Census Project Report

Data Analysis

This Project report analyses the population census of a town situated between two larger cities, to formulate decisions on future investments and decide necessary developments on an unoccupied plot of land. To formulate decisions based on the census, the data needs to be cleaned properly to rectify the errors and missing values.

The following sections of the report will highlight the main analyses carried out specifically for this purpose.

All the header fields are analyzed with respect to the data types, count, unique values, frequency to deeply understand the given data. These data are then cleaned to create visualizations which helps in the process of decision making. The following sections of the report will highlight the key analyses that were undertaken specifically to support the decision making.

Data Cleaning

The given Census data was cleaned up, to correct data errors and missing values. A complete record of all cleanup work undertaken can be accessed in the corresponding Jupyter Notebook.

The Blank values and some incorrect format entries were made into the House Number field. Some of the missing values were imputed by inferring the information of other members in the same household and those which cant be referenced, were imputed to number which is not present in the list. Since the house number should be integer, the data type is changed to int.

Using the same referential information technique, the blank fields in the First Name & Surname were imputed.

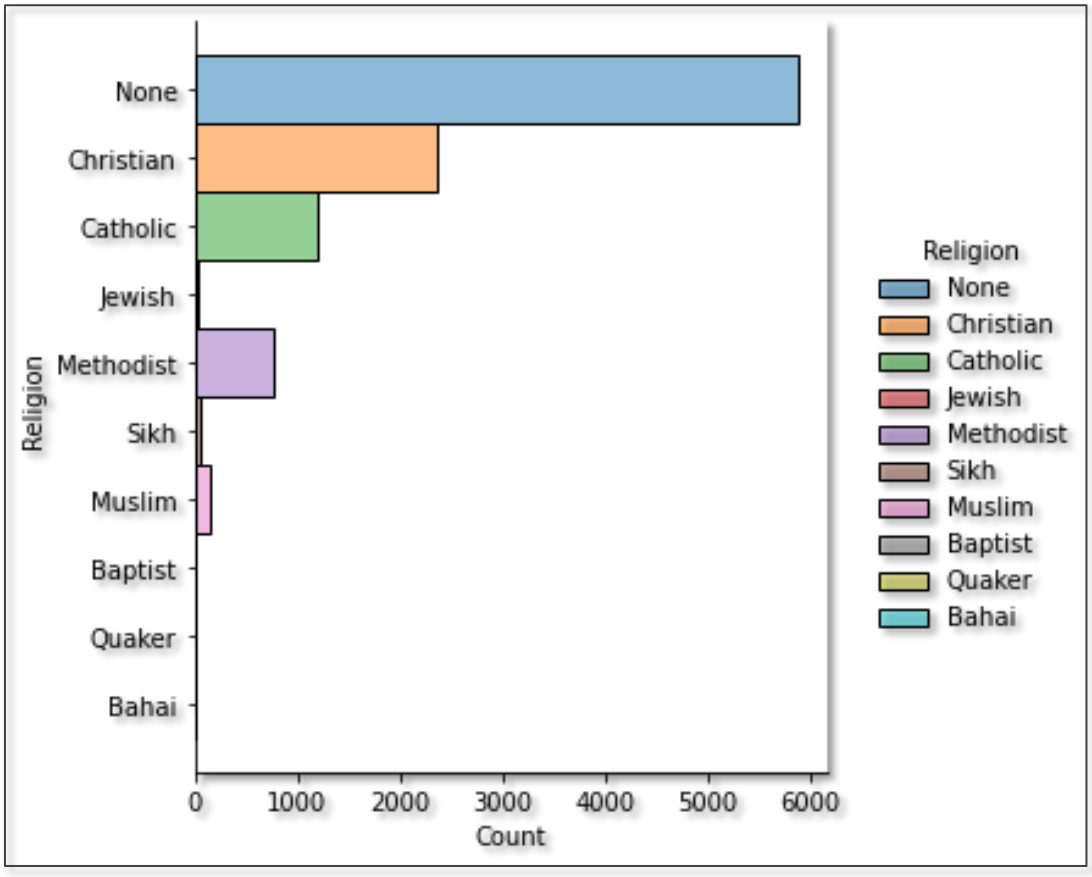
For the Age columns, the format of the given data wasn't same (like eight, six, 72.36415261) these all were imputed into single integer format for better understanding and visualizations. The blank fields were imputed to meaningful value (ie., 65) since the corresponding occupation contains Retired.

