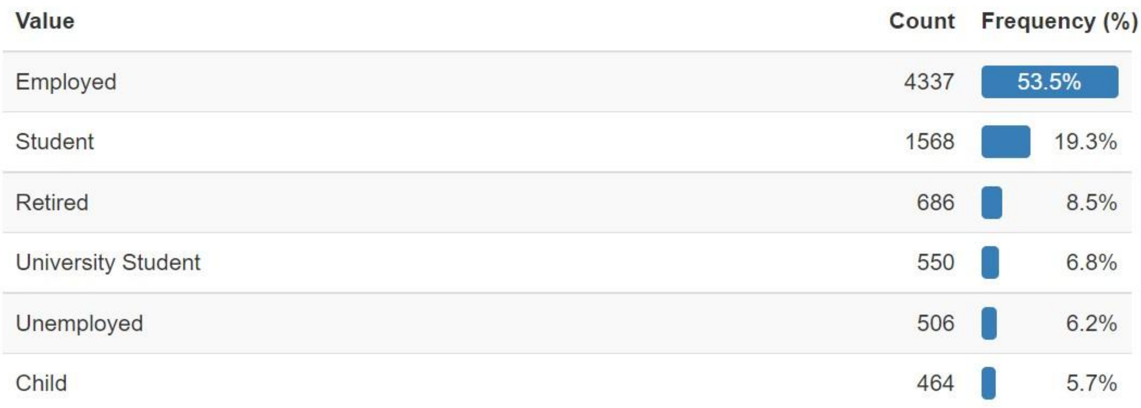
Fig 3: Population Pyramid

Employment Status

More analysis indicates that most of the census population are employed, followed by school children and retired population. Minority of the population are unemployed and babies. However, 6.2% of the total population are unemployed which according ONS (2020) indicates a high unemployment rate. The working age of the population makes up the majority of the unemployed populace.

The University students constitutes 6.8% of the population whereas those who are retired make up 8.5% of the populace. These can be used to determine the commuters as will be discussed further.

The limitation of the dataset is that the affluence of the community is not determined as the working hours or salary were not given.



Value	Count	Frequency (%)
Employed	4337	53.5%
Student	1568	19.3%
Retired	686	8.5%
University Student	550	6.8%
Unemployed	506	6.2%
Child	464	5.7%

Fig 4: Pandas Profile of Employment Status

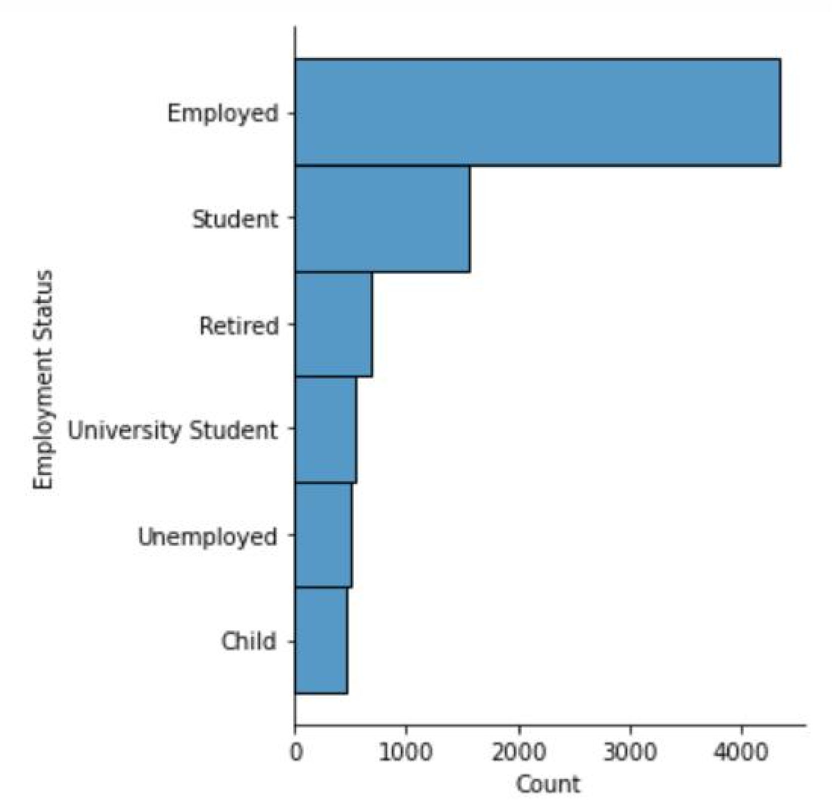


Fig 5: Employment Status Plot

Commuters

To determine the commuters, the University students and the upper percentile of the employed were used to calculate the percentage of the population that will commute.

