CENSUS PROJECT REPORT

ABSTRACT

This report illustrates various steps taken to clean and analyse a given census of a moderately sized town to make advice for investment and development in the future.

1.0 INTRODUCTION

A census is a survey of the number of people in a country, city, or town with a cumulative insight on the population. In this report, we will be looking at two main steps:

- Data Cleaning
- Data Visualization and Analysis

To clean the data, we will first look at the errors and missing data within the census data set, after which we will validate our proposed cleaned data by implementing some sniff testing functions. In subsequent section of this report, we will discuss and highlight key analytical techniques used in achieving our task for this project.

Before we get started with the process of data cleaning, it is necessary we create an extra copy of our data, this is to ensure we maintain an original of our data.

1.1 DATA CLEANING

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset (*Tableau, 2022).* The various steps taken to clean this data can be found in the corresponding Jupyter Notebook.

It is important to look at the data features which given us a summary of our data for proper guidance on the right structure on how to go about cleaning our data.

Few anomalies can easily be detected as shown in figure 1.1. Age ideally should be of type integer and that should first be investigated in our census data. It is important that the age column is first cleaned to utilize its numerical power for comparisons.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7260 entries, 0 to 7259
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   House Number                  7260 non-null   int64 
 1   Street                        7260 non-null   object
 2   First Name                    7260 non-null   object
 3   Surname                       7260 non-null   object
 4   Age                           7260 non-null   object
 5   Relationship to Head of House 7260 non-null   object
 6   Marital Status                5628 non-null   object
 7   Gender                        7260 non-null   object
 8   Occupation                    7260 non-null   object
 9   Infirmity                     7260 non-null   object
 10  Religion                      5596 non-null   object
dtypes: int64(1), object(10)
memory usage: 680.6+ KB
```

Figure. 1.1 *This illustration shows our census data feature which includes the columns in the data, their Non-Null Count and data types.*

Looking at each unique value for the age column, we detected float numbers, word numbers, integer numbers and a blank. The blank space was replaced with the mode of the age, considering the marital status in that row for a better outcome. All incorrect values and type format were corrected and converted to integer respectively.

The marital status column has null values, two format name representation and a blank. From investigation using our already cleaned age column, we observed that the null values and the blank were all ages below 18 years (minors). The character named type format was renamed appropriately in other to have a single format. Both the null values and the blank was replaced with minor.

The incorrect values in religion 'Sith', 'Nope' and a blank were replaced with None. Due to the volume of Christians, a proper classification was done for proper identification such as Christianity (Catholic), Christianity (Methodist), Christianity (Quaker), Christianity (Other Denominations). Other religions were then classified as follows Islam, Sikhism, Judaism Paganism, *(Chadbourn, 2013).*

The gender had three different formats in referring to the gender such as Female, Male, m, male, female, M, f, F. The character named type format was renamed appropriately in other to have a single format.

In the 'Relationship to Head of House' column, the word 'Niece' was misspelt as 'Neice', this was corrected accordingly.