The format for the Marital Status were simplified into four groups (Single, Married, Divorced, Widowed). All blank fields are imputed to Single. The marital status for the individuals below 18 years of age were imputed to Single. For a particular blank entry, were imputed to Single even the age is above 18 years since he was the only one in the household and were a student.

Similarly, there are no records of people under the age of 18 have Married as a marital status. Even though an exception has been made for those 16 or above, as it is legal to marry with parental consent (Marriage Act, 1949:s3).

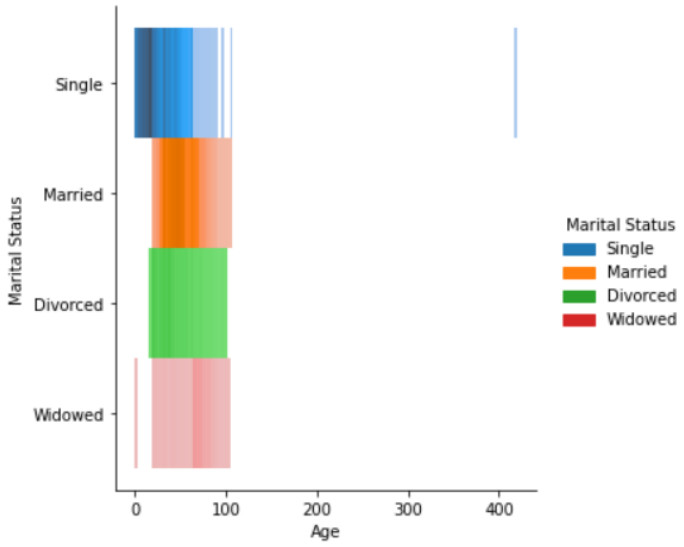
One household data is found to be lies from the data set. As it contains, a child of 16 were divorced, listed as Head of the House and had a child aged 0. There are three major inconsistencies present over here: One cannot be Head of the household until 18years and having a child at this age. There are no parents present in the household so as per the exception it would be illegal. Since it was not deemed significant (two records) to the overall analysis, the data were skipped.

Three religions were imputed to None – Undecided, Private, Nope. These are provenly individual errors. Sith were misleading information given in the data since it is not a religion. Since there is a chance of typo, Sith were imputed to Sikh. The estimation of people who didn't follow any religion is increased [2011 census]. Also like the previous census the people who stated Christian as religion is increased [2011 census].



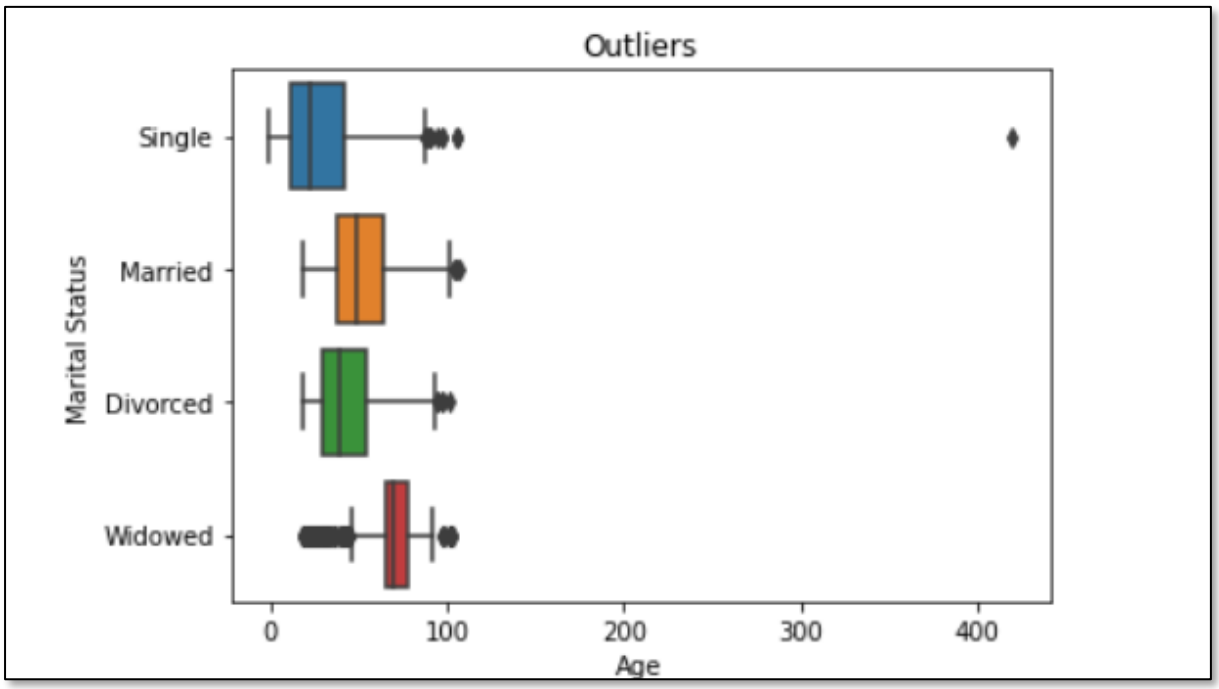
Outliers were removed for the individual of aged 420 (imputed to match the son of the same family, thinking it has typo error while entering the info) and similarly for the child aged -1 (imputed to match the daughter of the same family) and two individuals widowed at age 18.