From such inferences, it was determined that 46% of the populace will commute. This is significant as it is nearly half the census. The use of the upper percentile of the employed population was due to assumption that majority of those employed will commute, even if it is for shopping and it is assumed that there are no major malls in the small town.

Marital Status

The census shows a single population with a high number of married people as shown in the figure 6 below:

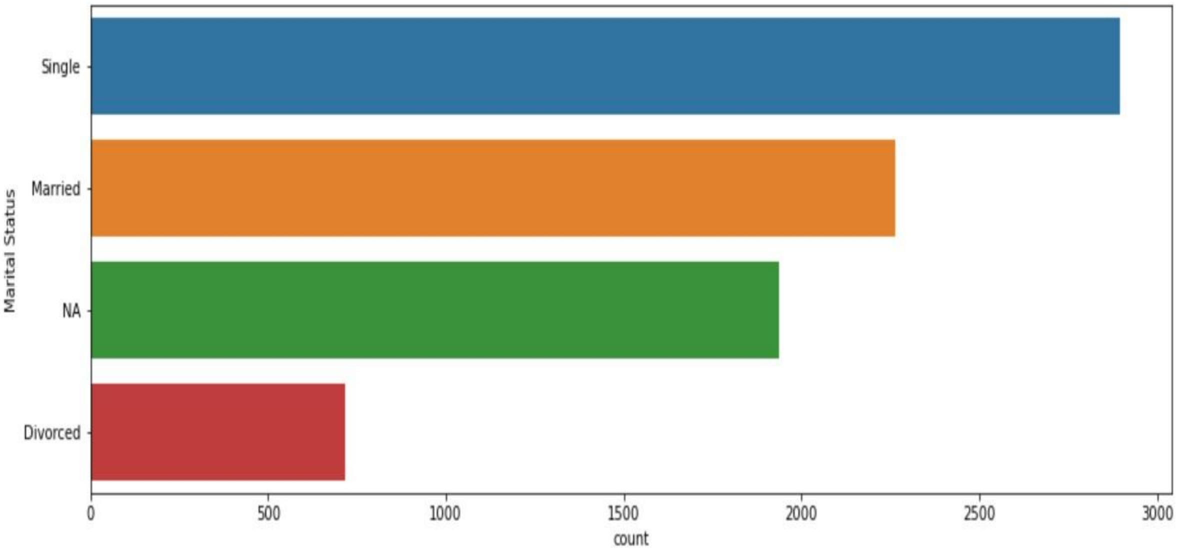


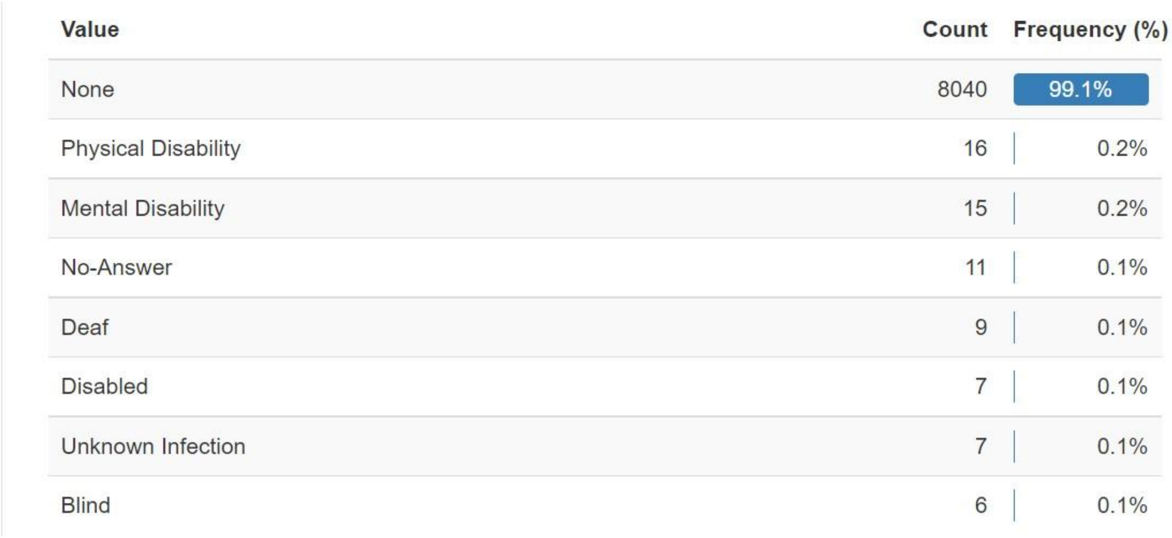
Fig 6: Marital Status Plot

Infirmity

The population is very healthy with less than 1% of the population with disabilities. This makes the population with disabilities insignificant. Those who are blind are between the ages 20 to 80 and are majorly male. The dataset also includes disabled members of the population who are only males between the ages 4 to 65 and the physically disabled also mostly masculine and within similar age range as the disabled. The deaf populace is significantly interesting, the middle-ages having good hearing with only one old man with deafness. The young population with this infirmity seems to either be born with it or had an accident causing their deafness.

Unknown infections seem to be more common among females. It affects the young population and those between the ages of 45 to 65. However, it is noticeable that there is a gap in this demographic from the ages 25 to 45. This may seem to indicate that the infection does not have effect on them due to immunity at such age, or they are rather more cautious.

The group with ‘No-Answer’ are sparse and the age bands are inconsistent.



Value	Count	Frequency (%)
None	8040	99.1%
Physical Disability	16	0.2%
Mental Disability	15	0.2%
No-Answer	11	0.1%
Deaf	9	0.1%
Disabled	7	0.1%
Unknown Infection	7	0.1%
Blind	6	0.1%

