CENSUS PROJECT REPORT

INTRODUCTION

This Mock Census Report is based on an imaginary modest town between two cities, in which I am part of the local government team, who will be making decisions on what to do and what to invest in with an unoccupied plot of land.

The purpose of this census is to compare different people across the nation to provide the government with accurate statistics of the population to enable better planning, develop policies, and allocate certain funding.

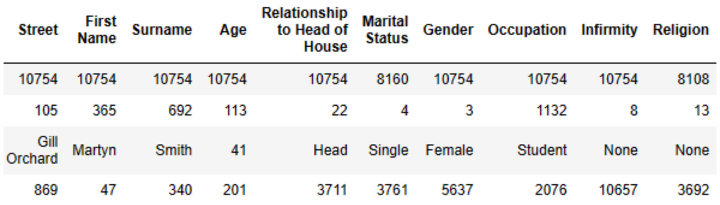
DATA CLEANING

From the mocked sample census data (CSV file) given, it is evident that there were errors in the data collated during the census. These errors could be typographical errors, transcription errors, or simply because people decide not to speak the truth. Hence, the census data will be cleaned, and it would involve.

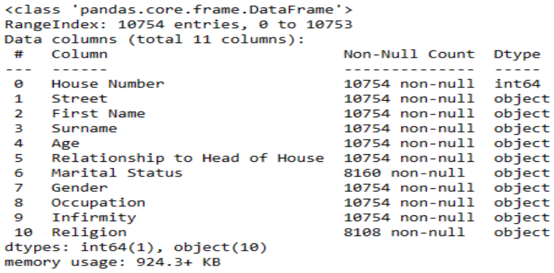
i. exploring each row and column
ii. replacing blank cells with the most appropriate and reasonable answer
iii. validating the census data to detect outright lies.
iv. for duplicate data and removing it
v. casting to appropriate data type where necessary.

Data cleaning was employed to correct the errors, then visualization and analysis were performed on the census data to give the best recommendations.

DATA DESCRIPTION



Street	First Name	Surname	Age	Relationship to Head of House	Marital Status	Gender	Occupation	Infirmity	Religion
10754	10754	10754	10754	10754	8160	10754	10754	10754	8108
105	365	692	113	22	4	3	1132	8	13
Gill Orchard	Martyn	Smith	41	Head	Single	Female	Student	None	None
869	47	340	201	3711	3761	5637	2076	10657	3692



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10754 entries, 0 to 10753
Data columns (total 11 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   House Number                   10754 non-null  int64
 1   Street                         10754 non-null  object
 2   First Name                     10754 non-null  object
 3   Surname                        10754 non-null  object
 4   Age                            10754 non-null  object
 5   Relationship to Head of House  10754 non-null  object
 6   Marital Status                 8160 non-null   object
 7   Gender                         10754 non-null  object
 8   Occupation                     10754 non-null  object
 9   Infirmity                      10754 non-null  object
 10  Religion                       8108 non-null   object
dtypes: int64(1), object(10)
memory usage: 924.3+ KB
```

From the above table, we can see that we have 10,754 data entries with attributes Street, First Name, Surname, Age, Relationship to Head of House, Marital Status, Gender, Occupation, Infirmity, and Religion.

DATA CLEANING

Jupyter Notebook with Pandas library was used in cleaning the data. The cleaning was performed across all rows and columns.

There was a record with a blank value in the First Name column. The data was replaced with the surname of the missing First Name due to the different names of people living in the house.

There was another record of blank value in the Surname column of a 17years old daughter, so an appropriate surname was given based on the people in the house.

There was a missing record in the age column for the Head of the house. The data were replaced with the average value of all Heads of the house.

The gender column has three cells with blank records, and appropriate gender (Male, Male, and Female) were given due to their Relationship to the head of the House (Son, Son, and Daughter).

There were N/A records for the Marital Status column for minors who are aged 17 and below. Validation was done against the Marital Status column to confirm if no one below the age of 17 has Married, Divorced, or Widowed as Marital Status.