Out[87]: <seaborn.axisgrid.FacetGrid at 0x2887c4f7fa0>




```
data['Age'].value_counts()
```

```
37     201
43     178
46     176
16     172
19     170
      ...
100      1
106      1
-1       1
420      1
107      1
Name: Age, Length: 109, dtype: int64
```



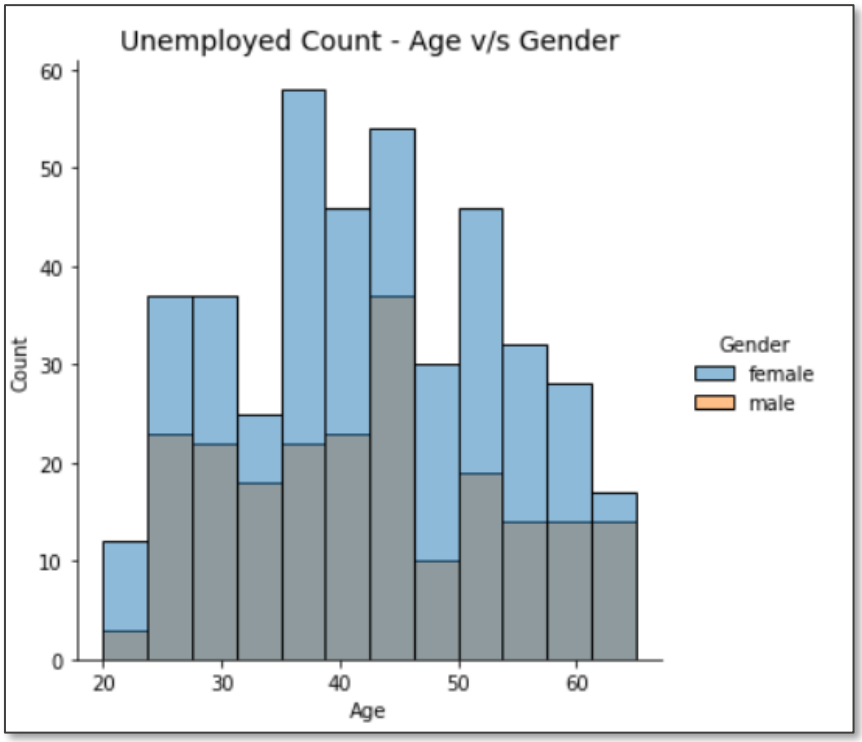
```
data['Marital Status'].value_counts()
```

```
Single      5939
Married     3000
Divorced     972
Widowed      529
Name: Marital Status, dtype: int64
```

It is uncommon for people to become widowed in age 18. Therefore, these records were considered too unlikely to be correct and imputed to ‘Single’ for the two records.

Data in Gender header comes with different format such as F or F-male for Female. So, these genders were imputed into single format like Male, Female and Not disclosed for the individuals who left the field blank (which means they aren’t interested to expose their gender identity). The two infirmity were imputed to None: Female and blank values.

