Census Project Report

This report analyses the census of a moderately sized town to make recommendations for investment in future services and use cases for developments on an unused plot of land. To make these recommendations, the census data has firstly been cleaned to correct data errors and missing records, which is detailed in the first section of this report.

Subsequent sections of the report will highlight key analyses undertaken specifically aimed at to support recommendations provided. This includes, initially, an overview of the town's population demographics, followed by detailed analysis of the town's predicted population growth, employment trends, commuters, and occupancy rates.

Data Cleaning

Census data was cleaned to correct the data errors; a full logbook of all cleaning undertaken can be found in the corresponding Jupyter Notebook.

Blank data were imputed by inferring information from an individual's record or others in their household (in the case of missing surnames). Records that could not be imputed with referential information were imputed to 'None' or 'Unknown' (in the case of Occupation).

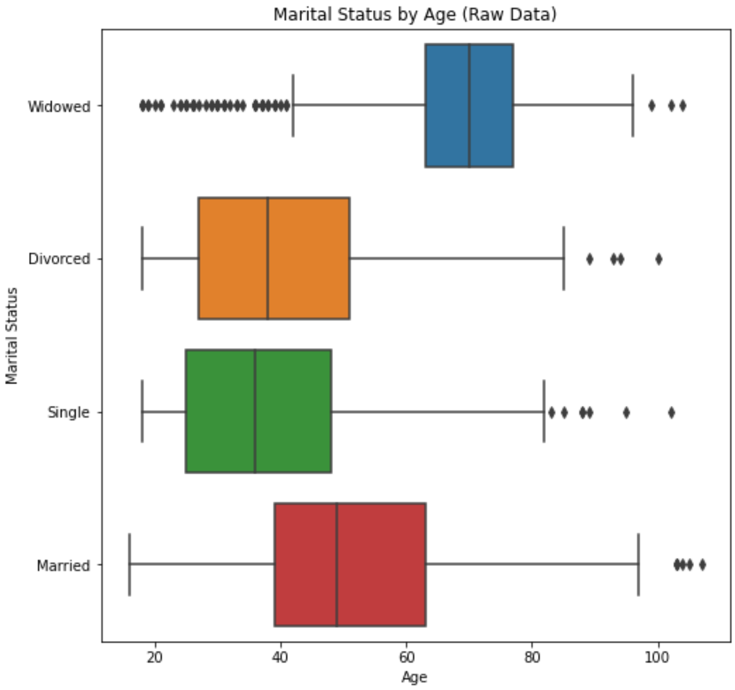
Four religions were converted to 'None' - Sith, Undecided, Housekeeper and Ironer. Sith was determined to be a deliberately misleading input due to only four entries and this has been known to be written as a 'joke' entry in previous censuses (BBC, 2016). Housekeeper and Ironer were deemed to be individual errors.

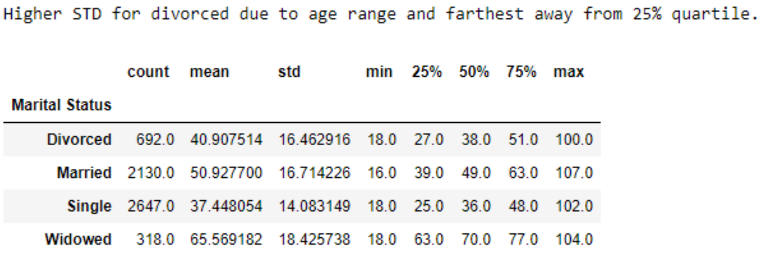
Religion for minors (under 18) were primarily NaN entries, aside from 7 which were a mixture of None or Triangulism. Given NaN was the most frequent value for under 18s, all those with an entry in this variable under 18 where converted to NaN. The impact of this is it will not be possibly to analyse religious transmission from parent to children – though this would have been difficult regardless, given most children had NaN as a religion.

Similarly, individuals under the age of 18 have NaN as a marital status. An exception has been made for those 16 or above, as it is legal to marry with parental consent (Marriage Act, 1949:s3). A further exception has been made for those under 18 who do not live with another over 18 in the household, whose records have not been imputed as it is possible to move out of a family home before 18 (NSPCC, 2020).

One household was removed from the dataset. In this household, a child of 15 was married with a child, listed as Head of the household, and had a husband over 18. As there are two major inconsistencies or a potentially illegal relationship, this household was dropped from the dataset; it was not deemed significant (three records) to the overall analysis enough to impute multiple counts of incorrect data across a household.