Fig 7: Infirmity Distribution

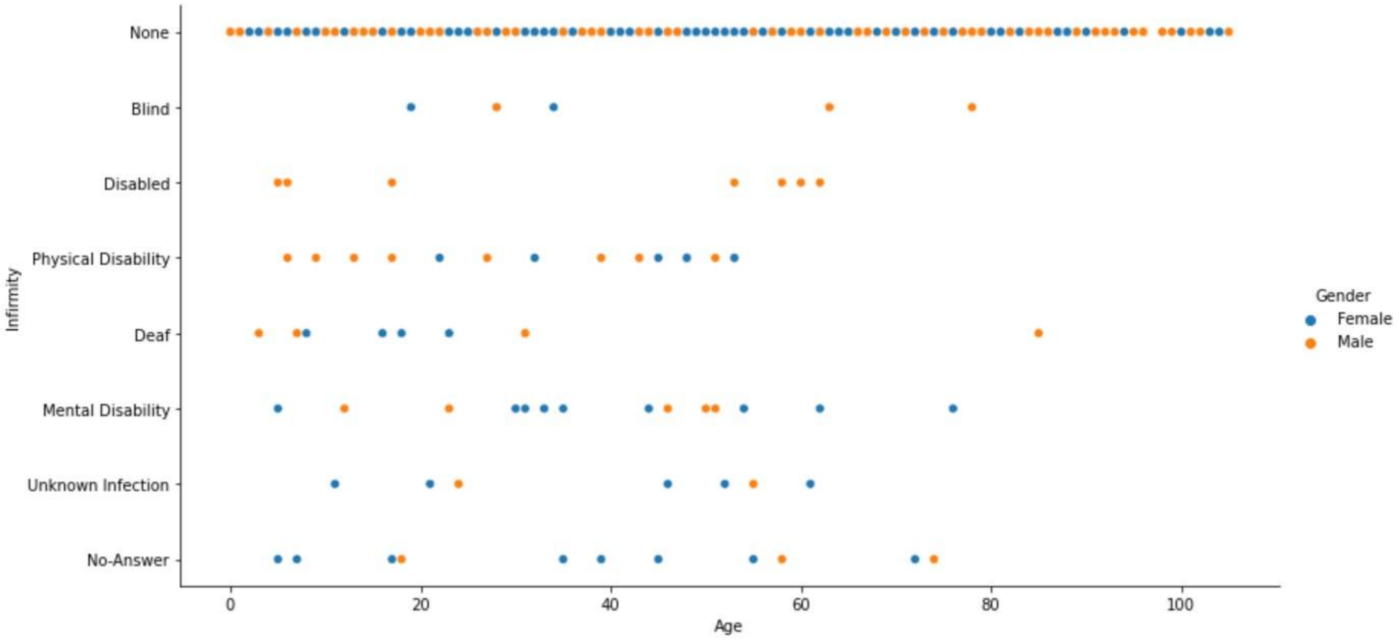


Fig 8: Infirmity Against Age Hued by Gender Plot

Religion

As shown in the figure 9 below, the census is majorly unaffiliated to a religious group. Those with religious affiliations are less than half of the census, with Christianity taking the lead as the highest and having a wide age distribution. It is uncertain at this point to determine which religion is growing, but seeing as the second highest occurring religion is Christianity, it is most certain that the parents of the children would play a major role in influencing their beliefs.

