CENSUS PROJECT REPORT

Outline

- Introduction
- Aim
- Data Cleaning
- Visualizations, Result Analysis and Discussion
- Recommendation
- Reference

Introduction

This report focuses on the examination of the statistics information of a relatively estimated town in the middle of two major urban towns connected by motorways, it is critical to take note that the town doesn't have a train station and university, yet students and employed class live and travel to the close-by urban areas. The Census information was randomly produced utilizing the python Faker Package, created incomparably, and intended to copy the arrangement of the 1881 census of the United Kingdom. The Fields recorded are as per the following: House Number, Street, First Name, Surname, Age, Relationship to the Head of the House, Marital Status, Gender, Occupation, Infirmity, and Religion.

Aim

This analysis aims to provide the most suitable recommendations to the following questions.

1. What should be built on an unoccupied plot of land that the local government wishes to develop?
   Choices are as follows.
   - High-density housing
   - Low-density housing
   - Train Station
   - Religious Building
   - Emergency medical building
2. Investment Options
   - Employment and Training
   - Old age care
   - Increase in spending for schooling
   - General Infrastructure

To give appropriate suggestions to the above questions, further examination of the following will be investigated in the report; Age distribution (population pyramid) typically, there are three trends in population pyramids: expansive, constrictive, or stationary. Unemployment trends such as age, gender, marital status will be examined.

Religion affiliation will be investigated to know whether certain religion is growing or shrinking, are there emerging religion to be investigated. The divorce and marriage rates will be examined to show the requirement for housing. Occupancy levels will be thoroughly investigated to show the number of persons per house. Commuters will be examined to determine the requirement for Infrastructure that helps traveling, Birth rate and death rate of the town will be investigated

Data Cleaning

Data Cleaning is a technique that improves erroneous data into meaningful information for further evaluation and decision making.

Age

Outliers such as three, twelve were replaced with integers 3, 12. Float values such as 80.21545403, 28.0, etc were converted to integers and rounded down into meaningful ages such as 80, 28. Blank spaces were investigated and discovered to be in loc 4594 and 8813 and imputed with to mean age. Negative age was investigated to be in loc 6730, House Number 29 in Gregory Rapids as the son to the Head of the house as male, having None infirmity, NaN religion, occupation as Child, marital status as NaN, hence this is a case of pregnancy or error in data entry, so it was imputed to zero (0).

Outlier a322 was investigated to be in loc 6991 House no 4, Burns Road and as Lodger single male, Health and safety adviser which portray a young male adult in his early 30s hence it could be inferred that the age attributed to Justin Howe is as a result of entry error hence 32 was imputed. Subsequently, the data frame is converted to integers data type from object data type.

House Number

Outlier nine was replaced with 9, nan imputed with zero (0) and that data frame was converted from object data type into Integer data type.

Street

Mode of the street was investigated, and the nan was replaced with Lee Stravenue

First Name

Individuals with B', 'S', 'E as the first name were investigated to be in loc 2680, 2703, 8117 replaced with nan, Jo-anne as a type of incorrect spelling was replaced with Jo-Anne, and subsequently, nan was filled with zero (0).

Surname

Individuals with surnames T, H were investigated to be in loc 8132, 8802, hence since the record could not be imputed with referential information, they were replaced with nan, and subsequently, nan was filled with zero (0).

Relationship to the Head of the House

The mode of the data frame was investigated to be Head, blank space investigated to be in loc 9463 Naomi Houghton 60, Marital status as divorced, which fits perfectly as the Head of the house hence the blank was replaced with Head. Subsequently, nan was replaced with Head

Marital Status

Mode of the marital status was investigated to be single, M, S as a case of wrong spelling were replaced with Married and Single, Marital status having one of the highest missing values of 2247, nan replaced with Single(mode) Marital Status

Gender

F-male, M, F were replaced with the appropriate spellings Female, Male. MALE, FEMALE as a form of Capitalization and duplication error were replaced with Male, Female. Blank was replaced with nan, subsequently, nan was replaced with Female(mode) Gender.

Infirmity

The mode of the data frame was investigated to be None, Female as an infirmity considered wrong information hence investigated to be in loc 2789, Marilyn Sheppard, 57 Head of house, hence replaced with None(mode). Blank investigated to be in 10 loc (1103, 1304, 1451 ,2761 ,3137 ,4054 ,4785 ,4985 ,5050 ,6821) hence replaced with nan. Nan was subsequently replaced with None(mode)Infirmity

Religion

Mode of the Religion was investigated to be None, Female as a form of Religion was investigated to be in loc 102 Female Rosemary Oliver 41 head of house single and Lecturer hence Female replaced with None(mode), Blank replaced with nan and Religion having 2298 missing values, Nan replaced with mode (None) Religion, also Undecided, Private was replaced to None and Buddist replaced with Buddhist.

The technique for managing empty cells is to embed another value instead. This way I don't need to erase entire rows due to some empty cells. The filling strategy permits me to supplant empty cells with a value (mode, zero), the value that shows up most is the mode of the column which I utilized in substitution during data cleaning of the columns above.

