THE CENSUS PROJECT

By Obinna Daniel Igboekwezeh

202118804

For   fundamentals of data science

MSc Artificial Intelligence and data science 2021/2022

ABSTRACT

An analysis of census data from a mini city between two megacities forms the basis of this project. The local government is looking to invest in unoccupied land in the town and wants recommendations on the best investment to make. A data clean-up was conducted to correct the errors seen in the data to make recommendations based on it. My observations are outlined in the first section of this report, and the subsequent section provides a detailed analysis. In light of these analyses, a recommendation was made to the local government that low-density houses and investments in general infrastructure should be the first priorities.

INTRODUCTION

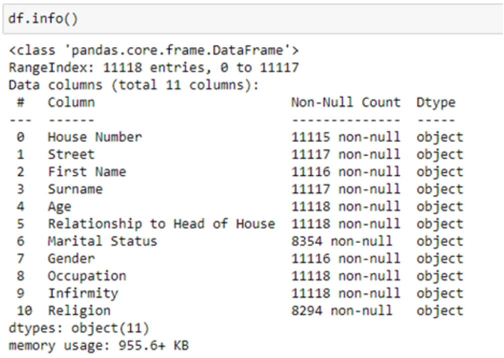
Data science involves the manipulation of data and generating insights from them. The process involves data collection, data cleaning, exploratory data analysis, interpretations, model building.

For this project, the programming language employed was python with jupyter notebook IDE. The data collection process was not employed because the data was provided for the project. So the first process was to read in the data in the IDE of choice and data explored using appropriate functions to be sure the data is of quality for analysis. This is where a large portion of the time for the project was spent. It was observed that there were various errors with the data which was cleaned to an excellent level.

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.

DATA CLEANING

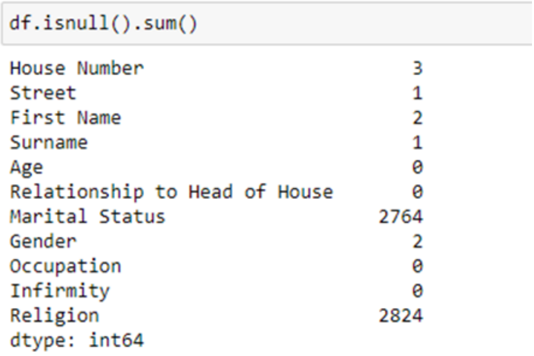
Exploratory Data Analysis carried out revealed that the type of all the columns were all “object” types and also the presence of null values in the data set as shown in figures below. Some blanks were not captured as null values. They were read in using “na_values” to convert them to null values. There were other types of errors observed in the data set which were cleaned using different methods. A record of all cleaning done on the data can be found in the associated jupyter notebook.



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11118 entries, 0 to 11117
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   House Number                  11115 non-null  object
 1   Street                        11117 non-null  object
 2   First Name                    11116 non-null  object
 3   Surname                       11117 non-null  object
 4   Age                           11118 non-null  object
 5   Relationship to Head of House 11118 non-null  object
 6   Marital Status                8354 non-null   object
 7   Gender                        11116 non-null  object
 8   Occupation                    11118 non-null  object
 9   Infirmity                     11118 non-null  object
 10  Religion                      8294 non-null   object
dtypes: object(11)
memory usage: 955.6+ KB
```

Figure 1: info command to check data types