Outliers were removed for one record of an age at 111 (imputed to match the spouse's age) and four individuals widowed at age 18 or below, as detailed in the boxplot distribution:

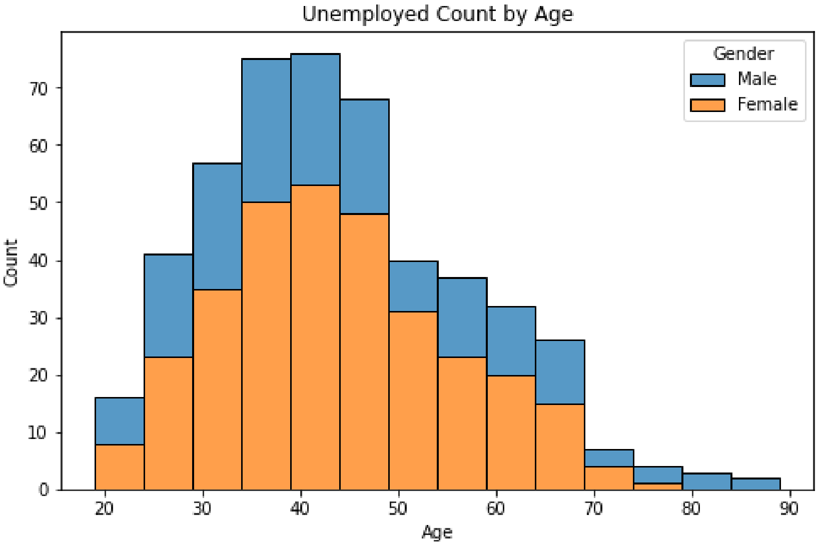




```
Higher STD for divorced due to age range and farthest away from 25% quartile.
```

Marital Status	count	mean	std	min	25%	50%	75%	max
Divorced	692.0	40.907514	16.462916	18.0	27.0	38.0	51.0	100.0
Married	2130.0	50.927700	16.714226	16.0	39.0	49.0	63.0	107.0
Single	2647.0	37.448054	14.083149	18.0	25.0	36.0	48.0	102.0
Widowed	318.0	65.569182	18.425738	18.0	63.0	70.0	77.0	104.0

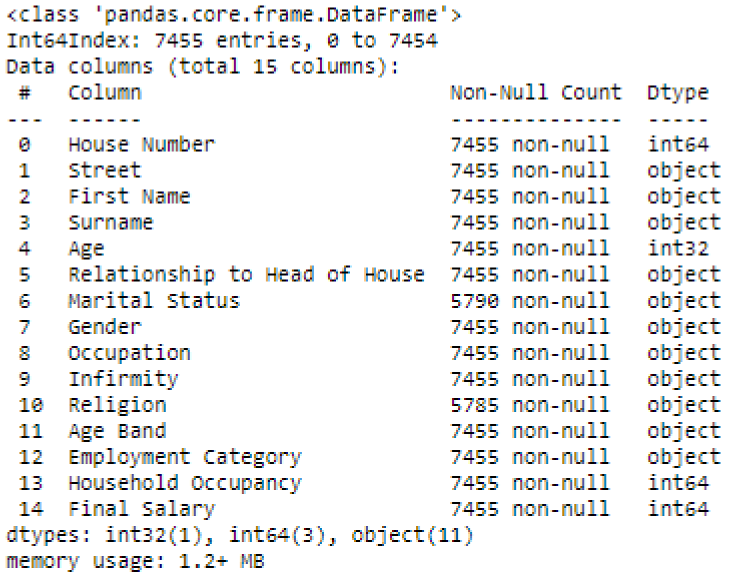
Although there are several outliers above, it is not unusual people to become widowed at old age as much as it for someone to become widowed at 18. Therefore, these records were considered too unlikely to be correct and imputed to 'Single' (if 18) or NaN (if under 18).



Records which listed Occupation as 'Unemployed' for individuals over 65 were imputed to 'Retired, unemployed'. Although it is possible to be unemployment upon reaching retirement age, these should not be considered in future analyses of unemployment as those over 65 would not usually be eligible for work (Gov.uk).

Population Demographics

Once data cleaning has been completed, the finalised census data will have the following features:



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7455 entries, 0 to 7454
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   House Number                  7455 non-null   int64 
 1   Street                        7455 non-null   object
 2   First Name                    7455 non-null   object
 3   Surname                       7455 non-null   object
 4   Age                           7455 non-null   int32 
 5   Relationship to Head of House 7455 non-null   object
 6   Marital Status                5790 non-null   object
 7   Gender                        7455 non-null   object
 8   Occupation                    7455 non-null   object
 9   Infirmity                     7455 non-null   object
 10  Religion                      5785 non-null   object
 11  Age Band                      7455 non-null   object
 12  Employment Category           7455 non-null   object
 13  Household Occupancy           7455 non-null   int64 
 14  Final Salary                  7455 non-null   int64 
dtypes: int32(1), int64(3), object(11)
memory usage: 1.2+ MB
```

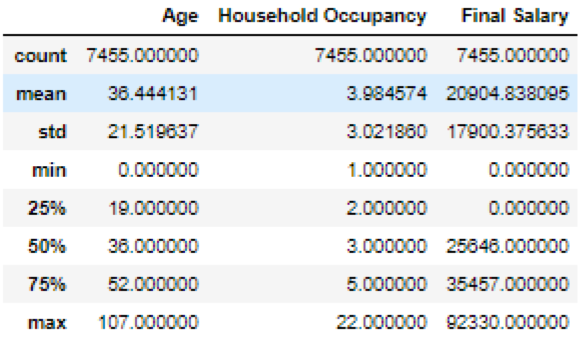
To aid analysis, the following has been added:

- **Age band:** Ages placed into 5-year age bands for population pyramid.
- **Employment category:** Simplified occupations with the values: Student (Child), Student, Employed, Unemployed, Retired.
- **Household Occupancy:** a count of all individuals in a household. *Note: this column should not be used for summation or counts*.
- **Final Salary**

Final Salary was calculated using a fuzzy match in spreadsheet software from occupations to Standard Occupation Classifications (ONS, 2016). All matches below 95% similarity were checked and non-matches were input manually. It was therefore possible to attach a median salary from the Office of National Statistics for census occupations based on SOC codes, including those who were retired (using former occupation). Those who were students, children, unemployed or unemployed upon retirement attracted a salary of £0.

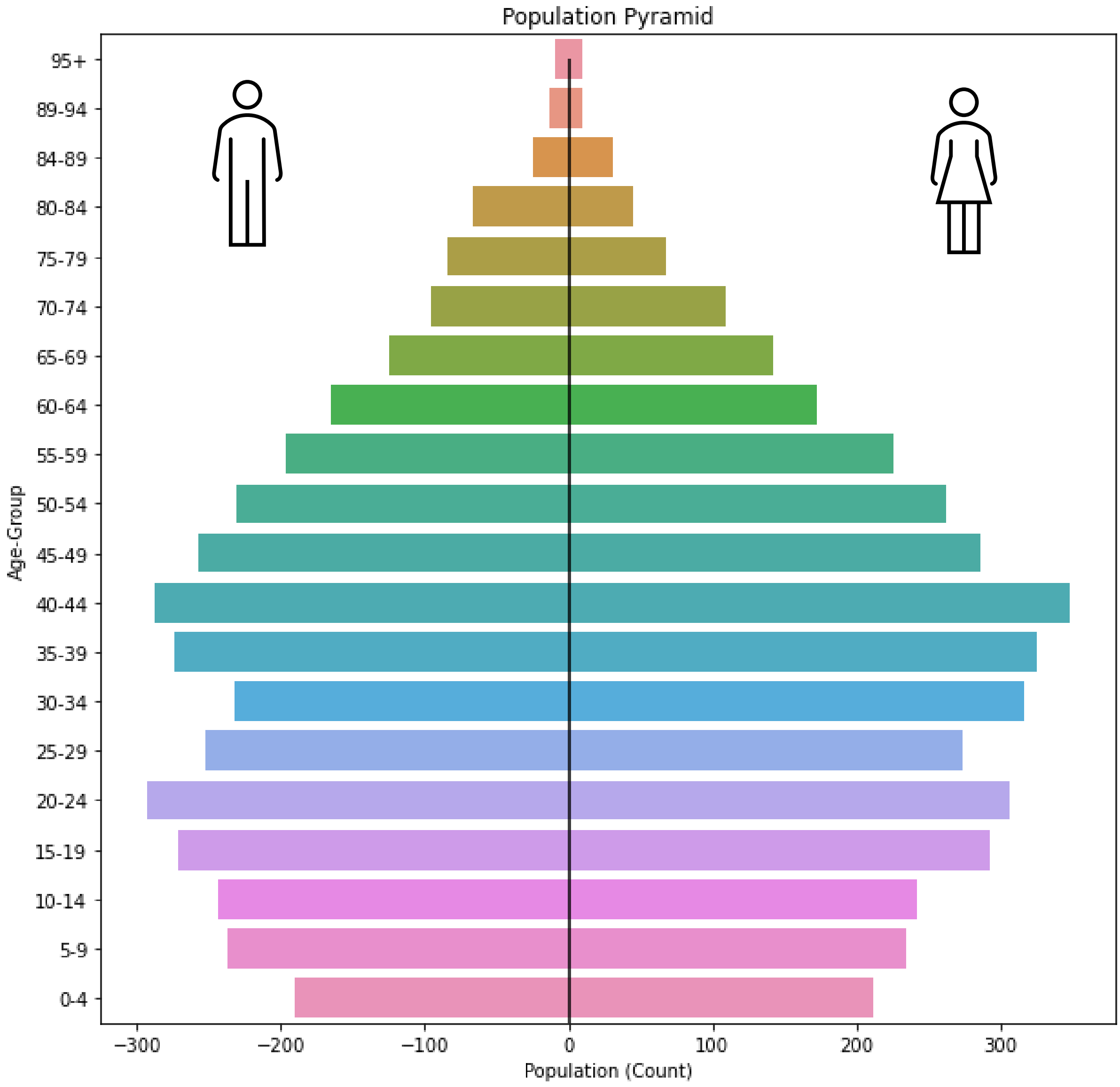
Demographics

Summary statistics describe a town of relatively good wealth, though large gaps between the lowest earners and highest; a mean age of 36 and a high occupancy rate for households, which will be examined in later sections.



