

CENSUS REPORT

FUNDAMENTALS OF DATA SCIENCE

CHIEDOZI ONYEWOTU

202220322



Table of Content

CENSUS PROJECT	2
Introduction	2
Data Cleaning	2
Age	3
Marital Status	3
First Name	3
Religion.....	3
Demographic Analysis and Visualization	4
Age distribution	4
Marital Status distribution	5
Religion Distribution	7
Occupancy Distribution	9
Relationship to head of house.....	10
Infirmity.....	11
Occupation Distribution	11
Birth Rate.....	12
Fertility Rate	13
Death Rate.....	13
Recommendation	14
Suggestions On What Is to Be Done On An Unoccupied Plot Of Land.....	14
Train Station	14
Suggestions For Investments.....	15
Reference list.....	16

CENSUS PROJECT

A census is an official periodic survey of a country or environment that is carried out in order to find out how many people live there and to obtain details such as age, occupation, sex, etc.

Every 10(ten) years the Office for National Statistics (ONS) undertakes a census to give an overview of all the people and households in England and Wales.

Introduction

This report shows data analysis carried out on an imaginary census data for a moderately sized town sandwiched between two much larger cities that are connected by motorways which commuters use to the nearby cities.

My aim is to clean the data using logical, and statistical method without bias, use the cleaned data to derive an insight as to what to do with an allocated plot of land in the town and what to invest in.

Data Cleaning

The Census data was cleaned to correct errors. Assumptions were made in the cleaning process as some columns required previous data in order to statistically arrive at a solution. Detailed cleaning undertaken can be found in the corresponding Jupyter Notebook.

Figure 1 shows an overview of the data set. the total number of columns (11), the number of rows (6000), non-null values, the data types, and memory usage. From observation, there are some irregularities; the Age Column has a data type of object which is supposed to be an integer, the marital status and religion column have missing entries.

- ▼ This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.

```
1 [10]: # This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.
census_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6000 entries, 0 to 5999
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   House Number    6000 non-null   int64  
 1   Street          6000 non-null   object  
 2   First Name      6000 non-null   object  
 3   Surname         6000 non-null   object  
 4   Age              6000 non-null   object  
 5   Relationship to Head of House 6000 non-null   object  
 6   Marital Status  4719 non-null   object  
 7   Gender           6000 non-null   object  
 8   Occupation       6000 non-null   object  
 9   Infirmitiy      6000 non-null   object  
 10  Religion         4681 non-null   object  
dtypes: int64(1), object(10)
memory usage: 515.8+ KB
```

Figure 1: Initial State of Dataset

Empty values were replaced by getting details from similar columns relating to the affected person. Below are the details of the cleaning process done on the affected columns.

Age

This column had a data type of object which needed to be converted to integer, age values are generally integers and not strings. The *unique* method identified entries, containing empty strings and float values. The affected person with an empty string for age was a married woman, for imputation the mode (48) of married women compared against the spouse's age to ensure that its was reasonable to be used.

Marital Status

This column had labels Married, Single, Divorced, Widowed. 1281 are missing values, 1121 of the missing values are minors (< 16) which was replaced with Not Applicable. For those 16 and above, it is legal to be married with parental consent according to the Marriage Act (Marriage Act, 1949: s3), 160 of the missing entries were 16 years of age and above, giving them the label undeclared.

First Name

This column had a value as empty string. Upon inspection for similarities like duplicates either first name or surname, affected person is a single man with no relative. I replaced with the mode first name for males above the age of 18.

Religion

This column had entries; 2106 'None', 407 'Methodist', 1406 'Christian', 620 'Catholic', 87 'Muslim', 1 8 'Jewish', 28 'Sikh', 3 'Orthodoxy', 1 'Baptist', 1 'Bahai', 1319 empty entries were minors that that no relation with which to draw inference to their religion which I replaced with undeclared. I specifically did not assign the mode religion to the adults as everyone has the right to choose religion or believe as a human right under the international law within the international Covent on civil and political rights (Article 18: UN Universal Declaration of human rights)

Demographic Analysis and Visualization

Age distribution

The Age distribution clearly shows a broad base in this census population and gradually flattens as the age becomes higher which indicates that the life expectancy rate in this population is low. This distribution shows a high percentage of young people, which indicates that over a period, the population will increase on a large scale.

