# Census Project Report

## 1 Introduction

This report presents census data cleaning and analysis for a moderately-sized town, concluding with recommendations for local government concerning land development and future service investment.

Data cleaning was performed to ensure technical correction of the data, handling missing values and data errors, detailed in section 2.1.

Recommendations (section 4) are based on the analysis of the cleaned data set, to include the demographics of the town (section 3.1), and predictions of population growth, affluency, employment and occupancy (section 3.2).

## 2 Methods

Census data was provided as a csv file and converted to a Pandas dataframe for cleaning and analysis (Appendix 6.1, Figure 1). Code for data cleaning and analysis is provided in the corresponding Jupyter notebooks (Census Data Cleaning.ipynb; Census Data Analysis.ipynb) and cleaning log (Appendix 6.3).

### 2.1 Data cleaning

A Pandas profiling report for the raw data identified instances of empty strings (YData profiling, n. d.). There were no duplicate rows or special characters. Age was converted to integer data type and renamed Age (years).

#### 2.1.1 Imputation

Where possible, empty strings were handled by imputation (detailed in Appendix 6.2, Table 1) using a combination of single-value estimates and inference from household records.

#### 2.1.2 Error handling

##### *2.1.2.1 Marital Status*

Default marital status for individuals aged 18 and under is NaN, with exceptions identified and corrected (Appendix 6.2, Table 2).

Two households (7 records) were removed from the dataset due to under-age marriages (GOV.UK, n. d.-c). There is one instance of a divorced 16 year old living alone with their child, possible due to previous legislation legalising marriage at 16 with parental consent (NSPCC, 2023; GOV.UK, n. d.-b). One instance of a single 17 year old with a child, living with others over 18, was updated to NaN.

Distribution of marital status by age (Appendix 6.1, Figure 2) displays the possibility of becoming a widow at any age, although less likely at younger ages indicated by outliers. As a result, those widowed age 18 were changed to single.

##### *2.1.2.2 Religion*

Three religions, Jedi, Housekeeper and Private were identified as outliers. Jedi is an illegitimate religion and, along with Housekeeper, were assigned ‘None’ as the household religion (GOV.UK, n. d.-a). Private was converted to NaN, interpreted as individual withholding religion.

Religion for under 16s was NaN in all instances except one (‘None’). All records were updated to ‘Undeclared’. As a result, calculation of religion inheritance within households is not possible.

#### *2.1.2.3 Occupation*

For individuals aged 66 and over and unemployed, occupation was updated to Retired, unemployed to align with state pension age (Age UK, 2023).

### 2.1.3 Additional data

Additional Age Group, Household Occupancy and Salary (GBP) columns were added for further analysis (Appendix 6.2, Table 3).

Salary was calculated by fuzzy matching with ONS national salary data for Standard Occupation Classifications (Pypi, 2020; Office for National Statistics, 2023c). A trade-off between similarity and accuracy of matching, coupled with manual checking and empty value insertion, ensured assignment of median salary to all employed census participants. Children, students, unemployed and retired were assigned zero salary.

The final features of the cleaned data are detailed in Appendix 6.1, Figure 3.

## 2.2 Data analysis

### 2.2.1 Assumptions

#### *2.2.1.1 Birth and death rates*

The birth rate is traditionally calculated using the total number of live births in a specific time period divided by the child-bearing population during that period, assuming demographic (mortality, fertility, and migration) and socioeconomic factors remain constant (Office for National Statistics, 2023a).

##### 2.2.1.1.1 Refined birth rate

Birth rate for child-bearing women (aged 15-44) assumes a uniform distribution of births across the specified age ranges, a constant birth rate, and an equal likelihood of women in each age range giving birth.

##### 2.2.1.1.2 Evolving birth rate

Assumes that women of child-bearing age are only aged 25-34, methodology has been used for cohort component population projection (Measure Evaluation, n. d.).

##### 2.2.1.1.3 Crude death rate

Based on the sum of the differential between age ranges for aged 65 and over, assuming that the number of people in each age range is constant over time. Population decline in over 65s is not assumed to be migration where more likely for younger age groups.

##### 2.2.1.1.4 Predicting number of deaths by religion

Assumes that the distribution of religions remains constant over the next 10 years.

##### 2.2.1.1.5 Predicting number of people requiring care

Assumes that 2.5% of the population aged over 65 need care (Office for National Statistics, 2023f), combined with birth rate assumptions.

### 2.2.2 Hypothesis testing

Hypothesis testing was conducted to determine statistical significance or correlation between variables, where appropriate (Newcastle University, n. d.-a). If the null hypothesis stands there is insufficient evidence to reject it based on the observed data.