```
df.isnull().sum()
```

```
House Number                        3
Street                              1
First Name                          2
Surname                             1
Age                                 0
Relationship to Head of House       0
Marital Status                   2764
Gender                              2
Occupation                          0
Infirmity                           0
Religion                         2824
dtype: int64
```

Figure 2: Checking for null values

Some of the errors were corrected by inferring from other columns. For example, missing surnames were inferred from family members living in the same street and house number. Missing gender was inferred from relationship to head of house and first name. There were errors regarded as typos that were corrected individually, some integers written in words, were also corrected.

For occupation, it was observed that the retirement age was 70 for the town and there were people above 70 years who were classified as unemployed, this was corrected to retired.

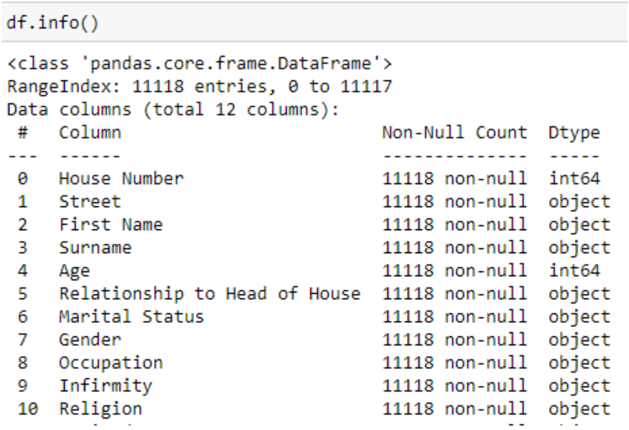
On the age column, there was missing data and an outlier where a person filled his age as 150 years which is untrue as the oldest person ever whose age has been independently verified was 122 years (Whitney, Craig R, 1997). The age was changed to 50 years based on the age of his son and wife, it is close to the mean ages of those in the same occupation, he isn’t retired, and might be a typographical error. A missing age was filled as 17years as the occupation was a student and the average age of parents having a 17year old was similar to his parents. A negative age (-1) was corrected to 1 year, as it was a newborn that has a name.

On the Marital status column, it was observed that all persons below 18 years had a null value, and some of the marital statuses were abbreviated. To clean this, all persons below 16 years were classified as “never married” as it is illegal for anyone under 16 to get married in the UK (Marriage Act 1949:s2). The same classification was used for the 2021 census in the United Kingdom (UK Government, 2021). Those between 16 -17 years living alone and the “relationship to head of house” is not “Head”, were assigned to be single as it is possible to get married with parental consent (Marriage Act 1949:s3). Those who were 18 and their "relationship to head of the house" is not "Husband" or "wife" were assigned single. The abbreviations were harmonized accordingly.

For religion, it was observed that all below 18 years had NaN values, these were assigned “None” because it was not possible to check transmission of religion from parents to children. Those above 18 years with NaN values were also assigned “None”. Jedi has been said to not be a religion in previous censuses (BBC, 2016). Buddhist, Hindu, Sikh, Jewish were converted to “None” due to very insignificant followership. Female and Nope, were determined to be wrong entries and regarded as "None". Undecided, Pagan, Agnostic was determined to mean no religion and were assigned "None”

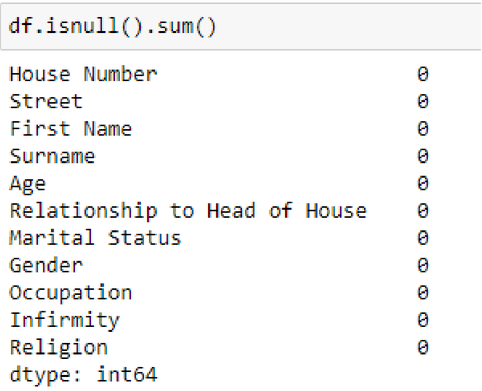
DESCRIPTIVE ANALYSIS

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



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11118 entries, 0 to 11117
Data columns (total 12 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   House Number                  11118 non-null  int64 
 1   Street                        11118 non-null  object
 2   First Name                    11118 non-null  object
 3   Surname                       11118 non-null  object
 4   Age                           11118 non-null  int64 
 5   Relationship to Head of House 11118 non-null  object
 6   Marital Status                11118 non-null  object
 7   Gender                        11118 non-null  object
 8   Occupation                    11118 non-null  object
 9   Infirmity                     11118 non-null  object
 10  Religion                      11118 non-null  object
```



