

Census Project Report

This report provides an analysis of a moderately sized town's census data to propose recommendations for future investments and development plans for an unoccupied plot of land. The first section details the cleaning of the mock census data to rectify missing and inaccurate records, following scientific and logical reasons followed by subsequent sections focusing on an overview of the town's demographic profile, predicted population growth, employment trends, commuter statistics, and occupancy rates, all aimed at supporting the investment and development recommendations.

Data Description

The data given was in a mock 1881 census data which consist of 8878 entries and 10 column figure 1.0 shows the data description in details and their original data types. The Age column data type was incorrect showing Float 64, the marital status had 6911 entries missing '1967' entries while the Religion column had 6857 Data entries with 2021 missing values. These incomplete columns were cleaned to ensure quality data analysis and representations.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8878 entries, 0 to 8877
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   House Number                        8878 non-null   int64
1   Street                             8878 non-null   object
2   First Name                         8878 non-null   object
3   Surname                           8878 non-null   object
4   Age                               8878 non-null   float64
5   Relationship to Head of House      8878 non-null   object
6   Marital Status                    6911 non-null   object
7   Gender                            8878 non-null   object
8   Occupation                        8878 non-null   object
9   Infirmary                         8878 non-null   object
10  Religion                           6857 non-null   object
dtypes: float64(1), int64(1), object(9)
memory usage: 763.1+ KB
```

Figure 1.0 *Dataset*

Data Cleaning Methodology

The Jupyter Notebook submitted, provides a comprehensive report of the steps undertaken to clean the census data, detailing the removal of missing and inaccurate values, detection of lies, and the correction of data errors.

The Age data type was converted from float to integer, further investigation was done to check the Head of house age to check for lies if ages are below 18. This revealed three cases where 2 individuals aged 16 and one aged 17 were heads of households; two were single parents and one was divorced, this was a possibility because British Genealogy (n.d.) explained boys from the age of 16 could get married from 1850 till 1929 as civil law. The maximum age observed was 105 without outliers, which falls within the UK's life expectancy range. Gerontology Wiki (n.d.).

Spelling errors ‘Neice’ to ‘Niece’ was changed in the Relationship to the Head of House Column and four blank entries were resolved by investigating household information and filling in the appropriate information, with additional investigation for student-related entries.

Two blank entries were resolved on the Gender Column by referencing data from the related relationship to the head of the house data entries.

Marital Status Column Addressed 1967 blank entries, categorizing under 18 as 'Minor' and fixing their values as 'NA', and corrected a false entry (lie) from 'Widowed' status for a student named Donald to 'Single'.

Two empty fields on the Occupation column were filled based on the mode occupation of their respective age group. Occupation was also grouped into 'Grouped Occupation' for better future analysis.

8 blank entries in the Infirmary Column were replaced with 'None', which was identified as the most frequent (mode) value in the column.

Non recognized entries like ‘Sith’, ‘Private’, ‘Nope’ (Jahl, 2023) were changed to 'None'. Also, children's below the age of 18, their religion was aligned with their family heads. This is possible because the parents can determine the religion of a child until 18 and above (ONS, 2022)

Two entries were removed on the Surname Column that couldn't be traced back to any household details, as they were marked as 'Lodger' and 'None', with no clear way to determine their surnames except to physically ask the Head of house. For a much cleaner data set for analysis, this was removed.

Population Demographics

After the cleaning of the Dataset is done, the census data contains these details:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8874 entries, 0 to 8873
Data columns (total 13 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   House Number                          8874 non-null   int64
1   Street                                8874 non-null   object
2   First Name                            8874 non-null   object
3   Surname                               8874 non-null   object
4   Age                                    8874 non-null   int64
5   Relationship to Head of House          8874 non-null   object
6   Marital Status                        8874 non-null   object
7   Gender                                8874 non-null   object
8   Occupation                            8874 non-null   object
9   Infirmary                             8874 non-null   object
10  Religion                              8874 non-null   object
11  Grouped Occupation                    8874 non-null   object
12  Age Group                             8874 non-null   category
dtypes: category(1), int64(2), object(10)
memory usage: 841.4+ KB
```

Figure 2.0 *Cleaned Census Dataset*

To further improve the analysis, Grouped Occupation consisting of a categorized occupation in the dataset and Age Group which is a 5-year age population was further added to improve the analysis done on the report.

Using the Age Group category distribution, a population pyramid was made to illustrate the structure of the population. As seen in Figure 3.0 below, the structure suggests a larger number of younger people compared to older age groups, which is indicative of a growing population. It also shows a relatively balanced gender distribution but with more females in the population. We can also make some more social and economic inferences from this plot.