The infirmity column has 12 blank cells with N/A, and these values were replaced with None because it is the Infirmity with the highest record.

Two cells had Sith values as a religion. This was updated to None because of the validity of the Sith religion.

DATA VISUALIZATION

Figure 2 below shows the frequency distribution of the population. The data was grouped with a class interval of 5 and a histogram was used to show the frequency distribution.

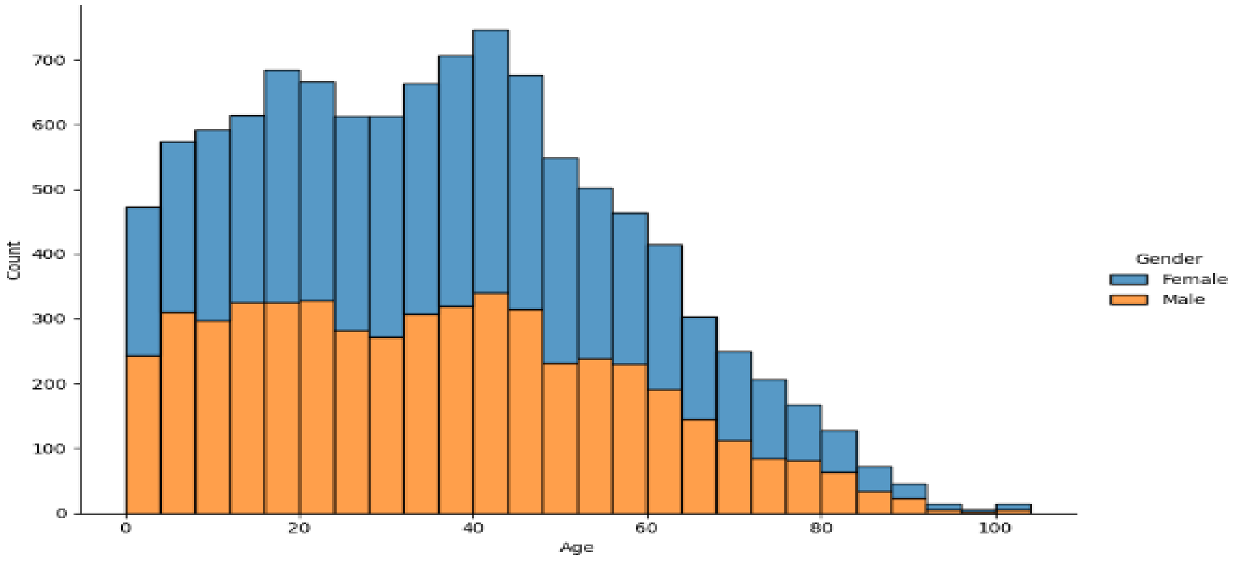