#### *2.2.2.1 Statistical significance*

The independent samples t-test determines significant difference between the means of two independent groups. The p-value, typically set at a significance level of 0.05, indicates the probability

of observing a significant difference, rejecting the null hypothesis when the p-value is below the significance level (Statistics How To, 2023).

#### *2.2.2.2 Correlation analysis*

The Pearson correlation coefficient quantifies the degree and direction of the linear association between two continuous variables, assuming normal distribution (SciPy, n. d.-a).

The Spearman rank correlation coefficient indicates monotonic relationships between variables, with an increasing variable, the other may increase or decrease at varying rates (Newcastle University, n. d.-b; SciPy, n. d.-b).

Increased correlation is indicated the nearer to 1 or -1 for both coefficients. Correlation does not imply causation, and the lack of a significant correlation in the data does not confirm the absence of a real-world relationship.

# 3 Results and discussion

## 3.1 Demographics

Summary statistics for the town are displayed in Appendix 6.2, Table 4. The mean age of the town is 36.3 years, with the oldest individual aged 116. The mean salary across a wide range of salaries for the population, indicates affluency. Mean household occupancy is 4.1, with maximum occupancy of 22.

The age pyramid (Appendix 6.1, Figure 4) of the population indicates a lower number of 0-4 year olds compared to school-aged children, suggesting a lower birth rate in the last five years. Those aged 65 and above live well into old age. Aside from the increased population in their early twenties likely due to the student population, the largest number of people are middle-aged (particularly females), suggesting a potential requirement to provide additional care and facilities in the future for an aging population.

Much of the town is employed (Appendix 6.1, Figure 5). Distribution of marital status indicates that most of the population are single or married (Appendix 6.1, Figure 6). The population has a low rate of infirmity at less than 1%. The predominant religions within the population are Christian, Catholic and Methodist (Appendix 6.1, Figure 7).

## 3.2 Detailed analysis

### 3.2.1 Infirmity and religion

Those with an infirmity represent less than 1% of the population, and therefore further analysis was forgone due to lack of significance.

The dominant religions are Christian, Catholic and Methodist (Appendix 6.1, Figure 8), having a higher median age and greater number of deaths predicted over the next ten years (Appendix 6.1, Figure 9). Therefore, justification of a new house of worship for Christians or Methodists requires determination of inheritance within families/households, not possible with this data.

For younger generations under forty there are emerging religions Buddhist, Sikh, Muslim and Jewish (Appendix 6.1, Figure 8). This is represented by the lower median ages and comparatively narrower IQR. However, these religions represent only 2.1% of the population, deemed insignificant to warrant new houses of worship at this time.

### 3.2.2 Marriages and divorces

The majority of the divorced population are in middle age. The range of the data indicates that divorce occurs from legal age of marriage to old age (Appendix 6.1, Figure 2). Divorce by gender (Appendix 6.1, Figure 10) indicates there are more female divorcees, indicating some male divorcees may potentially leave the town. Divorced women may remain due to care responsibilities for children.

Divorce rate, marriage rate and divorce to marriage ratio are calculated based on the count of divorced women (Appendix 6.2, Table 5). Number of marriages is calculated by dividing count of married individuals by two, assuming that each marriage contains two individuals. The divorce rate is 42.7%, equalling the national divorce rate (Harbour Family Law, 2023).

### 3.2.3 Births and deaths

Birth rate exceeded the death rate indicating population growth (Appendix 6.2, Table 6). Birth rate was calculated on the number of child-bearing women (also known as fertility rate), a method used commonly for population prediction, in addition to crude birth rate generalising for the total population (Office for National Statistics, 2023a; World Health Organisation, 2023).

The birth rate for childbearing women is lower than the national rate (Office for National Statistics, 2022). Evolving birth rate indicates decline in number of births compared to five years ago, correlating with the age distribution plot where there are a lower number of 0-4 year olds compared to school-age children (Appendix 6.1, Figure 4).

Death rate was calculated on the sum of the differential of the number of individuals in age groups aged 65 and over (divided by the age group range (five years)). Differences in these age groups is more likely due to increased deaths compared to lower age groups, where differences could be (and are more likely) due to migration. Death rate was also slightly lower than the national rate (Office for National Statistics, 2023b).

The predicted percentage of the population needing care is 4.7%, calculated assuming 2.5% of over 65s may need care (the national statistic), coupled with the expected number of births (Office for National Statistics, 2023f). The predicted percentage increase of over 80 year olds in ten years is 34.6%, correlating with an increasingly aging population.