The records, which indicated that the occupation was "unemployed" for individuals over the age of 65, were imputed as "Retired" as those over 65 would not usually be eligible for work and won’t be considered as unemployment analysis (GOV.UK). The blank fields in the occupation records were imputed based on their age. For the age < 5, the occupation would be imputed as ‘Child’. For the age<18, the occupation would be imputed as ‘Student’. For the age<30 the occupation would be imputed as ‘University Student’ and for the age>30 the occupation would be ‘Unemployed’.



Population Demographics:

Once data cleaning is completed, we will get the finalised census data with the following features:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10440 entries, 0 to 10439
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   House Number                    10440 non-null  int32
 1   Street                          10440 non-null  object
 2   First Name                      10440 non-null  object
 3   Surname                         10440 non-null  object
 4   Age                             10440 non-null  int64
 5   Relationship to Head of House   10440 non-null  object
 6   Marital Status                  10440 non-null  object
 7   Gender                          10440 non-null  object
 8   Occupation                      10440 non-null  object
 9   Infirmity                       10440 non-null  object
 10  Religion                        10440 non-null  object
dtypes: int32(1), int64(1), object(9)
memory usage: 856.5+ KB
```

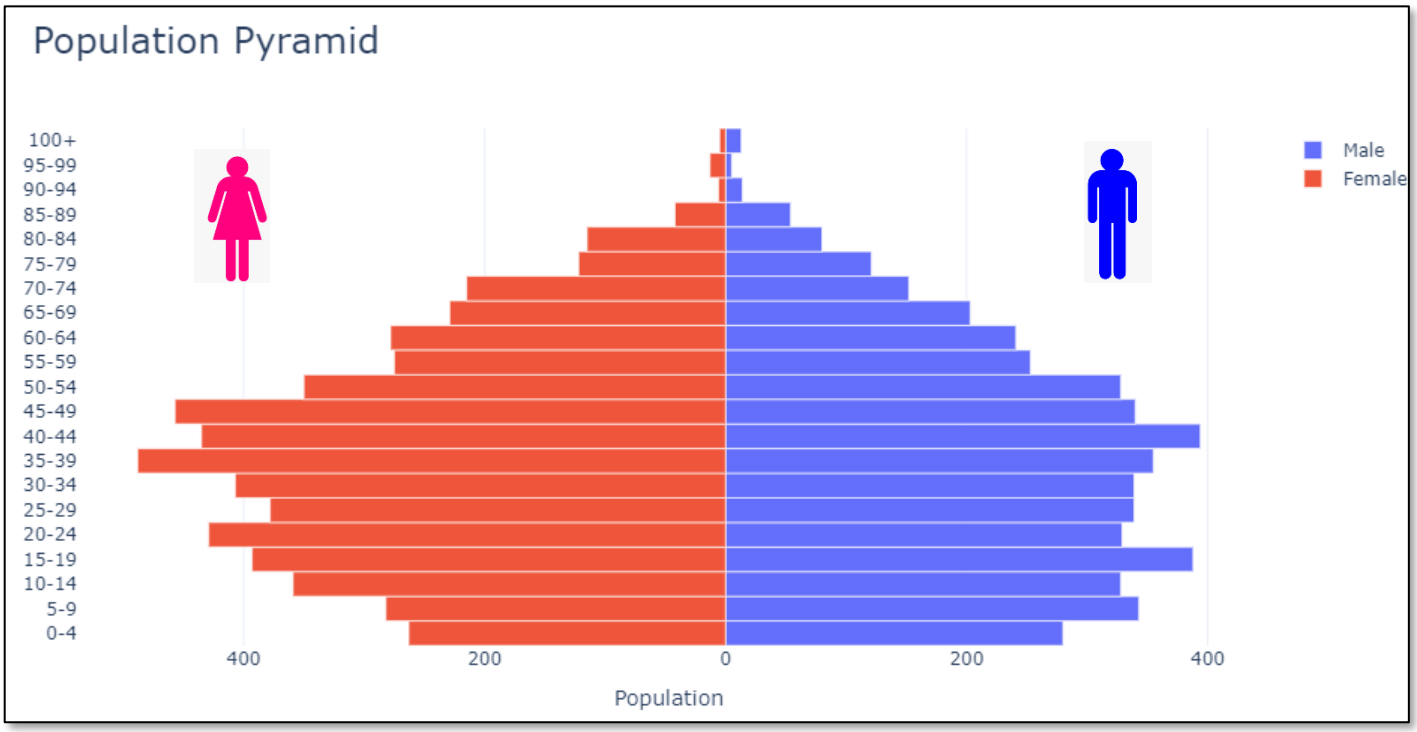
	House Number	Street	First Name	Surname	Age	Relationship to Head of House	Marital Status	Gender	Occupation	Infirmity	Religion
count	10440.000000	10440	10440	10440	10440.000000	10440	10440	10440	10440	10440	10440
unique	NaN	104	371	673	NaN	23	4	3	1154	7	10
top	NaN	Thomas Avenue	Katie	Smith	NaN	Head	Single	Female	Student	None	None
freq	NaN	920	47	265	NaN	4251	5939	5541	1895	10362	5871
mean	58.241092	NaN	NaN	NaN	37.510536	NaN	NaN	NaN	NaN	NaN	NaN
std	61.088536	NaN	NaN	NaN	22.053013	NaN	NaN	NaN	NaN	NaN	NaN
min	1.000000	NaN	NaN	NaN	0.000000	NaN	NaN	NaN	NaN	NaN	NaN
25%	12.000000	NaN	NaN	NaN	19.000000	NaN	NaN	NaN	NaN	NaN	NaN
50%	31.000000	NaN	NaN	NaN	37.000000	NaN	NaN	NaN	NaN	NaN	NaN
75%	91.000000	NaN	NaN	NaN	53.000000	NaN	NaN	NaN	NaN	NaN	NaN
max	242.000000	NaN	NaN	NaN	107.000000	NaN	NaN	NaN	NaN	NaN	NaN

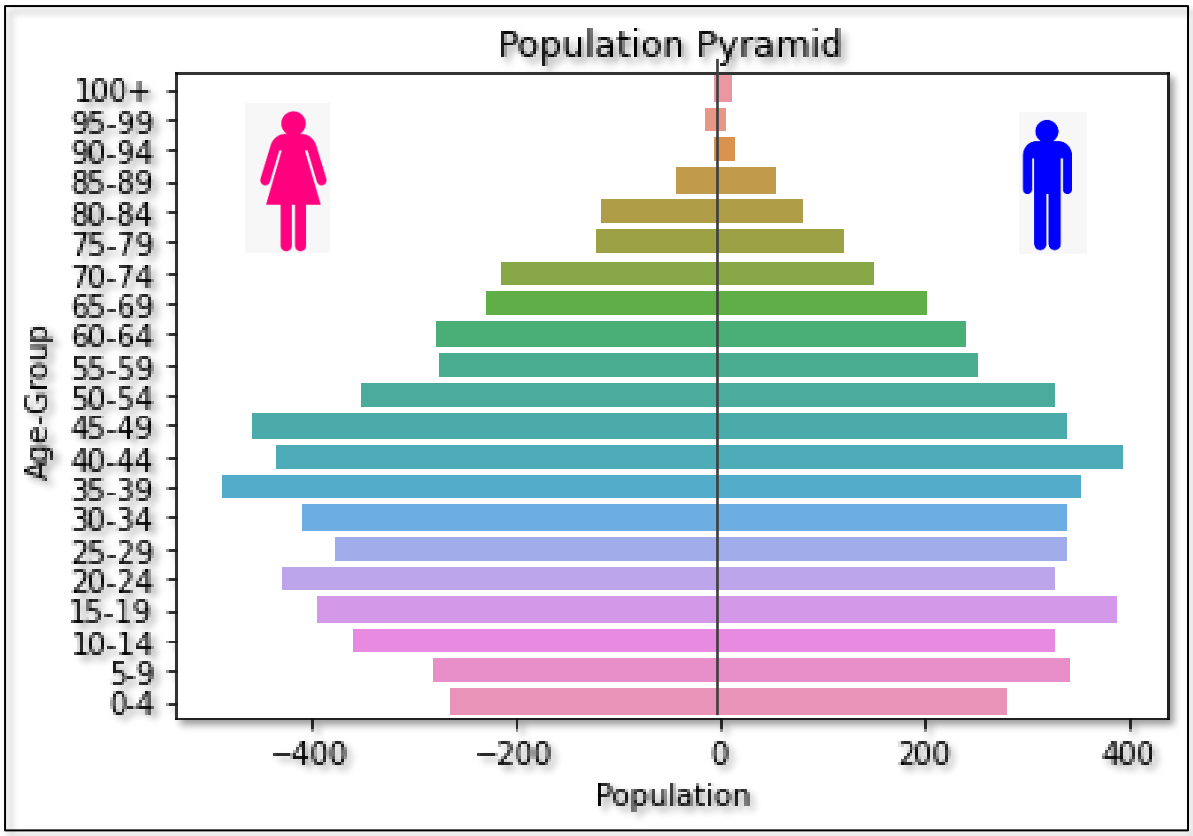
And final check for the blank values present after the data cleaning