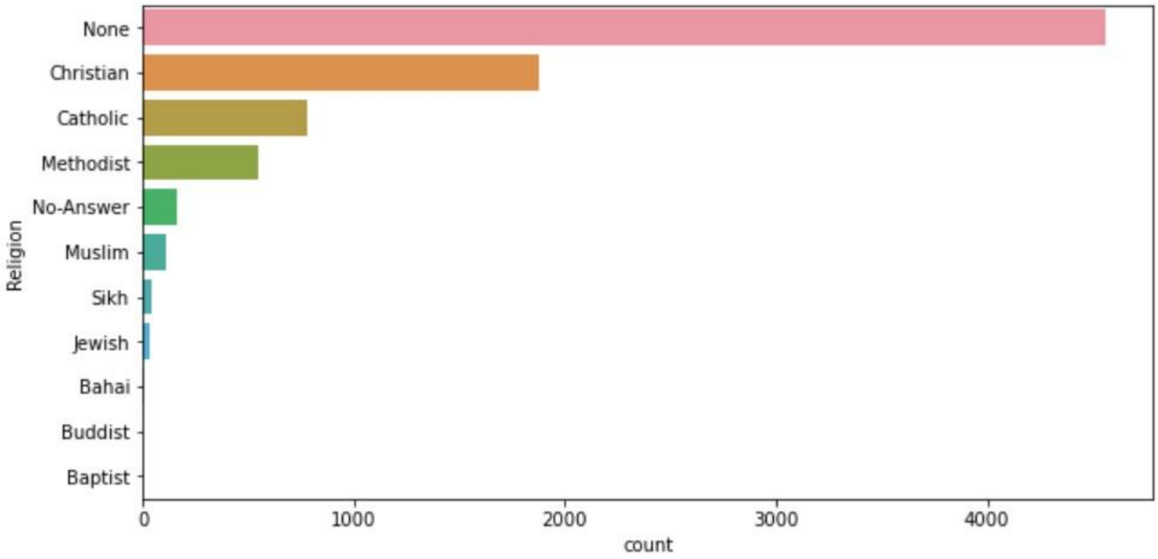


Fig 9: Religion Count

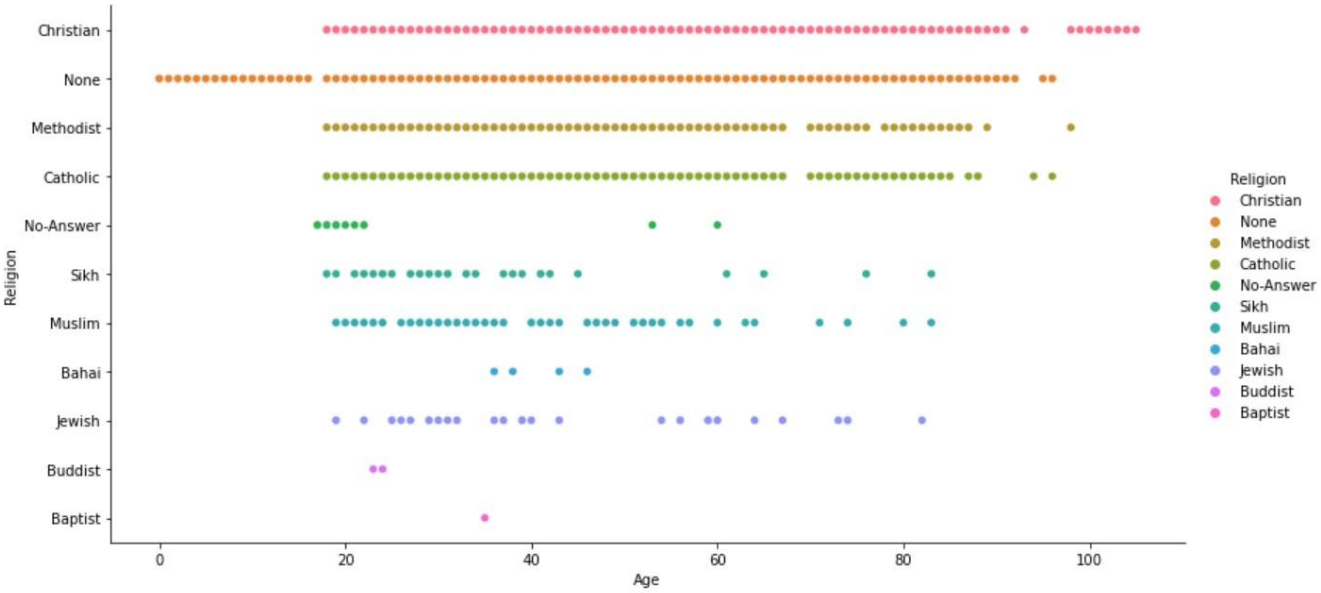


Fig 10: Religion Against Age Plot

Married and Divorce

The dataset indicates that the individuals that are divorced range from an underaged age up until old age and the demography has more divorced females than males. This may go to suggest that the difference in the male vs female divorced count makes up part of the number of emigrants from the census and they happen to be male. Underaged married people are perhaps consented by their parents, although they are very few; the dataset indicates that the populace has a high number of marriage rate as compared to the divorce rate.

The divorce and marriage rates are calculated by the finding half the count of divorcees and married couples per thousand in the population. Therefore, the ratio of the divorce rate to marriage rate in a round figure is 44 to 139.