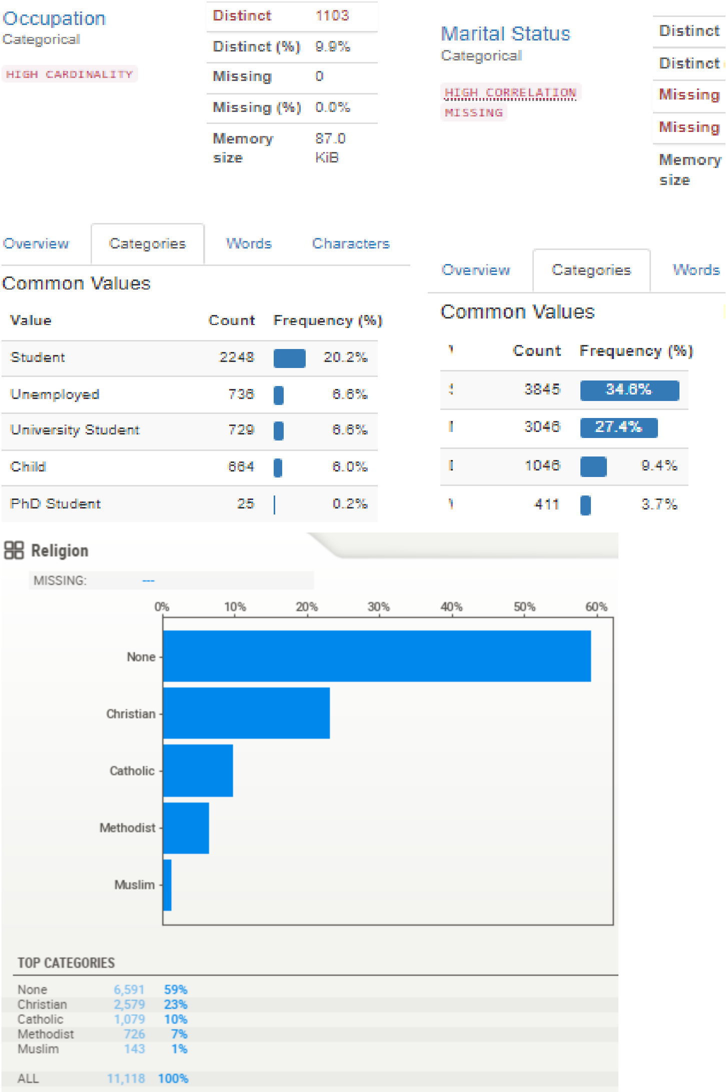
```
df.isnull().sum()
```

```
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
```

Figure 3: checking info of the data after cleaning

Figure 4: checking null values after cleaning

An overview of the data statistics shows that the general population is healthily represented by less than 1% infirmity rate, a highly irreligious town where 59% 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 or married.



Occupation
Categorical

HIGH CARDINALITY

Distinct	1103
Distinct (%)	9.9%
Missing	0
Missing (%)	0.0%
Memory size	87.0 KiB

Common Values

Value	Count	Frequency (%)
Student	2248	20.2%
Unemployed	738	6.6%
University Student	729	6.6%
Child	664	6.0%
PhD Student	25	0.2%

Marital Status
Categorical

HIGH CORRELATION
MISSING

Distinct
Distinct
Missing
Missing
Memory size

Common Values

	Count	Frequency (%)
	3845	34.6%
	3046	27.4%
	1046	9.4%
	411	3.7%

Religion

MISSING: —

TOP CATEGORIES

None	6,591	59%
Christian	2,579	23%
Catholic	1,079	10%
Methodist	726	7%
Muslim	143	1%
ALL	11,118	100%

Figure 5: Statistical overview

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 15-19, 40-44, these are as a result of migration into the town of new university students and lodgers who moved into the town as a result of employment respectively. The decrease between ages 25-29 could be attributed to graduating students leaving the town.