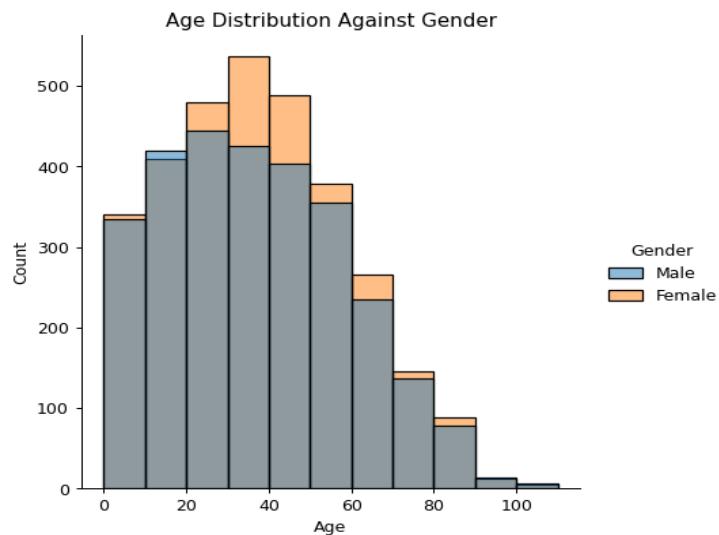


Figure 2: Age Distribution Against Gender

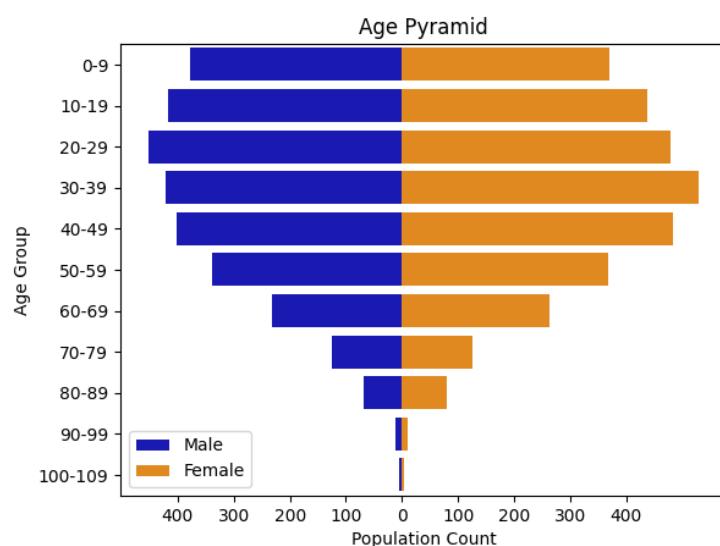


Figure 3: Age Pyramid

The figure 2 shows that there is a higher percentage of females in the overall population of the city. The males only have a greater number between the ages of ten (10) and twenty (20).

The age pyramid in figure 3 shows a clear distinction between the different age categories and gives an overview of how the population is distributed across the ages. A clear indication shows that most of the population are below age 60.

Marital Status distribution

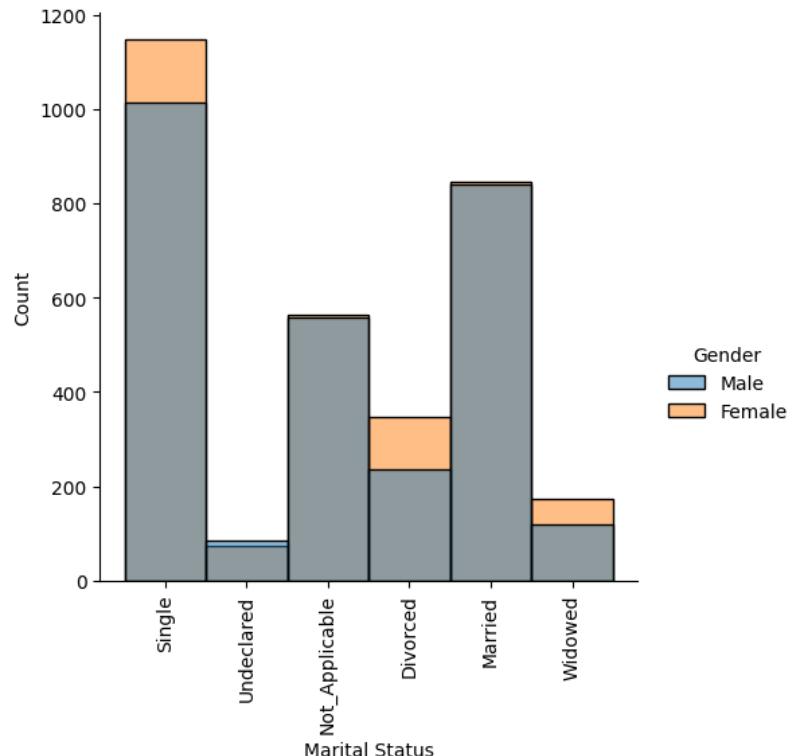


Figure 4: Marital Status

The figure 4 shows that two thousand one hundred and sixty-three (2,163) which is 36.1% of people in the population are single which constitutes most people in the marital status analysis, followed by the married people comprising of one thousand six hundred and eighty-three (1,683) which is 28.1%, then the Not applicable (minors) which constitute one thousand one hundred and twenty-one (1,121) 18.7% of people in the population as they decline from divorced (582 people, 9.7%), widowed (290 people, 4.8%), and unknown (160 people, 2.7%) respectively. The married to divorced ratio in the census is 16.84: 5.82.