Mode is utilized on both Numerical and Categorical Data and slanted dissemination (one tail longer than the other) as displayed in the histogram plot of Age, Religion, Gender

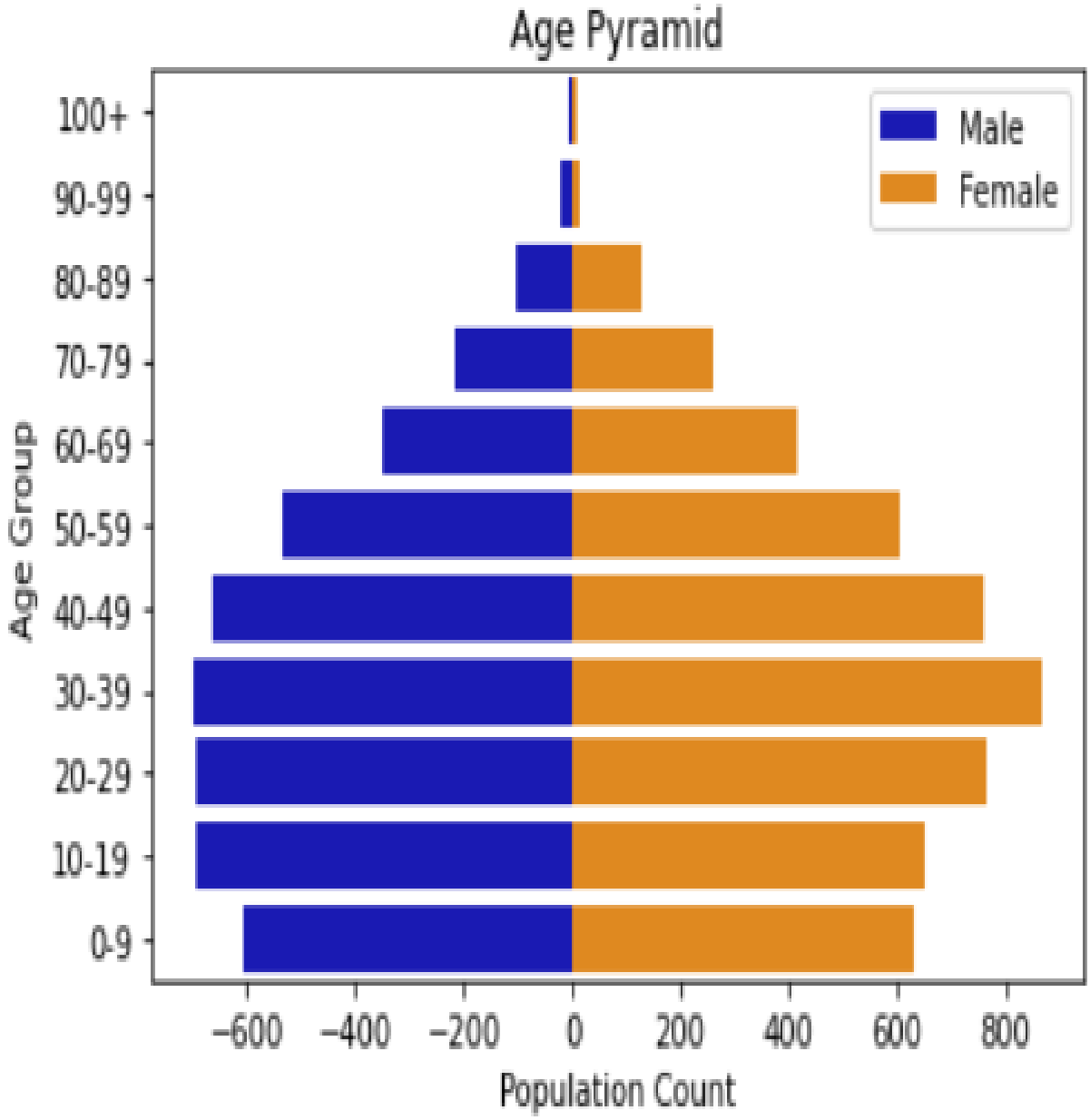
Features of the cleaned data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9687 entries, 0 to 9686
Data columns (total 20 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   House Number                            9687 non-null   int32
 1   Street                                  9687 non-null   object
 2   First Name                              9687 non-null   object
 3   Surname                                 9687 non-null   object
 4   Age                                     9687 non-null   int32
 5   Relationship to Head of House           9687 non-null   object
 6   Marital Status                          9687 non-null   object
 7   Gender                                  9687 non-null   object
 8   Occupation                              9687 non-null   object
 9   Infirmity                               9687 non-null   object
 10  Religion                                9687 non-null   object
 11  House Number_ismissing                  9687 non-null   bool
 12  Street_ismissing                        9687 non-null   bool
 13  First Name_ismissing                    9687 non-null   bool
 14  Surname_ismissing                       9687 non-null   bool
 15  Relationship to Head of House_ismissing 9687 non-null   bool
 16  Marital Status_ismissing                9687 non-null   bool
 17  Gender_ismissing                        9687 non-null   bool
 18  Religion_ismissing                      9687 non-null   bool
 19  num_missing                             9687 non-null   int64
dtypes: bool(8), int32(2), int64(1), object(9)
memory usage: 908.3+ KB
```

Visualizations, Result Analysis and Discussion

Age Pyramid

