CENSUS PROJECT REPORT

771766_A21_T1: Fundamentals of Data Science

001_CWRK:Project Report

INTRODUCTION

This report focuses on various steps taken to make deductions based from a given set of mock data. The data is collected from a moderately modest town in between 2 cities connected by motor ways. The aim of this report is to analyse the data collected and use the patterns or correlations in the data to decide on what to build on an unoccupied plot of land that the local government wishes to develop and also what should be invested in.

During the course of several workshops, several analyses has been made on the data in order to determine the above. The report focuses on the various steps taken from cleaning, labelling, visualising, plotting, etc that has been done on the data over a series of workshop.

DATA CLEANING

The census data collected contained a lot of errors viz missing values, nan, misspelling, etc which has to be cleaned before any analysis can be carried out. A log book which contained all the errors found was made from which the jupyter notebook was found.

1. **NAN and Missing Values**:

Only 3 columns contained nan values which are house number, marital status and religion with marital status and religion containing the highest number of nan values due to children aged below 18. These were changed to 'minors' in the case of marital status and 'Undeclared' in the case of religion since it will be inappropriate to assume the religion of the head of house. In the case of 'House number' it was changed to the appropriate number by inferring from the data.

2. **EMPTY SPACES:**

'HOUSE number', street, first name, surname, relationship to head of house, gender, occupation, infirmity and religion all contained blank spaces. This was a thorough part of the data cleaning process. By inferring from the data, they were all filled with the appropriate values. Empty spaces in infirmity were converted to missing since it cannot be inferred from the data if one is sick or not and unknown was already a unique value.

Also, missing values in religion was converted to unknown as this cannot be inferred from the data.

3. **Misspellings:**

Unique values that contained misspellings were corrected by visual inspection of the unique values and corrected.

4. **Data Type:**

The age column was casted to integers since it would be inappropriate for someone's age to be a float or a string. Another reason this was done is because of the calculations and comparisons that will be done with the age column.

5. Changes

Unknown infection was converted to unknown and physical disability and mental disability was converted to disabled since they connote the same meaning and also to reduce variations in the infirmity column for better visualisation.

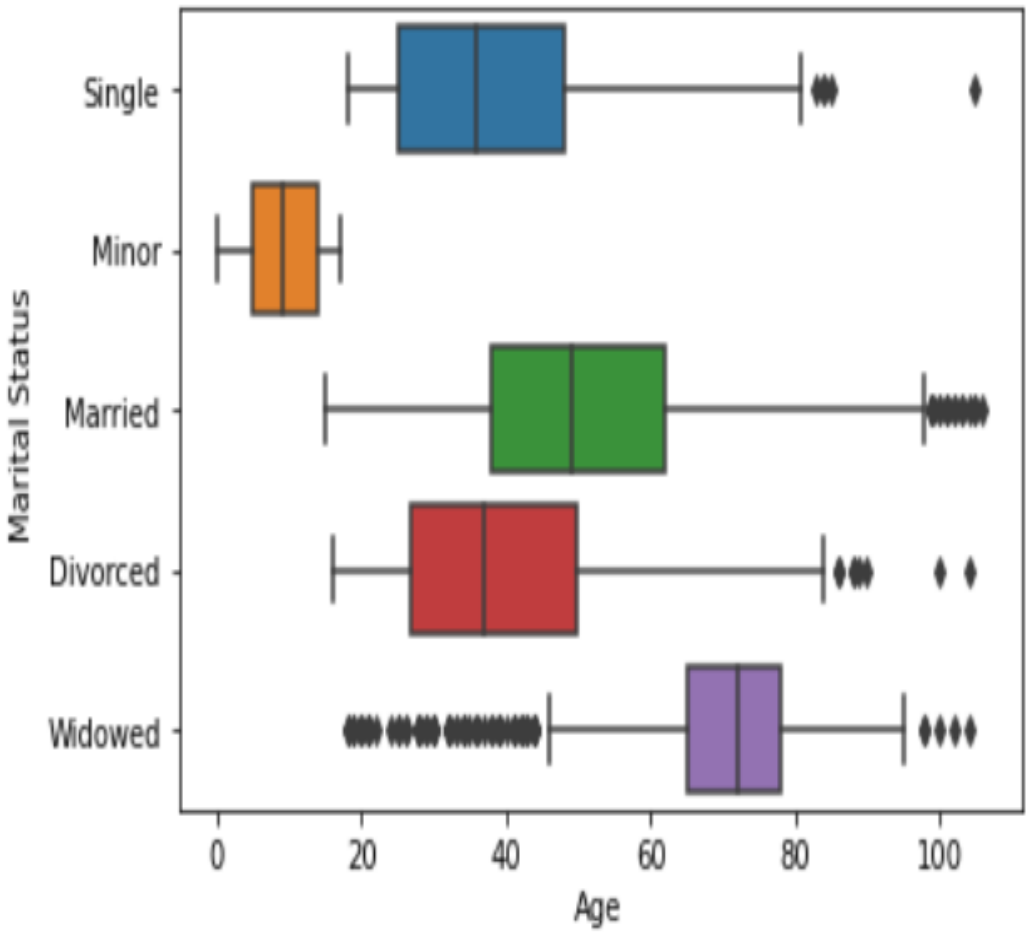
In the religion column, ‘nope’ and ‘none’ were changed to Agnostic since they are closely related in meaning.