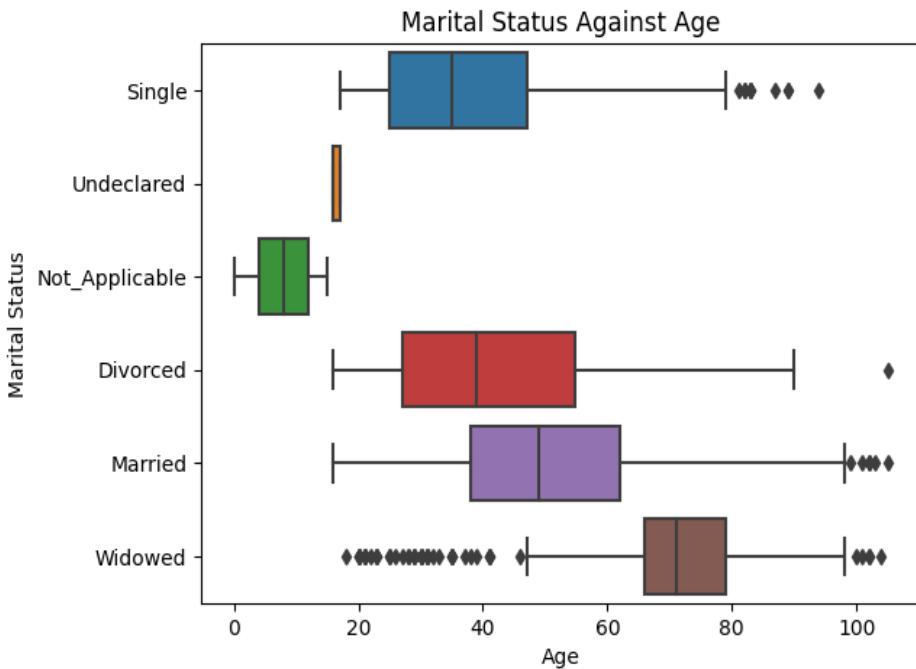


Figure 5: Marital Status Against Age

The figure 5 shows the marital status and age variation in the population. The older they get the higher the chances of being widowed, married and divorced. While the younger population falls with the single category.

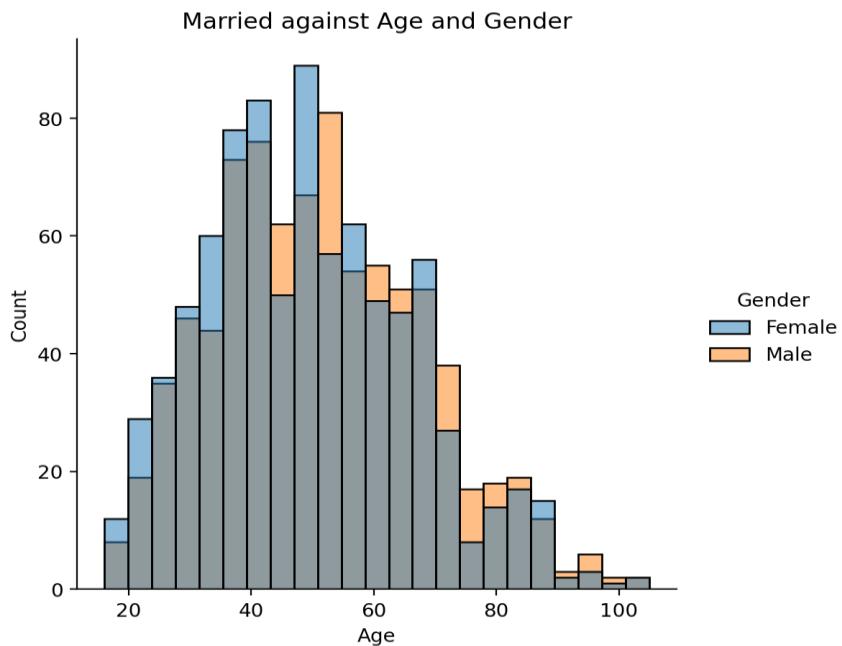


Figure 6: Married Against Age and Gender

The figure 6 shows the category of married, gender and age. From statistical analysis we have a total number of eight hundred and forty-four (844) married women and eight hundred and thirty-nine (839) married men

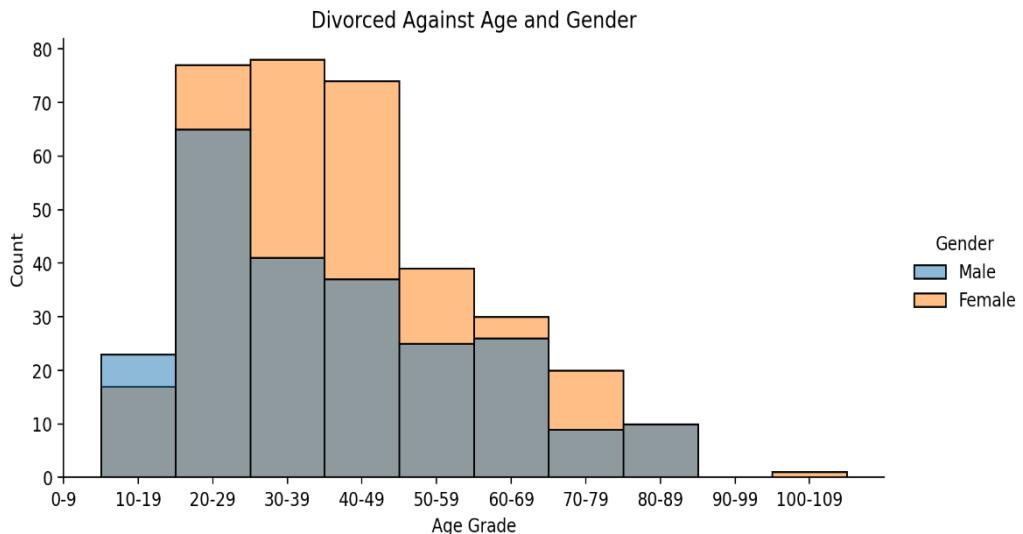


Figure 7: Age Grade

The figure 7 shows the category of divorced, gender and age. From statistical analysis carried, I was discovered that the dataset contained a higher number of divorced females at three hundred and forty-six (346), while the number of divorced males at two hundred and thirty-six (236)