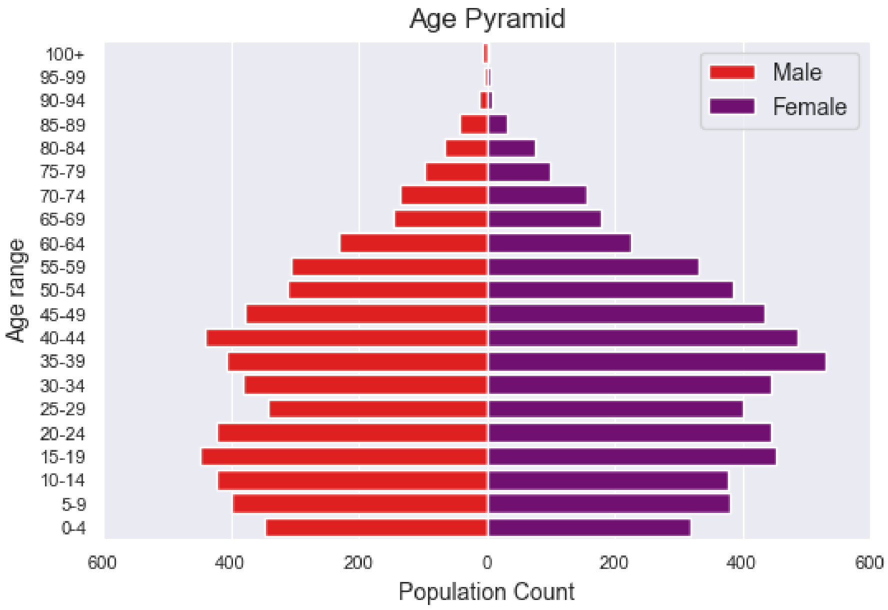
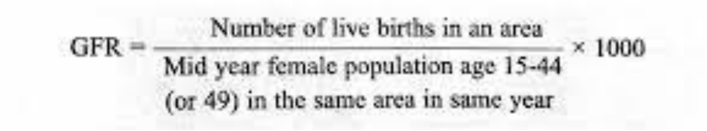


Figure 6: Age pyramid of the town

DETAILED ANALYSIS

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(theOECD. 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.



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

Figure 7:GFR (pinterest.com)

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). The crude birth rate for the town is 10 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.

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

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

Figure 8 CBR and CDR (geog100.org)

The death rate is calculated by estimating deaths by the difference in age bands for those over 55 years. From the age pyramid figure, you would notice there were decreases in other age groups, these are attributed to migration. However, a decrease for those above 55years hints at death. The death rate was calculated as 6 deaths per thousand. This was calculated by summing the death rates per annum for each age range above 55years.

There was a decrease in birth rate from 5 years ago to 10 births, but when compared to the death rate for the year, it signifies the population is still growing.

Infirmity

There are less than 1% of the population with infirmity which indicates a very healthy population. The rate of people above 55years with infirmity was calculated as 0.1% of the population. These indicate that there is an excellent health care system available for the town and an extended care program for the older population. Building another health care system would not be a priority for the plot of land based on this evidence of the health status of the population.