```
print(data.isnull().sum())
```

```
House Number                    0
Street                          0
First Name                      0
Surname                         0
Age                             0
Relationship to Head of House   0
Marital Status                  0
Gender                          0
Occupation                      0
Infirmity                       0
Religion                        0
dtype: int64
```

Population Pyramid:





In the above population pyramid, Population is depicted on the horizontal X axis and Age is aligned on Vertical Y axis. The resultant plot is a series of bars, each representing an age category (5-year age groups). The youngest age (0-4) is represented by the bottom bar and oldest age is represented by upper bar.

Examining the above population pyramid, there exist lower mortality rate with fertility rate remaining constant which makes population pyramid wider in the middle of the population has highest number of middle-aged people but fewer birth rates.

Also, in the middle and older age groups (40+) in the population, the number of females is greater than the number of males, this is reflected in the shape of the pyramid as the bars on the left-hand side(female) is longer than the bars on the right-hand side(male).

Further descriptive analysis shows that Most of the population are single or married and infirmity rate in the town is low also most of the population are employed.

All these analysis shows that there would be a need of medical and age services in future.

Infirmity:

```
None                    10362
Physical Disability        18
Disabled                   16
Mental Disability          14
Blind                      14
Deaf                       12
Unknown Infection           4
Name: Infirmity, dtype: int64
```

Religion:

```
None          5871
Christian     2354
Catholic      1204
Methodist      759
Muslim         144
Sikh            66
Jewish          36
Quaker           4
Baptist          1
Bahai            1
Name: Religion, dtype: int64
```

Marital Status:

```
Single      5939
Married     3000
Divorced     972
Widowed      529
Name: Marital Status, dtype: int64
```

Occupation:

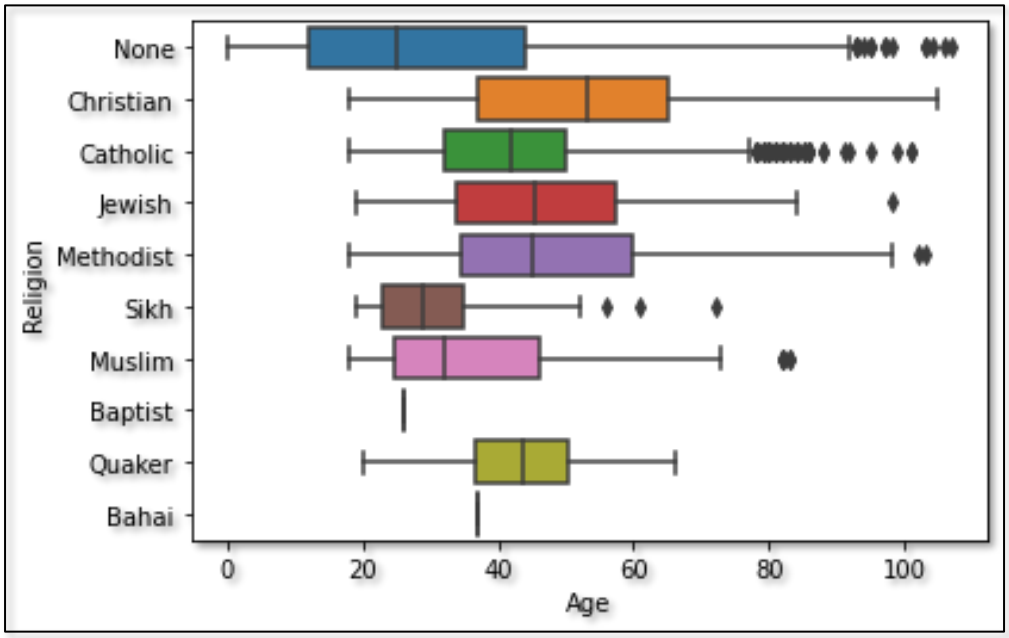
```
Student                                  1895
Unemployed                                643
University Student                        621
Child                                     541
Retired                                    53
                                          ...
Retired Counselling psychologist            1
Retired Television floor manager            1
Retired Research officer, trade union       1
Retired Accountant, chartered management    1
Retired Database administrator              1
Name: Occupation, Length: 1154, dtype: int64
```

Detailed Analysis:

- **Religion and Infirmity**:

There won't be any recommendation based on the infirmity as it comprises of small percentage compared to the whole population data.

On detailed analysis on Religion, it is found that one of the religions are in the state of growing compared to other religions (Methodist).



Religion Per Age Count

