Report of Population Census Conducted in a UK Town

4/28/2023

Data Cleaning, Analysis

&

Recommendations for the Government Officials

OLIVIA CHIZOBA OKORO

MSC. AI & DATA SCIENCE 2023/2024

1.0 Introduction

This is a thorough account of a census faker data that was taken in 1881 in a small British town that was wedged between two considerably larger towns.

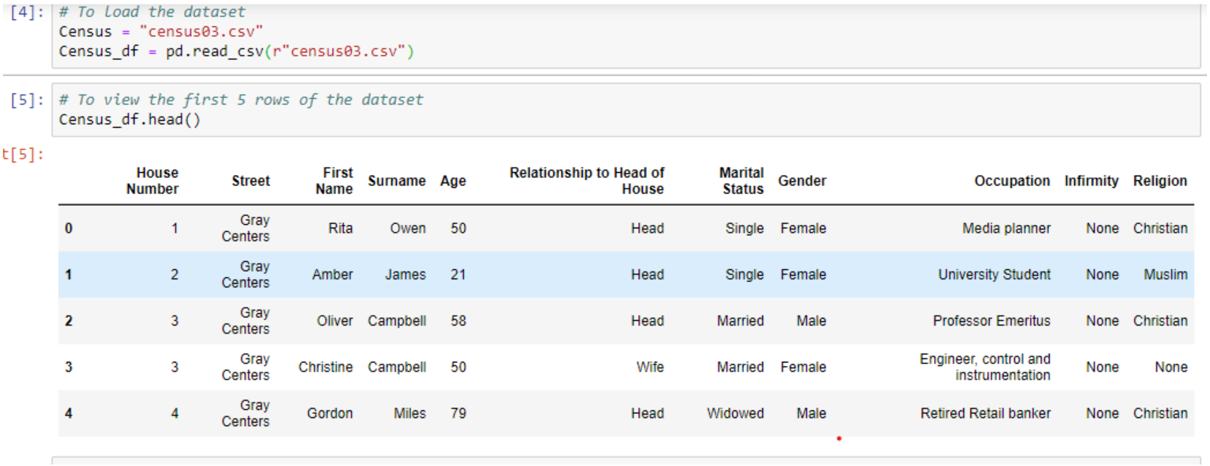
In order to ensure sufficient quality for analysis, the first step was to read the data into the IDE of choice and investigate it using the relevant tools. This is where most of the project's time was spent. It was found that the data, contained several errors which will be cleaned thoroughly.

To facilitate in-depth investigation and interpretation, visualizations were created after data cleansing. Data-driven suggestions were provided to the local government regarding the optimal use for the vacant parcel of land and additional services that ought to be funded based on insights gained from the study.

2.0 Data Preparation and Cleaning

This section includes the dataset's data exploration processes as well as the cleaning procedure used to produce a cleaned dataset.

On reading in the file, the head function was used to check the datasets features and replaced the spaces in the names with an underscore (_) as seen in the picture extract below.



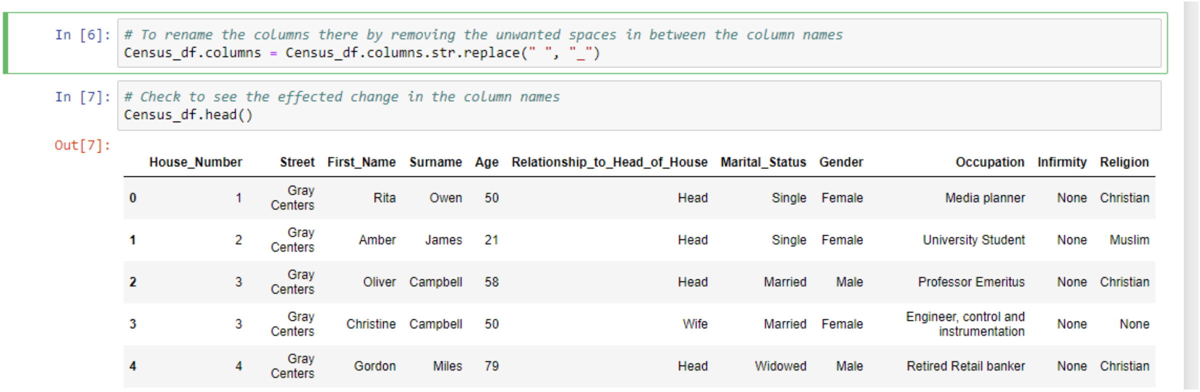
```
[4]: # To load the dataset
     Census = "census03.csv"
     Census_df = pd.read_csv(r"census03.csv")
```

```
[5]: # To view the first 5 rows of the dataset
     Census_df.head()
```

t[5]:

	House Number	Street	First Name	Surname	Age	Relationship to Head of House	Marital Status	Gender	Occupation	Infirmity	Religion
0	1	Gray Centers	Rita	Owen	50	Head	Single	Female	Media planner	None	Christian
1	2	Gray Centers	Amber	James	21	Head	Single	Female	University Student	None	Muslim
2	3	Gray Centers	Oliver	Campbell	58	Head	Married	Male	Professor Emeritus	None	Christian
3	3	Gray Centers	Christine	Campbell	50	Wife	Married	Female	Engineer, control and instrumentation	None	None
4	4	Gray Centers	Gordon	Miles	79	Head	Widowed	Male	Retired Retail banker	None	Christian

Fig. 1: A view of the dataset after loading



```
In [6]: # To rename the columns there by removing the unwanted spaces in between the column names
        Census_df.columns = Census_df.columns.str.replace(" ", "_")
```