Sith which is not a religion was converted to unknown. (BBC, 2016).

Catholic, Methodist, quaker and Baptist were all converted to Christian since they are all denominations of Christianity.

6. Outliers

Two households were removed from the data as they were aged 15 and 16, were head of the house (Marriage Act, 1949:s3) and one was married while one was divorced. No outlier was found in the age column as the max age found was 106.




```
count      10480
unique         5
top       Single
freq        3705
Name: Marital Status, dtype: object
```

Although there are several outliers above, it is not unusual people to become widowed at old age as much as it for someone to become widowed at 20. Therefore, these records were considered too unlikely to be correct and imputed to ‘Single’ if 18 or ‘minor’ if under 18 ((NSPCC, 2020).

POPULATION DEMOGRAPHICS:

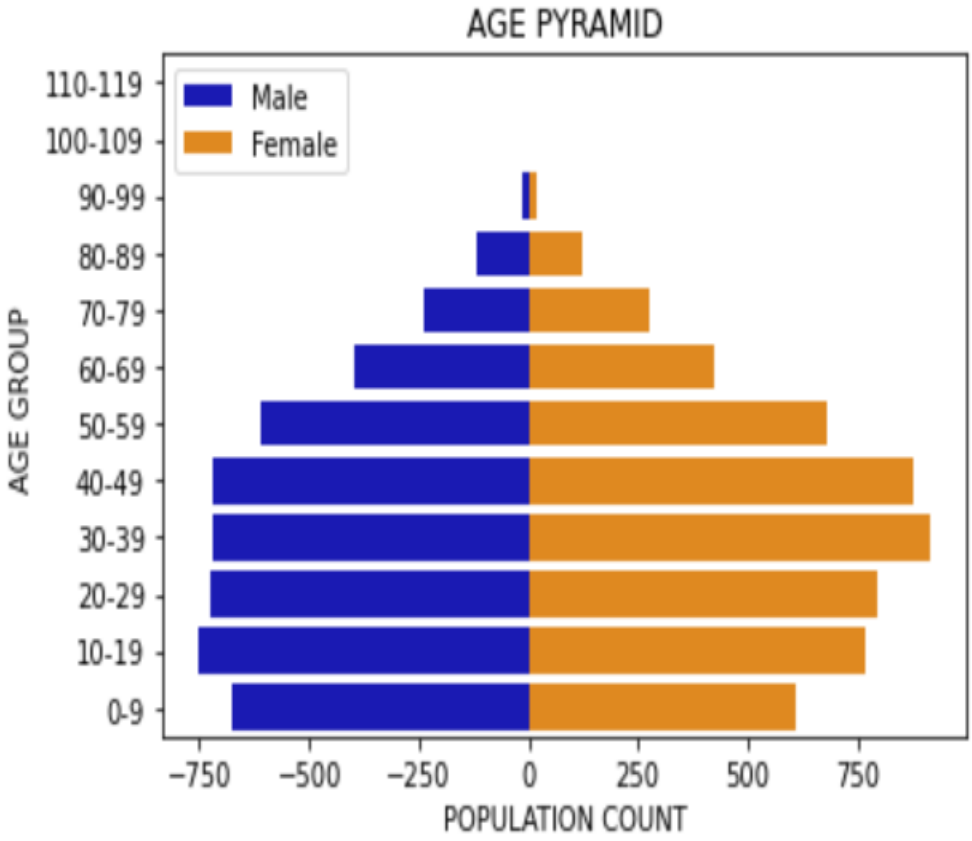
After data cleaning has been done a new column *Employed* was added to get the required number of retired people.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10480 entries, 0 to 10479
Data columns (total 12 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   House Number                    10480 non-null  object
 1   Street                          10480 non-null  object
 2   First Name                      10480 non-null  object
 3   Surname                         10480 non-null  object
 4   Age                             10480 non-null  int32
 5   Relationship to Head of House   10480 non-null  object
 6   Marital Status                  10480 non-null  object
 7   Gender                          10480 non-null  object
 8   Occupation                      10480 non-null  object
 9   Infirmity                       10480 non-null  object
