Census 20 Data Report

Introduction

The United Kingdom conducts a census of its population every ten years. This census aims to compare various groups of people across the country and provide the government with accurate population statistics. This information is used to aid in planning, developing policies, and allocating funding more effectively.

Aim and Objective

The aim of the census20 data project is to clean and analyze a provided mock census dataset using pandas. The objective is to gain insights and make informed decisions about an empty plot of land and potential investments based on the analyzed data.

Data Information

The Census 20 contains 8577 entries with 11 columns consisting of various data types such as object, float, and integer. The House number column contains integer data type while the Age contains the float data type and the rest of the columns; Street, First name, Surname, Relationship to head of house, Marital status, Gender, Occupation, Infirmity and Religion contains the object data type. There are two columns with NaN values in the dataset, which are Marital status and Religion. So the number of NaN values in the Marital status is 2085 citizens while the Religion is 2132 citizens where 3 columns have missing rows in which the columns are; Relationship to head of house, Infirmity and Religion

```
In [8]: #Dimension of the Data
        df.shape
Out[8]: (8577, 11)
```

- It shows that the census20 data contains 11 columns a

```
In [9]: # To check the data types of each columns in the
        df.dtypes
Out[9]: House_Number                  int64
        Street                      object
        First_Name                  object
        Surname                     object
        Age                        float64
        Relationship_to_Head_of_House object
        Marital_Status              object
        Gender                      object
        Occupation                  object
        Infirmity                   object
        Religion                    object
        dtype: object
```

Figure 1

```
In [11]: #To check for the missing values
         df.isnull().sum()
Out[11]: House_Number                  0
         Street                        0
         First_Name                    0
         Surname                       0
         Age                           0
         Relationship_to_Head_of_House 0
         Marital_Status             2085
         Gender                        0
         Occupation                    0
         Infirmity                     0
         Religion                   2132
         dtype: int64
```

Figure 2

The data reveal a number of fascinating facts about the citizens. It shows that Anna is the most occurring name while Smith is the most common occurring surname and the majority of the citizens are single and more of the citizens are head of their household. Also, a large proportion of the citizens are female, with students also occupying more of the population, many of the citizens were not identify with any religion and most do not have any infirmities. Furthermore, it can be concluded that Faith Street is the most inhabited street among the citizens.

```
In [12]: df.describe(include = 'all').T
```

Out[12]:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
House_Number	8577.0	NaN	NaN	NaN	37.082896	44.790872	1.0	9.0	21.0	42.0	220.0
Street	8577	104	Faith Street	708	NaN	NaN	NaN	NaN	NaN	NaN	NaN
First_Name	8577	364	Anna	41	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Surname	8577	673	Smith	223	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	8577.0	NaN	NaN	NaN	35.594086	21.377013	0.0	18.0	35.0	50.0	107.0
Relationship_to_Head_of_House	8577	22	Head	2991	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Marital_Status	6492	4	Single	2831	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	8577	2	Female	4501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	8577	1070	Student	1680	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Infirmity	8577	8	None	8496	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Religion	6445	14	None	2776	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3

Data Cleaning

The data cleaning was done with the help of Jupyter Notebook with a tool called Pandas which provides a powerful data manipulation and analysis toolkit. Checking through each column missing values, NaN, the wrong imputation or lies was a check for all through the column. The process by which each column was cleaned was explained below.

Age: The Age in the census data was in the float data type but for the Age to be fully represented, the Age column was converted from a float to an integer data type. This is because, for a census dataset, Age is represented as the number of years since birth, which is a discrete value and can allow integer values. Using a float data type potentially could cause confusion in the data analysis process

Relationship to head of house: This column contains two empty rows, in which the two citizens are students and also have their gender as male so the missing rows were replaced with (Son) since as a student their relationship to the head of house will be dependent and their gender is a male