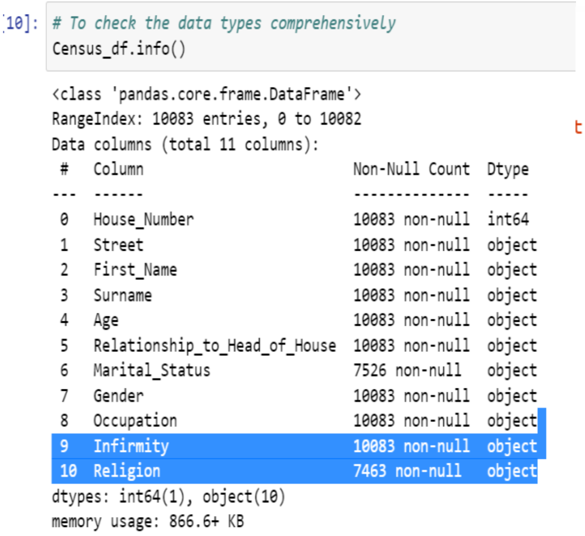
```
In [7]: # Check to see the effected change in the column names
        Census_df.head()
```

Out[7]:

	House_Number	Street	First_Name	Surname	Age	Relationship_to_Head_of_House	Marital_Status	Gender	Occupation	Infirmity	Religion
0	1	Gray Centers	Rita	Owen	50	Head	Single	Female	Media planner	None	Christian
1	2	Gray Centers	Amber	James	21	Head	Single	Female	University Student	None	Muslim
2	3	Gray Centers	Oliver	Campbell	58	Head	Married	Male	Professor Emeritus	None	Christian
3	3	Gray Centers	Christine	Campbell	50	Wife	Married	Female	Engineer, control and instrumentation	None	None
4	4	Gray Centers	Gordon	Miles	79	Head	Widowed	Male	Retired Retail banker	None	Christian

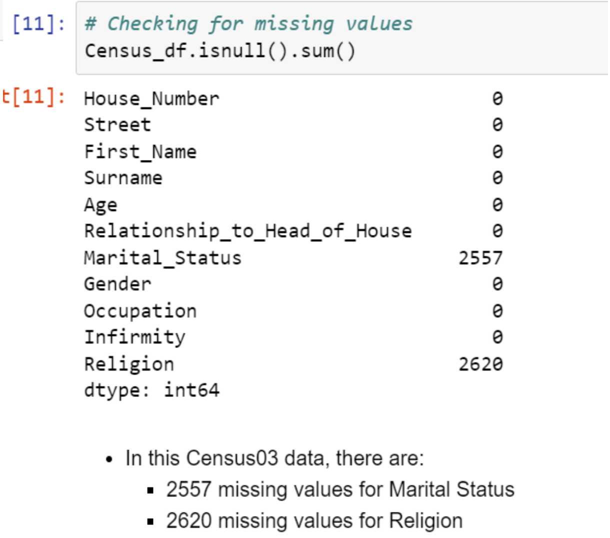
Fig 2: showing Column name change and the first five rows

Random columns were selected using the sample function to check/ familiarize with the different contents in the output. The info function was used to determine the data types at a glance. The census data contains eleven (11) columns which related to issues like incorrect casting of the data types, empty strings, null values as described with the (.info) function.



```
10]: # To check the data types comprehensively
     Census_df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 10083 entries, 0 to 10082
     Data columns (total 11 columns):
      #   Column                         Non-Null Count  Dtype
     ---  ------                         --------------  -----
      0   House_Number                   10083 non-null  int64
      1   Street                         10083 non-null  object
      2   First_Name                     10083 non-null  object
      3   Surname                        10083 non-null  object
      4   Age                            10083 non-null  object
      5   Relationship_to_Head_of_House  10083 non-null  object
      6   Marital_Status                 7526 non-null   object
      7   Gender                         10083 non-null  object
      8   Occupation                     10083 non-null  object
      9   Infirmity                      10083 non-null  object
      10  Religion                       7463 non-null   object
     dtypes: int64(1), object(10)
     memory usage: 866.6+ KB
```



```
[11]: # Checking for missing values
      Census_df.isnull().sum()

t[11]: House_Number                        0
       Street                              0
       First_Name                          0
       Surname                             0
       Age                                 0
       Relationship_to_Head_of_House       0
       Marital_Status                   2557
       Gender                              0
       Occupation                          0
       Infirmity                           0
       Religion                         2620
       dtype: int64
```

- In this Census03 data, there are:
  - 2557 missing values for Marital Status
  - 2620 missing values for Religion

Fig. 3: A view of datatypes

Fig. 4: A view of the null features

The Age column was casted as type Object even though it contained floats and integers. This was changed initially to a float and subsequently casted to its appropriate type using the lambda function. Although some of the ages was converted to floats, it returned an error that strings cannot be converted to floats.

Apparently, there were some empty strings contained which was initially replaced with nan to enable the calculation of the descriptive statistics after which the ages were replaced with the median value instead of the mean due to the presence of outliers that made the mean value larger than the median. Before these ages were replaced, their marital status and relationship to head of house was considered.

Also, the Infirmity column contains some empty strings which was replaced with the most occurring feature which is “None”. This column was re-categorized to Disabled and Nondisabled.

The Relationship to the Head of House contains twenty-one unique entries with no empty strings and Nan values. Neice was recorded instead of Niece, this was corrected.

The marital status column does not contain any empty strings but with 2557 Nan values. It was discovered from analysis that all the nan entries were for individuals who were less than 17 years of age and not of the legal age to get married, so their marital status was replaced with ‘Minors’.