In the Occupation column, it was observed that some of the individuals who are unemployed were over the retirement age of 66 years. It was deemed necessary to classify sure categories to retired since they are not employed and can benefit from pension in accordance with the UK government. The figure 1.2 below clearly shows that the unemployed age is within 18 to 66 years. In addition, we can also see that the ages between 35 to 45 are the most unemployed, (*Department for Work and Pensions, 2017).*

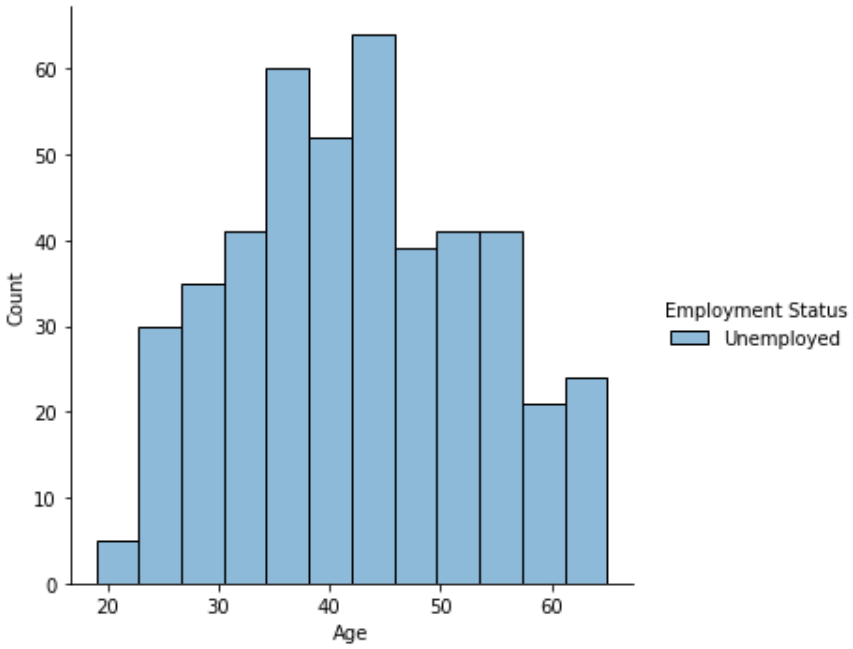


Figure. 1.2 *This is a histogram plot of the ages for the unemployment individuals after data cleaning.*

Three Sniff testing functions were created to aid us in data cleaning and data analysis.

The address checker function: This function takes in two arguments which are the house number and the street. This function will filter out all the information in the census data of a specific household. This will provide a better insight of the kind of people leaving in a particular household and aid to fill in blanks with realistic information or used to detect wrong data while analysing the household.

The blank-nan checker function: This function takes in a single argument which will be a column name, this function is mainly to confirm if there are blanks or nan on a specified column.

The underage-comparison function: This function takes in a single argument which will be a column name. The purpose of this function is to compare a given column in our census data to the minor/underaged category. This function will help differentiate between realistic data and unrealistic data. For example, using Marital Status column as an argument it will be improper for a minor under the age of 18 to be married or divorced. In the figure 1.3 below we can see that the minor age range does not overlap with other categories.

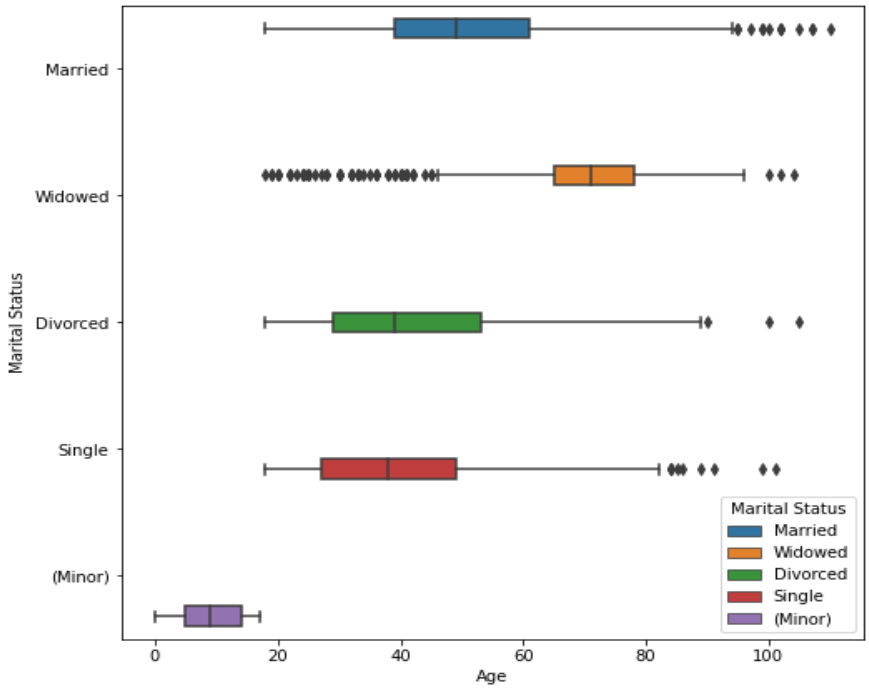


Figure. 1.3 *This shows a boxplot of the marital status against the age.*

We can also look at other columns in comparison with the minor such as Relationship to Head of House, Occupation and Religion to investigate any anomaly like a minor being head of house or a minor being retired or having a religion.

These functions help to revalidate and to give a certain degree of confidence about our cleaned data before proceeding further.

Also referring to figure 1.3, outliers are easily observed in some of the categories in the marital status column. It is reasonable to say that people in their early years of marriage are unlikely to me widowed, therefore it is understandable to have a lot of outliers. We can also observe outliers over the age of 80, this is likely to be as a result of the death rate which causes a decline in population within that age group.

1.2 DATA VISUALIZATION AND ANALYSIS

After data cleaning, additional columns were created to best aid our analysis such as:

- **Address:** This is the concatenation of both the house number and street column.
- **Employment Status:** This gives a better classification of the occupation.
- **Age Group**: This categorizes the age column into 5-year age bands.
- **Number of Occupants:** This shows the number of people in a particular household.

In figure 2.1, we can see the features of our cleaned data including the added columns.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7260 entries, 0 to 7259
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House Number                  7260 non-null   int64
 1   Street                        7260 non-null   object
 2   First Name                    7260 non-null   object
 3   Surname                       7260 non-null   object
 4   Age                           7260 non-null   int64
 5   Relationship to Head of House 7260 non-null   object
 6   Marital Status                7260 non-null   object
 7   Gender                        7260 non-null   object
 8   Occupation                    7260 non-null   object
 9   Infirmity                     7260 non-null   object
 10  Religion                      7260 non-null   object
 11  Address                       7260 non-null   object
 12  Employment Status             7260 non-null   object
 13  Age Group                     7260 non-null   object
 14  Number of Occupants           7260 non-null   int64
dtypes: int64(3), object(12)
memory usage: 907.5+ KB
```