Marital status: The marital status contains many NaN values so after filtering the dataset to check the age of the citizens in relation to their marital status, it was shown that those who are 17 years and below were the ones with the NaN value. The Marriage and Civil Partnership (Minimum Age) Act 2022 in the UK, states that the minimum age for marriage is 18 years. Hence, citizens who are 17

years and below are considered minors and not eligible for marriage. Furthermore, based on the filtering in Figure 4 below the ages who are 18 years and marital status is widowed, inputting the marital status as widowed will be false and a lie since it is uncommon for an 18-year-old to be widowed. Therefore, the widowed who are 18 years old were replaced with single.

```
#To Ages = 18 and Marital status = Widowed

df[(df['Age'] == 18) & (df['Marital_Status'] == 'Widowed')]
```

	House_Number	Street	First_Name	Surname	Age	Relationship_to_Head_of_House	Marital_Status	Gender	Occupation	Infirmity	Religion
1308	21	Singh Forge	Gerald	Jordan	18	Son	Widowed	Male	Student	None	Catholic
3843	30	Corporation Road	Sharon	Edwards	18	Head	Widowed	Female	Student	None	NaN
6095	5	Bishop Lane	Joe	Rogers	18	Son	Widowed	Male	Student	None	None

Figure 4

Religion: In the religion column a lie or a wrong imputation was discovered as there is no religion called Sith. Sith is a cult organization according to Wikipedia, not a religion so the Sith was replaced with the Sikh religion because the Sith might also be a wrong input of Sikh. Also, an empty row was discovered the row was replaced with the NaN value. Furthermore, upon filtering the Age in relation to Religion, it was discovered that citizens who are 18 years and above have missing values (NaN) and the same with citizens who are 17 years and below. For citizens who are 18 years and above the NaN value was replaced with unknown because they are adults and they have freedom of religion so a particular religion cannot be forced upon them by law. (Human Rights Act 1998). Moreover, for citizens of Age 17 years and below and having a religion missing value (NaN), the NaN was replaced with the religion of the same members of their family.

Infirmity: The total empty row in the infirmity column is 7, which was replaced with “Not available” since it was not given in the dataset

Occupation: There are no missing values and empty rows in the occupation column but since the professionals who have retired in different filed have been stated in the column it is advisable to group them under a common category called “Retired” to avoid redundancy in information.

Population Demography

After the process of the data cleaning the Census data has the following features and characteristics

To check the if the dataset is clean

```
[72]: df.isnull().sum()

[72]: 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
```

Figure 5

```
In [108]: #Checking the clean dataset info
          df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8577 entries, 0 to 8576
Data columns (total 12 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House_Number                  8577 non-null   int64
 1   Street                        8577 non-null   object
 2   First_Name                    8577 non-null   object
 3   Surname                       8577 non-null   object
 4   Age                           8577 non-null   int32
 5   Relationship_to_Head_of_House 8577 non-null   object
 6   Marital_Status                8577 non-null   object
 7   Gender                        8577 non-null   object
 8   Occupation                    8577 non-null   object
 9   Infirmity                     8577 non-null   object
 10  Religion                      8577 non-null   object
 11  range_of_age                  8577 non-null   object
dtypes: int32(1), int64(1), object(10)
memory usage: 770.7+ KB
```

Figure 6

To get more insight on the analysis a new column was introduced called the range of age.

Insight Analysis and Results

Age Distribution

```
In [88]: #Plot for the Distribution Age

         histogram_boxplot(data = df, feature = 'Age')
         plt.title("Age Distribution")

Out[88]: Text(0.5, 1.0, 'Age Distribution')
```

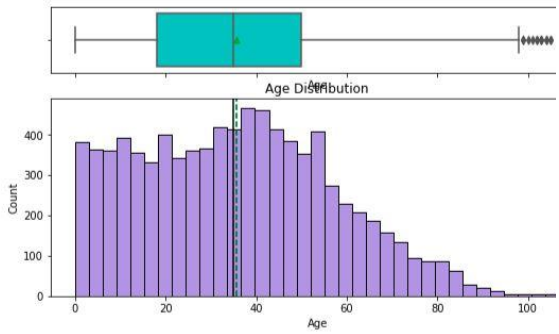

Figure 7

```
In [114]: #To get the statistics describtion for the Ag

          df["Age"].describe()

Out[114]: count    8577.000000
          mean       35.593098
          std        21.376178
          min         0.000000
          25%        18.000000
          50%        35.000000
          75%        50.000000
          max       107.000000
          Name: Age, dtype: float64
```

Figure 8

Figure 7: Showing the Age distribution. Figure 8: Showing the statistics description of the Age

Figure 7 and 8 above shows the age distribution slight right skew and an approximately normal distribution while also showing the mean and median age to be 35.5 and 35.0 respectively. The dataset contains a small number of individuals who are older.