Fig 2: Population Histogram

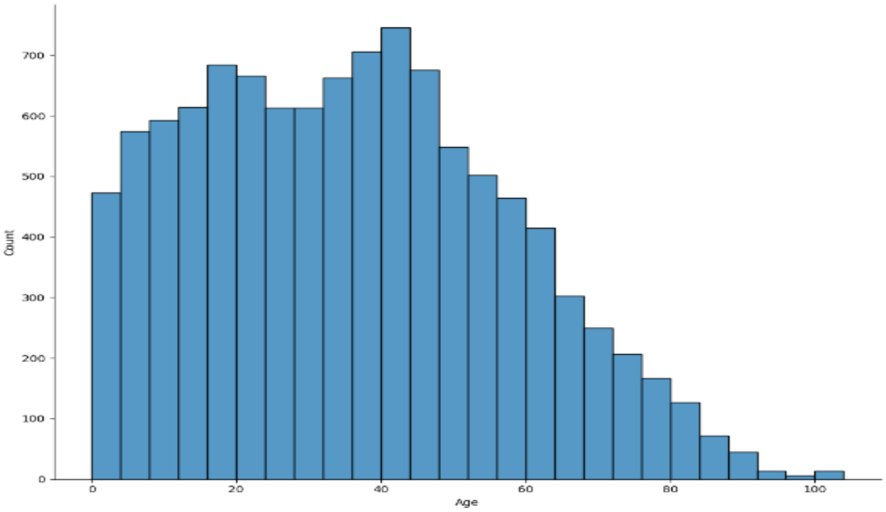


Fig 3: Age Histogram

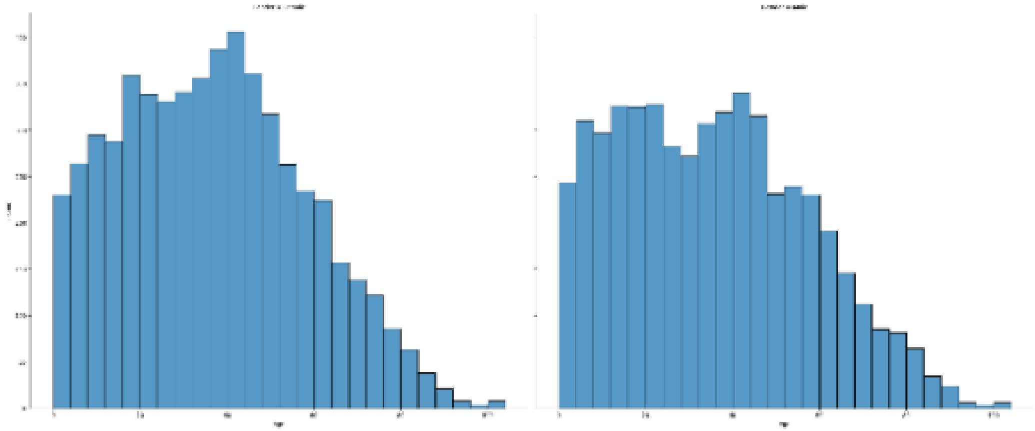


Fig 4: Female and Male Age Histogram

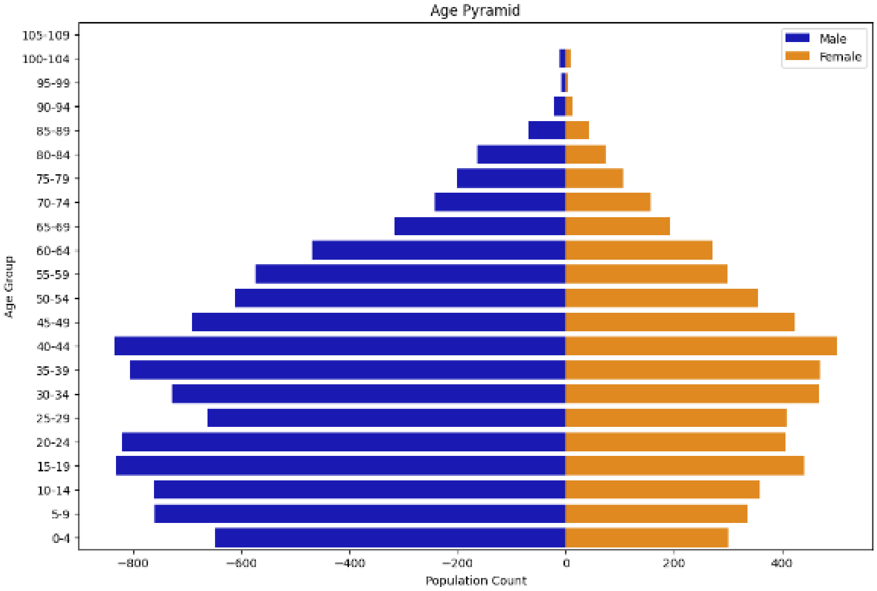
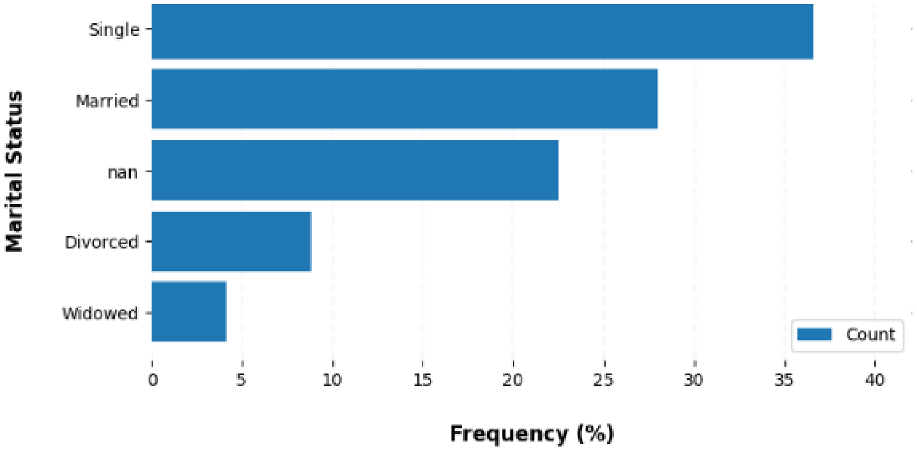


Fig 5: Population Age Pyramid



The Population Age Pyramide above shows that the majority of the population is single (37%), 28% are married, 23% are single and are underage(i.e. less than 17 years old), 9% are divorced and 4% are widowed.