Also, age and marital status was conditioned on each other to determine if there was anyone less than 18 years and married. There is a sixteen (16) year old who was married and from in-depth analysis to discover if it was error, but it was discovered that she was married to a 19-year-old and has a child already.

The gender comprises two unique arrays “Male and Female” with neither a Nan value nor an empty string.

The house number contains several unique values with no empty strings or Nan values. This was later merged with the street column to form an address to retrieve occupancy level.

There are quite a lot of entries in the occupation column which was grouped according to their common names e.g., all occupations containing teacher was assigned as Teacher likewise; Retired Engineer, Analyst, Consultant, Lecturer, Student for those less than 17 years (i.e., Primary and Secondary), etc. While others with ages higher than 18 with occupation slated as university student or PHD were grouped as university student.

There were no empty strings in occupation but a single Nan value. The occupation of this individual could not be determined from the relationship to the head of house hence the need to replace with the mode. The mode of the Occupation is Student which has already been conditioned to ages17 and below, but the age associated with the Nan is 41 years hence the Nan value was replaced with ‘No Occupation’ because dropping that row will mean dropping data from other features. All occupations were later re-categorized into Employed, Unemployed and Retired.

There were some nonexistence religions in the census dataset like Sith, Nope and None which was converted to None. There was also an empty string and Nan values which was traced to the religion of other members and assigned. Religion was later re-categorized as Religious and Non-Religious.

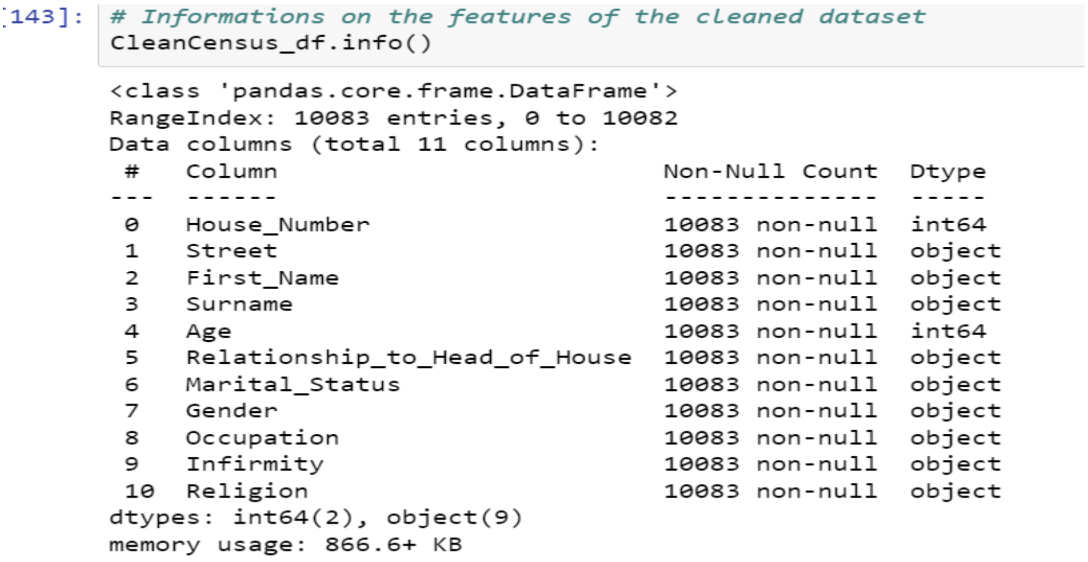
The first name column was also cleaned and there was no way an individual's first name could be determined from other features and was replaced with no first name.

The surname column also contained empty strings which was traced to members of the family and assigned the household name.

Following data cleaning, visualizations were made, and statistical methods were applied to aid detailed analysis and interpretation. Insights on the project were generated and data-led recommendations were made to the local government on the best use for the unoccupied plot of land, and what other services should be invested in.

2.1 DESCRIPTIVE ANALYSIS

After the data has been cleaned, the data now has the following features:

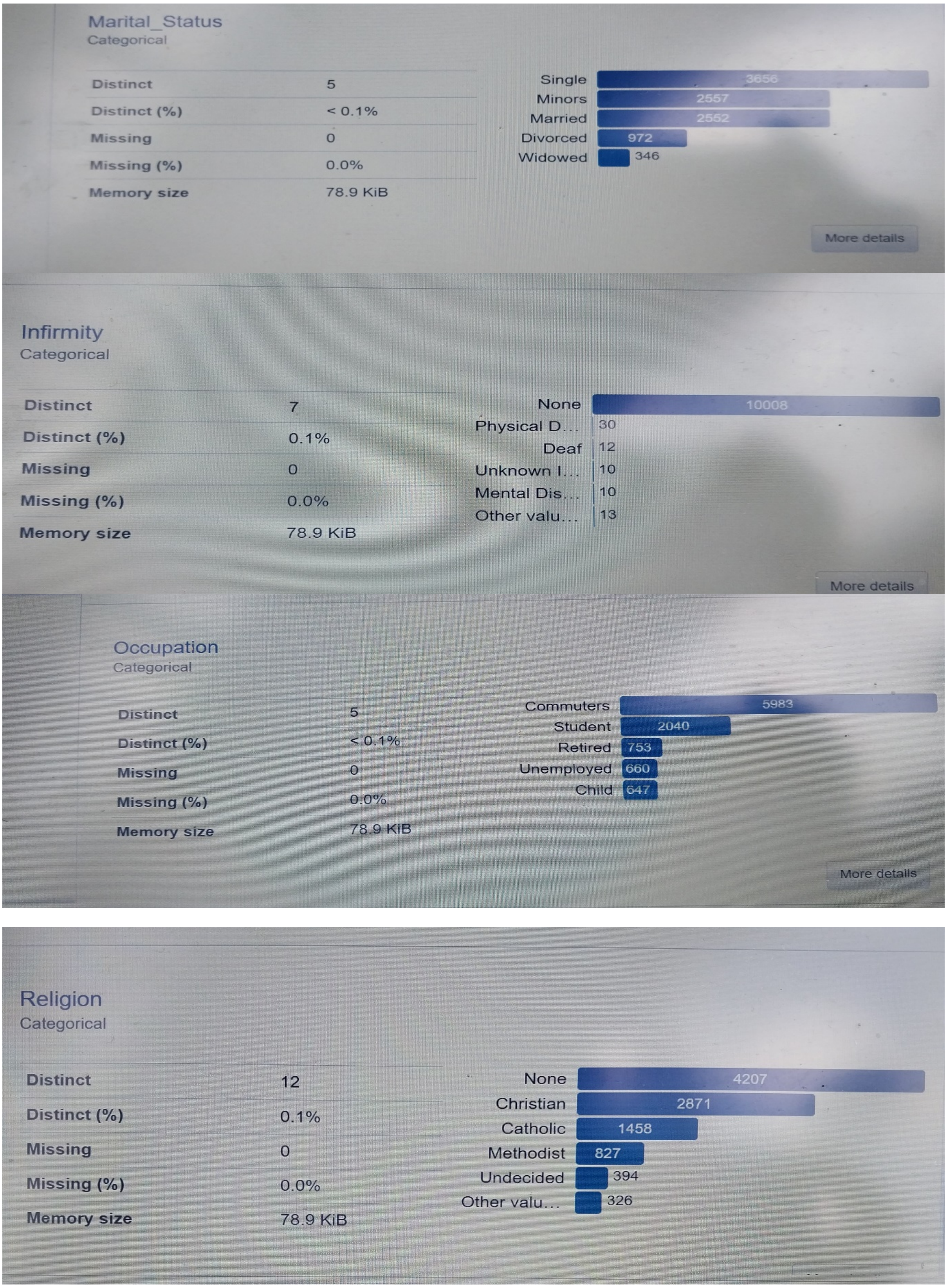


```
[143]: # Informations on the features of the cleaned dataset
        CleanCensus_df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 10083 entries, 0 to 10082
        Data columns (total 11 columns):
         #   Column                         Non-Null Count  Dtype
        ---  ------                         --------------  -----
         0   House_Number                   10083 non-null  int64
         1   Street                         10083 non-null  object
         2   First_Name                     10083 non-null  object
         3   Surname                        10083 non-null  object
         4   Age                            10083 non-null  int64
         5   Relationship_to_Head_of_House  10083 non-null  object
         6   Marital_Status                 10083 non-null  object
         7   Gender                         10083 non-null  object
         8   Occupation                     10083 non-null  object
         9   Infirmity                      10083 non-null  object
         10  Religion                       10083 non-null  object
        dtypes: int64(2), object(9)
        memory usage: 866.6+ KB
```

Fig. 5: Information on the cleaned data

Given the statistics shown below, the general population is healthily represented by less than 1% infirmity rate, a highly irreligious town where 41.7% of the population do not identify with any religion. Most of the population are employed, there are high school children and an unemployment rate of 6%. Also, most of the population is either single, married or minors.



Marital_Status
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.9 KiB

Single	3656
Minors	2557
Married	2552
Divorced	972
Widowed	346

Infirmity
Categorical

Distinct	7
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.9 KiB

None	10008
Physical D...	30
Deaf	12
Unknown I...	10
Mental Dis...	10
Other valu...	13

Occupation
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.9 KiB

Commuters	5983
Student	2040
Retired	753
Unemployed	660
Child	647

Religion
Categorical

Distinct	12
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	78.9 KiB

None	4207
Christian	2871
Catholic	1458
Methodist	827
Undecided	394
Other valu...	326

Fig. 6: Statistical overview

3.0 DETAILED ANALYSIS

3.1 Age Population Pyramid

From the age pyramid of the town, the structure of the population shows that the population of the younger people, particularly those between 0-4 years, seems lesser than that of middle age which suggests a low birth rate. Also, the population seems to grow well into old age. The population increases for age band 15-20, 35- 40 and 40- 44, these are as a result of migration into the town for new university students and lodgers who moved into the town as a result of employment respectively. The decrease between ages 20-25 could be attributed to graduating students leaving the town.