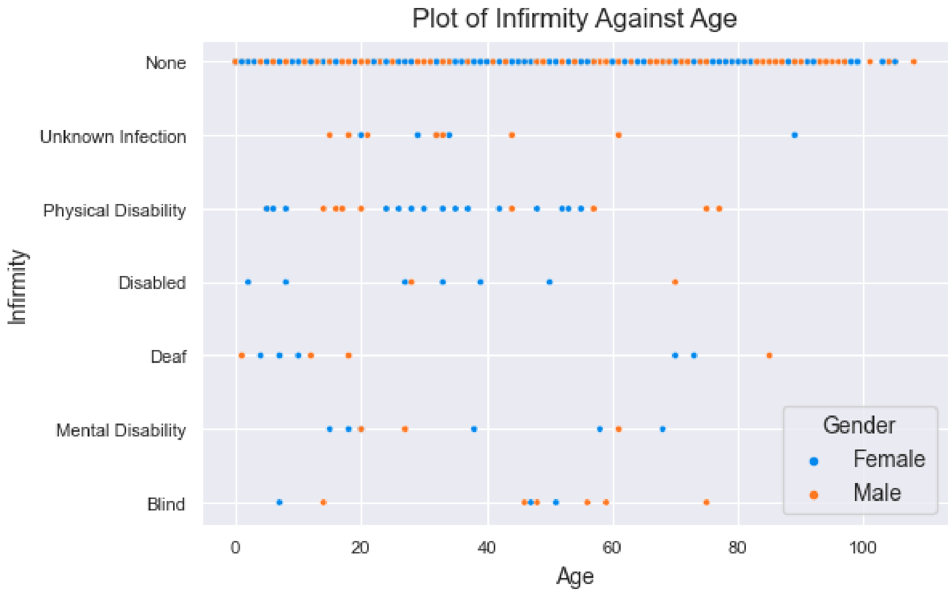


Figure 9: A scatterplot of Infirmity against age

Religion

From the data, it is seen that a larger proportion of the population (59%) 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

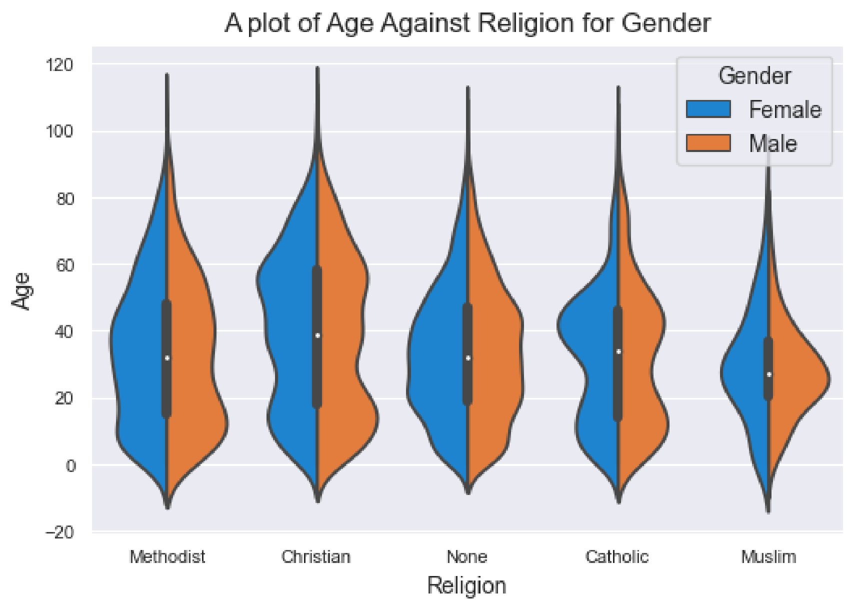


Figure 10: violin plot of Age and Religion

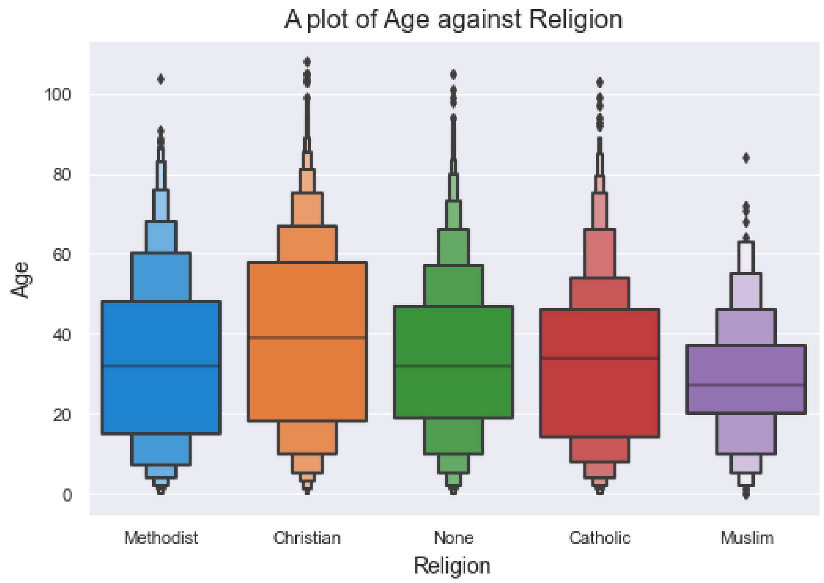


Figure 11:whiskerplot of Age and Religion

Employment and 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.