```
[Text(0.5, 0, 'Population Count'), Text(0, 0.5, 'Age Group'
```



age_p data frame

	Age	Male	Female
0	100+	-4	8
1	90-99	-24	13
2	80-89	-102	126
3	70-79	-215	262
4	60-69	-350	413
5	50-59	-535	605
6	40-49	-662	758
7	30-39	-698	866
8	20-29	-696	763
9	10-19	-697	649
10	0-9	-607	629

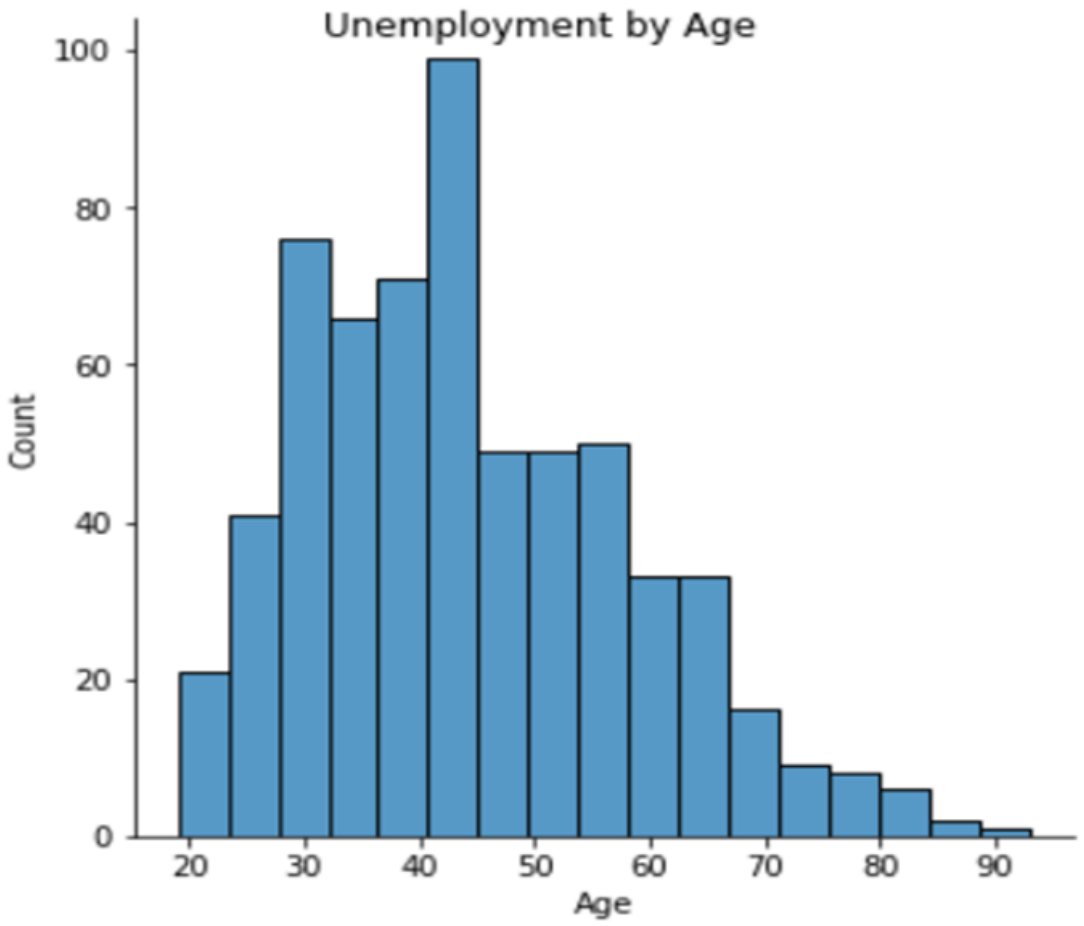
The population pyramid reveals the composition of a population by comparing relative counts in different age groups. Usually, pyramids are illustrated with the counts of the male population on the left and counts of the female population on the right. It provides us knowledge regarding birth and death rates as well as life expectancy. A population pyramid tells us about dependants in the population as well as the aging class. There are two classes of dependants; young dependents (aged below 15) and elderly dependents (aged over 65). Dependants depend on the economically active class for support, aged 15-65.

Examining the population pyramid, the shape of the pyramid reveals a smaller number of young people aged 0-9 compared to middle-aged group 30-49, implying a low birth rate, there is a high number