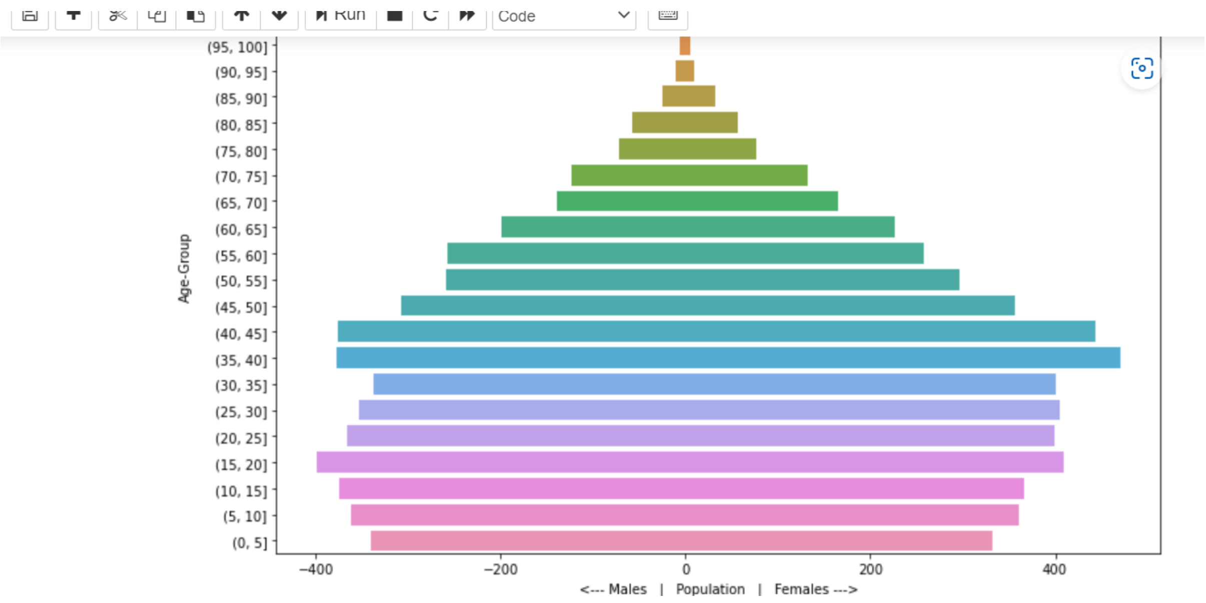


Fig. 7 Age Pyramid of the Town

3.2 Fertility rate

The General fertility rate was used to estimate the fertility rate of the town. This was calculated by dividing the total number of live births by the total number of women of childbearing age per 1000 (the OECD. 2021). This was the basis for calculation, and it was observed that the fertility rate was 40 per thousand, while 4 years before, it was 55 per thousand. This indicates a decrease in the fertility rate in the town. The total fertility rate was also calculated by summing all the age-specific fertility rates as 2.1 children per woman of childbearing age.

$$\text{GFR} = \frac{\text{Number of live births in an area}}{\text{Mid year female population age 15-44 (or 49) in the same area in same year}} \times 1000$$

3.3 Crude Birth rate and Death Rate

The crude birth rate is the number of resident live births for a specified geographic area during a specified period (usually a calendar year) divided by the total population (usually mid-year) for that area and multiplied by 1,000 (DOH,2021).

$$\frac{\text{\# deaths in 1 year}}{\text{\# thousand total population}} = \text{Crude Death Rate}$$

$$\frac{\text{\# births in 1 year}}{\text{\# thousand total population}} = \text{Crude Birth Rate}$$

The crude birth rate for the town is 12 births per thousand. Five years prior, the crude birth rate was estimated at 15 births per thousand. The birth rate has therefore fallen by 5 children per thousand in five years.

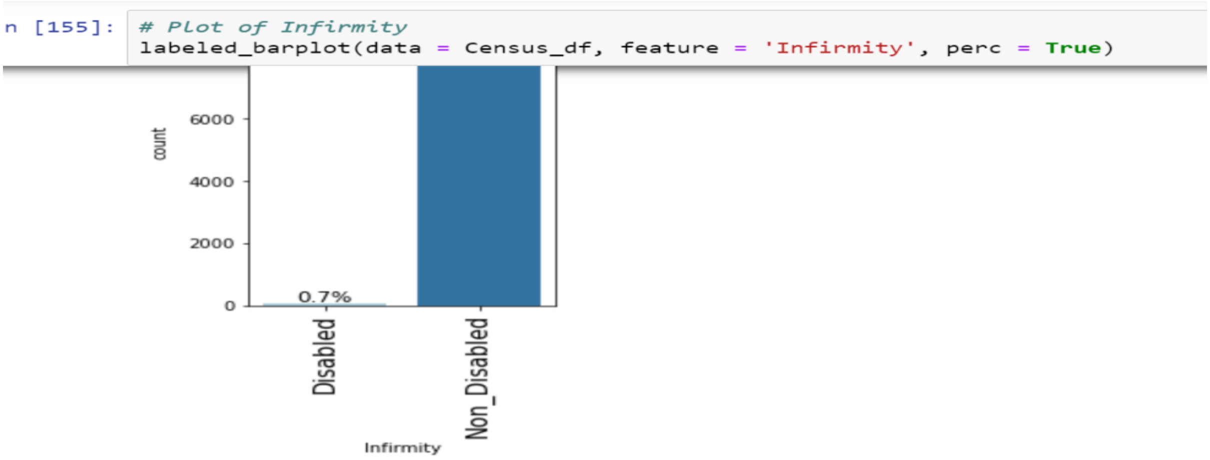
When assessing mortality for people above 55, the death rate was determined by the disparity between the age bands. You can see from the age pyramid that some age groups have decreased, which migration is to be blamed for. For individuals over 55, a decline suggests that they may be dying, nevertheless. Six fatalities per 1,000 people were estimated to be the mortality rate. This was determined by adding the annual death rates for each age group above 55 years old.

Although there were 10 births this year as opposed to fifteen five years ago, this indicates that the population is still expanding when compared to the death rate for the year.

3.4 Infirmity

Less than 1% of the population is infirm, which shows that the population is relatively healthy. The percentage of the population over 55 who are infirm was estimated to be 0.1%. These show that the town has a top-notch healthcare system and an extended care program for the senior population.

Based on this proof of the population's health, the plot of land would not prioritize the construction of a new health care system.



```
n [155]:  # Plot of Infirmity
          labeled_barplot(data = Census_df, feature = 'Infirmity', perc = True)
```

Fig. 9: A plot of Infirmity

3.5 Religion

From the data, it is seen that a larger proportion of the population (41.3%) do not have a religion, this is followed by Christians making up of the religion. Based on a survey in the Guardian 2021, where the number of irreligious people has been on the increase and Christianity reducing, we expect the number of people identifying their religion as "None" to increase and that of Christians to decrease. The Christians consist of different denominations lumped together giving it a high count.