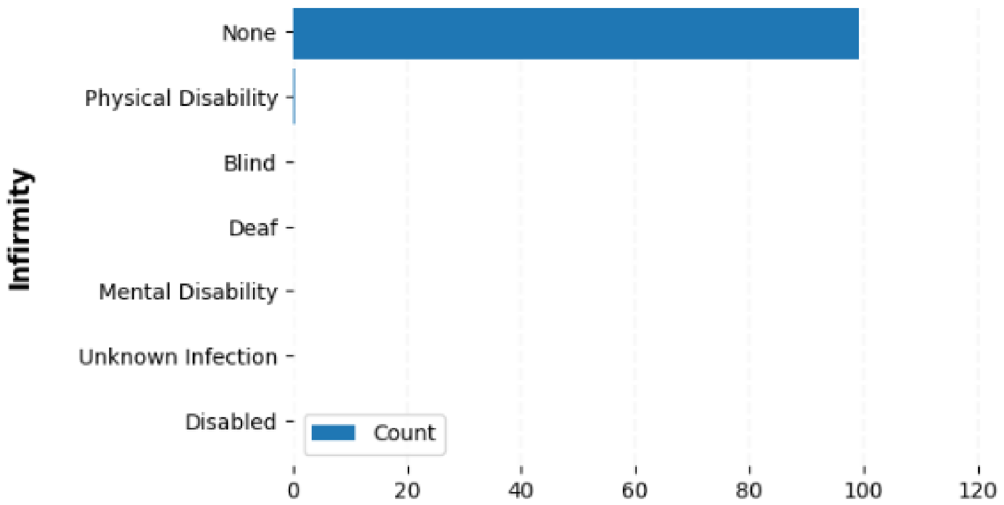


Fig 6: Infirmity frequency

From the Infirmity Frequency above, we can conclude that approximately 99% of the population has no disability.