Due to a high number of young populations between age 0-24 more educational infrastructure and resources may need to be built such as daycare, primary, secondary, and high education. Additionally due the population structure, in future years the community would need to make provision for care homes because of the growing rate of the population the younger ones will be the aged group in the future years.

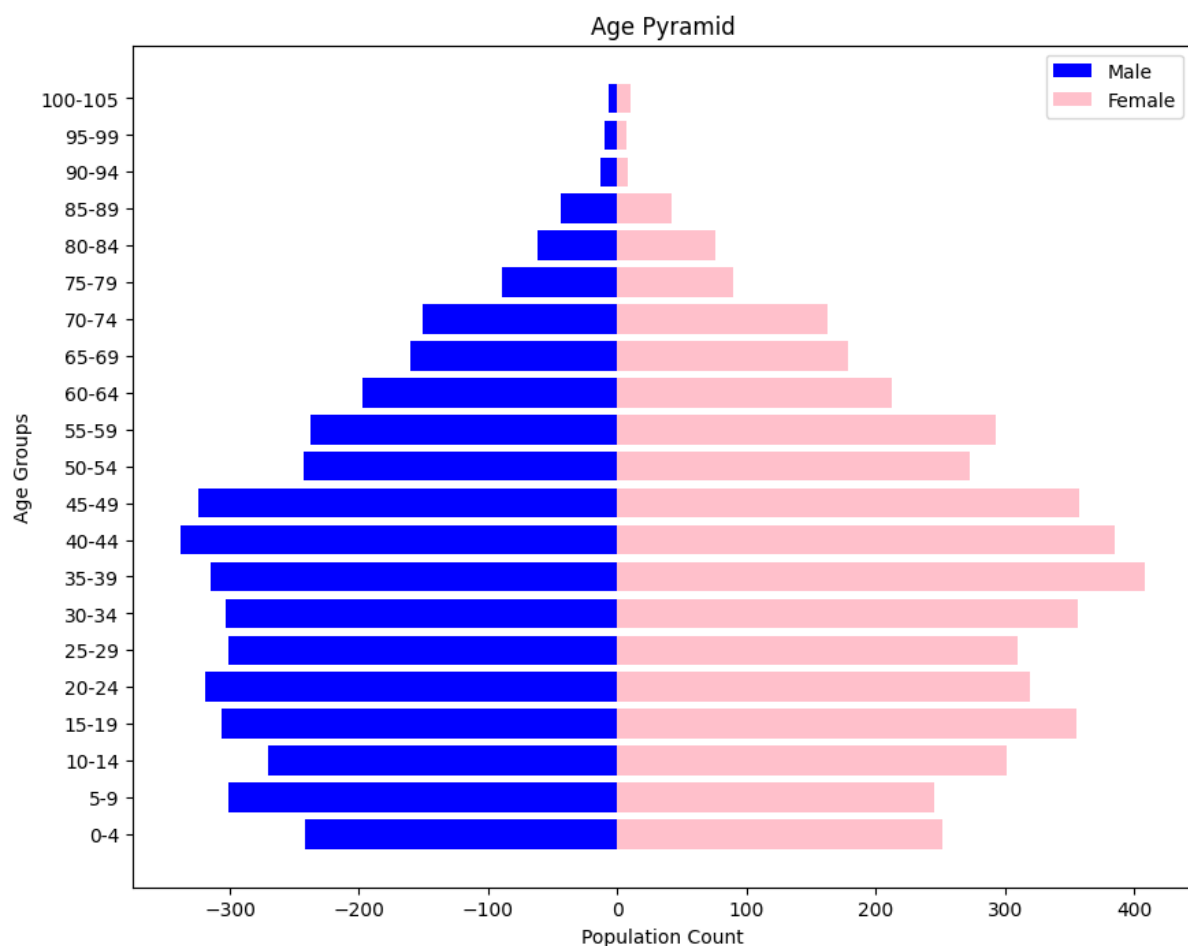


Figure 3.0 *Age Pyramid*

Detailed Analysis

Marital Status and Gender:

Seeing as we have the 2 acceptable genders in the census data collection (Male, Female) in the Gender column, after cleaning the marital column, we see that more single persons that take 34.9% of the population, the married take 28.7%, 9.7% are Divorced and 4.5% are Widowed. The remaining 22.2% are the minors or those below the age of 18 whose marital status couldn't be counted. Further analysis using the Age tells us mainly that between age 40 and 70 are married individuals, ages 20 and 50 are single, ages 20 and 60 divorced, ages 60 and 80 widowed and ages 0 to 18 are the minors.