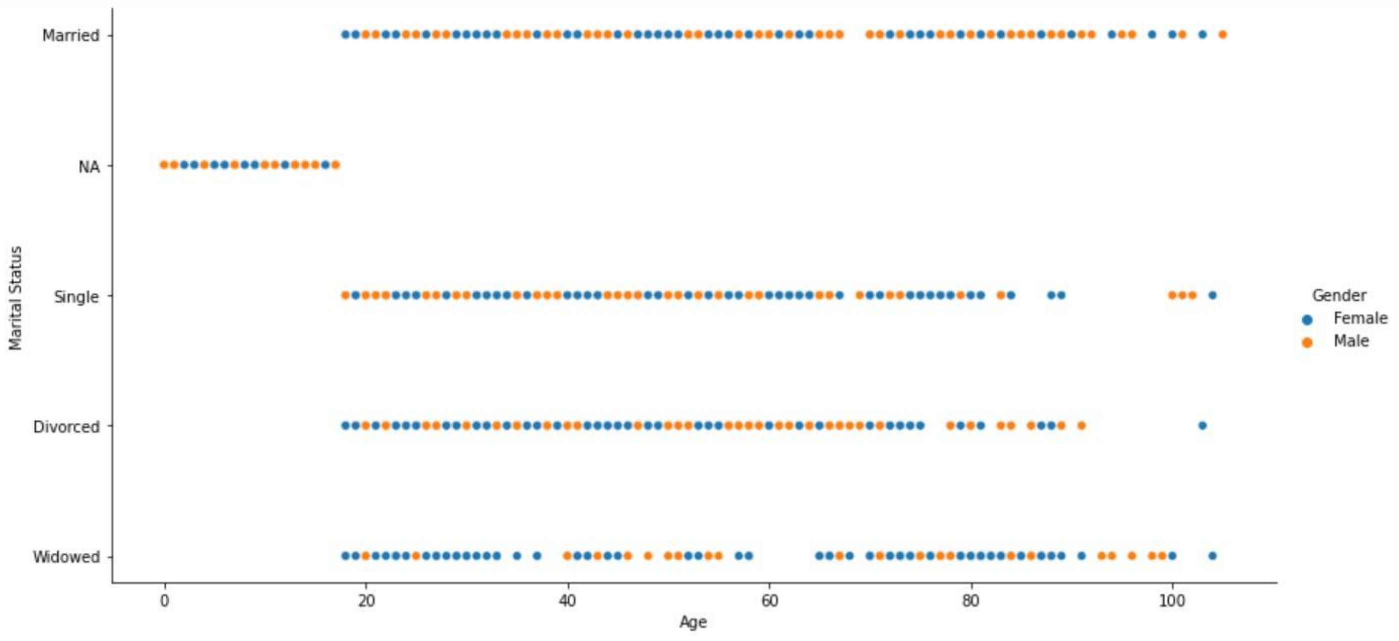


Fig 11: Marital Status Against Age Hued by Gender Plot

Migration

The migration demographics is determined majorly by the 'University Students' and the 'Visitors'. The 'Lodgers' are more difficult to determine because members of the adult indigent population may or may not have moved out of their parents or family house to be independent.

The students should be considered a constant in the community as it seems to suggest cheaper housing close to the University of their study, which would also signify cheap commute to school. The 'University Students' constitute of a young populace of the age bracket 18 to 35. They also include 'PhD Students' which is why the age bracket extends into the mid 30's. But this not sufficient to calculate the net migration statistics because the students leave after studies and their vacant addresses are offered up again for rent.

The inferred method for determining the gross migration is gotten by calculating the count of the lodgers and visitors to get the assumed immigration and subtracting the emigration from it. The emigration is assumed to be the difference in male and female divorcees. By this, the net migration comes a value of 218 in the whole population and migration rate of 27 per thousand.

Therefore, there are approximately 47 immigrants per thousand in the town and the emigrants are 20 per thousand.

Occupancy Rate

The maximum occupancy for any household is 22. Although it is not easily known what type of housing they occupy, it could be assumed that it is over-occupied. Extremely high occupancy could be an attribute of an affluent family living together or students living in block of flats, but these are meagre compared to the other occupancy levels. This however shows the overall household size in the population which seems to suggest a need for high-density housing.