of young dependants in the population. When a population is growing (more babies are being born than people are dying), hence the population pyramid of the town as shown above indicates a shrinking population because fewer babies are being born than people are dying. The Age pyramid shows a high number of retired/aging population.

Unemployment by Age

Unemployment is a term referring to individuals who are employable and actively seeking a job but are unable to find a job. Included in this group are those people in the workforce who are working but do not have an appropriate job. Usually measured by the unemployment rate, which is dividing the number of unemployed people by the total number of people in the workforce, unemployment serves as one of the indicators of a country's economic status

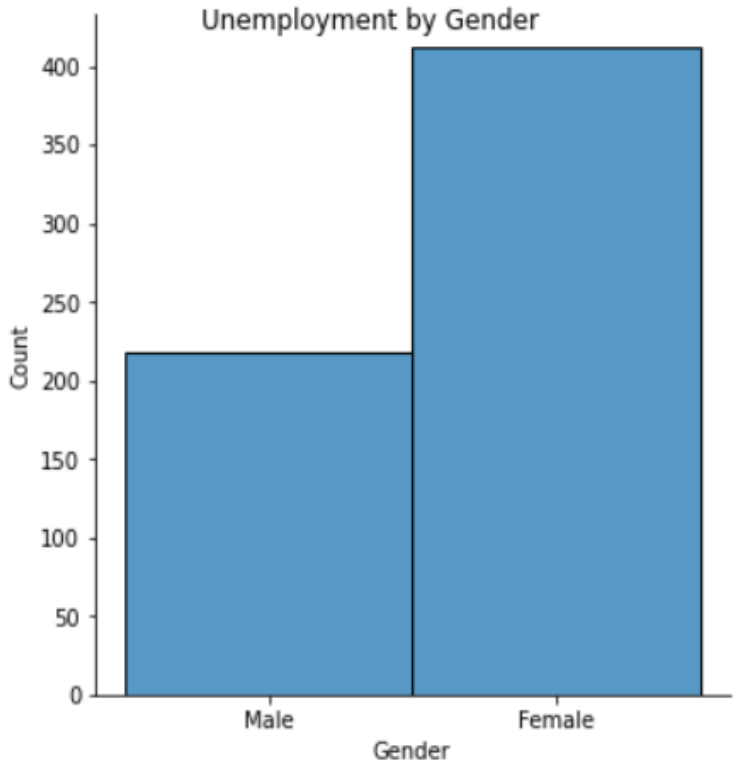
```
Text(0.5, 0.98, 'Unemployment by Age')
```



The histogram indicates that the unemployment count is highest at age group 40-45, followed by age 30. The economically active age group 20-65 has a high unemployment count.

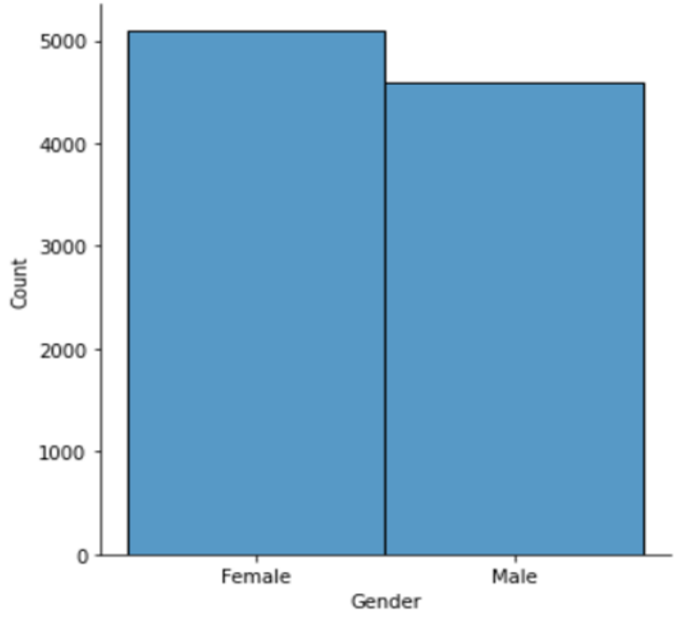
Unemployment by Gender