Value	Count	Frequency (%)
Female	4641	52.3%
Male	4233	47.7%

Figure 4.0 *Gender Column*

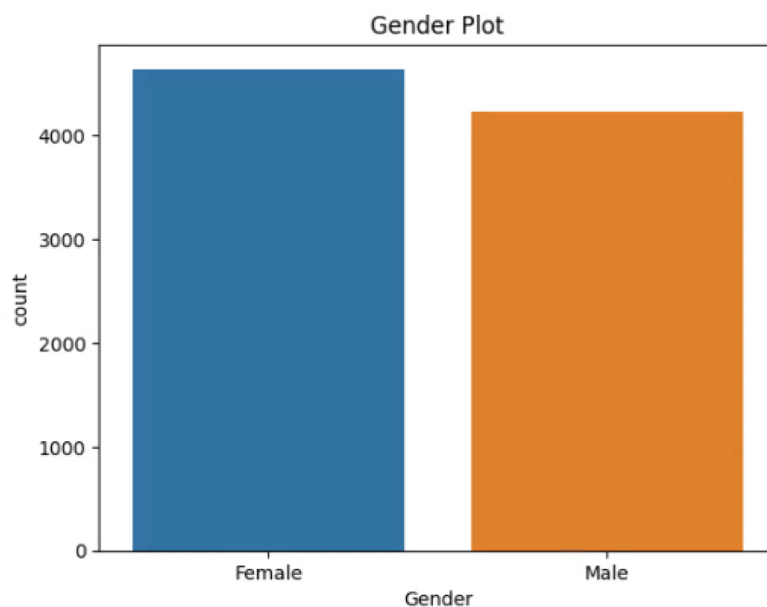


Figure 5.0 *Gender Plot*

Value	Count	Frequency (%)
Single	3093	34.9%
Married	2551	28.7%
NA	1967	22.2%
Divorced	861	9.7%
Widowed	402	4.5%