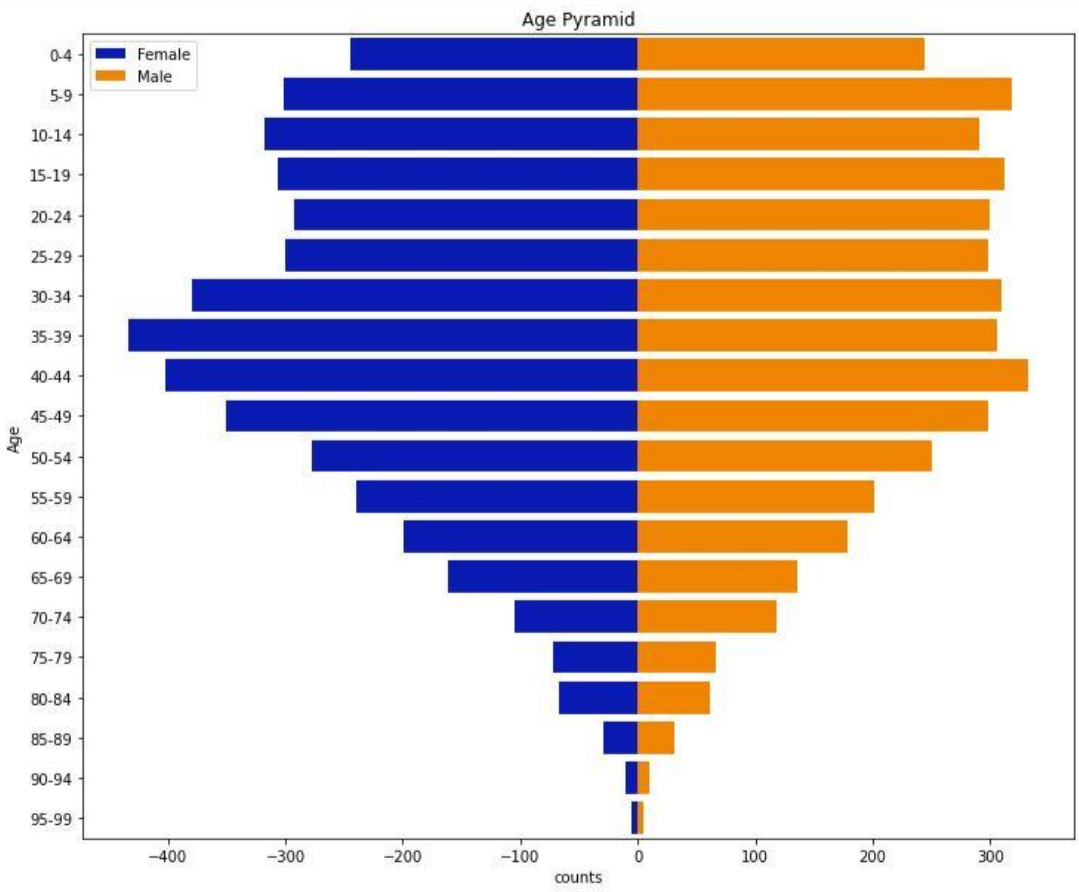

Figure 9

Figure 9 showing the Age Pyramid

From the Age pyramid it shows that there is slightly lower number of citizens who are children in the range of age 0 – 4 which indicates a low birth rate. More also the highest population of citizens are in the range of age 35 – 39 and 40 – 44 which is a great advantage to the workforce for productivity and this will also lead to more retired age people in the future. The older the citizens in the Age pyramid the lower the population of the citizens which clearly shows that the people aged well. It can be said that the age distribution is declining after the age range of 40 – 44.