 10  Religion                        10480 non-null  object
 11  Employed                        10480 non-null  object
dtypes: int32(1), object(11)
memory usage: 941.7+ KB
```

AGE PYRAMID

An age pyramid was constructed to better visualise the various age distribution between the groups. The ages of the population were divided into 12 categories with an age band of 10, the minimum being 0-9 and maximum being 110-119.

By inspecting the age pyramid below, it significantly shows a steady rise in count between ages 20-49 suggesting a rise in the middle class and a decline as the pyramid moves upwards from 50-109 as expected. It also shows a slightly lower number of young people compared to middle-aged, especially those aged 0-4, suggesting a low birth rate perhaps. The population also tends to live well into old age, for both male and female.

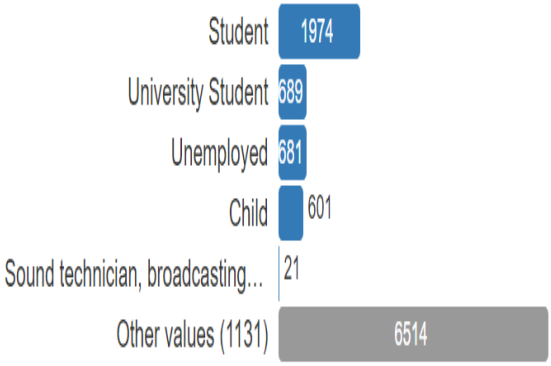


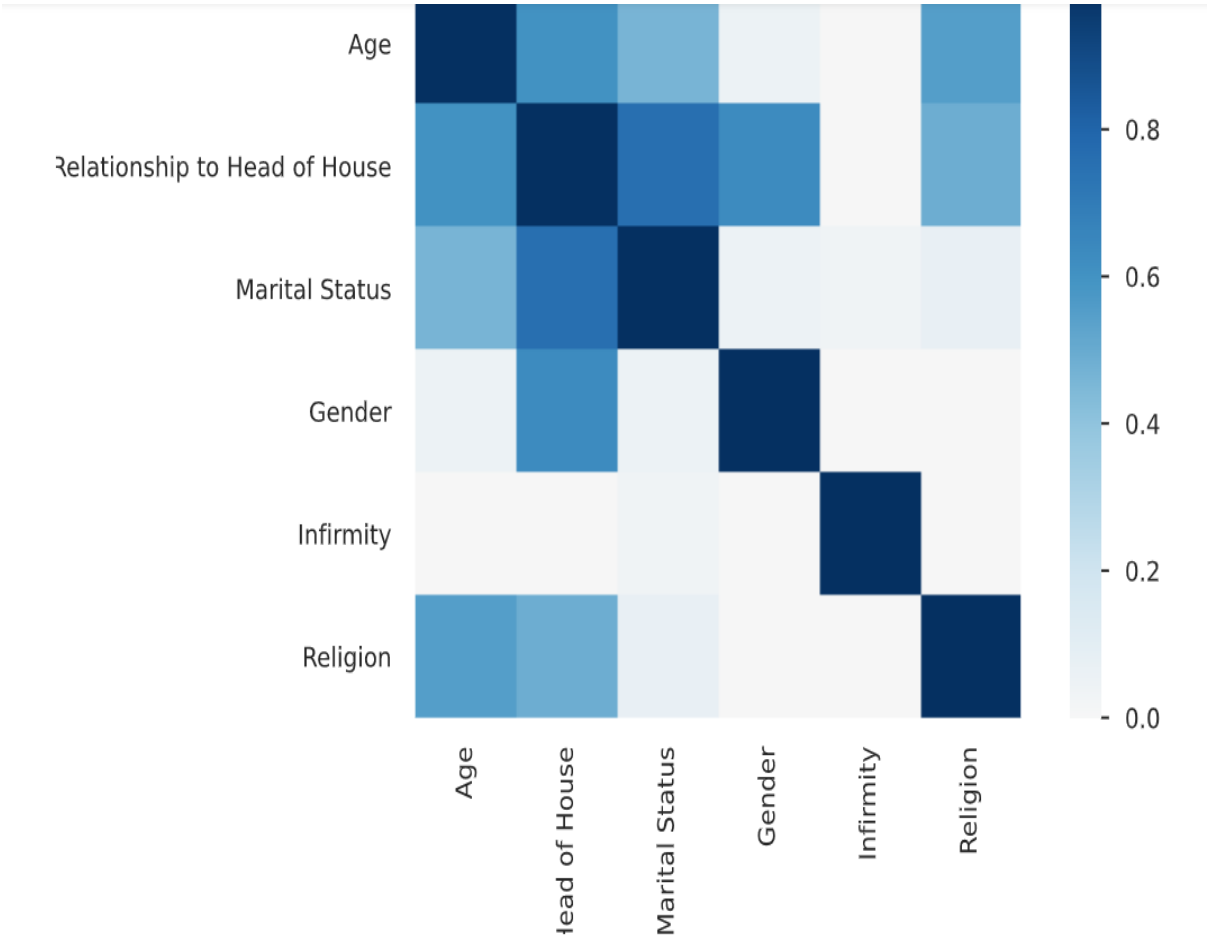
Occupation

Categorical

HIGH CARDINALITY

Distinct	1136
Distinct (%)	10.8%
Missing	0
Missing (%)	0.0%
Memory size	82.0 KiB

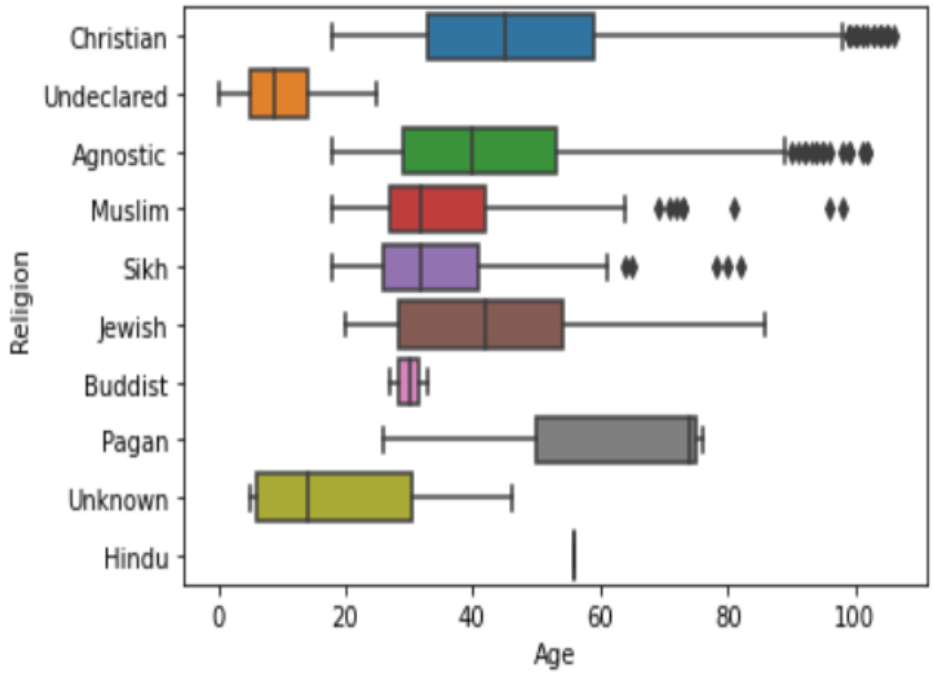


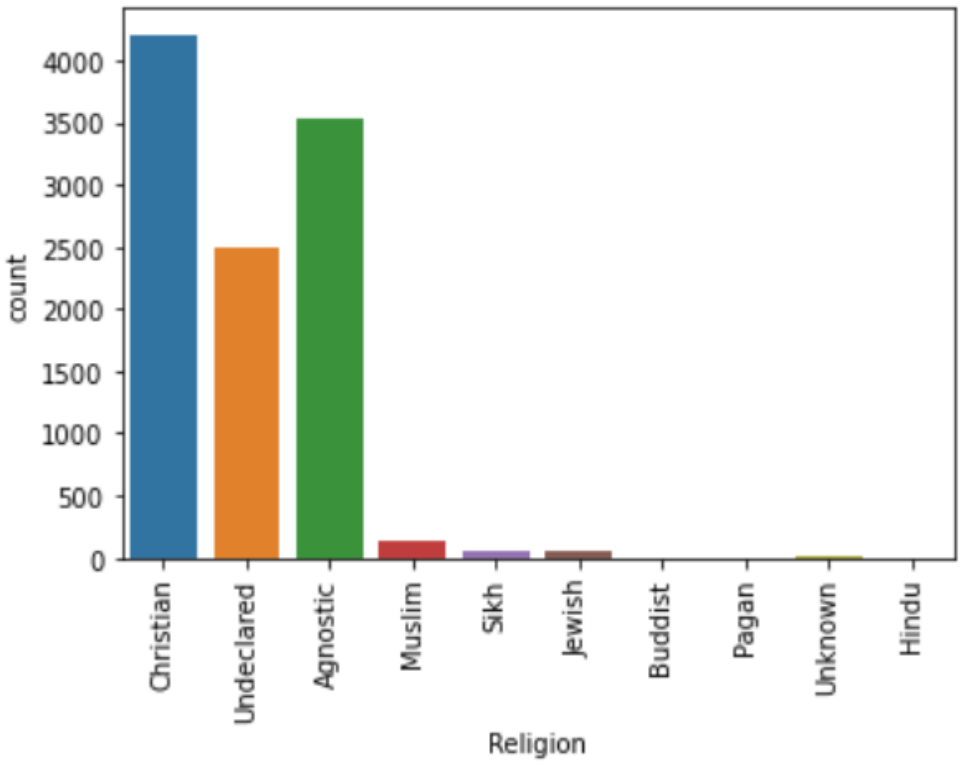

Age
Relationship to Head of House
Marital Status
Gender
Infirmity
Religion
Age
lead of House
Marital Status
Gender
Infirmity
Religion
0.8
0.6
0.4
0.2
0.0

DATA ANALYSIS:

Here I will look at correlations that exist between different columns and draw conclusions on what should be built on an unoccupied plot of land that the local government wishes to develop.

A. RELIGION





From the above we can see that Christians has the highest number of counts, the followed by agnostic. The average number of each religious group was calculated per 1000 to see if there is a growing need for a religious building to be built.

B. POPULATION DENSITY

	House Number	Street	Occupancy Count