```
Text(0.5, 0.98, 'Unemployment by Gender')
```



The Histogram shows a high unemployment count among Female population distribution, meaning most of the population are unemployed as females constitute the highest count of the population distribution as shown in the histogram below

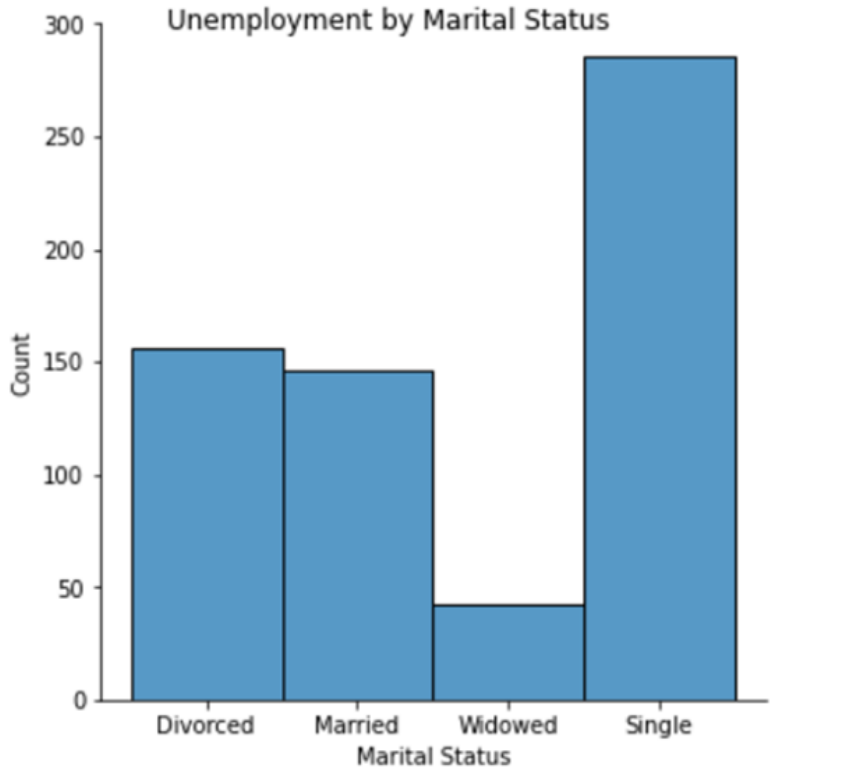
```
<seaborn.axisgrid.FacetGrid at 0x27c06eacf70>
```



Gender distribution Histogram

Unemployment by Marital Status

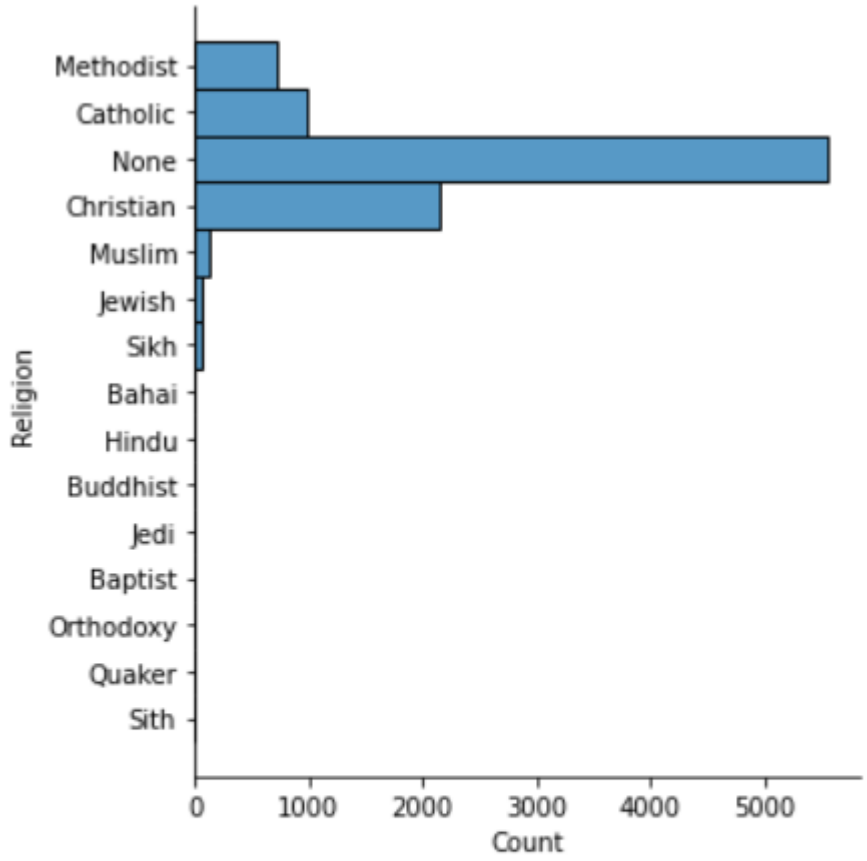
```
Text(0.5, 0.98, 'Unemployment by Marital Status')
```



The histogram indicates a high level of unemployment among the Single population, followed by the Divorced population, Married population, and the widowed population having the lowest unemployment count in the population distribution. From the graphical representation above on unemployment (Unemployment by Age, Unemployment by Gender, Unemployment by Marital status) it is evident that most of the population are unemployed.

Religion Affiliation