### 3.2.4 Migration

Net migration is positive based on the proportion of immigrants (number of single lodgers and visitors) compared to emigrants (difference between divorced males and females) (Appendix 6.2, Table 7). Students are considered a constant in terms of migration due to temporary occupancy and regular turnover in the town. Divorced lodgers and visitors were excluded from the immigration calculation, not classified as migrants, after separating from their partner.

A significant proportion of immigrants (71.1%) are living with families with children. For 25.7% of families with lodgers and visitors, the head of household is single, divorced or widowed. These families may be accommodating immigrants for affordability reasons, or to avoid downsizing.

### 3.2.5 Employment and commuters

#### *3.2.5.1 Unemployment*

The mean age for unemployment is 42.5. The unemployment rate of 5.8% of the total population is slightly higher than the rate for England and Wales (Office for National Statistics, 2023e). The

unemployment rate for those eligible to work is 9.8%. This higher rate may be due to the location of jobs outside of town. Or could be due to a skills gap where positions require relatively high skill for high pay, correlating with the mean salary of the town.

The unemployment rate for females is higher than for males (the male rate equals the UK unemployment rate) (Appendix 6.1, Figure 11). A higher proportion of people in their mid-thirties to early fifties are unemployed. At every age, except for 18 year olds and early twenties, a higher proportion of females are unemployed compared to males.

Reasons for the gender disparity may include occupation choice where some industries may have different unemployment rates, differences in educational attainment, family responsibilities, and gender discrimination.

#### *3.2.5.2 Salary*

The majority of the employed population earn £10000 up to £40000 (Appendix 6.1, Figure 12). Over 40% of employed earn £20000 up to £30000. There is very little difference in median age for those earning above or below £50000. However, the proportion of employed earning over £50000 is significantly less (7.8%) than those earning £50000 or lower (Appendix 6.1, Figure 13). Median salary equals national median annual earnings, indicating the relative affluency of the town (Office for National Statistics, 2023d).

There is no significant difference in salary by gender and/or marital status, determined with independent samples t-test (Statistics How To, 2023). No significant correlation (using Pearson and Spearman rank) exists between salary and age (SciPy, n. d.-a; n. d.-b).

#### *3.2.5.3 Commuters*

An estimated 57% of the population commute. The commuter population is composed of university students and employed individuals, excluding those in potential non-commuting occupations like teaching (except higher education), public, youth, community and local roles, food and shop workers and firefighters. Of the employed population, 93% commute. This high number of commuters provides a case for providing improved transport links to the neighbouring cities (like a train service).

### 3.2.6 Occupancy

Occupancy summary statistics are displayed in Appendix 6.2, Table 8. The positive skewness (2.7) and kurtosis (10.0) suggest that the distribution is not perfectly normal and is right-skewed (Appendix 6.1, Figure 14). The median (4.0) is a more robust measure of central tendency compared to the mean, accounting for outliers. A mode of 3 indicates a concentration of smaller household sizes. 45% of houses in the town are occupied by families.

Widowers tend towards lower occupancy households (Appendix 6.1, Figure 15). Occupancy is between 2 and 5 primarily for married individuals, a large number with occupancy of 2, showing the spread of couples with and without children. Divorced and single occupancy has a wider range, with these individuals also potentially living with others in shared households (rented accommodation, student accommodation) and flats.

No significant correlation exists between occupancy and median salary or age.

Over or under-occupancy was determined using median occupancy by street, instead of using the more generalised median for town. This approach can be used to identify localised regions of over-occupancy should this analysis be required for future development. The sum of the difference in over-

occupancy versus under-occupancy by street indicated that overall, the town is over-occupied. The percentage of houses with immigrants that are over-occupied is 53.5%.

## 4 Recommendations

The over-occupied town requires additional housing, driven by a growing population with birth rate surpassing the death rate, positive net migration, and projections of a progressively aging demographic.

The town is experiencing positive net migration, with incoming individuals contributing to population growth. Family households, a quarter with an unmarried parent, accommodate the majority of immigrants. It is plausible that families are choosing to host immigrants to increase affordability or avoid downsizing.

High-density housing would accommodate immigrants, families wishing to downsize, divorced people with or without children, and older people in an increasingly aging population.

The prevalence of commuters indicates potential need for a train station. This could enhance transport accessibility and potentially attract more immigrants, exacerbating housing pressures. While investing in a train station may be a viable option in the future, prioritising housing development will potentially benefit more of the population. It is advisable to consider investing in council services concurrently with housing development. Additionally, allocating resources for road maintenance is crucial to support commuters.