Religion Distribution

```
In [96]: #Plot of Religion

labeled_barplot(data = df, feature = "Religion", perc = True)
```

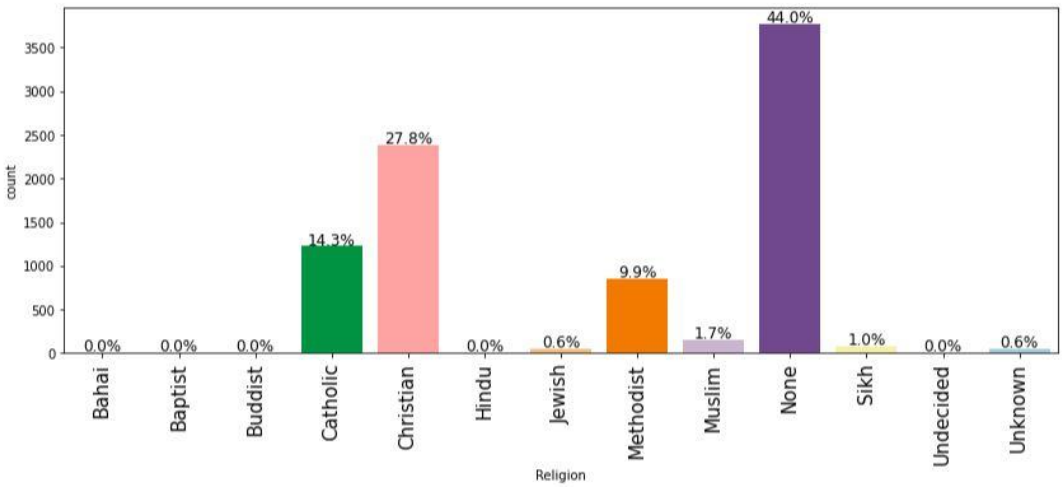

Figure 10

Figure 10. showing the Barplot of the religion

Figure 10 shows majority of the population occupying 44.0% does not adhere to any religion. Among those who identify with a religion Christians dominate with 27.8%, followed by Catholics at 14.3% and Methodists at 9.9%

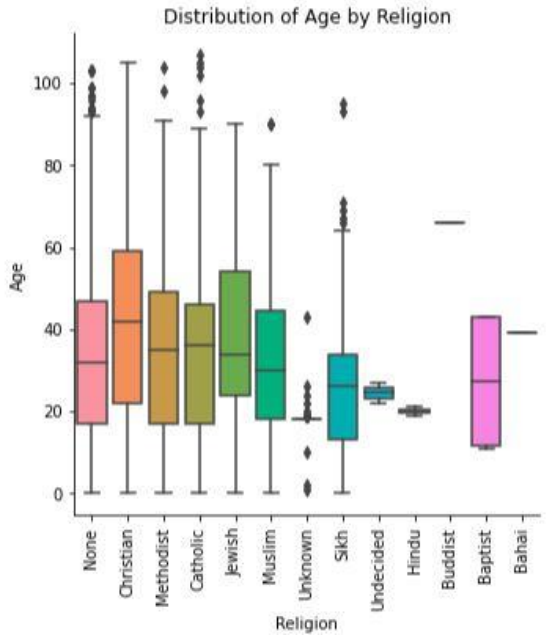

Figure 11

```
: Religion
Bahai        39.0
Baptist      27.5
Buddist      66.0
Catholic     36.0
Christian    42.0
Hindu        20.0
Jewish       34.0
Methodist    35.0
Muslim       30.0
None         32.0
Sikh         26.0
Undecided    24.5
Unknown      18.0
Name: Age, dtype: float64
```

Figure 12

Figure 11. showing the Box plot of Age by religion. Figure 12. Showing the median Age by Religion

Looking through Figure 11 and 12 Christian, Catholic, Methodist and Jewish had a slightly median age of followers indicating their growth as religions but they constitute 53.7% of the population.

Christianity is the dominant religion so there is a need for a second church for the Christians. Though there be an increase in the dominance of other religions in the future they do not have many followers that warrant a new church building now.