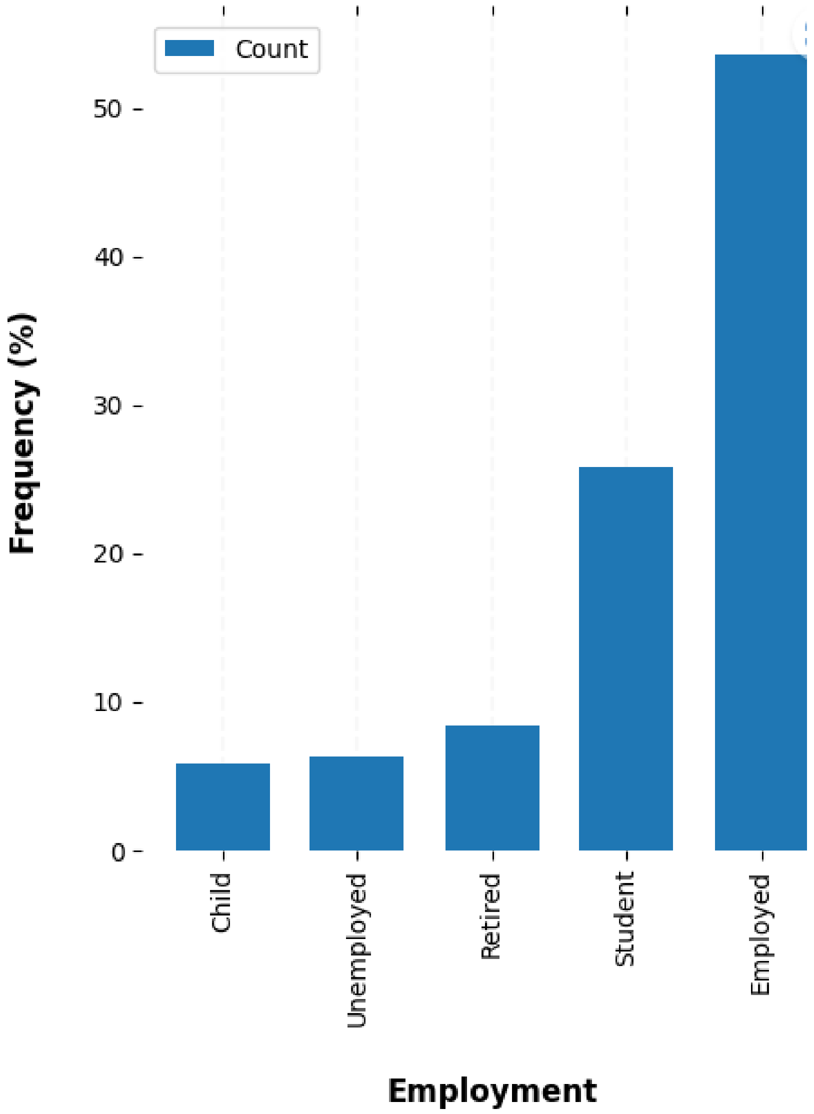


Fig 7: Employment Frequency

Fig 7 above shows a distinct frequency between the rate of employed and unemployed. The above shows that the majority of the population is employed (53%), 27% are students(this can be both university and Ph.D. students), are 8% retired, are 7% unemployed and 5% are children.

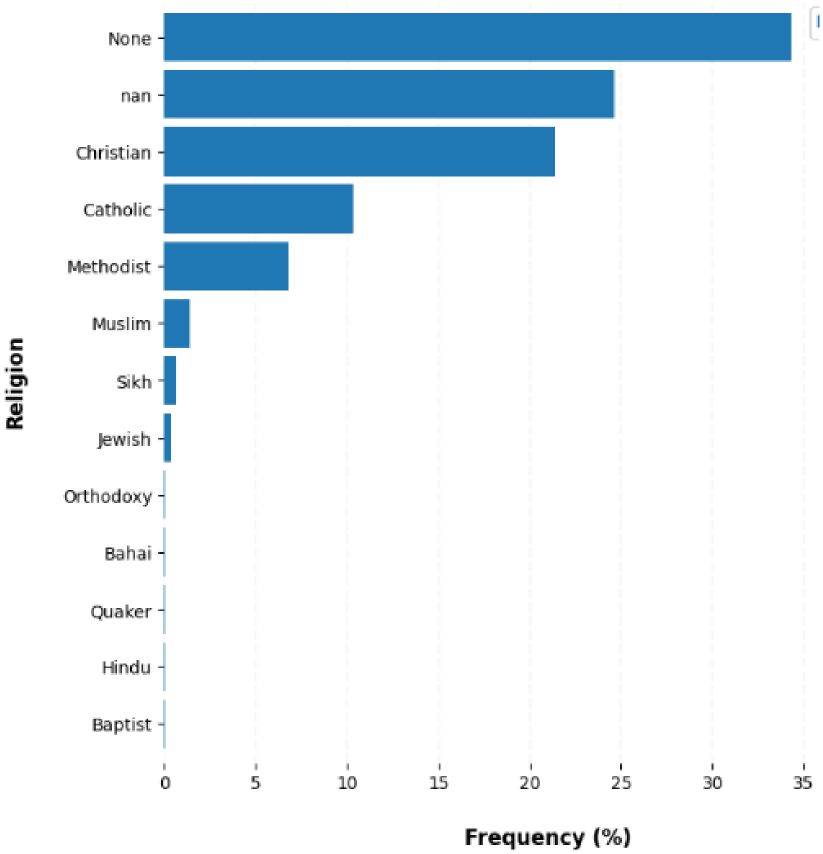


Fig 8: Religion Frequency

Fig 8, Religion frequency shows the various beliefs within the population. 34% of the population have no religion, 21% are Christian and 24% of the population are children that cannot choose their religion yet.