Religion Distribution

Value	Count	Frequency (%)
None	2563	42.7%
Christian	1590	26.5%
Catholic	762	12.7%
Methodist	488	8.1%
Undeclared	436	7.3%
Muslim	99	1.7%
Sikh	33	0.5%
Jewish	24	0.4%
Orthodoxy	3	0.1%
Baptist	1	< 0.1%

Figure 8: Religion Distribution count

The figure 8 shows that 50% people in the population are religious, 42.7% of the population filled their religion as None, while 22.1% of the population have unknown (Undeclared) as religion.

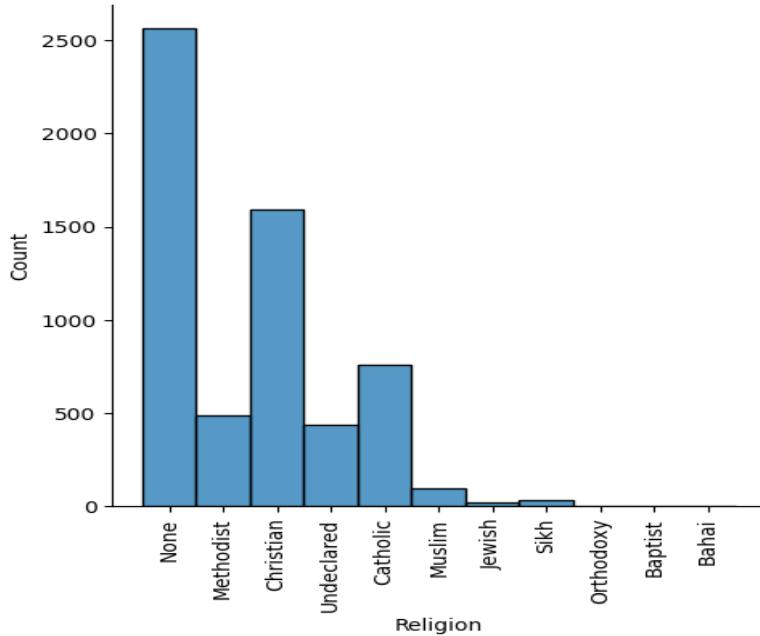


Figure 9: Religion Distribution

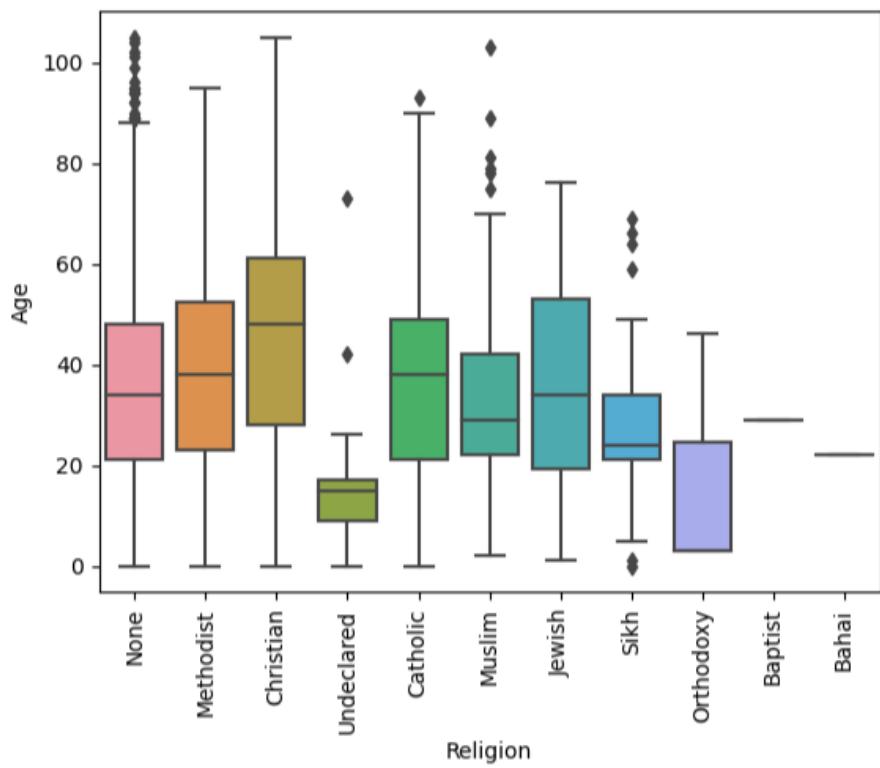


Figure 10: Median Age in Religion

Figure 9 indicates an increase in the Christian religion and shrink in the Bahai, Baptist, and orthodoxy religion. Figure 10 highlights the age spread of individuals across the religions, None and Christian are spread evenly across elderly and children. This also shows the mean age across the different religions.