```
Religion
Bahai          1
Baptist        1
Catholic    1204
Christian   2354
Jewish        36
Methodist    759
Muslim       144
None        5871
Quaker         4
Sikh          66
Name: Age, dtype: int64
```

Religion Per Age Median

```
Religion
Bahai       37.0
Baptist     26.0
Catholic    42.0
Christian   53.0
Jewish      45.5
Methodist   45.0
Muslim      32.0
None        25.0
Quaker      43.5
Sikh        29.0
Name: Age, dtype: float64
```

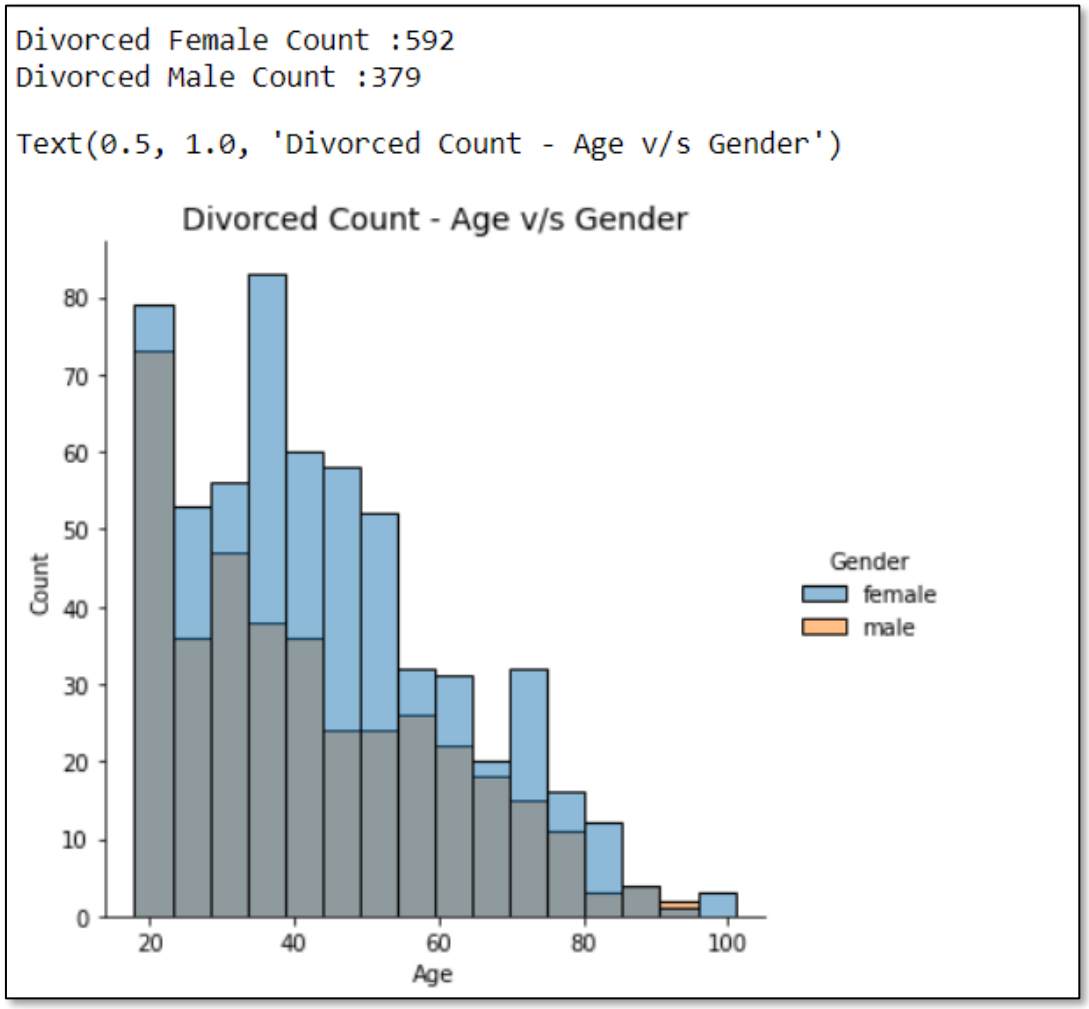
Christian was still dominant religion when compared to the rest of the religion. Since Methodist have significant followers in the town and currently in growing state which will require building of new church above the other needs of the population.

Divorce and Marriage:

As per the plot provided in data cleaning section, the divorce occurs from younger age to old age ones. The gender breakdown of marital status indicates that there are more divorced women than men, indicating that divorced men could leave the city.