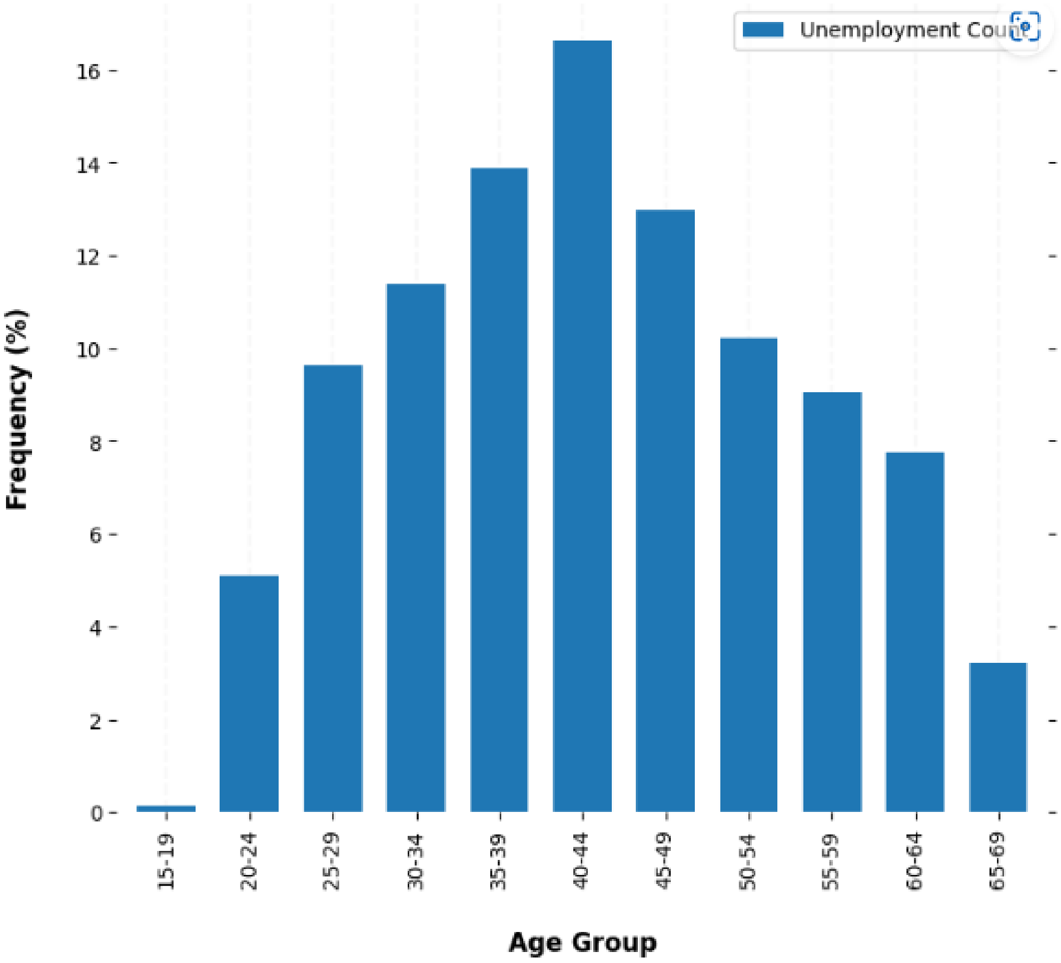


Fig 9: Unemployment

Birth Rate
The fertility and mortality rate of the population is important, but due to insufficient data provided, assumptions will be made on the current population to determine both the birth rate and death rate. To determine the birth rate, assumptions were made that women that can give birth within the population were between 25 to 29 years of age. The birth ratio for the current year is calculated per hundred thousand using the formula below.