Figure 6.0 *Marital Dataset*

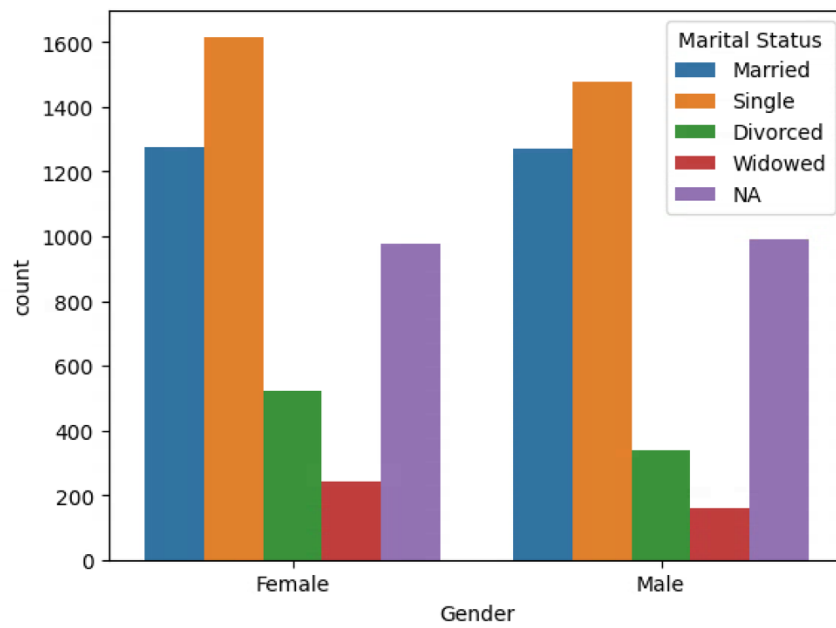


Figure 7.0 *Gender plot with Marital category*

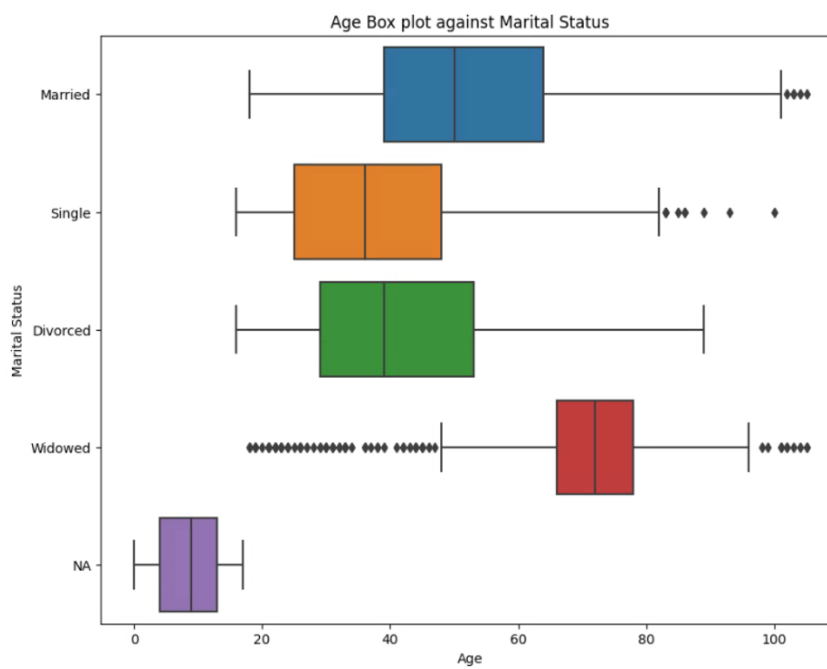


Figure 8.0 *Marital and Age Box Plot*

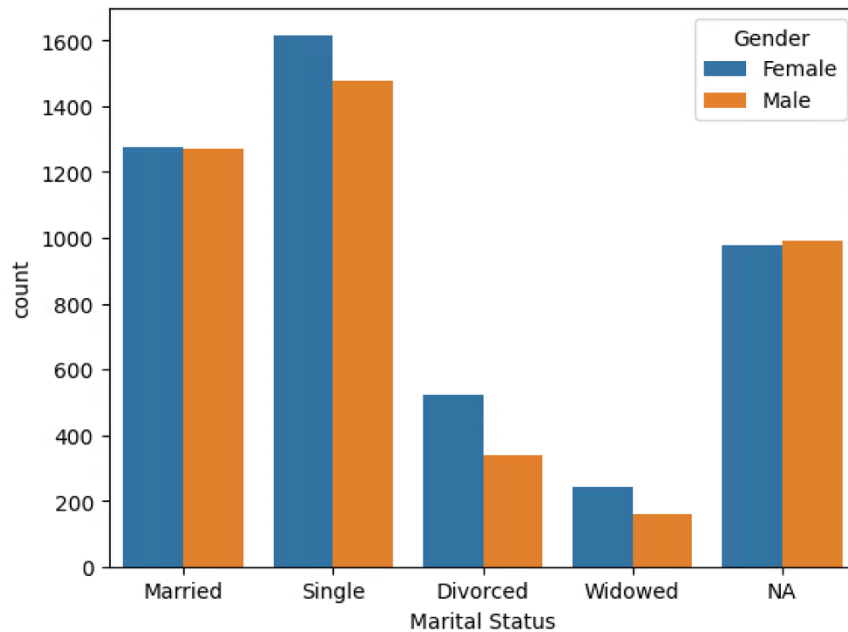


Figure 9.0 *Gender Plot*

This further tells us that almost an equal number of male and female are Married, the larger percentage of the 54% are single ladies, more of the Divorced 9.7% are females, more males die in the population signifying the widowed bars, slightly more males are also contained in the minors than the females.

Religion:

After cleaning the data set, we can infer that a large portion the population are religious having 3 classifications of the Christian faith with a cumulative percentage of 48.03% containing 3 denomination Christians, Catholics, Methodist. We can also see that the age practicing the Christian faith are between 10 to 60 years. Based on the number and Age group religion category, we can also infer that in the future years the Christians will still be dominant identified religion in this community.

Value	Count	Frequency (%)
None	3915	44.1%
Christian	2570	29.0%
Catholic	1354	15.3%
Methodist	740	8.3%
Muslim	157	1.8%
Sikh	73	0.8%
Jewish	58	0.7%
Hindu	5	0.1%
Bahai	2	< 0.1%