0	1	Alba Keys	2
1	1	Arch Avenue	3
2	1	Ash Estates	1
3	1	Autumn Shoal	3
4	1	Bailey Islands	5
...	...	...	...
3744	99	Coventry Causeway	4
3745	99	Horton Drive	2
3746	99	Kelly Pass	1
3747	99	Newfound Corner	2
3748	99	Thomas Greens	5

3749 rows × 3 columns

A new column called 'Occupancy Count' was calculated which added to the data frame which gives us an idea of the number of people per street living in a particular household. Assuming that all houses on a given street are built similarly (i.e., same number of bedrooms per house on a given street), then the mode will be the average number of people per house. This was calculated using the formula:

Household Density= mode(occupancy count)- mean(occupancy count)

The household density was found to be -1.79 which is an indicator that there is pressure on housing.

C. **DIVORCED AND MARRIAGE**:

As seen from the wisker plot between Religion and Age, a lot of people were divorced between the ages of 25 -50 with a mean age of 30. Marital status split by gender identifies there are more female divorcees than there are male, indicating male divorcees potentially leave the town. This will be particularly important when calculating emigrants.

D. Birth and Death Rate

The birth rate and death rate are a powerful indicator whether the town is growing or shrinking. Since the data for the previous year census was not given, it is impossible to calculate the birth and death rate. Instead, we calculate the Crude birth and death rates.

$$\text{CBR} = \frac{\text{No. of live births occurring in the year}}{\text{Average population in the year}} \times 1000$$

The current birth rate is 9 births per 1000. This is calculated by identifying live biths of children aged 0 and applying the formula above. Four years ago the birth rate was estimated to be 13 births per 1000. This is an indicator that the population is shrinking.