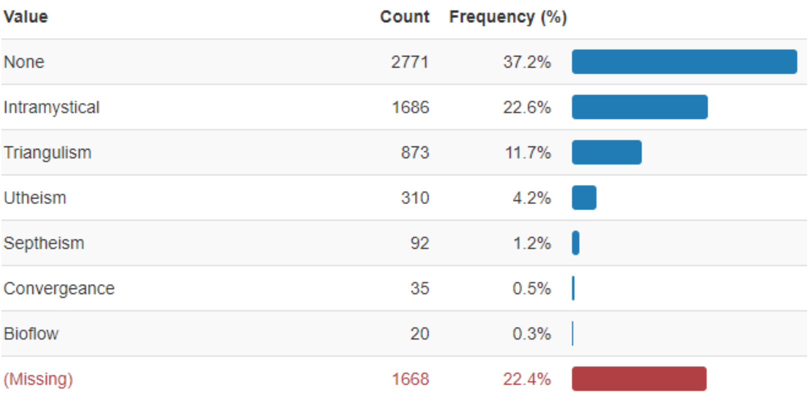
	Age	Household Occupancy	Final Salary
count	7455.000000	7455.000000	7455.000000
mean	36.444131	3.984574	20904.838095
std	21.519637	3.021860	17900.375633
min	0.000000	1.000000	0.000000
25%	19.000000	2.000000	0.000000
50%	36.000000	3.000000	25646.000000
75%	52.000000	5.000000	35457.000000
max	107.000000	22.000000	92330.000000

Examining the population pyramid, the structure of the population shows a slightly lower number of young people compared to middle-aged, especially those aged 0-4, suggesting a low birth rate. The population also tends to live well into old age, for both male and female.



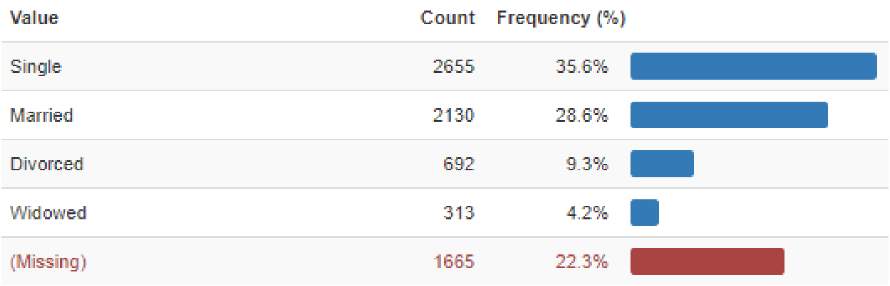
Further descriptive analysis shows most of the town are employed, with a high number of school-age children and around 6% of the population unemployed. Most of the population are married or single and infirmity in the town is low compared to other variables; those with an infirmity constituted less than 1% of the overall population:

Religion



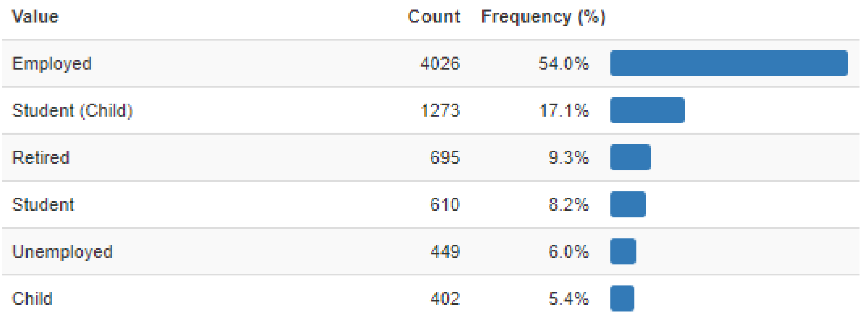
Value	Count	Frequency (%)
None	2771	37.2%
Intramystical	1686	22.6%
Triangulism	873	11.7%
Utheism	310	4.2%
Septheism	92	1.2%
Convergeance	35	0.5%
Bioflow	20	0.3%
(Missing)	1668	22.4%

Marital Status



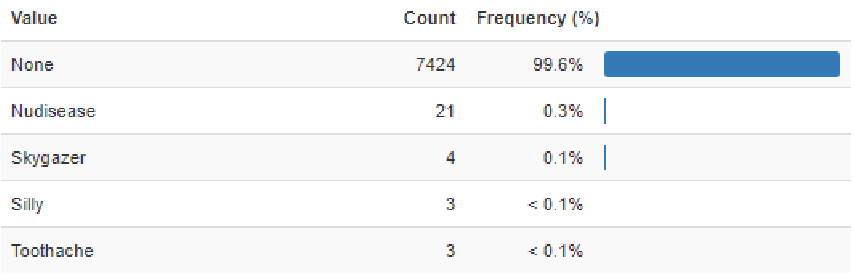
Value	Count	Frequency (%)
Single	2655	35.6%
Married	2130	28.6%
Divorced	692	9.3%
Widowed	313	4.2%
(Missing)	1665	22.3%

Employment



Value	Count	Frequency (%)
Employed	4026	54.0%
Student (Child)	1273	17.1%
Retired	695	9.3%
Student	610	8.2%
Unemployed	449	6.0%
Child	402	5.4%

Infirmity



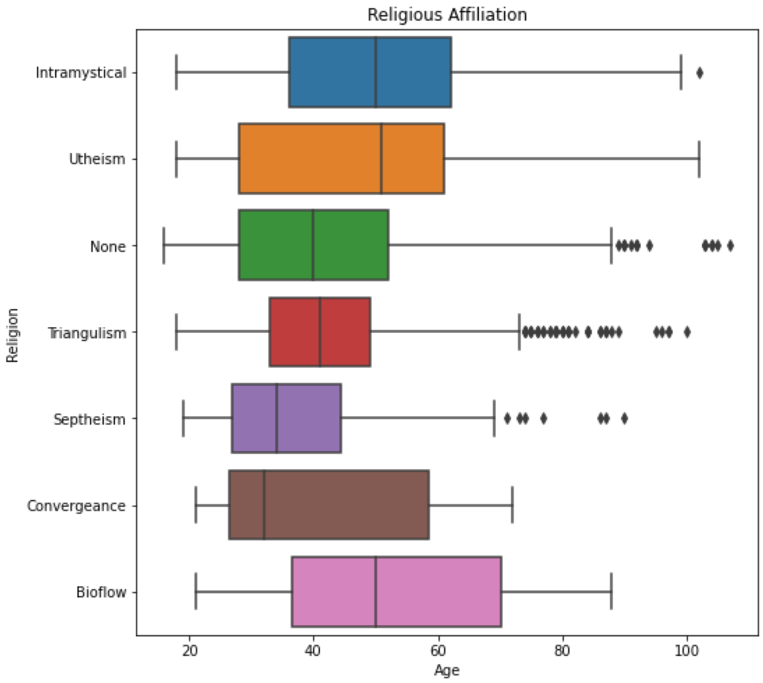
Value	Count	Frequency (%)
None	7424	99.6%
Nudisease	21	0.3%
Skygazer	4	0.1%
Silly	3	< 0.1%
Toothache	3	< 0.1%

Detailed Analysis

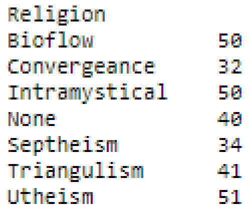
Religion and Infirmity

As infirmity comprised such a small percentage of the population, it was not deemed significant to warrant extensive analysis – thus, there will be no recommendations based on infirmity.

There were some growing religions, identified by the slightly lower median age of followers (Bioflow and Convergeance). However, these equally comprised a small percentage of the population.



Median Age Per Religion