```
<seaborn.axisgrid.FacetGrid at 0x21ce3bf2a00>
```




```
df['Religion'].describe()
```

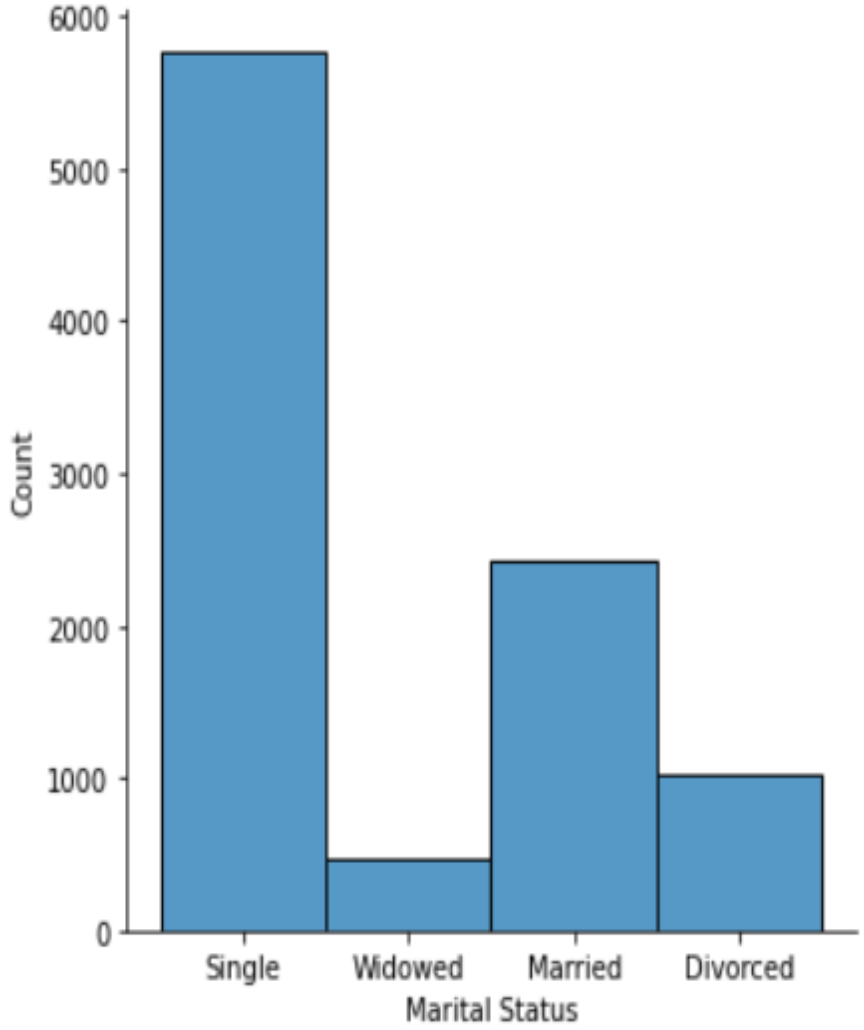
```
count      9687
unique       15
top        None
freq       5560
Name: Religion, dtype: object
```

Analyzing the Religion chart above, the Majority of the population have no Religion Affiliation, meaning they don't belong to any Religion group, Christians have the highest Religion affiliation in the population distribution, Sikh, Jewish, and Muslim are growing religion while Bahai, Hindu, Buddhist, Jedi, Baptist, Orthodoxy, Quaker and Sith can be classified as low-frequency Religion and maybe a lie people tell during the population Census of the town, hence building a religion center is not a priority at this time. Catholic has one already.

Divorce and Marriage rate

Marital Status Histogram

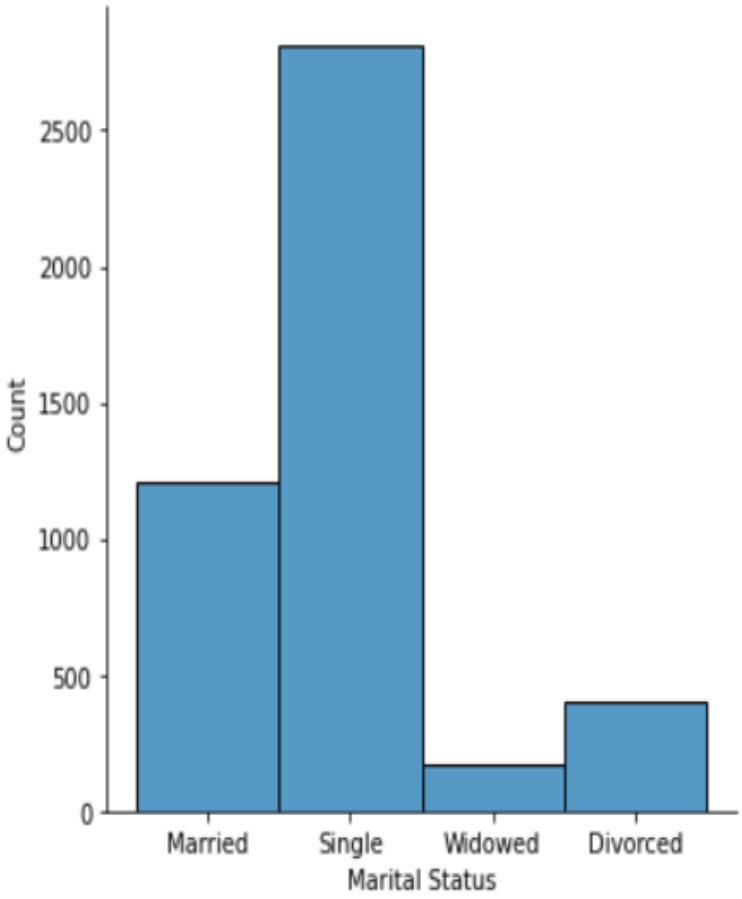
```
<seaborn.axisgrid.FacetGrid at 0x21ce5113160>
```



Examining the marital status histogram, it is evident that there are more Married groups in the population than Divorced groups in population distribution, pointing towards the need for more family-oriented housing i.e., low-density housing in the town

Male Marital Status Histogram