Figure 2.1 *This illustration shows our cleaned census data feature which includes additional columns to aid analysis.*

1.2.1 AGE DISTRIBUTION (AGE PYRAMID) OF THE POPULATION

Examine the age distribution of the population for the male and female, we can observe in figure 2.2 that the population is somewhat evenly distributed for both genders, and the age range between 35 and 55 is the largest in population. We can also observe a low birth rate looking at the size of the population between 0-4 in comparison to the middle-aged category. The population with ages above 90 years tend to be very few and could be considered as outliers. This is normal since life expectancy in the UK is approximately 80 years *(Office for National Statistics, 2021).*

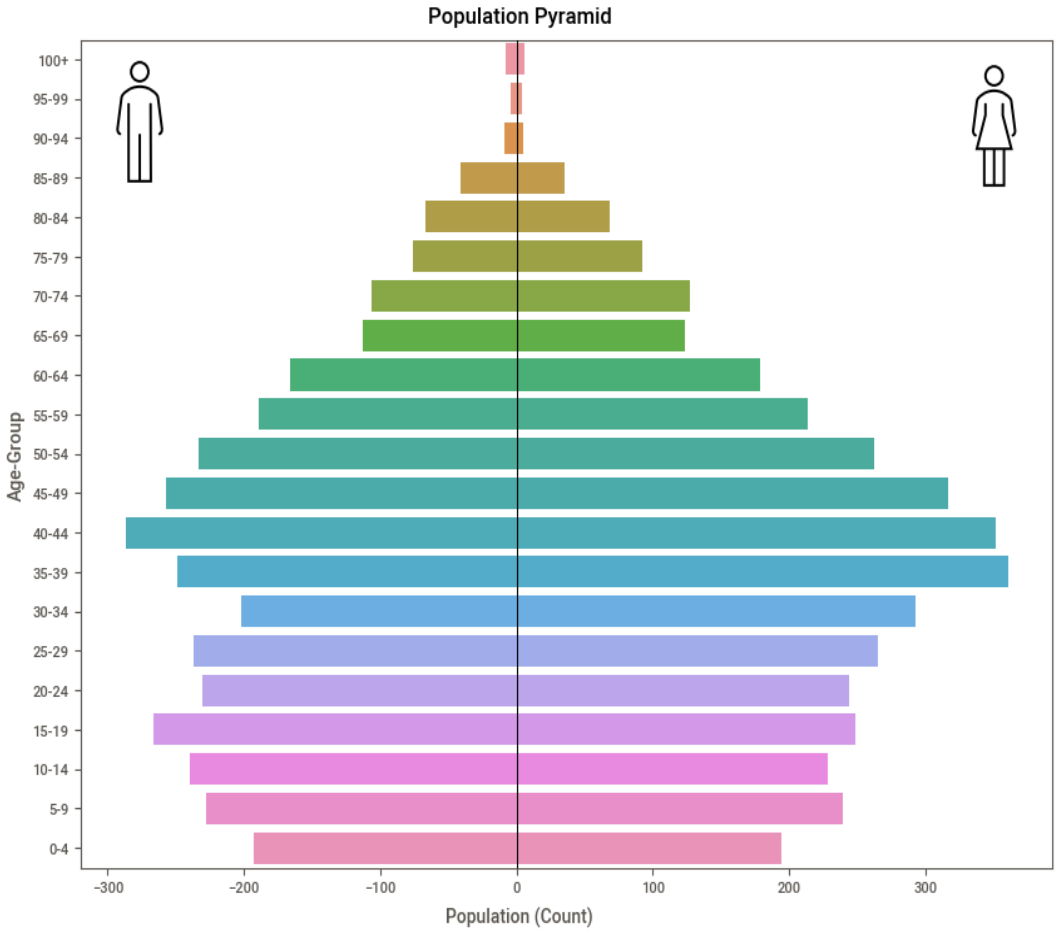
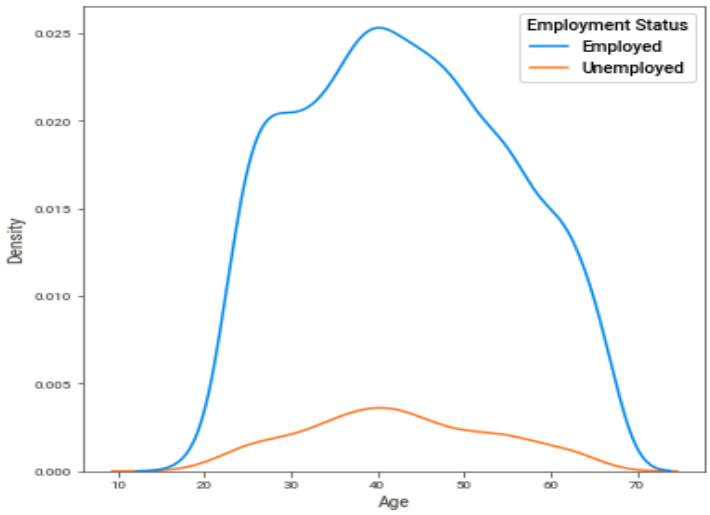