Occupations such as booksellers, tutors teachers (excluding university lecturers), lawyers, and retailers were considered noncommuting jobs. This means over 75% of the employed population commute to work.

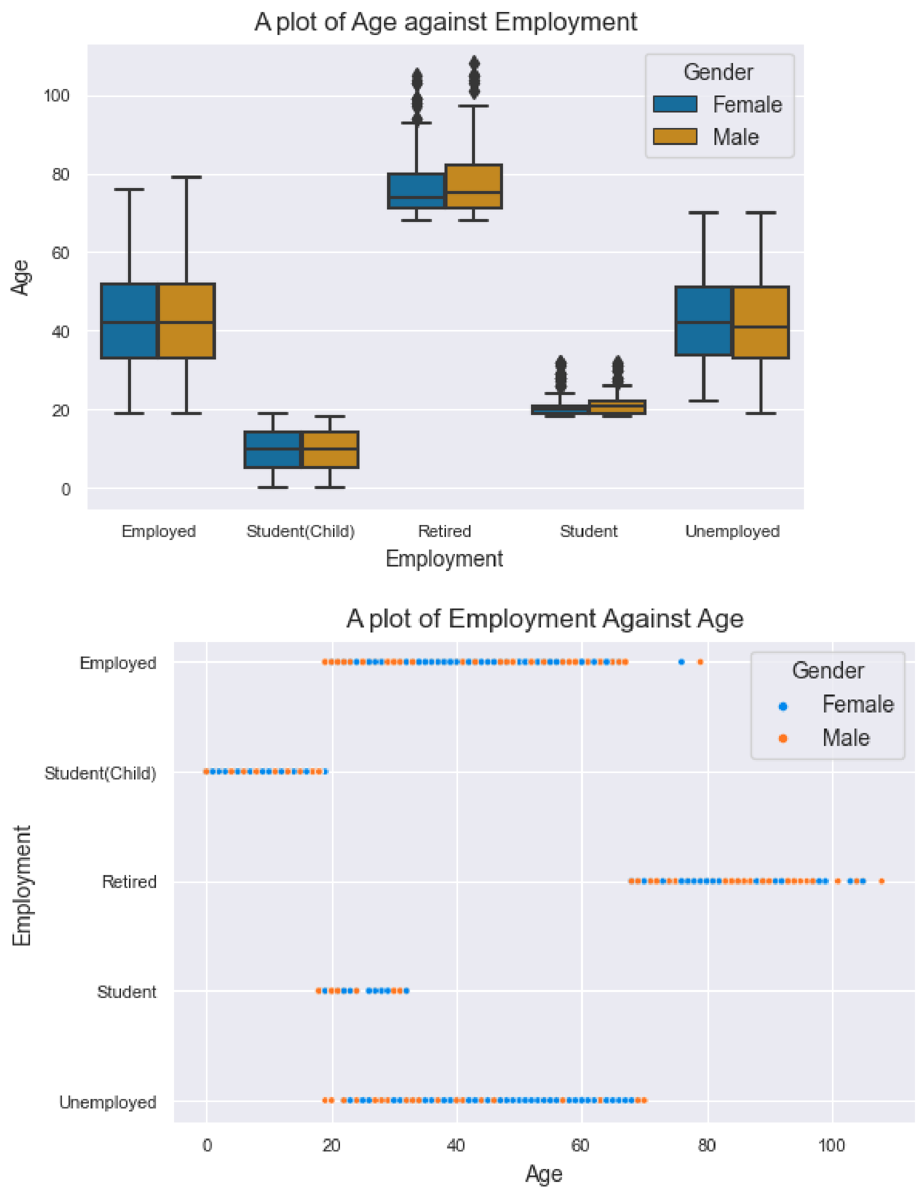


Figure 12:box and scatterplot of Age and employment

Divorce and marriage

As seen from that, 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.