```
In [122]: sns.displot(data=df, x=Male['Marital Status'])
Out[122]: <seaborn.axisgrid.FacetGrid at 0x25b5ee2c370>
```

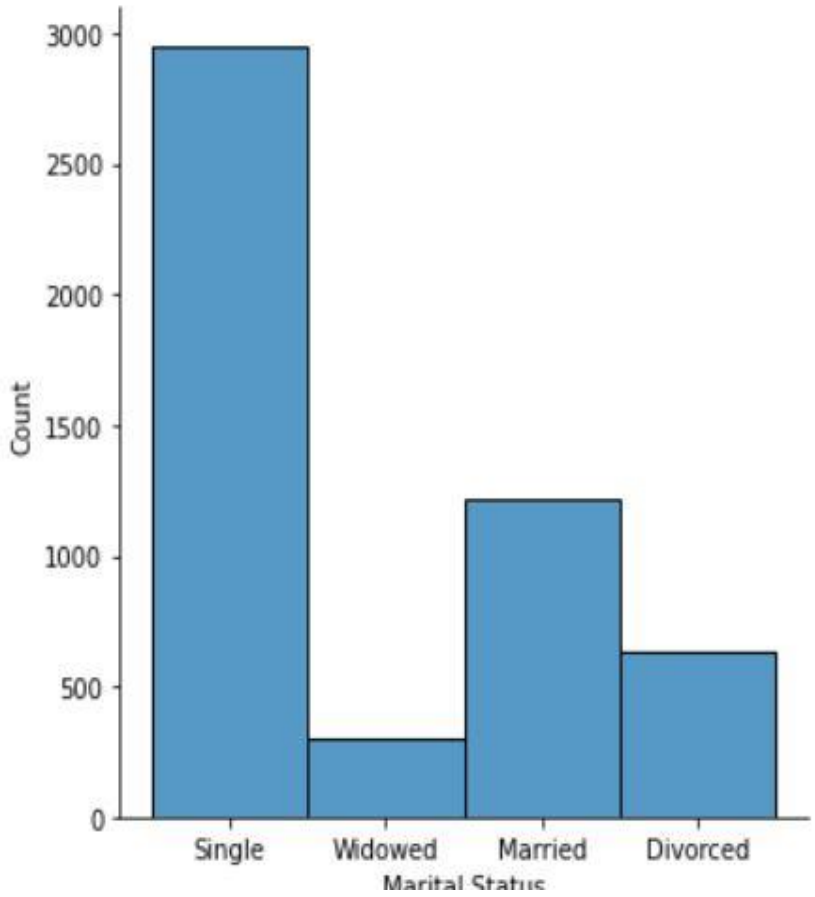


Histogram showing male marital status with the single having the highest count, then married, divorced, and widowed having the lowest count in the male population distribution of the town, the key interest here is the divorced Male below 500.

Female Marital Status Histogram

```
: #histogram of Female Marital Status
  sns.displot(data=df, x=Female['Marital Status'])

: <seaborn.axisgrid.FacetGrid at 0x25b5c4cb8e0>
```



Histogram showing female marital status with the single having the highest count, then married, divorced, and widowed having the lowest count in the female population distribution of the town, the key interest here is the divorced female, above 500.

Hence analyzing the three Marital Status Histogram above it is evident that there are more Divorced Women in the population than men, indicating that divorced men move away from the town, reducing the population count and increasing the demand for low-density housing in the town for the sole purpose of down-sizing.

Occupancy level

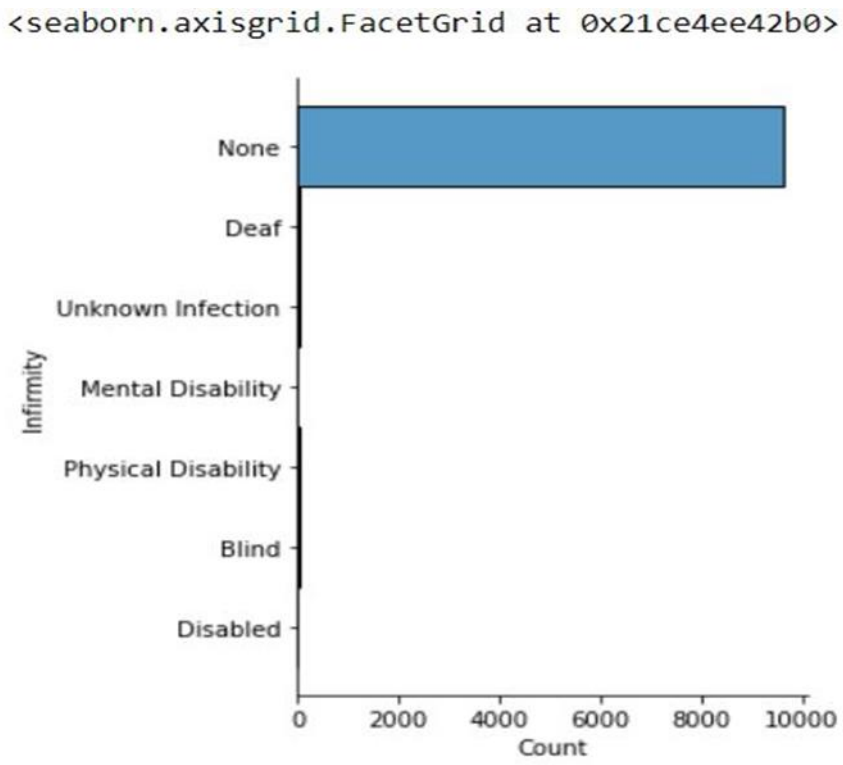
```
df['Street'].describe()
```