Where N. I am the number of infants in the population and
N.W. is the number of women between the ages of 25 to 29.

The birth ratio (per hundred thousand) of the population was calculated to be 23529. It is necessary to determine if the birth rate increased or decreased when compared to the past year. Since there is no record of the previous year's census, a logical assumption was made to get the past year's birth ratio. This assumption was to use the number of children aged 4 as a fraction of women between the age range of 30 to 34 as the birth rate. The birth ratio (per hundred thousand) of the population as of five years ago was calculated to be 32479. This difference in birth rate shows there was a decrease in the birth rate of the population. Hence, the population birth rate per annum is approximately -5511 per hundred thousand.

Death rate

Due to the lack of previous census data, an assumption is also made for death rate calculation. For this calculation, the differential in the number of people in an age bracket will be compared to the younger one. For instance, the number of people in the age range 56-60 was compared to the number of people in the age range 61-65. The assumption was that the number of people in the two age ranges should be the same. With this, we can get the number of people who might have died over 5 years. The result would be divided by 5 to get the rate per year. This process is also repeated for older generations. For instance 66-70 vs 71-75, 76-80 vs 81-85 etc. The average death rate (per hundred thousand) calculated for the population is 3766.

Immigration and emigration rate

Immigrants and emigrants are not indicated in the available data and this has led to making assumptions about whom to categorize as immigrant or emigrant. For this report, the immigrants are assumed to be lodgers and visitors who are still single at the time of taking the census, and the emigrants are assumed to be divorcees that must have probably left town. The number of immigrants calculated per hundred thousand is 3097 and the number of emigrants calculated per hundred thousand is 8806. This difference shows there is a significant difference between the number of immigrants and emigrants in the population.

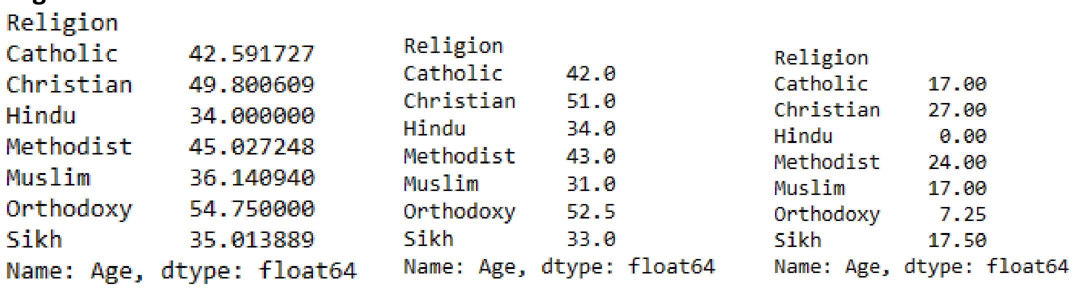
Population growth

population growth rate can be calculated using the formula below

Population growth rate = (Birth rate + Immigration rate) - (Death rate + Emigration rate)

The population growth rate per hundred thousand was calculated to be -14422

Religion



```
Religion
Catholic      42.591727
Christian     49.800609
Hindu         34.000000
Methodist     45.027248
Muslim        36.140940
Orthodoxy     54.750000
Sikh          35.013889
Name: Age, dtype: float64
```

```
Religion
Catholic      42.0
Christian     51.0
Hindu         34.0
Methodist     43.0
Muslim        31.0
Orthodoxy     52.5
Sikh          33.0
Name: Age, dtype: float64
```

```
Religion
Catholic      17.00
Christian     27.00
Hindu          0.00
Methodist     24.00
Muslim        17.00
Orthodoxy      7.25
Sikh          17.50
Name: Age, dtype: float64
```

Fig 10: Mean, Range, and Interquartile Range of the Religion Population

Fig 10 above shows the various religions and their mean, range, and interquartile values. The Christian religion shows to have the highest value with a mean value of 49, high age range of 51, and also a high range for Interquartile range of 27.