Figure 2.2 *This shows the age distribution (age pyramid) of the population for our census.*

1.2.2 UNEMPLOYMENT TRENDS

As seen earlier during data cleaning, we observed in figure 1.2 that the most affected age group who are unemployed were between the ages of 35 to 45 years. This might be influenced by other factors.



We can have a better insight for the trend in unemployment by comparing both the employed and unemployed.

As shown in figure 2.3, there is a similar structure for both employed and unemployed which is somewhat evenly distributed and something to consider.

Figure 2.3 *This shows a KDE plot of the ages for both employed and unemployed category.*

1.2.3 RELIGIOUS AFFILIATIONS

In figure 2.4, many young people between the ages of 18 to 25 are non-religious. The 'None' group even with the minors, still have a majority of the group falling within the early 20's. In a general sense people are more likely to be less religious.

We can observe from the median for both Islam and Sikhism that the younger people of age less than 40 years are the majority. This shows that these sets of religion are not likely to shrink.

Most Christian groups fall within the middle age which does not show promises for growth given the rate of young people who are not religious.

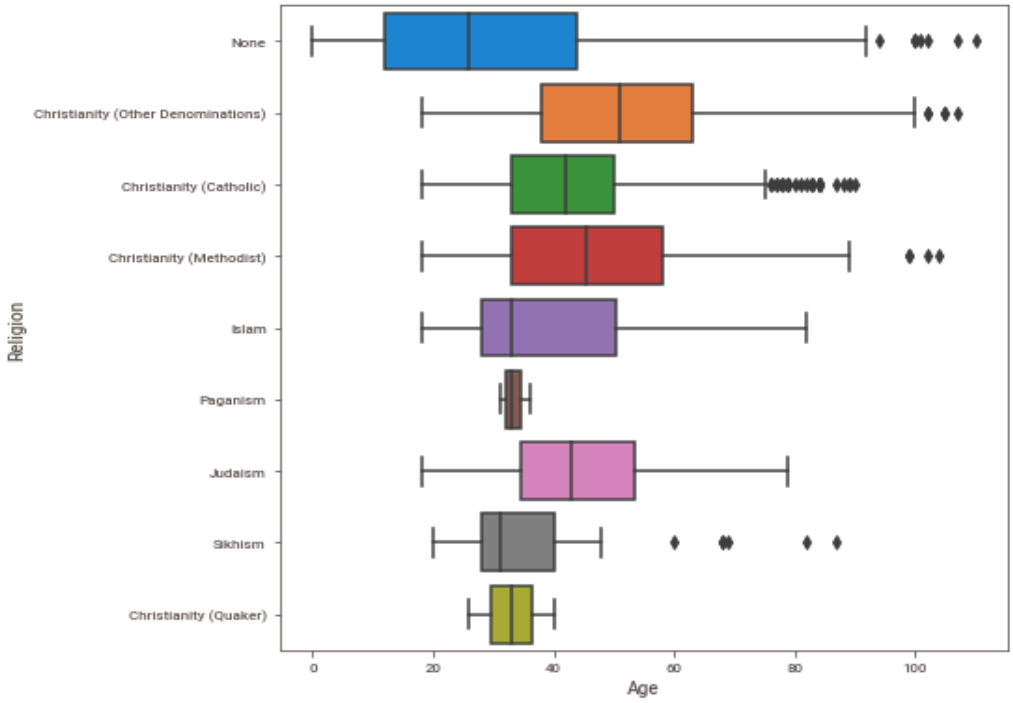
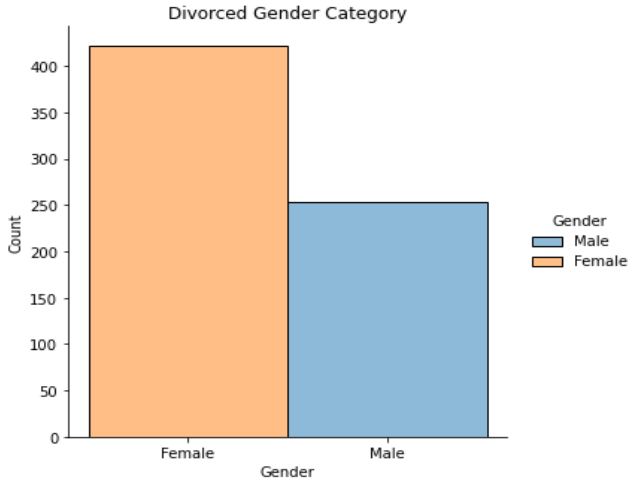


Figure 2.4 *This shows a boxplot of people religion against their ages in our census.*



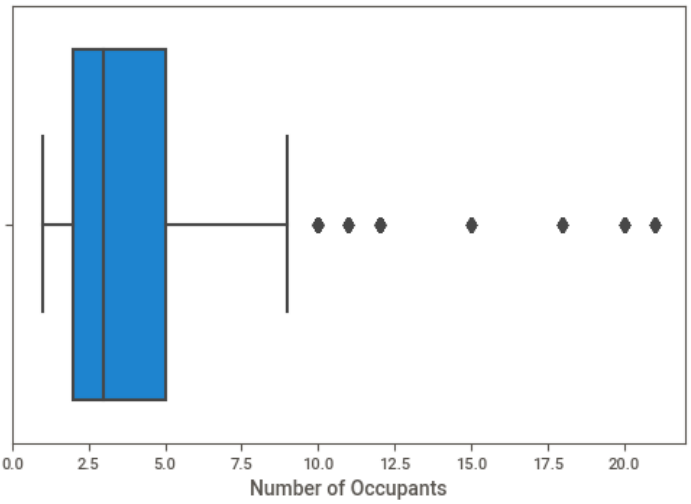
1.2.4 DIVORCE AND MARRIAGE

Considering the marital status by gender as seen in figure 2.5, This shows that there are more female divorcees than there is male. Also referring to figure 1.3, most of the divorcees are in their mid-age. Comparing this insight to housing, it can be said that male divorcee has the potential to leave the town.

Figure 2.5 *This is a histogram plot of the divorced gender category*

1.2.5 OCCUPANCY RATES

From figure 2.6(a) below we can see in the box plot that the median is approximately 3, and majority of the occupants fall within the range of two to five people per household. We can also see a strong similarity in the marital status, the majority of the married category falls within the same range of two to five people per household as we can see in figure 2.6(b). The minor category also has similar tends which tells us that most minors are with their parents.