Occupancy Distribution

This town has 105 streets. Figure 11 indicates that Threpenny Road street had the most occurring instance, meaning a lot of houses but Figure 11 tells that not a lot of people on the street, while Evans Hills Street had the greatest number of occupants on it even though with less houses as shown in figure 12.

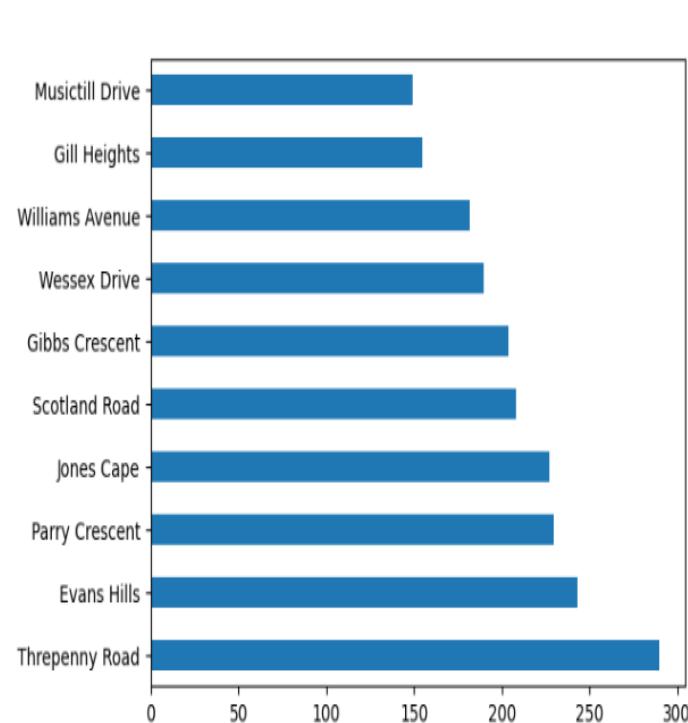


Figure 11: Street with Most Occupants

Street	occupants
0	Evans Hills
1	Jones Cape
2	Jones Avenue
3	Threpenny Road
4	Parry Crescent
...	...
100	Bradford Cottage
101	King Observatory
102	Blue Granary
103	Woodward Haven
104	Young Cabin

105 rows x 2 columns

Figure 12: 10 Most Occurring Streets

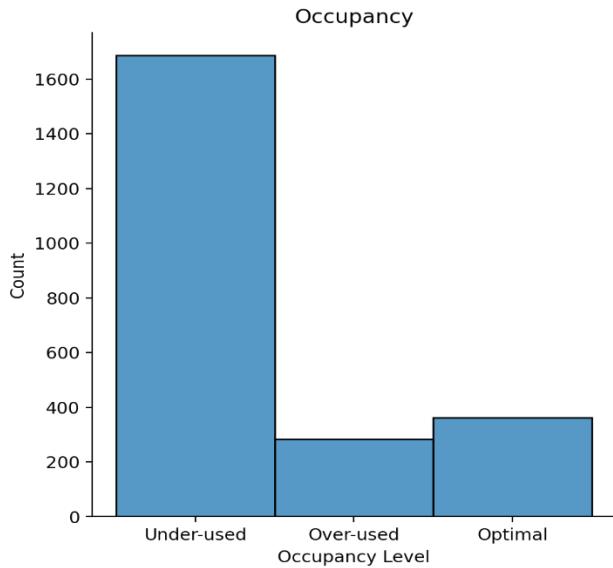


Figure 13: Occupancy Level

Figure 13 shows that the occupancy in the town per house is under-used having one thousand six hundred and eighty-six (1686) as majority of the population fall into the category of under-used house with a number less than three (3) occupants in a building or apartment as the case may be as compared to the number of people in the population that optimize the house use with a number of 4 people in the building / apartment having three hundred and sixty two houses optimized and two hundred and eighty-two cases where the occupants per house is alarming, having more than 4 persons in the building /apartment.

Relationship to head of house

The relationship to the head of house column did not have empty strings or null values but contained a wrong spelling. ‘Niece’ was spelled as ‘Neice’. It also contained three (3) entries that seemed like an outlier, having two (2) sixteen and one seventeen-year-old as head of house. After analysis it was discovered that they aren’t as it is legal for 16-year-olds to marry parental consent.