Marital Status (Marriage and Divorced)

The plot below depicts the marital status distribution in the dataset, with singles accounting for the biggest percentage (33.0%), followed by married people (30.03%). Children, who are considered minors, account for 24.3% of the population. The number of widowed people is quite low at 3.6%, whereas the divorce rate is 8.8%.

```
[107]: #Plot of Marital Status

       labeled_barplot(data = df, feature = "Marital_Status", perc = True)
```

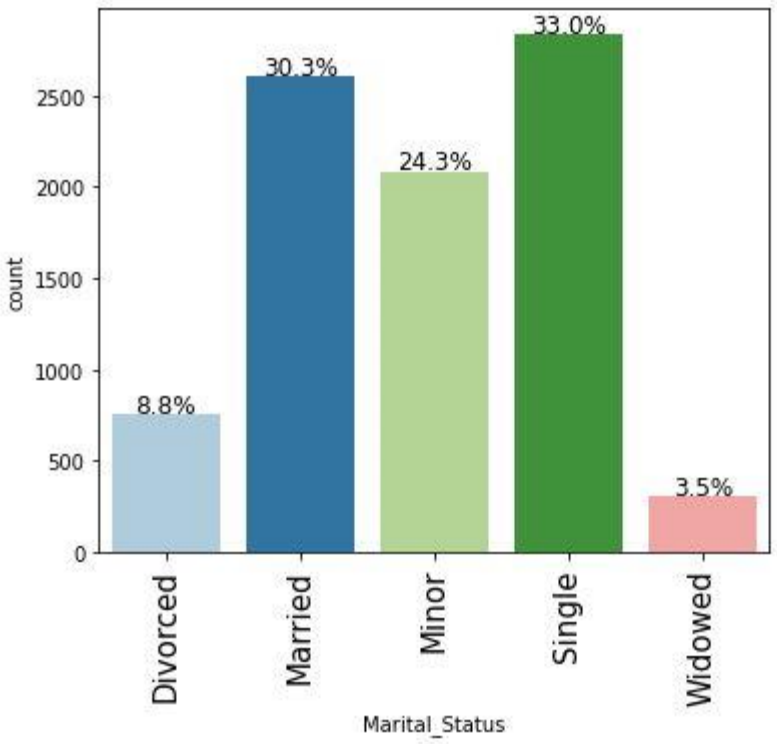

Figure 13

As seen from the data cleaning divorce occurred across the age range of the citizen from young to old. When the marital status is broken down by gender. It shows that there are more female divorces than the male

```
: #To check for number of male and female divorced

  df.loc[df['Marital_Status'] == 'Divorced'].groupby('Gender')['Marital_Status'].count()
```

```
: Gender
  Female    456
  Male      297
  Name: Marital_Status, dtype: int64
```

Figure 14

Also, more females married than the male

```
#To check the number that are married from the male and female

df.loc[df['Marital_Status'] == 'Married'].groupby('Gender')['Marital_Status'].count()
```

```
Gender
Female    1312
Male      1290
Name: Marital_Status, dtype: int64
```

Figure 15

And the Figure 16 below shows most of the most of the population are single, note that the figure 16 covers those who are 18 years and above. If the singles keep increasing with addition to those divorced and also the widowed this can affect the rate of increase in the population.

To show for Marital Status of citizens whose Age >= 18

In [105]:
```
# filter the dataset for individuals whose age is >= 18

citizen_18 = df[df['Age'] >= 18]


marital_status_counts = citizen_18.groupby(['Marital_Status']).size()


marital_status_counts
```

Out[105]:
```
Marital_Status
Divorced     753
Married     2602
Single      2834
Widowed      303
dtype: int64
```

Figure 16

Birth Rate and Death Rate

From the calculation of the crude birth rate, it was shown that the birth rate is 160 per 1000 which is 16%. This shows the town is slightly low population growth. Furthermore, the figure below shows the death rate in the different age ranges.