```
Divorced Female Count :592
Divorced Male Count :379

Text(0.5, 1.0, 'Divorced Count - Age v/s Gender')
```



Birth Rate:

After the calculation of birth rate, it proves that the town has lower fertility rates and increasing aging population. The crude birth rate are calculated based on below:-

CBR = (No of births in year/total population) ×1,000

The current crude birth rate was estimated at 4 births per 1000 people which is significantly low in birth rate.

Employment & Commuters:

Unemployment percentage is relatively low compared to the overall population. Commuters are classified based on :

- Anyone who's occupation provided as a University Student.
- Count of the number of Occupations which would require commuting like Teacher, Pilot, Actor, Nurse, Engineers, Journalists and Survivors.

Occupations such as teaching (except Higher Education), Social Workers, retired peoples, working in food service department roles, community workers were considered non-commuting occupations.

Using this classification method, 60 % of the working population are likely to commute. Rest 40 % falls under non commutators.

Recommendation:

Due to the small number of lodgers emigrating to the town, not much overpopulation and most of them are employed, it will be ideal invest in low-density housing and a train station for the commuters (due to the high number of commuters). This also help the people, those who are divorced young children can move into housing.

It will also help in future from house shortage relief as the houses of the retired persons would be free at that time. Also, there would be a demand for a new second church as the followers for the Methodist are increasing gradually compared to other minor religions. Which will also attract other people (who believes in Methodist) from nearby cities thus inviting more immigrants in future.

As predicted in the age population pyramid the overall count of middle aged and old aged peoples are more than the younger ones. So, investing in Old age care should be prioritised than any other services.

Other services to invest in, such employment and training, is not required at present due to the unemployment percentage is relatively less when compared to the overall population. Also investing in Schooling as the population has small growth on year. And the town is not expanding now with the slow population growth, investing in general infrastructure would not be necessary at the moment. However there occurs a significant increase in the middle aged and old aged people, which require care facilities and it is pertinent to invest in Old age care than any other services.

Bibliography

Crude Marriage and divorce rates:

Available at:
https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Marriage_and_divorce_statistics#Fewer_marriages.2C_more_divorces

Marriage Laws:

Available at:

https://www.parliament.uk/about/living-heritage/transformingsociety/privatelives/relationships/overview/lawofmarriage-/

Families and households statistics:

Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/articles/familiesandhouseholdsstatisticsexplained/2019-08-07

Summary of risks to children's safety:

Available at:

https://learning.nspcc.org.uk/research-resources/2020/social-isolation-risk-child-abuse-during-and-after-coronavirus-pandemic

Religion error in census:

Available at: https://saspac.org/ew-sept-2011/2011-census-religion-error/

Table population and median age of religious groups

Available at:

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011census/2013-05-16#religion