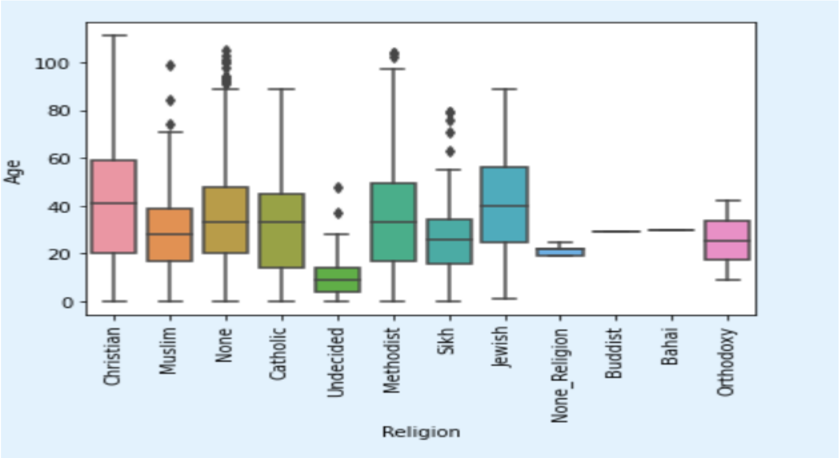
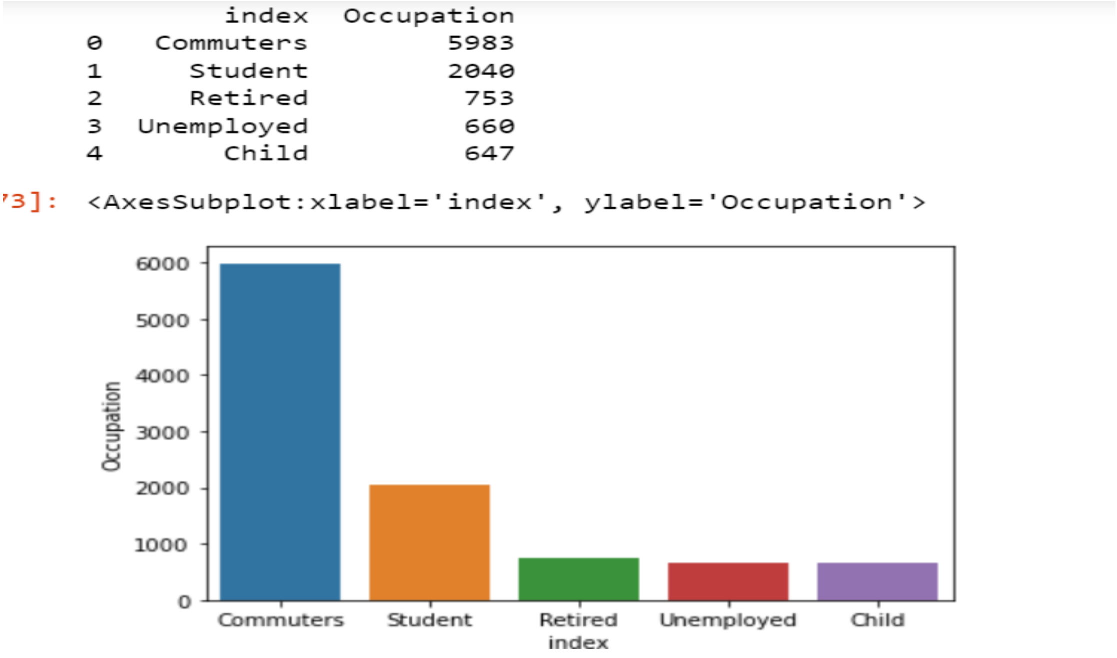