Figure 10.0 *Religion Dataset*

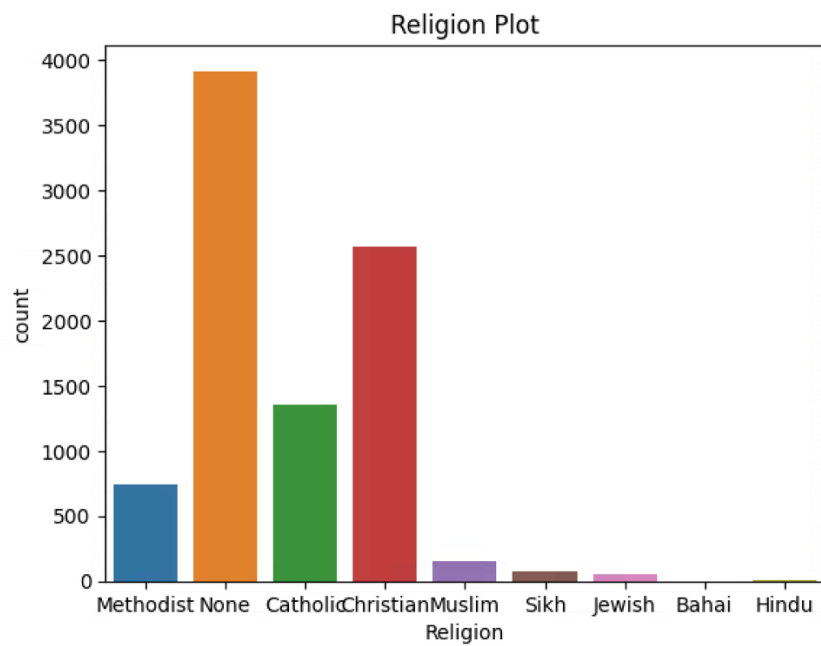


Figure 11.0 *Religion Plot*

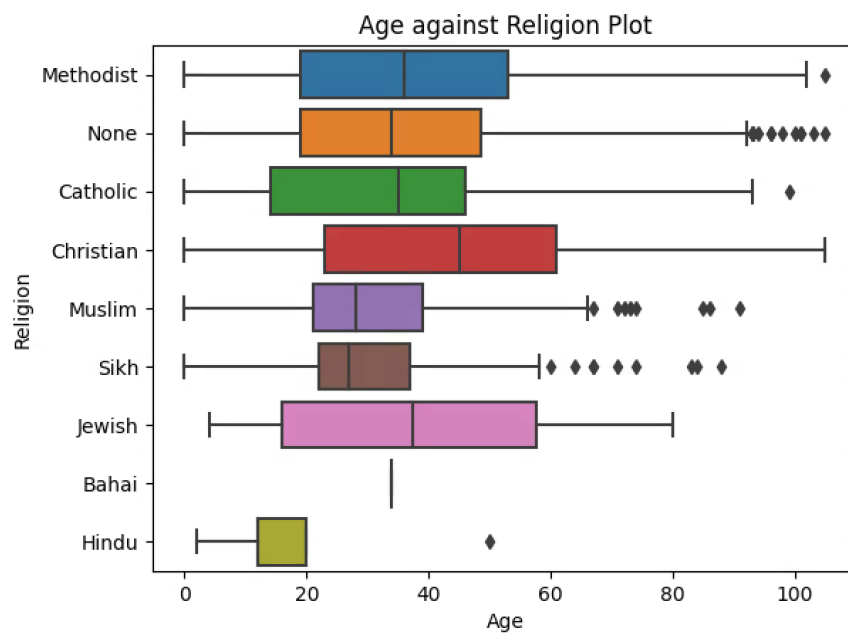


Figure 12.0 *Religion and Age Plot*

Infirmity:

Based on the population, about 8809 persons which is about 99.3% of the population do not have any Infirmity which shows a healthy and thriving population. However, it's a growing population and in future a little clinic can be developed for the community.

Value	Count	Frequency (%)
None	8809	99.3%
Physical Disability	18	0.2%
Disabled	11	0.1%
Mental Disability	10	0.1%
Deaf	9	0.1%
Unknown Infection	9	0.1%
Blind	8	0.1%

Figure 13.0 *Infirmity Dataset*

Occupation:

Using our Grouped Occupation data, 53.2% of the population are employed. From (ONS, 2022) workers in the UK contribute to the commuting and travel rate. With the high employment rate, we can infer that this would contribute to the commuting rate within the town, validating the investment to be recommended in the transportation sector within the community.