```
count              9687
unique              105
top       Lee Stravenue
freq               1199
Name: Street, dtype: object
```

The mode of the Street is Lee Stravenue, meaning the street with the highest number of occupants, hence it is evident that the housing facility in the street is over-utilized, hence a need for additional low-density housing to address the issue of overcrowded street/housing facility in the town.

Infirmity

Histogram of Infirmity



The histogram above shows the low level of Infirmity in the population distribution of the town, with the mode as None. Deaf, Unknown Infection, physical disability, blind, mental disability, and disabled are in very low frequency, building a medical facility is not a priority.

Commuters, Student and Occupation

```
In [124]: Student=df[df['Occupation']=='Student']
          Student['Occupation']
```

```
Out[124]: 7       Student
          17      Student
          48      Student
          49      Student
          50      Student
                   ...
          9675    Student
          9676    Student
          9682    Student
          9683    Student
          9684    Student
          Name: Occupation, Length: 1806, dtype: object
```

```
In [126]: Student['Occupation'].count()
```

```
Out[126]: 1806
```

The above figure indicates those that identify as a student in the population, although this group will consist of school children, high school, and university students. The student having 1806 counts all through the population, hence the university student only can be categorized as commuters since the information provided from the census indicates that there is no university in the town and university students' commuters are in very low frequency.

```
In [125]: df['Occupation']

Out[125]: 0                              Producer, radio
          1                Retired Administrator, sports
          2                          Retired Bookseller
          3                  Retired Industrial buyer
          4            Retired Scientist, audiological
                                      ...
          9682                                 Student
          9683                                 Student
          9684                                 Student
          9685                                   Child
          9686     Retired Research officer, government
          Name: Occupation, Length: 9687, dtype: object
```

People who identify as Retired are non-commuters in the population, some professions don't necessarily commute like primary and high school teachers, retailers, restaurant workers (people in the food industry), city council waste management services, bar owners, generally high level of unemployment and high level of retired groups as evident in the data suggest a low level of commuters in the town.

Birth rate and death rate

From the data, the current crude birth rate is 9.8 births per 1000. The previous year, the crude birth rate was estimated at 11.5 births per 1000, calculated by adjusting the population to what it may have been in the previous year, the birth rate has therefore declined by 1.7 births.

From the data, the calculated Death rate is 14.3 deaths per 1000. The crude Growth rate is calculated by subtracting the Current Crude birth rate from the Death rate as thus estimated as -4.5 and as -0.45%.

The Birth rate, Death rate, and Crude Growth rate show that the town is not growing, thus classified as shrinking according to the population pyramid (age pyramid) assuming it was compared to previous years' assumptions data.

Recommendation

1. What should be built on an unoccupied plot of land that the local government wishes to develop?

Low-density housing will be of huge benefit to the family, divorced family and those who intend to live in moderately size housing and to cut down the overcrowded Lee Stravenue Street with the high number of occupancies, although train station will benefit university students and working force who needs to commute to work regularly, they are low in number because of very high unemployment and the high number of retired groups in the population distribution, also university students are low in number, else weighing the need for both low-density housing and train station, it is evident that the need for low-density housing outweighs that of the train station, on another thought having a train station for an economically vibrant town will boost the town's revenue in so many ways but the town is not doing so well economically considering the high rate of unemployment and high rate of retired/aging groups in the town.

2. Investment Options?

Employment and Training are highly recommended due to the high level of unemployment in the population, men and women need to be trained and retrained to acquire new skills most especially the economically active group to bring about development and economic growth to the town, afterward the thought of building a train station can be a welcomed idea because there will be an increment in the employment rate and skilled individuals which in turns need to regularly travel to nearby towns to offer their skills and services thereby boosting the town's revenue from the export of highly trained men and women into other towns. Hence Offering employment and training will be a good investment option that will lead to the huge economic growth of the town in the nearest future.

Furthermore, Old age care can be considered next if the issue of unemployment is fixed to a reasonable level because there is also a high level of retired/aging group in the town, but the priority is still the high level of unemployment. Fixing unemployment is directly proportional to high revenues for the town, which can be channelled into other areas where the town is lacking.

In the future, if the level of unemployment goes down, there will be an inflow of people and visitors in the town, hence the idea of other infrastructure like waste collection, road maintenance will require more investment.

In addition, high-density housing can be considered if the population of the town is expanding in the future because with huge economic growth and earnings of individuals there will be an inflow of people into the town and an increase in the number of babies being born.

Reference

Available Online

https://www.simplybusiness.co.uk/knowledge/articles/2021/08/what-is-the-retirement-age-uk/

https://www.nationalgeographic.org/encyclopedia/population-pyramid/

https://www.bbc.co.uk/bitesize/guides/ztr2w6f/revision/2

https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

https://corporatefinanceinstitute.com/resources/knowledge/economics/unemployment/