Fig. 10: Plot of Age against Gender

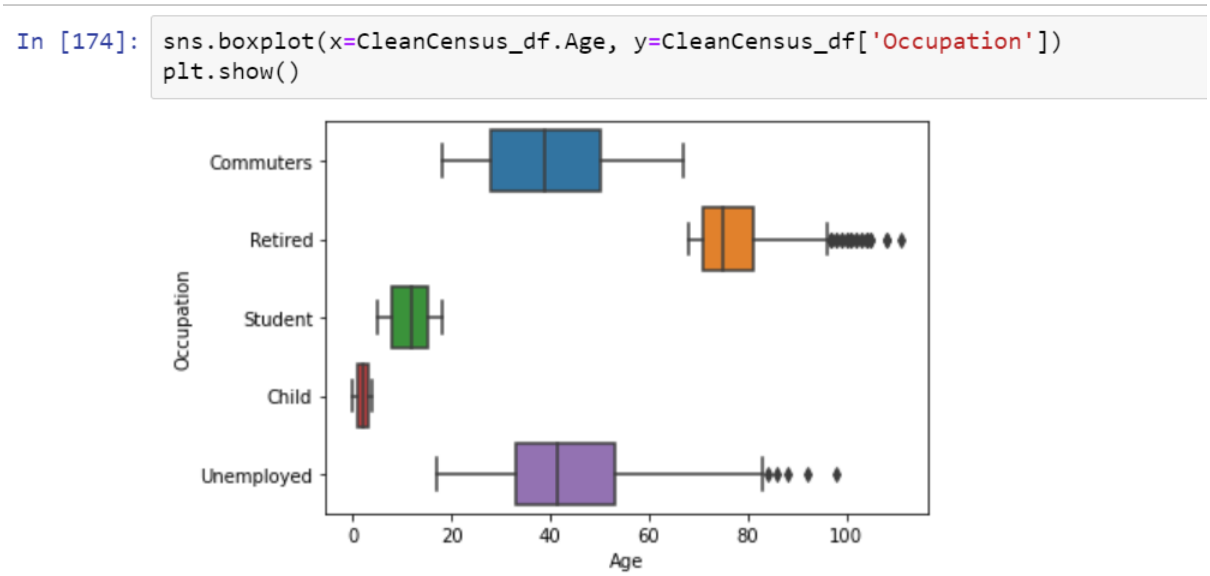
3.6 Commuters

From the data, it was deduced that the retirement age was 70 years and the majority in their active years were gainfully employed. Commuters were identified as; University students who attend school in the nearby city, Employees who go to work. This means over 75% of the employed population commute to work.



```
         index  Occupation
0    Commuters        5983
1      Student        2040
2      Retired         753
3   Unemployed         660
4        Child         647
```

```
'3]:  <AxesSubplot:xlabel='index', ylabel='Occupation'>
```

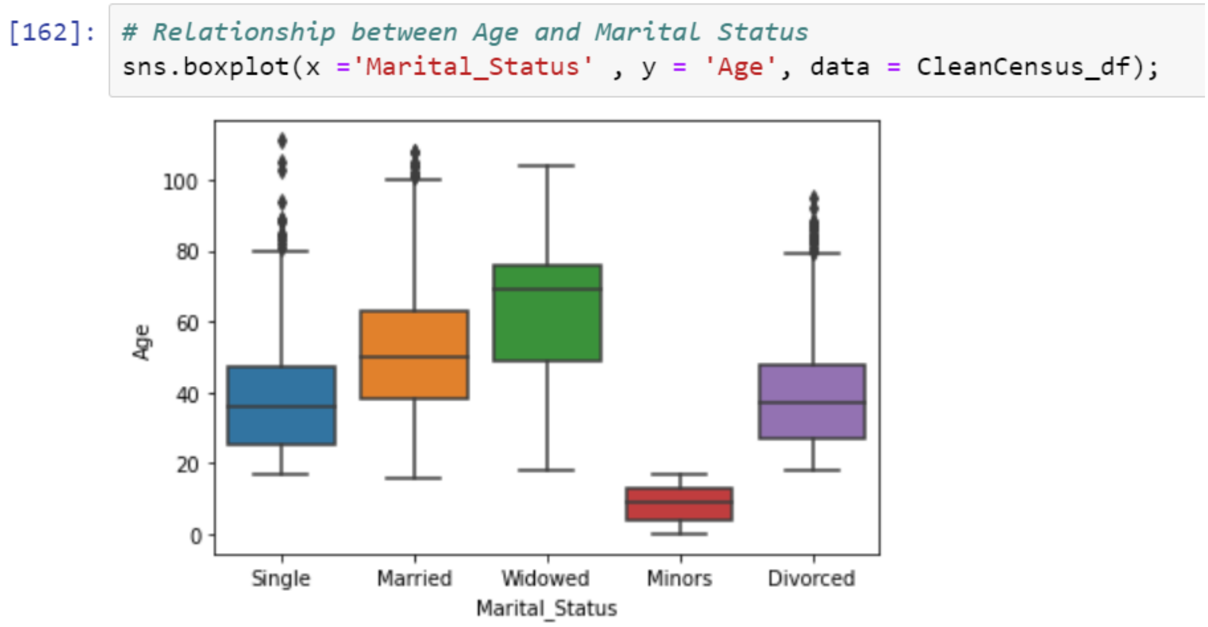


```
In [174]:  sns.boxplot(x=CleanCensus_df.Age, y=CleanCensus_df['Occupation'])
           plt.show()
```

Fig. 11 & 12: Plot of Commuters

3.7 Divorce and Marriage

As seen from the analysis, divorce occurs through all marriageable ages from young to old. From the data set, there are more female divorcees than the male which suggests that the male divorcees leave the town. The divorce to marriage ratio is 1:3. The crude divorce rate was calculated by dividing the number of divorces by the total population and multiplied by 1000. This was estimated to be 94 divorcees per 1000.



```
[162]: # Relationship between Age and Marital Status
       sns.boxplot(x ='Marital_Status' , y = 'Age', data = CleanCensus_df);
```

Fig. 13: Plot of Age on Marital Status

3.8 Occupancy

The occupancy level was calculated by concatenating the house number with the street. The mode of "House Number" per street which represents the average occupancy for each street. This is with the assumption that all houses on a street are built similarly and has the same number of bedrooms. Based on this, it was discovered that, of the 3,105 apartments in the town, 2,362 houses were under occupied with occupants less than 5 individuals which made up 76% of the housing population.

256 were over-occupied with occupants higher than 5. This represents 8% of the apartments in the town. The source of this could be attributed to the fact that not all the commuters do sleep over in the town.