Value	Count	Frequency (%)
Employed	4718	53.2%
Student	1577	17.8%
Retired	925	10.4%
University Student	611	6.9%
Unemployed	549	6.2%
Child	494	5.6%

Figure 14.0 *Occupation Dataset*

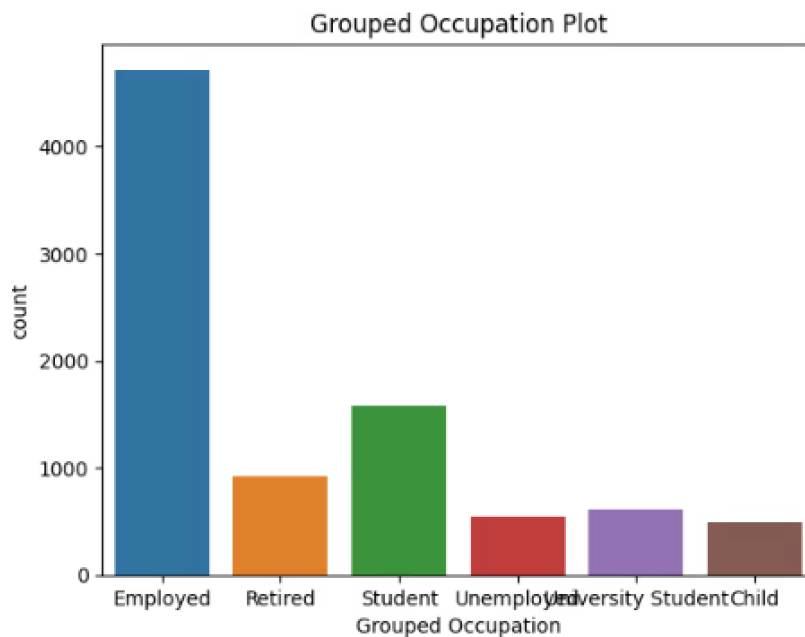


Figure 15.0 *Grouped Occupation Plot*

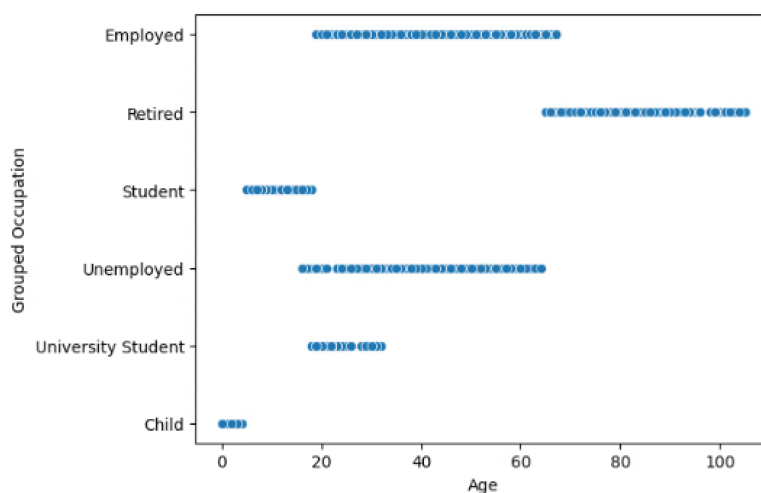


Figure 16.0 *Grouped Occupation Scatterplot*

Unemployment Rate:

The number of employed and unemployed individuals were gotten, and the unemployment rate was calculated based on the population. The formula used is:

$$\text{Unemployment Rate} = \left(\frac{\text{Total Work Force}}{\text{Number of Unemployed People}} \right) \times 100$$
The calculation shows 11.12 % persons are unemployed in the population and from the graph plot below, more females between Ages 30 – 60 are unemployed than the males. With this we can infer that a portion of the investment by the government can be invested into skills training and employment programs to improve workforce skills especially among the females.