```
Religion
Bioflow          50
Convergeance     32
Intramystical    50
None             40
Septheism        34
Triangulism      41
Utheism          51
```

Intramystical is still the dominate religion, which already has a church; though other religions may grow in the future, they do not have a significant following in the town to warrant building a new church above other needs of the population.

Divorce and Marriage

As seen in the data cleaning section, divorce occurs from a young age up until old age. Marital status split by gender identifies there are more female divorcees than there are male, indicating male divorcees potentially leave the town.

Therefore, to calculate crude divorce rates, marriage rates and divorce to marriage ratios, a baseline count of ‘divorces’ must be based on divorced women. Marriage is calculated by counting the number of ‘married’ individuals and dividing by two. Thus, the divorce to marriage ratio is 40 divorces per 1065 marriages – close to the UK average (Eurostat, 2017). This ratio will become crucial in later sections examining occupancy and migration.

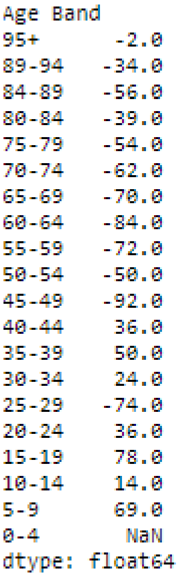
Birth and Death Rate

Calculating the birth and death rate indicates the town is not growing, though there may be an increasingly aging population. The crude birth rate and death rate are calculated as:

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

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

The current crude birth rate is 11 births per thousand. Five years prior, the crude birth rate was estimated at 13 births per thousand, calculated by adjusting the population to what it may have been in the previous given year. The birth rate has therefore fallen by 2 children per thousand in a five-year period.