There is evidence indicating a rise in retired individuals in the future. Given the affluence of the town, life expectancy may increase. Currently, less than 1% of the population has an infirmity, and the health of the population is robust. However, this may change with anticipated increases in life expectancy. The town needs to allocate funding for end-of-life care to meet the needs of an aging population.

*Word count 2448*

# 5 References


Age UK (2023) *Changes to state pension age.* Available online: https://www.ageuk.org.uk/information-advice/money-legal/pensions/state-pension/changes-to-state-pension-age/ [Accessed 26/11/2023].

GOV.UK (n. d.-a) *Decision of the Charity Commission on the temple of the Jedi order.* Available online: https://www.gov.uk/government/publications/the-temple-of-the-jedi-order [Accessed 26/11/2023].

GOV.UK (n. d.-b) *Marriage act 1949.* Available online: https://www.legislation.gov.uk/ukpga/Geo6/12-13-14/76/section/3 [Accessed 26/11/2023].

GOV.UK (n. d.-c) *Marriage and civil partnership (minimum age) act 2022.* Available online: https://www.legislation.gov.uk/ukpga/2022/28/enacted [Accessed 26/11/2023].

Harbour Family Law (2023) *How many marriages end in divorce in the UK?* Available online: https://www.harbourfamilylaw.co.uk/how-many-marriages-end-in-divorce-in-the-uk/ [Accessed 08/12/2023].

Measure Evaluation (n. d.) *Lesson 8: the cohort component population projection method.* Available online: https://measureevaluation.org/resources/training/online-courses-and-resources/non-certificate-courses-and-mini-tutorials/population-analysis-for-planners/lesson-8.html [Accessed 08/12/2023].

Newcastle University (n. d.-a) *Introduction to hypothesis testing (psychology).* Available online: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/psychology/introduction-to-hypothesis-testing.html [Accessed 26/11/2023].

Newcastle University (n. d.-b) *Strength of correlation.* Available online: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html [Accessed 26/11/2023].

NSPCC (2023) *Moving out.* Available online: https://www.nspcc.org.uk/keeping-children-safe/in-the-home/moving-out/ [Accessed 26/11/2023].

Office for National Statistics (2022) *Births in England and Wales: 2021.* Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2021 [Accessed 08/12/2023].

Office for National Statistics (2023a) *Births in England and Wales: summary tables.* Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/birthsummarytables [Accessed 08/12/2023].

Office for National Statistics (2023b) *Deaths registered in England and Wales: 2021 (refreshed populations).* Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2021refreshedpopulations [Accessed 08/12/2023].

Office for National Statistics (2023c) *Earnings and hours worked, occupation by four-digit SOC: ASHE Table 14.* Available online: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/occupation4digitsoc2010ashetable14 [Accessed 26/11/2023].

Office for National Statistics (2023d) *Employee earnings in the UK: 2023.* Available online: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2023 [Accessed 08/12/2023].

Office for National Statistics (2023e) *Employment in the UK: 2023.* Available online: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/employmentintheuk/september2023 [Accessed 08/12/2023].

Office for National Statistics (2023f) *Older people living in care homes in 2021 and changes since 2011.* Available online: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/olderpeoplelivingincarehomesin2021andchangessince2011/2023-10-09 [Accessed 08/12/2023].

Pypi (2020) *Fuzzywuzzy 0.18.0.* Available online: https://pypi.org/project/fuzzywuzzy/ [Accessed 26/11/2023].

SciPy (n. d.-a) *Scipy.stats.Pearsonr.* Available online: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html [Accessed 08/12/2023].

SciPy (n. d.-b) *Scipy.stats.Spearmanr.* Available online: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html [Accessed 08/12/2023].

Statistics How To (2023) *T test (Student's t-test): definition and examples.* Available online: https://www.statisticshowto.com/probability-and-statistics/t-test/ [Accessed 26/11/2023].

World Health Organisation (2023) *Crude birth rate.* Available online: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1139 [Accessed 08/12/2023].

YData profiling (n. d.) *YData profiling.* Available online: https://docs.profiling.ydata.ai/4.6/ [Accessed 26/11/2023].