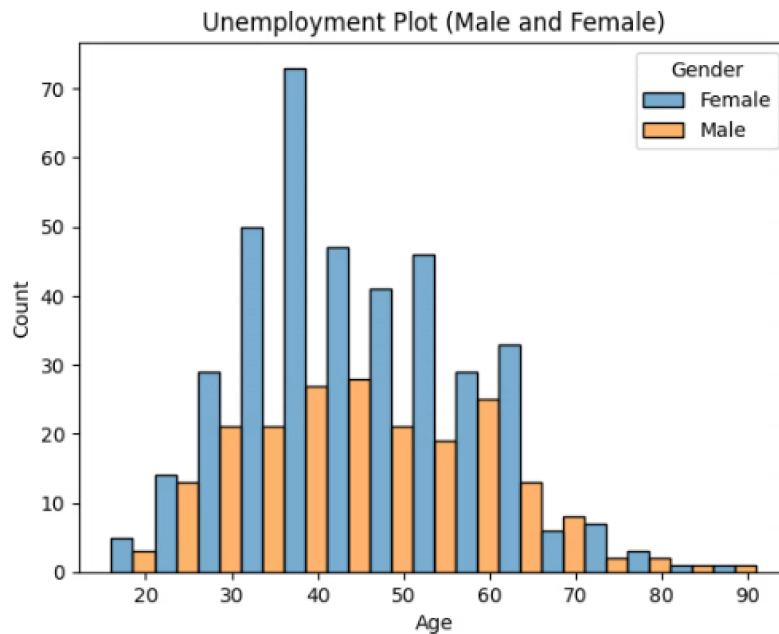


Figure 17.0 *Unemployment Plot*

House Occupancy:

The figure below, shows the median occupancy level of the house is 2 withing the population and based on the mean and median values, most households seem to have a normal occupancy level which depends on the different family sizes and house capacities. However, the first quartile shows only 1 person per house which shows the house is underused. The third quartile (75%) also shows possible house over use with an occupancy of 4 per house depending on the house capacity. The max shows the highest occupancy of 21 per house which assumes house over-use. Though we know there are several students within the population so we can assume a hotel or hostel communal living in some environment based on that we can say the house occupancy by the population is both under and overused.

	House Number	Occupancy
count	3384.000000	3384.000000
mean	45.520095	2.622340
std	50.553041	1.661705
min	1.000000	1.000000
25%	10.000000	1.000000
50%	25.000000	2.000000
75%	57.000000	4.000000
max	236.000000	21.000000

Figure 18.0 *House Occupancy*

Commuters:

Using the Grouped Occupation and the Relationship to the head of house entries, The Lodgers, Visitors and University Students was used to calculate the commuter's percentage within the commuters.

Formular: $(\text{Lodgers} + \text{Visitors} + \text{University Students} \div \text{Total Population}) * 100$.

Using this method, a 28.16% (2499) of commuting persons travelling in and out of the city was gotten, in addition to the 53.2% employed people will also be commuting from home to their various place of work, this which implies that there are potentially a lot of commuters in this growing town and Investing in the transportation system by buying more public busses and building a train station could take some pressure off the roads.

	Grouped Occupation	Count
0	Child	494
1	Employed	4718
2	Retired	884
3	Student	1577
4	Unemployed	590
5	University Student	611

Figure 19.0 *Commuters (University Students, Employed)*

	Relationship to Head of House	Count
0	Adopted Daughter	18
1	Adopted Granddaughter	1
2	Adopted Son	12
3	Cousin	26
4	Daughter	1349
5	Granddaughter	87
6	Grandson	83
7	Head	3383
8	Husband	901
9	Lodger	294
10	Nephew	14
11	Niece	5
12	None	619
13	Partner	247
14	Sibling	31
15	Son	1389
16	Step-Daughter	14
17	Step-Son	14
18	Visitor	17
19	Wife	370

Figure 20.0 *Commuters (Lodgers, Visitors)*

Divorce Rate:

From the Divorce rate plot under the marital status section, we have more females divorced persons and so we can infer that their male counterparts leave the town while the women remain. Also, the data cleaning the divorce rate was calculated using this formular:
 $\text{divorce_rate} = (\text{divorced_number} / \text{total_marital_status}) * 1000$. We could then say that in every 1000 marriages 97 of them were divorced.

Marriage Rate:

The marriage rate plot under the marital status section in relation to the gender column, shows we have relatively the same number of married females and male in the population and to calculate the marriage rate, the marital number of the population was gotten which is the addition of the total number of divorces, married people divide by the total number of marital statuses multiplied by 1000.

$(\text{marriage_rate} = (\text{marital_number} / \text{total_marital_status}) * 1000)$

We could then say that in every 1000 persons 384 of them were married.

	count	mean	std	min	25%	50%	75%	max
Marital Status								
Divorced	861.0	41.912892	16.589730	16.0	29.0	39.0	53.0	89.0
Married	2551.0	51.702470	16.727378	18.0	39.0	50.0	64.0	105.0
NA	1967.0	8.868327	5.101857	0.0	4.0	9.0	13.0	17.0
Single	3093.0	37.711607	14.347931	16.0	25.0	36.0	48.0	100.0
Widowed	402.0	67.681592	18.251174	18.0	66.0	72.0	78.0	105.0

Figure 21.0 *Marital Status Insights*

Birth and Fertility rate:

The birth rate is calculated as the total number of registered births in a census year, while supposing that an infant's age should be zero. To determine the birth rate in the UK, (ONS, 2021), the number of women between 15 and 49 years old is determined. This is then divided by the total population and multiplied by 10 to attain birth rate as $(\text{birthrate} = \text{no_of_new_born} / \text{total_population} * 1000)$. The number of women bearing age is then multiplied by 1000 to get the Fertility rate per 1000 women.