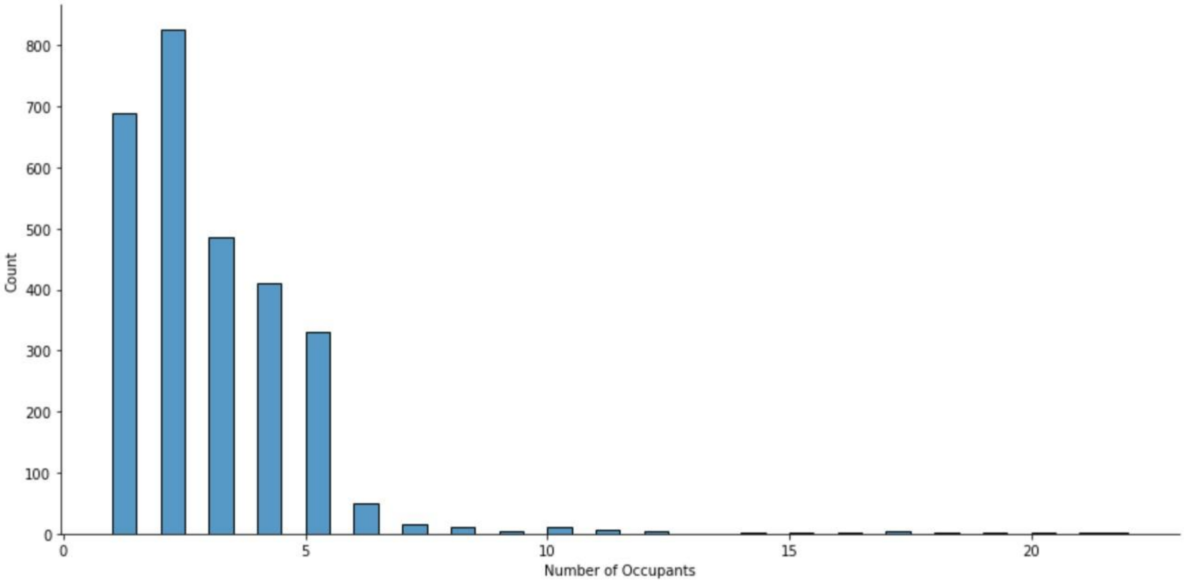


Fig 12: Number of Occupants Count

Birth and Death Rate

From the calculation undertaken for the birth and death rates, it shows a higher birth to death ratio. However, from the population pyramid, it indicates a decline in the population growth. The current birth rate at 11 per thousand as opposed to the death rate at 1.4 per thousand seem to indicate a healthy community. The number of children of age 1 are less than those of age zero (0), but the number of children from the youngest generation are less than the number of young adults which portrays a decline in population according to the population pyramid. This disparity can only be explained with an assumption that the decline/inconsistency is probably be due to child mortality rate at birth, reduce fertility of the population or the high number of the single population.

$$\frac{births}{population} \times 1000 = birth\ rate$$

$$\frac{deaths}{population} \times 1000 = death\ rate$$

source: https://www.watt-watchers.com/activity/population-math/

The number of births was calculated by counting the number of babies in the census while the deaths was gotten from the population difference across every generation.

Hence, the population growth derived from the difference in birth and death rate is 9.6 per thousand. This describes a slow population growth which translates to a decline in the number of the population.

RECOMMENDATIONS

As indicative of the analysis, the occurrence of the number of occupants per household suggests high-density housing as the population seems to stagnate. The dataset does not provide the analysis with a way to determine the true affluence of the town but, concluding from the assumption of the number of occurrence of occupants per household, the demography shows only a handful of the population are over-occupied or have a large family and need low density housing. Therefore, it is recommended that high-density housing be built. The altering variable will be the number of lodgers, visitors and university students who would immigrate and contribute to the increase in population.

The percentage of the population that would be commuting is about half the town. This is indicative of a need for a train station, thereby taking pressure off the roads and reducing the mortality rate of the population due to road accidents.