Commuters

There will be people whose universities or workplace is in nearby cities. All university students living in the town are commuters as they would have to go to schools for lectures and other school-related activities. Some other people can be categorized as commuters based on their occupation. Professionals like pilots, cabin crew, scientists, technologists, engineer immigration officers, surveyors, etc will likely work in cities. However, people with the occupation of type Child, Student (not university or Ph.D.

student), teacher, and community workers are categorized as non-commuters. Also, retired people are categorized as noncommuters. With this assumption, 6271 of the population are commuters while 4483 are noncommuters. This will make the road busy as many people will have to travel down to work every day. Another means of transportation will take off the pressure on the road.

Health care

The population has a very low number of infirmities recorded which is less than one percent of the population. The number of people that might need care is old people aged 80 and above. This shows there will be an increase in the number of old aged people over time. The birth rate analysis done above shows there is a reduction in the current birth rate when compared to the birth rate of the assumed previous census. This means the government need not invest more in health care.

Occupancy level

There would be a different number of people living in a specific house in the town. However, an assumption will be made for the number of people expected per house using one of the measures of central tendency (mode of houses per street). Each of the houses in a particular street is expected to have the same number of occupants which will be the mode of all occupants living in each of the houses in the same street. With this assumption, 1268 houses are oversized and 534 houses are underused. The difference between the oversized and underused houses indicates the existing housing is being overused. This could have happened if house owners allow students or others to lodge more than expected or even if the extended family lived together in a small house. Any of these could also happen due to a financial issue in the family.

Recommendation

The current housing in the town is currently oversized and there is also a high number of people emigrating from the town compared to the number of immigrants. The population has a very low record of infirmities (less than one percent of the population has a disability) and the birth rate of the population also decreased. This means an emergency medical building is not necessary at the moment. There are more commuters in the town and this would affect the road pressure. Investing in train stations will take pressure off the road but this will majorly be during the hours when people go to work and when they close from work. Train stations should not be prioritized at the moment since the population is shrinking and not significantly expanding. The population growth is decreasing (decreased birth rate and high emigration rate) and the population has more singles. This means there would be more demand for low-density housing. Also, the majority of the population is employed. The majority of the population will be able to afford low-density housing. Therefore, the government should build low-density housing on unoccupied plots of land as this would benefit the population. The majority of the population is employed and very few are unemployed. This means the government does not have to invest in re-training people for new skills any time soon. Also, the birth rate of the population decreased which means the number of school-aged children will not increase any time soon, and investing in schooling would not be necessary at the moment. The government does not have to invest in general infrastructure at the moment as the town is not significantly increasing. Furthermore, there are more people in the working force, especially between the ages of 30 to 59 which means the number of retired people will increase over time. Hence, the government should invest in old-age care.

Bibliography

BBC (2016) *Jedi is not a religion, Charity Commission rules.*

Available online: https://www.bbc.co.uk/news/uk-38368526 [Accessed 04/20/2023]

Eurostat (2017) *Marriage and Divorce Statistics*

Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php/Marriage_and_divorce_statistics [Accessed 04/20/2023]

Gov.uk (No Date) *Universal Credit Eligibility*.

Available Online: https://www.gov.uk/universal-credit/eligibility [Accessed 04/18/2023]

*Marriage Act (*1949*)* Section 3

Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3 [Accessed 04/27/2023]

National Society for the Prevention of Cruelty to Children (2020) *Moving Out*

Available online: https://www.nspcc.org.uk/keeping-children-safe/in-the-home/moving-out/ [Accessed 04/15/2023]

Office for National Statistics (2016) *Standard Occupation Classification*

Available online: https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010 [Accessed 04/20/2023]

Office for National Statistics (2022) *Labour market in the regions of the UK: July 2022*

Available online: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/regionall abourmarket/july2020#:~:text=Local%20labour%20market%20indicators,-Indicators%20from%20the&text=For%20the%20period%20April%202019,Middlesbrough%2C%20both%20at%206.9%25. [Accessed 04/12/2023]