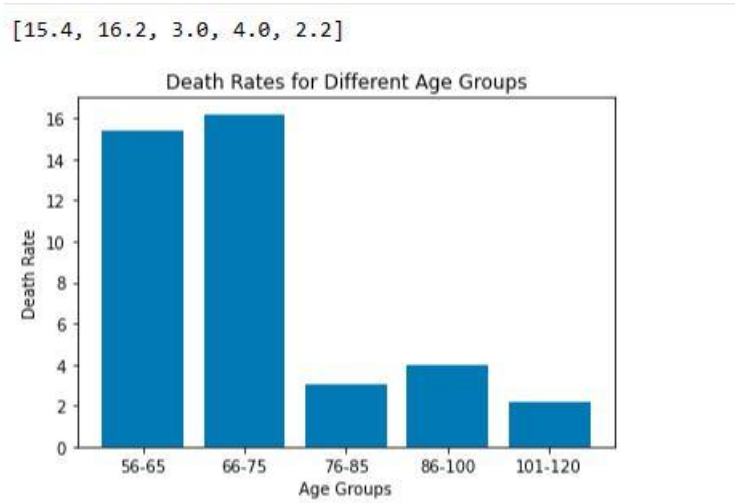

Figure 17

The range of age 56 – 65 years has a death rate of 15.4, 66 -75 years have a death rate of 16.2, 76-85 years have a 4.0 death rate, and 101 – 120 years have a 2.2 death rate.

Migration

From the analysis calculation it shows there is low immigration and emigration in the town

Employment and Unemployment

From the data analysis the proportion of the unemployed is 5.98% in which the employed is 94.01% this shows there is low employment rate in the town, so the training of people for new skills might not be really needed in the town.

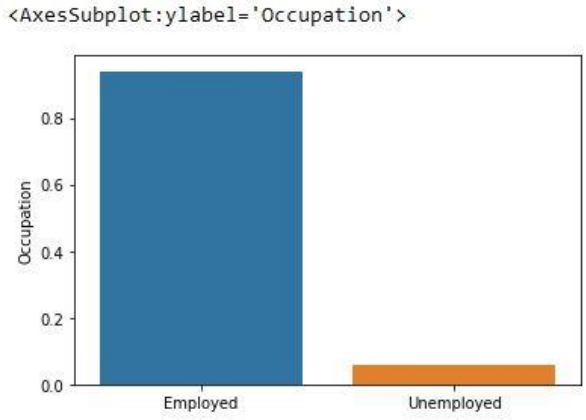

Figure 18

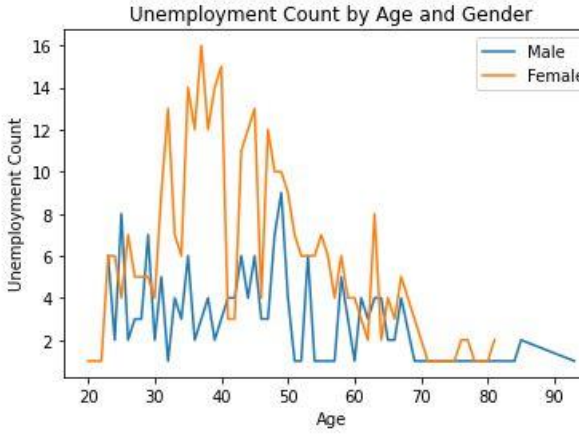

Figure 19

Figure 19. shows that unemployment is dominant among the male from the ages of 25 – 40 years than the female.

Occupancy of House

The houses in the town are not overpopulated as the Figure 20. below shows that 1,2, and 3 occupants majorly are the ones who occupy the houses in the town. Therefore, the existing houses are underutilized and not used to its capacity so no need for more housing.