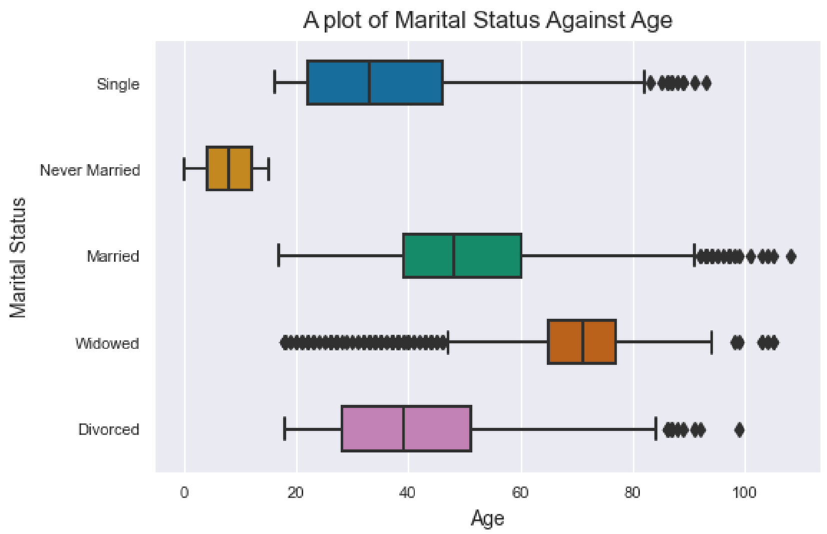


Figure 13: Box plot of marital status and age

Occupancy

The occupancy level was calculated by getting 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 bedroom. Based on this, it was discovered that, of the 3,747 apartments in the town, 1,360 were over-occupied. This represents 36% of the apartments in the town. The source of this could be attributed to University students sharing an apartment meant for one person, families, and divorcees subletting their houses to lodgers.

Migration

Migration was calculated based on the number of visitors (lodgers) in the town and divorcees who have left the town. Students make up a large number of 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 the immigration rate was calculated to be 34 immigrants per thousand. For emigration, this was calculated as 18 per thousand. This implies almost twice as many people are moving into the town than leaving. Matching these figures with corresponding birth and death rate signifies that the population is increasing this would lead to increasing pressure on existing infrastructural facilities.

RECOMMENDATION

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 population 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 population 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 decreasing birth and fertility rate. The population is largely employed and there’s no urgent need for training.

CONCLUSION

From analysis, it was deduced that the subject town is a growing town that requires investment in low-density housing and general infrastructure to cater to population growth. A data-led decision making would reduce the probability of making wrong decisions however, the accuracy relies heavily on the quantity and quality of available data. One of the limitations encountered in this project was insufficient data, more data would be required to be able to make trends and comparisons for more accurate results. This would greatly improve insights generated and recommendations being made from the data set.

REFERENCES

BBC News. 2006. Jedi is not a religion, Charity Commission rules. [online] Available at: https://www.bbc.com/news/uk-38368526 [Accessed 29 November 2021].

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 2
Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/2 [Accessed 28/11/2021]

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

The Guardian. 2021. Less than half of Britons expected to tick 'Christian' in UK census. [online] Available at: https://www.theguardian.com/uk-news/2021/mar/20/less-that-half-of-britons-expected-to-tick-christian-in-uk-census [Accessed 9 December 2021].

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/doh-shad/view/sharedstatic/CrudeBirthRate.pdf [Accessed 25 November 2021].