```
Age Band
95+       -2.0
89-94    -34.0
84-89    -56.0
80-84    -39.0
75-79    -54.0
70-74    -62.0
65-69    -70.0
60-64    -84.0
55-59    -72.0
50-54    -50.0
45-49    -92.0
40-44     36.0
35-39     50.0
30-34     24.0
25-29    -74.0
20-24     36.0
15-19     78.0
10-14     14.0
5-9       69.0
0-4        NaN
dtype: float64
```

The death rate is calculated by estimating deaths by difference in age-bands for those over 65. Although there is decline across other age groups, the migration section details that these are likely individuals moving from the town (students or divorcees) as opposed to deaths. However, those over 65 are most likely to be retired and settled, and differences in these age bands are more likely to be deaths.

Thus, by summing the difference of the left table for those above 65 and dividing by 5 (to account for the age-banding), the death rate is calculated as 8.5 deaths per thousand.

Migration

University Students constitute most emigration and immigration in the town. The age-band table above shows a corresponding increase of ages 15-19, and corresponding decrease at 25-29 when students are likely to finish their studies. Students should be considered a constant in the town's growth – the homes they leave will be re-occupied by students each year, and they are not likely to use employment services or benefit from future aging care services since they vacate the town after their studies.

Instead, 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 28 immigrants to the town per thousand people. Using only the difference in divorced males and females, emigration from the town is 23 per thousand.

Calculating immigration and emigration based on the above age bands (including students, but excluding 65+), the rate of emigration is higher at 39 per thousand and rate of immigration 30 per thousand.

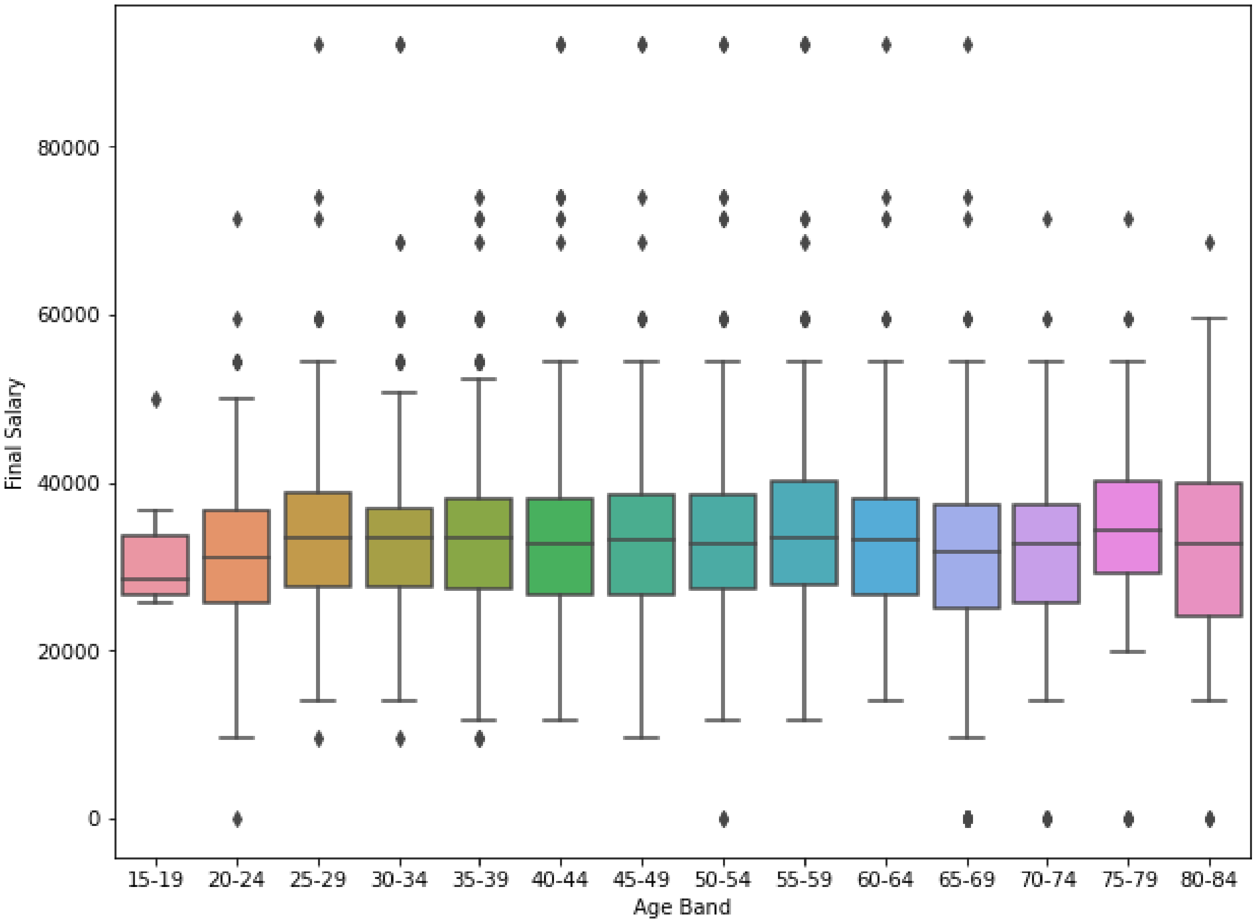
Employment and Commuters

Commuters are identified based on the following methodology:

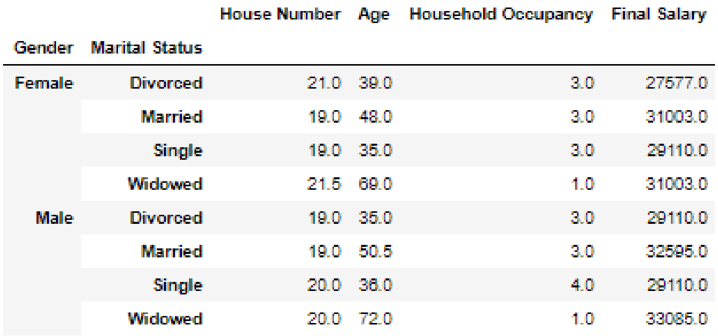
- Anyone who identifies as a University Student (including PhD Students);
- A sample (100 records) of those in employment and a count of the number of Occupations which would require commuting.

Occupations such as teaching (except Higher Education), baristas, retail workers, community roles or food service industry roles were considered non-commuting occupations. Using this method, roughly 45% of the population were commuters – of those employed and not students, 70% of the working population were estimated to commute.

The addition of the ONS codes to the dataset provided some insight into the average salary of those employed in the town. Overall, the town seemed to have a wealthy population:



Broken down by marital status, two things become apparent: women divorce later than men, and women earn marginally less (particularly if divorced):



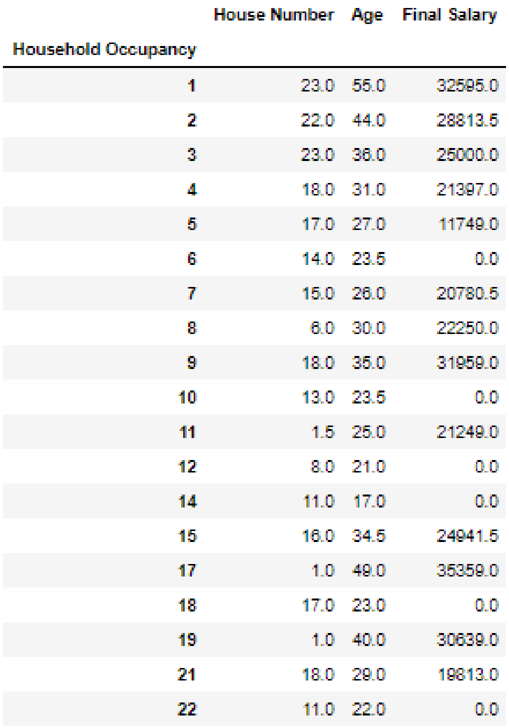
		House Number	Age	Household Occupancy	Final Salary
Gender	Marital Status				
Female	Divorced	21.0	39.0	3.0	27577.0
	Married	19.0	48.0	3.0	31003.0
	Single	19.0	35.0	3.0	29110.0
	Widowed	21.5	69.0	1.0	31003.0
Male	Divorced	19.0	35.0	3.0	29110.0
	Married	19.0	50.5	3.0	32595.0
	Single	20.0	36.0	4.0	29110.0
	Widowed	20.0	72.0	1.0	33085.0

A limitation of the Final Salary column is not knowing hours worked; as potential care givers (determined by the higher rate of women remaining in the town after divorce), it may be the case that women do not work full-time hours.

Overall, the unemployment rate of the town was 6% of the total population, or 9% of the population qualified to work. This is a somewhat high unemployment rate (ONS, 2020), likely because jobs are outside of town and, judging for the salary of most of the town, relatively high-skilled and high paid.

Occupancy Rates

The median occupancy rate of the town is two. This does not necessarily mean each home has a median two bedrooms, but instead operates as a proxy to determine over-occupancy. Using this median, there are 1259 homes over-occupied – increasing the median to three shows 698 homes over occupied.



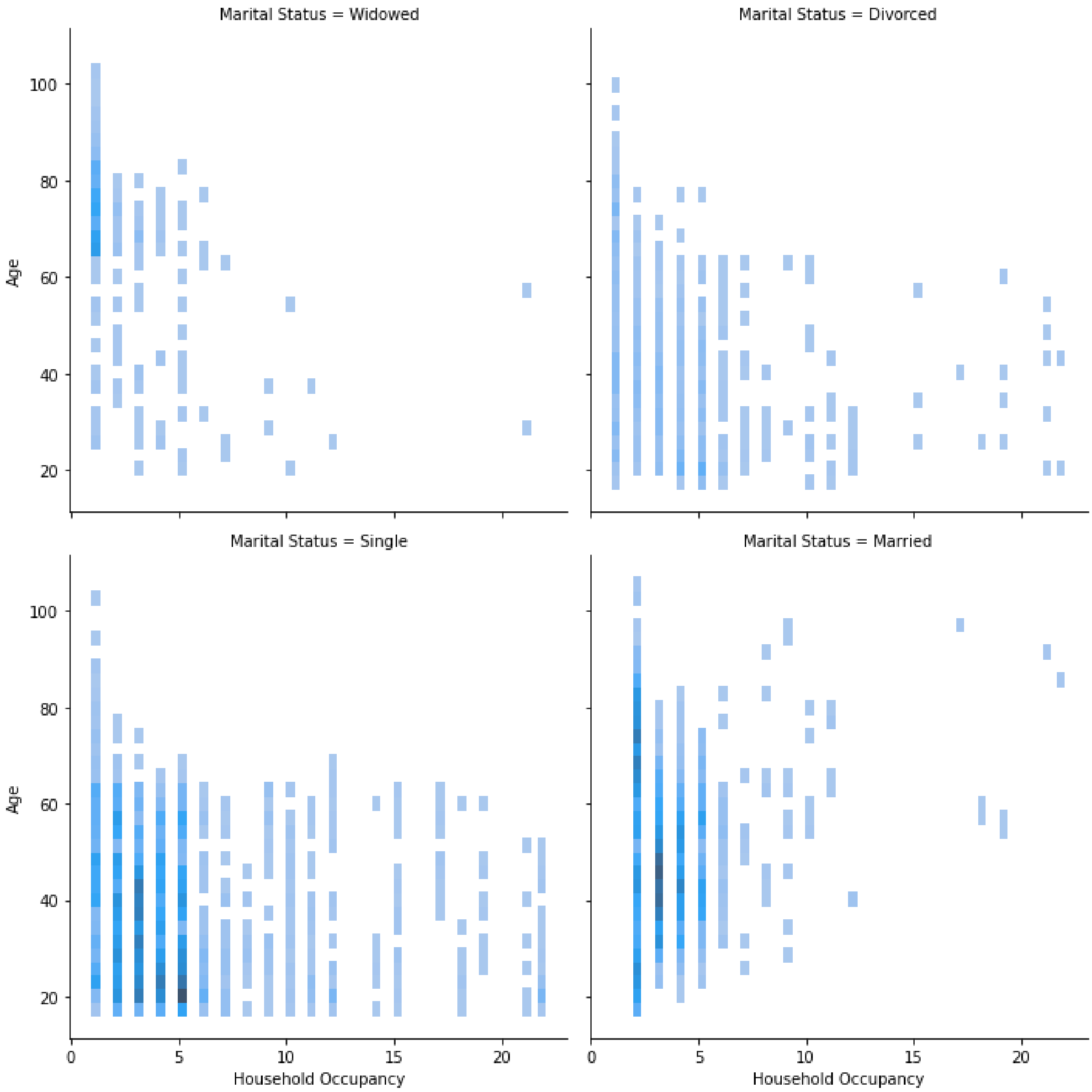
	House Number	Age	Final Salary
Household Occupancy			
1	23.0	55.0	32595.0
2	22.0	44.0	28813.5
3	23.0	36.0	25000.0
4	18.0	31.0	21397.0
5	17.0	27.0	11749.0
6	14.0	23.5	0.0
7	15.0	26.0	20780.5
8	6.0	30.0	22250.0
9	18.0	35.0	31959.0
10	13.0	23.5	0.0
11	1.5	25.0	21249.0
12	8.0	21.0	0.0
14	11.0	17.0	0.0
15	16.0	34.5	24941.5
17	1.0	49.0	35359.0
18	17.0	23.0	0.0
19	1.0	40.0	30639.0
21	18.0	29.0	19813.0
22	11.0	22.0	0.0

Additionally, there is a tendency for household occupancy to be higher as the median age drops (left).

Extremely high occupancy (10+) could be attributed to families with large salaries living together and students, though these could also be blocks of flats.

However, there are an estimated 181 lodgers living with families who have children, and more in other family homes (e.g., a husband and wife only).

There are three potential reasons for this: families are over-occupying homes and sharing rooms to allow lodgers; families cannot downsize and are instead renting rooms to lodgers or; divorced women are renting rooms to lodgers to afford their mortgage or rent rather than downsize.



The above pair plot visualises occupancy levels by marital status – older widowers and married individuals tend to occupy median-sized homes of two occupants. Divorced and single individuals occupy a spread of household sizes, suggesting cohabitation with non-relatives (the median family size is calculated as 4, but there is a small concentration of divorcees living with 5 people or more).