The death rate is calculated by estimating deaths by difference in age-bands from 65-99. Age bands were created from 65 and above and the crude death rate was calculated.

Crude Death Rate

$$\text{Crude birth rate} = \frac{\text{Number of births during the year 2009}}{\text{Mid-year population}} \times 1000$$

The death rate was found to be approximately 14 deaths per 1000.

E. MIGRATION

University Students, sea workers and pilots constitute most emigration and immigration in the town. Immigration statistics are calculated from lodgers and visitors who are single. This is to exclude those who are divorced and are lodging after leaving their spouse and would not classify as immigrating to the town. On the other hand, emigration statistics are calculated from the difference in male and female divorcees. By this method, there are approximately 38 immigrants to the town per thousand people. Using only the difference in divorced males and females, emigration from the town is 18 per thousand.

F. Employment and Commuters

Commuters are identified based on the following methodology:

- Anyone who identifies as a University Student (including PhD Students)
- Anyone with occupation in aviation or water since the town is between two larger towns.

Looking at the population, anyone aged from 65 is considered retired. The average number of retired people per 1000 is 90. The average number of university students per 1000 is 65 while the average number of unemployed people is approximately 60. This means the town has a lot of retired people aged 65 and above and should be looked into.

Recommendations

From the inspection of the data, since the immigration rate is higher than the emigration rate, it means lots of people are coming into the town hence and there are a high number of commuters there is need to build a train station since there are potentially a lot of commuters in the town and building a train station could take pressure off the roads. Also, building more high-density houses should be considered as the house hold density of the population is not up to one meaning there is a lot of pressure on the existing houses built.

Also, old aged care should be considered as there are a lot of retired people in the population. Since the number of aged people is increasing across the age 65 in the next 10 years there will be more retired people as the population change decreases.

Also, as the unemployment rates is high, avenues for employing and training people with skills should be also considered. If this can be implemented, it will help curb the decrease in population as there will be more available jobs thereby increasing emigration and reducing immigration. The more people are employed lesser the unemployment rates.

Other services to invest in include road and house maintenance as this will help the growing immigration rate. If these can be implemented, then there will be a rise in population growth and a decline in household density thereby creating better opportunities for immigrants and occupants of the town.

REFERENCES

BBC (2016) Jedi is not a religion, Charity Commission rules. Available online: https://www.bbc.co.uk/news/uk-38368526 [Accessed 09/12/2021]

Marriage Act (1949) Section 3 Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3 [Accessed 10/12/2021]

National Society for the Prevention of Cruelty to Children (2020) Moving Out Available online: https://www.nspcc.org.uk/keeping-children-safe/in-the-home/moving-out/ [Accessed 8/12/2021]