Statistics

count	7260.000000
mean	3.687052
std	2.506830
min	1.000000
25%	2.000000
50%	3.000000
75%	5.000000
max	21.000000

Figure 2.6(a) *This shows a boxplot and a statistical representation of the number of occupants in a particular household.*

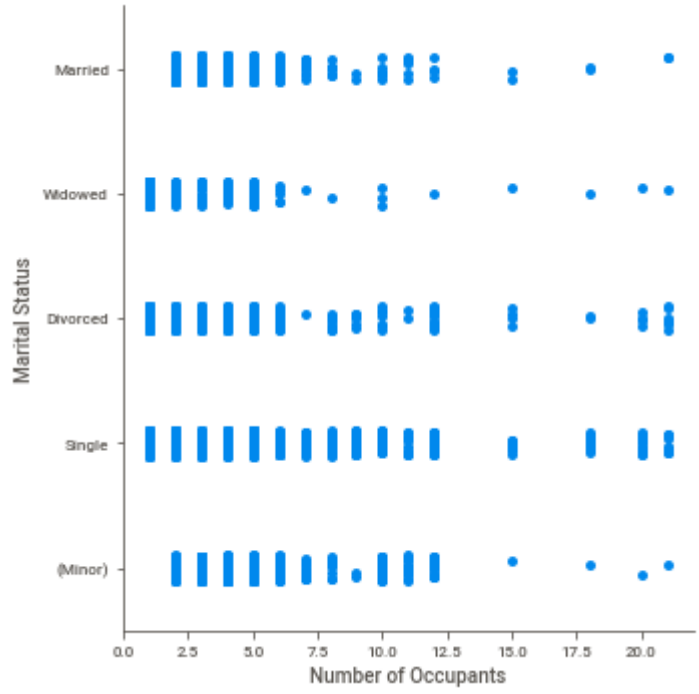
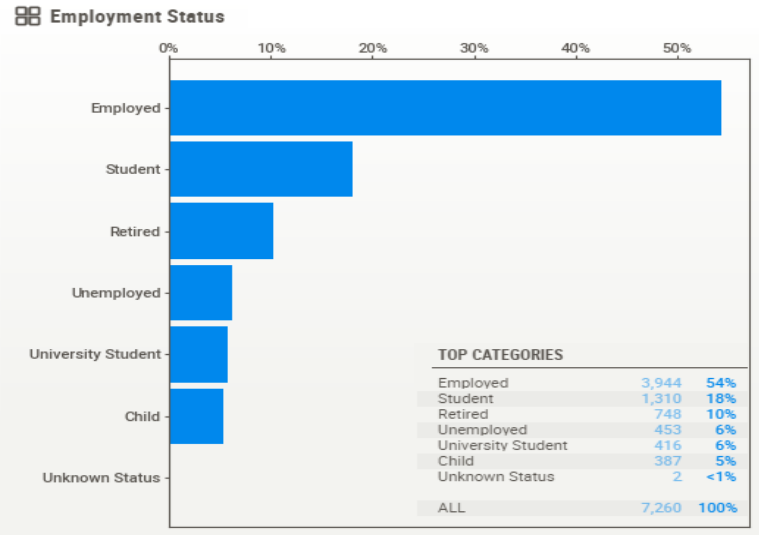


Figure 2.6(b) *This a cat-plot showing of the number of occupants against marital status.*

1.2.6 COMMUTERS

In the employment status column of our data, we clustered both university and PhD students as university students. This will give us a better insight on the total number of students that are commuters.



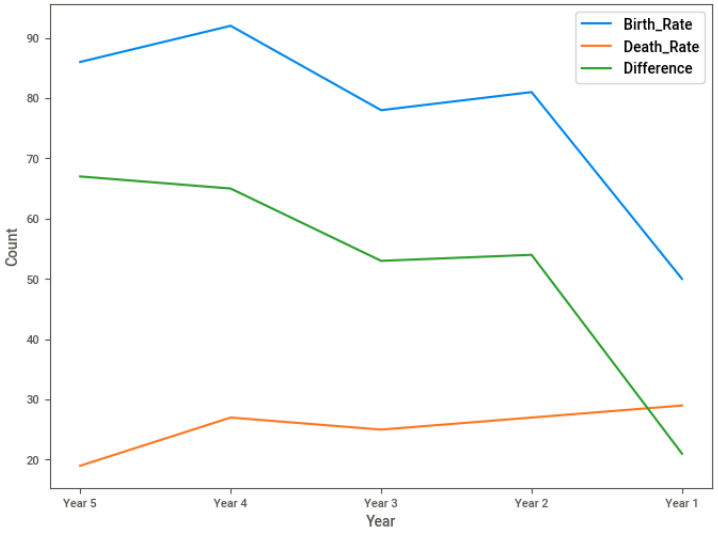
From the statistics shown in figure 2.7, over half of the population are employed but due to the limited data we have the number of employed individuals that are commuters cannot easily be estimated. On the other hand, university student account for only account for 6% of the total population.