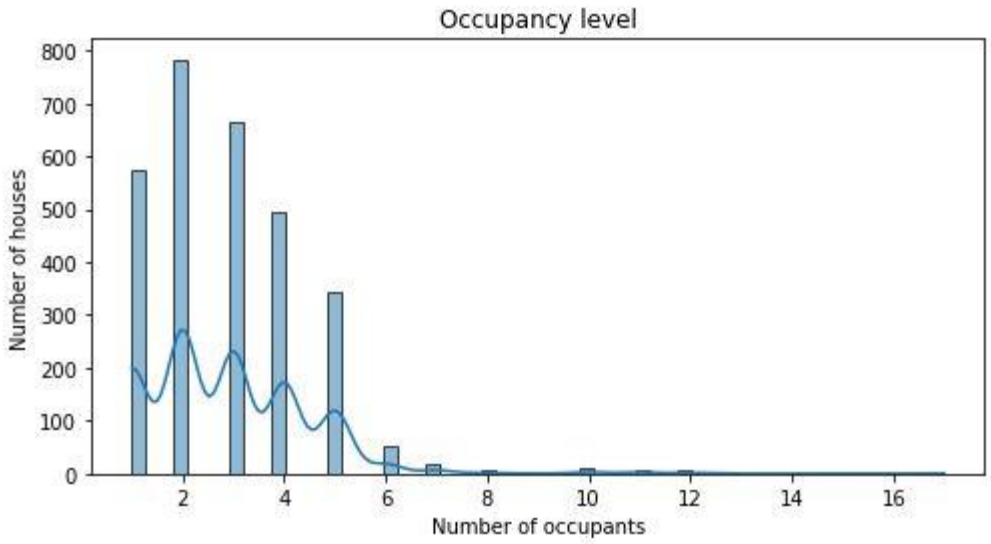

Figure 20

Infirmity

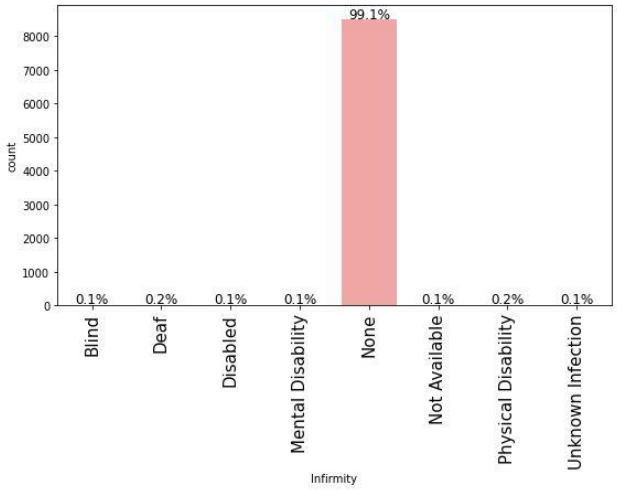

Figure 21

From the Figure 21 it shows majority of the population in the dataset of about 99.1% have no infirmity

Recommendation

After the detail analysis of the census 20 data given that the population has the highest fatality rate, as shown in figure 17, between the ages of 56-65 and 65-75, I will recommend a medical facility so that this age group has access to appropriate emergency medical facilities, though the figure 21 shows most of the population have no infirmity but this might change has the population grow older. Furthermore, the Age pyramid in Figure 9 illustrates that there will be a larger population that will retire in the future, therefore old age care and more financing should be supplied for their retirement stage.

According to Figure 10, there is also a necessity for Christians to have their own church in town because they are the most common religion with which the people identify. Given the town's high employment rate, as shown in Figure 18, there is a need to construct a train station for workers to easily move in and out of town. Although it has been demonstrated that the age range of 35-39 is the highest in the population, high-density residences may be required in the future due to the population's tendency to give birth on a greater scale.

Infrastructure like roads might be worth investing in since the highest population are the workforce to aid the ease of transportation, elementary and secondary school also worth investing in since this age range have a slightly higher population

Bibliography

Gov. uk (Marriage and Civil Partnership (Minimum Age) Act 2022
Available online: https://www.gov.uk/government/news/implementation-of-the-marriage-and-civil-partnership-minimum-age-act- [Accessed 15/04/2023]

Legislation.gov.uk (Human Rights Act 1998)
Available online: https://www.legislation.gov.uk/ukpga/1998/42/contents [Accesed 22/04/2023]

Office for National Statistics

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021

Stackoverflow

Available online: https://stackoverflow.com/questions/67329095/when-should-you-convert-age-column-in-float-or-int#[Accessed 14/04/2023]

Wikipedia.org (Sith)

Available online:https://en.wikipedia.org/wiki/Sith [Accessed 20/04/2023]