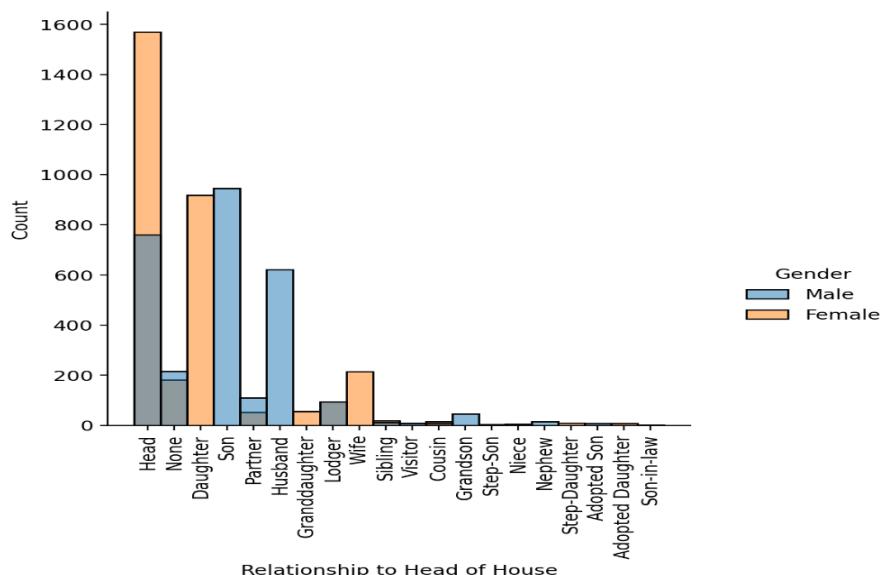


Figure 14: Relationship to Head of House

Infirmity

The infirmity contained empty values which were replaced with None (mode of infirmity). Figure 15 indicates that 99% of the population are healthy.

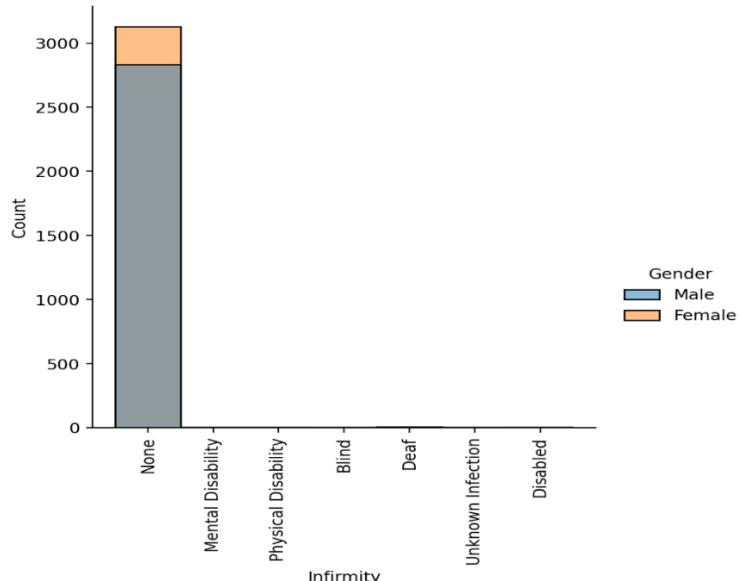


Figure 15: Infirmit

Occupation Distribution

Figure 16 shows that the major occupation in the population are students and a high number of employments which includes skilled and unskilled workers.