Figure 2.7 *This is a visual and statistical representation of the employment status.*

1.2.7 BIRTH AND DEATH RATE

A function called '**birth death rate'** was created to give us a clear analysis of the birth and death rate of our census data. This function takes in a single argument the number of years span and returns a graphical representation of both birth and death rate as shown in figure 2.8. This function is found in our Jupyter Notebook.

We can see a significant reduction in birth of approximately 42% within a 5-year span. For the death rate, there is an increase in the number of deaths from 19 people five year ago to 29 people today. From the difference of both the birth and death rate, we can see that the population is shrinking.



Year	Birth	Death	Difference
Year 5	86	19	67
Year 4	92	27	65
Year 3	78	25	53
Year 2	81	27	54
Year 1	50	29	21

$$\frac{\text{population(n)} \times 1000}{\text{total population}} = \text{(birth/death) rate}$$

n = number of births or number of deaths

Figure 2.8 *This shows change in both birth and death rate within the last five years*

1.3 RECOMMENDATION

People at age 55 years to under State Pension age earn 25 times higher than those aged 16 to 24 years and are categorised to be the most affluent age group *(Mainwaring, 2022).* From our analysis, the ratio of employment and unemployment is 9 to 1 and they both have similar age trend. In our population pyramid, the largest population falls within the age range of 35 and 55 years which tells us that there will be more earnings the town. Married couples are 30% of the total population and most of them are in the affluent age group with an occupancy rate of 2-5 people per household. I will recommend that the local government should build low-density housing since people are more likely to earn more and the housing structure is ideal for families.

Since the population tend to be shrinking and there is a rise in the number of young people over the age of 18 who are non-religious, I will not recommend for High-density housing or religious building to be built. Building a train station may seem reasonable but from the number of proven commuters which are university students (including PhD students), they are only 6% of the total population which is not sufficient to build a train station considering cost. Less than one percent of the total population suffer from infirmity and there is a decline in birth rate, an emergency medical building will not be ideal at this time.

From our analysis, unemployment is 6% of our total population. Reducing unemployment may be important but given the low infirmity rate and high number of elderly people living well above the life expectancy age, I would recommend that the local government should invest in old age care *(Office for National Statistics, 2021)*. Looking at our population pyramid, we can argue that there will be many elderly people in this category between now and the next few decades.

Increase spending for schooling will not be the best decision as there is no reasonable evidence of a growing population of school aged children. General infrastructure will also not be recommended since the population is on a decline.

BIBLIOGRAPHY

Chadbourn, B.C. (2013). *Is Paganism a Religion? Exploring the Historical and Contemporary Relevance of Paganism. Inquiries Journal, [online] 5(12). Available at*: http://www.inquiriesjournal.com/articles/828/is-paganism-a-religion-exploring-the-historical-and-contemporary-relevance-of-paganism.

Department for Work and Pensions (2017). *Proposed new timetable for State Pension age increases. [online] GOV.UK. Available at*: https://www.gov.uk/government/news/proposed-new-timetable-for-state-pension-age-increases.

Mainwaring, H. (2022). *Household total wealth in Great Britain - Office for National Statistics. [online] www.ons.gov.uk. Available at:* https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/totalwealthingreatbritain/april2018tomarch2020.

Ministry of Justice (2022). *Implementation of the Marriage and Civil Partnership (Minimum Age) Act 2022. [online] GOV.UK. Available at*: https://www.gov.uk/government/news/implementation-of-the-marriage-and-civil-partnership-minimum-age-act-2022.

Office for National Statistics (2021). *National life tables – life expectancy in the UK - Office for National Statistics*. [online] www.ons.gov.uk. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2018to2020.

Tableau (2022). *Data cleaning: The benefits and steps to creating and using clean data*. [online] Tableau Software. Available at: https://www.tableau.com/learn/articles/what-is-data-cleaning.