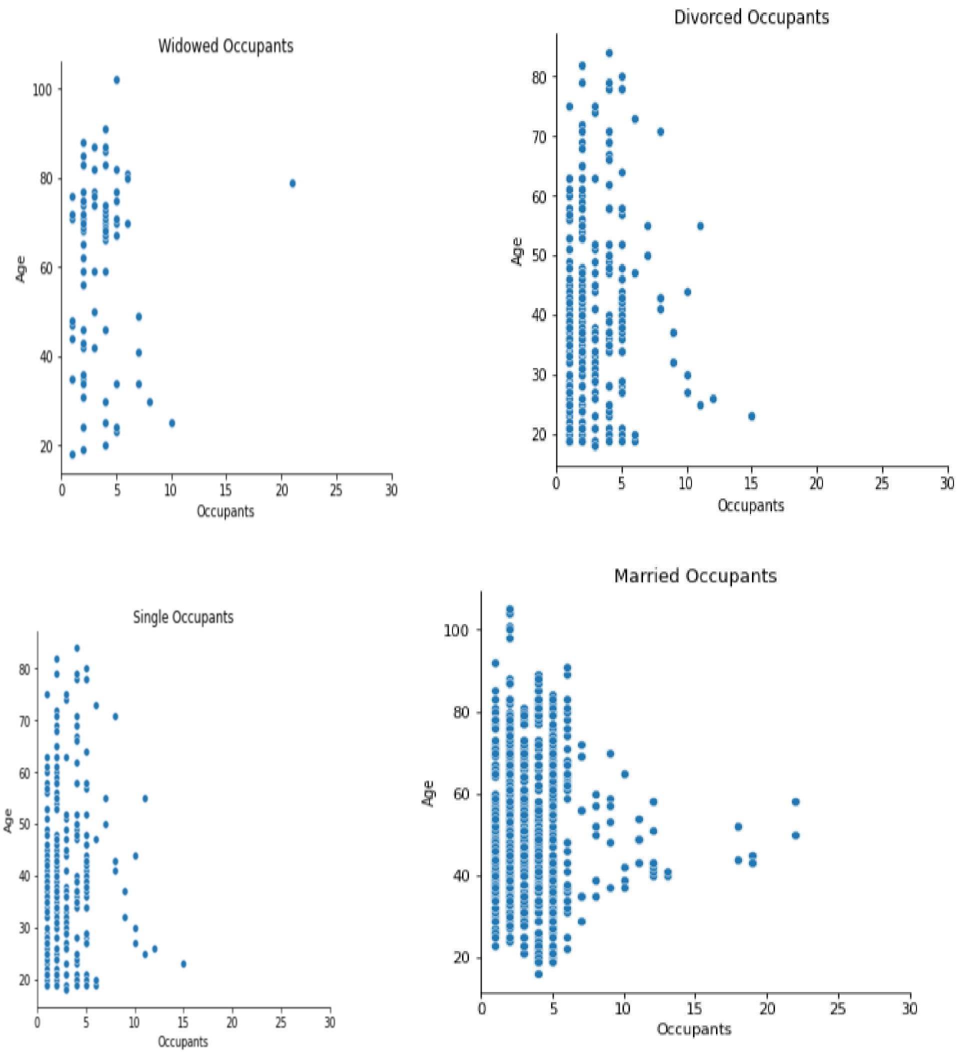


Fig. 14: Plot of Age on Occupants

Although there are extreme occupancy rates that are above 10 with a maximum number of people in a household as 22 and the potential reasons for this could be more people renting their homes to lodgers, those who are into agriculture might have their farm workers living with them and this shows that the government needs to invest in housing as there is likely more lodgers will move into the town, the single people will get married and have more children.

3.9 Migration

Migration was calculated based on the number of visitors (lodgers) in the town and divorcees who have left the town. Students make up many visitors in the town, but they are replaced yearly which doesn't have much effect on immigration. For divorcees, the difference between male and female divorcees could be attributed to more male divorces leaving the town compared to female divorcees. The number of lodgers who are not university students was used to compute as those who immigrated into the town and is about 5.5% of the population. Matching the figure with corresponding birth and death rate signifies that the population is increasing. These would lead to an increasing pressure on existing infrastructural facilities.

4.0 RECOMMENDATIONS AND CONCLUSIONS

The population of the town increased based on the data available, and this will create additional stress on existing infrastructures. Most affected would-be housing and transportation. Based on the high number of immigrants and higher number of commuters, there is a need to invest in low-density housing and a train station. My recommendation based on the data would be to prioritize the building on low-density housing to reduce the pressure on existing housing as this would benefit more of the population than the train station in the interim. Proper maintenance of existing means of transport could assist with transportation challenges till housing issues are solved. Building a religious house was not considered important due to the high number of irreligious populations which might increase in the future. Also with a decreasing birth rate, very low infirmity, a new emergency medical building is not needed. Given the high rate of immigration, it is important to invest in general infrastructures to support the existing ones. This would reduce maintenance costs and serve as a source of income to the authorities. Despite the large number of working populations which would age in the future, investing in old age care was not considered a priority. This is because based on the infirmity rate and death rate, there is a higher chance of the population dying than getting sick at old age. Schooling was not considered important due to the decline in birth and fertility rate. The population is largely employed and there's no urgent need for training.

REFERENCES

Census 2021. On 21 March 2021, what is your legal marital or registered civil partnership status? - Census 2021. [Online] Available at: https://census.gov.uk/help/how-to-answer-questions/paper-questions-help/on-21- march-2021-what-is-your-legal-marital-or-registered-civil-partnership-status [Accessed 22 November 2021].

Marriage Act (1949) Section 3 Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3 [Accessed 05/12/2021]

Schürer, K., Garrett, E.M., Jaadla, H. and Reid, A. (2018). Household and family structure in England and Wales (1851–1911): continuities and change. Continuity and Change, 33(3), pp.365–411. doi:10.1017/s0268416018000243.

theOECD. 2021. Demography - Fertility rates - OECD Data. [online] Available at: https://data.oecd.org/pop/fertility-rates.htm [Accessed 27 November 2021]. Whitney, Craig R. (5 August 1997). "Jeanne Calment, World's Elder, Dies at 122". The New York Times. ISSN 0362-4331. Www-doh.state.nj.us. 2021. [online] Available at: https://www-doh.state.nj.us/dohshad/view/sharedstatic/CrudeBirthRate.pdf [Accessed 25 November 2021]