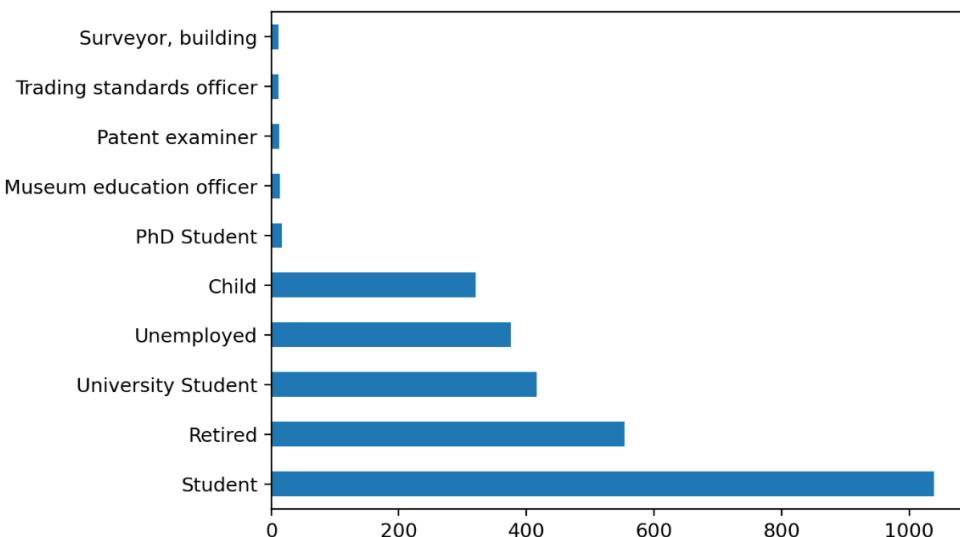


Figure 16: 10 Most Occurring Occupation

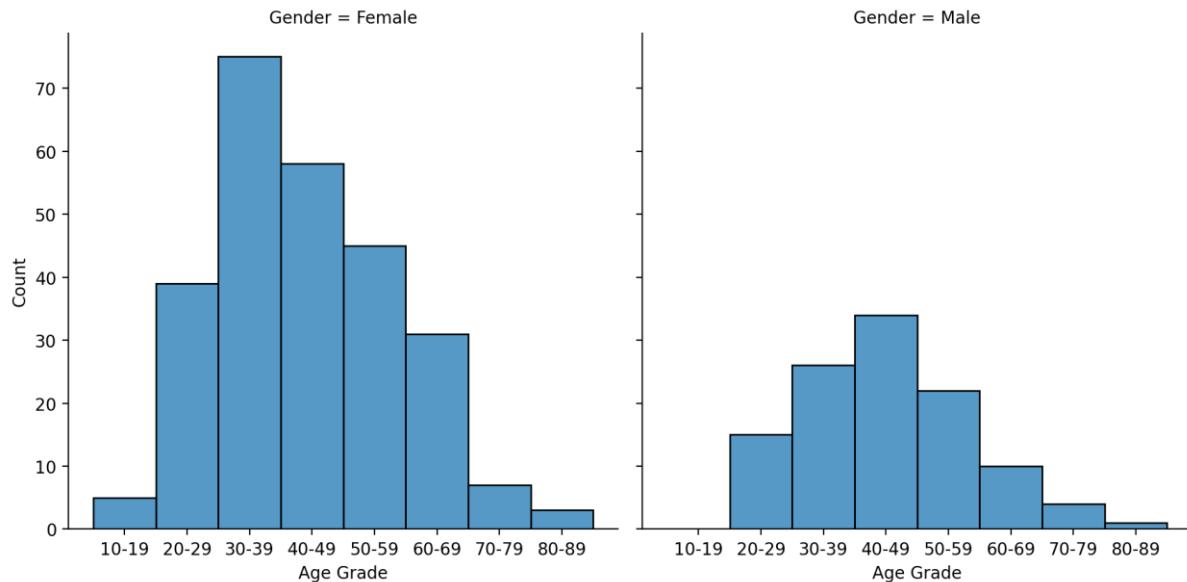


Figure 17: Unemployment Distribution

The level of unemployed in figure 17 shows greater number in the female category. The highest age range of unemployed amongst the female gender is between the age of 20 and 69, while the male category has a low level of unemployed and the age ranges from 30 to 59.

Birth Rate

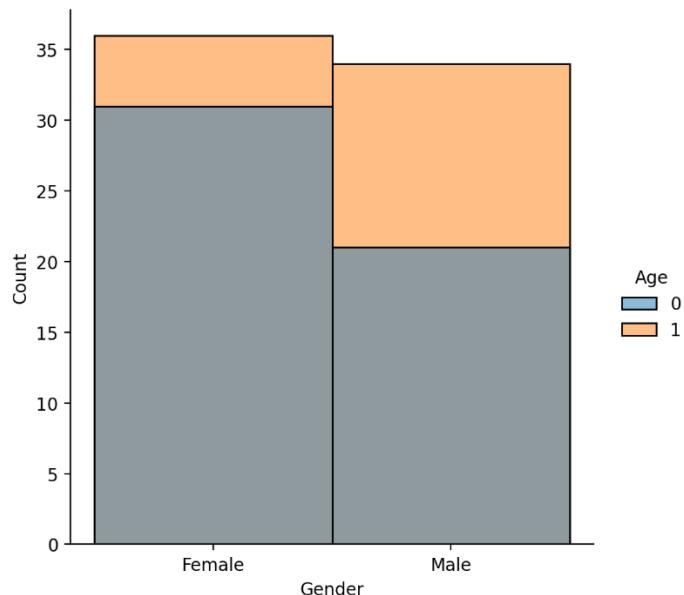


Figure 18: Birth rate Against Age and Gender

The figure 18 shows the number of children born in census data per 1000 people in the population as it has been categorized between the ages of zero (0) / one (1) and gender. Calculating the birth rate

showed an increase in the population within the last four (4) years as the birth rate of children fours ago was at 12.34 per 1000 people in the population as the current birth rate of children in the population has increased to 20.33a children per 1000 people in the population. An addition of 8 more children compared to 4 years ago. These statistics were gotten by getting the number of newborns within the age of zero (0) and one (1) and dividing by the total number of people in the census then multiplying by 1000.

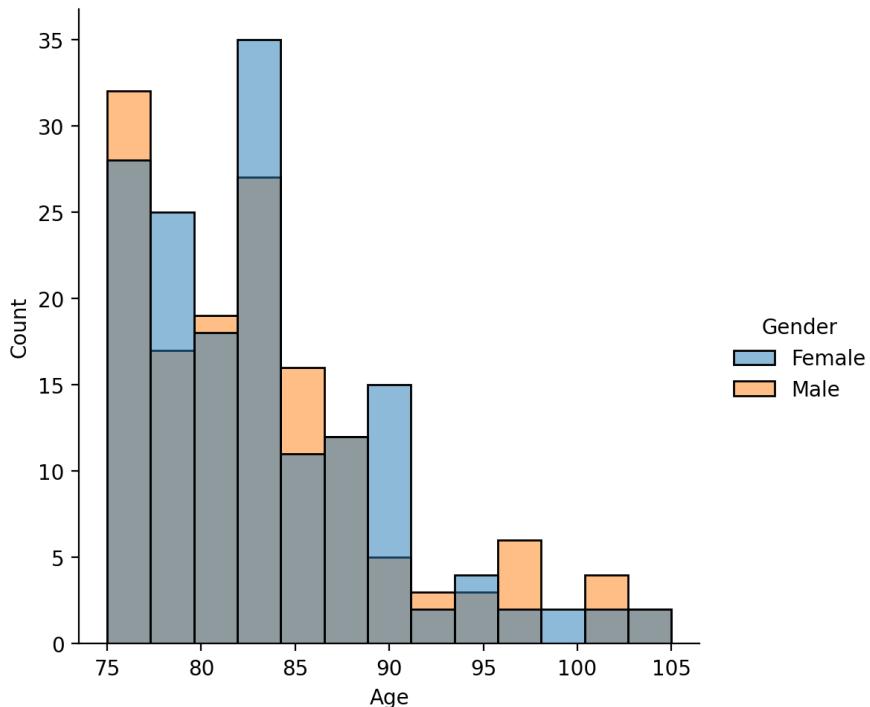
$$\text{Birth rate} = \frac{\text{New borns}}{\text{Total Population}} * 1000$$