With the census depicting a potential growth in ‘Christianity’, the need for another religious building for the Christians is advisable. It is probable that the Christian parents of the children would influence their children to their beliefs.

The census seems healthy with most of them not having infirmities. Although this does not provide the same guarantee as the population ages, it may come of need to invest in an old people’s care home in 20 to 30 years’ time seeing that majority of the current census are in their middle-ages and may constitute a large portion of the future population should the mortality rate remain very low for their generation. Nevertheless, it is imperative to invest in a minor clinic to cater for the minority of the population with infirmities, especially those with unknown infections in order to curb further potential spread.

According to ONS (2020), the population is plagued by high unemployment. This suggests investing in training individuals for new skills in order that they re-join the work force. Since there is no indication of a growing population of school-aged children, the need for increased spending for schooling could be revisited in the future.

Other general infrastructure such as waste collection, broadband, open space and sporting facilities do not require more investment since the population growth seems stagnant due to the relatively small yearly growth.

REFERENCE

Gov.uk (2022) *Implementation of the Marriage and Civil Partnership (Minimum Age) Act 2022*
Available online: https://www.gov.uk/government/news/implementation-of-the-marriage-and-civil-partnership-minimum-age-act-2022 [Accessed 04/12/2022]

Legislation.gov.uk (2022) *Marriage And Civil Partnership (Minimum Age) Act 2022*
Available online: https://www.legislation.gov.uk/ukpga/2022/28/notes/division/6/index.htm [Accessed 05/12/2022]

Office for National Statistics (2020) *Labour market in the regions of the UK: July 2020*
Available online: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/regionallabourmarket/july2020#:~:text=Local%20labour%20market%20indicators,-Indicators%20from%20the&text=For%20the%20period%20April%202019,Middlesbrough%2C%20both%20at%206.9%25 [Accessed 05/12/2022]

Watt Watchers of Texas (2022) *Activity: Population Math*
Available online: https://www.watt-watchers.com/activity/population-math/ [Accessed 05/12/2022]

Wikipedia (2022) *Data Integrity*
Available online: https://en.wikipedia.org/wiki/Data_integrity [Accessed 04/12/2022]