Recommendations

Given the large number of lodgers emigrating to the town, and the high number of commuters, there is both a need to build a train station and potentially invest in low-density housing. However, a train station may not benefit families who are renting rooms to lodgers. In this case, low-density housing would benefit lodgers moving to the town, families, those who are divorced with young children who may need to downsize.

It is likely in future years those smaller houses occupied by retired individuals will eventually become available. At this point, there is more opportunity for the town population to grow and develop, inviting more immigration from those looking for work and further relieve the housing shortage. In the future, then, a train station may be worth investing in, but at this point low density housing will benefit more of the population.

Given the high ages of some individuals, and the relatively good income of most in the town, it is likely they will live longer. Very few suffered from an infirmity, however this may not be the case as the population ages. Thus, investing in age old care should be a priority above other services.

Other services to invest in, such as schooling or the general infrastructure, is not warranted at present due to the population's relatively small growth year on year. However, the aging population will only increase and require more care, so it is pertinent to invest in this before that happens.

Bibliography

BBC (2016) *Jedi is not a religion, Charity Commission rules.*
Available online: https://www.bbc.co.uk/news/uk-38368526 [Accessed 05/12/2020]

Eurostat (2017) *Marriage and Divorce Statistics*
Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php/Marriage_and_divorce_statistics [Accessed 01/12/2020]

Gov.uk (No Date) *Universal Credit Eligibility*.
Available Online: https://www.gov.uk/universal-credit/eligibility [Accessed 05/12/2020]

*Marriage Act (*1949*)* Section 3
Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3 [Accessed 05/12/2020]

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 20/11/2020]

Office for National Statistics (2016) *Standard Occupation Classification*
Available online:
https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010
[Accessed 10/11/2020]

Office for National Statistics (2020) *Labour market in the regions of the UK: July 2020*
Available online:
https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/regionallabourmarket/july2020#:~:text=Local%20labour%20market%20indicators,-Indicators%20from%20the&text=For%20the%20period%20April%202019,Middlesbrough%2C%20both%20at%206.9%25.
[Accessed 01/12/2020]