Fertility Rate

The fertility rate is calculated by the total number of newborns divided by the total number of women under conceivable age, multiplied by 1000 *Natality - Birth Records Documentation* (2022). Based on this calculation, the fertility rate of the population per 1000 is at 80.42

$$\text{Fertility rate} = \frac{\text{New borns}}{\text{Number conceiveable Women}} * 1000$$

Death Rate



The demographic analysis the census data shows a clear decrease in the number of people aged 76 and above the census data. Statistical analysis carried out on the census data illustrate that the number of people broken down into the different age category on the age pyramid decrease in this sequence; Age grade (70 – 79) = 281 people, Age grade (80 - 89) = 166 people, Age grade (90 - 99) = 25 people, and Age grade (100 - 109) = 11 people. This numbers between the age grade of potential death range are summed and divide by 10 to take account of migration and unforeseen possibilities of unavailability during the census. This value is then divided by the total population, multiplied by 1000 which is 8.05.

Commuters

based on the analysis carried on the census data, university students, PhD students, make up the category of commuters in the population as they require to move from the city to the university regularly which makes up 7% of the population. The visitors in the town which are 18 in number also add to the number of commuters. The population has an employment rate of 54.55% excluding students, university students, retired and children. occupations like sales, engineers, lecturers, architects and surveyors who are potential commuters add up to the number of commuters as their job require them to move from In and out of the town.

Migration

From the statistical analysis carried out on the census data, most of the students are from the town and they stay back in the town as the numbers did not decrease drastically within the age range (20 -35) of students that have finished their studies. In other words, most of them do no emigrate but make up the workforce of the population. The immigration statistics were determined by the lodgers and visitors who were single excluding the widowed and divorced as they could have left their partners to live elsewhere after separation. The total percentage of migrant constituted of just 2% of the population on the census data.

Recommendation

Suggestions On What Is to Be Done on An Unoccupied Plot Of Land

Train Station

The construction of a train station based on the current demand of the town, is paramount as the greatest number of commuters are the people that fall into the university category which includes lecturers, university students and PhD students, the town also a high employment percentage of 54.55 % of the population whose jobs that require them to commute often like sales, surveyors, engineers just to mention a few.

The economic benefits of building a train station out weighs the economical contributions from foods, drinks, tobacco manufacturing, chemical and pharmaceutical industry in the UK. Provides more job, generates more tax revenue and income for the government as researched by the Oxford Economics. The train construction will give rise to more job opportunity both temporary jobs (construction) and permanent jobs like renovations, cleaning, maintenance, train management etc. the benefits of a train station for the town is most preferred as it provides more opportunities than housing, which the town clearly shows a high under-used rate from the occupancy analysis, it also provides more benefits than the construction of a Religious building and construction of an emergency building as the revenue that can be generated from the train station can be used to develop other sectors of the economy for the town.

Suggestions For Investments

From the analysis carried out on the census data, it is advised that the town should invest in building a university/college as they currently have one thousand and thirty-nine (1039) potential university students, which is greater than the current university students in the town which is at four hundred and sixteen that go through the stress of commuting from the town to the bigger cities for higher education.

A university or college equips the town with students who have skill to compete on a global scale and makes the town have an increase in employment percentage as the presence of a university doesn't only create skilled people that are employable but also employs a great deal of people for it to properly function. A university also can change the face of town as it will bring people from other cities with different diversities.

Reference list

Office for National Statistics (2022). *About the census - Office for National Statistics*. [online] www.ons.gov.uk. Available at: <https://www.ons.gov.uk/census/aboutcensus/aboutthecensus>.

Tableau (n.d.). *Guide To Data Cleaning: Definition, Benefits, Components, And How to Clean Your Data*. [Online] Tableau. Available at: <https://www.tableau.com/learn/articles/what-is-data-cleaning#:~:text=tools%20and%20software->.

William, C. (2023). *Census definition and meaning / Collins English Dictionary*. [online] www.collinsdictionary.com. Available at: <https://www.collinsdictionary.com/dictionary/english/census>.

Natality - Birth Records Documentation (2022) Wonder.cdc.gov. Available online: <https://wonder.cdc.gov/wonder/help/natality.html>.

93digital (2021) The economic contribution of UK rail Oxford Economics. Available online: <https://www.oxfordeconomics.com/resource/the-economic-contribution-of-uk-rail/>

National Archives (2020a) *Marriage Act 1949 Legislation.gov.uk*. Available online: <https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/contents>.

Addie, J.-P. (2017) *Seven ways universities benefit society* The Conversation. Available online: <https://theconversation.com/seven-ways-universities-benefit-society-81072>.