In this population, 310 women have been recorded as having given birth. Regarding the census year, there were 84 births in total translated into a rate of nine per one thousand women. The fertility rate of childbearing women in this population is 271 per thousand.

Death rate:

The death rate for this population was calculated using an estimated death rate for a specific age range within the population because no information about death was given and we are forced to make some assumptions. Statista (2023) shows the life expectancy rate from 66 and this will be used to calculate the death rate for the population. The death rate per 1000 persons is about 16 persons. Which indicates a very low mortality rate.

Recommendation:

Based on the insights gotten from the census data, and justifications made, from the growing population, High-density housing based on occupancy levels and the age pyramid structure shows a growing population with a large number of youths and young adults, therefore high-density housing can provide affordable living space for individuals starting a family or students. Also, with 28.16% of commuters and a high employment rate of 53.2%, the government can improve the transportation of these individuals by investing in building a train station which would help reduce the road transport pressure. Though the population seems healthy with a low infirmity rate, provision for emergency medical buildings/ clinics in case of future emergencies like childbirth or maternity/ normal injuries/ accidents cases would be handled properly.

Given that the population has about 11.12% unemployment rate, especially among females between 30 and 60, the government should invest in skill training programs to help reduce the unemployment rate to empower a portion of the growing population. The presence of a young population and the potential for an increasing number of school-aged children warrant an increase in educational spending to build and expand schools and educational facilities. For future development, old age care facilities should also be looked into ahead though this is not an immediate need. As the population grows there will be a need for a general infrastructure service within the town to ensure proper maintenance. However, this is also not an immediate need. These strategic investments are anticipated to pave the way for a well-rounded community development that aligns with both current demands and future growth projections.

References

British Genealogy (n.d.) *The age of marriage under Scots Law*, Available online: <https://www.britishgenealogy.com/forum/archive/index.php/t66084.html#:~:text=The%20age%20of%20marriage%20under,Wales%20and%20Scotland%20until%201929>

Accessed: [12/23/2023].

Gerontology Wiki (n.d.) *List of oldest living people in the United Kingdom*.

Available online:

https://gerontology.fandom.com/wiki/List_of_oldest_living_people_in_the_United_Kingdom

Accessed: [6/01/2024].

Jahl, J. (2023) *HISTORY OF SITH RELIGIOSITY-Jedi and Sith religion in the offline world*.

Available online: <https://www.researchgate.net/publication/368713271>. Accessed:

[6/01/2023].

Citizens Advice (2021) *Young people's rights*

Available online:

<https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/religion/bulletins/religionenglandandwales/census2021>. Accessed: [8/01/2024].

Office for National Statistics (2022) *Religion, England and Wales: Census 2021*

Available online:

<https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/religion/bulletins/religionenglandandwales/census2021>. Accessed: [8/01/2024].

Child Law Advice (2016) *Child Employment*

Available online: <https://childlawadvice.org.uk/information-pages/child-employment/#:~:text=Young%20people%20can%20start%20full,they%20turn%2016%20years%20old>.

Accessed: [11/01/2024].

Office for National Statistics (ONS) (2022) *Travel to work (England and Wales)*

Available online:

<https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/traveltoworkenglandandwales/census2021>. Accessed: [14/01/2024].

Age UK Cymru (2024) *State Pension*

Available online: <https://www.ageuk.org.uk/cymru/information-advice/money-legal/pensions/state-pension/#:~:text=phased%20out%20completely.->

[.When%20will%20I%20reach%20State%20Pension%20age%3F,the%20government%27s%20State%20Pension%20calculator](#). Accessed: [28/12/2023].

Office for National Statistics (2022) *Childbearing for women born in different years, England and Wales: 2020*

Available online:

<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/bulletins/childbearingforwomenbornindifferentyearsenglandandwales/2020#:~:text=A%20woman%27s%20childbearing%20is%20assumed,not%20affect%20the%20overall%20patterns>. Accessed: [14/01/2024].

Statista (2023) *Death rate in the United Kingdom (UK) by age*. Available online:

<https://www.statista.com/statistics/1125118/death-rate-united-kingdom-uk-by-age/>
Accessed: [14/01/2024].