# 6 Appendix

## 6.1 Figures

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8377 entries, 0 to 8376
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   House Number                    8377 non-null   int64
 1   Street                          8377 non-null   object
 2   First Name                      8377 non-null   object
 3   Surname                         8377 non-null   object
 4   Age                             8377 non-null   object
 5   Relationship to Head of House   8377 non-null   object
 6   Marital Status                  6402 non-null   object
 7   Gender                          8377 non-null   object
 8   Occupation                      8377 non-null   object
 9   Infirmity                       8377 non-null   object
 10  Religion                        6365 non-null   object
dtypes: int64(1), object(10)
memory usage: 720.0+ KB
```

**Figure 1.** Pre-cleaned features

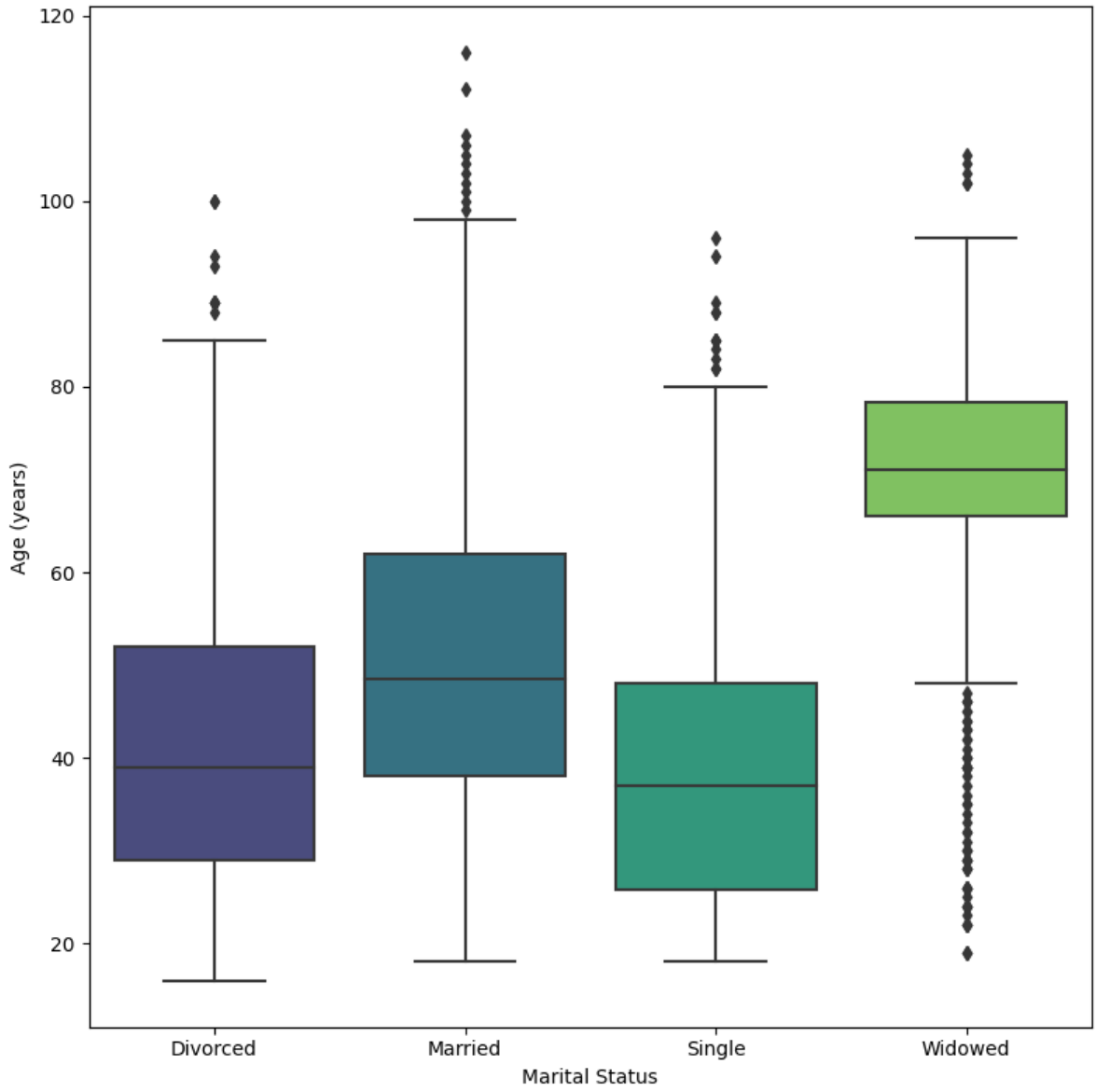


**Figure 2.** Marital status by age

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8370 entries, 0 to 8369
Data columns (total 15 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   House Number                    8370 non-null   int64
 1   Street                          8370 non-null   object
 2   First Name                      8370 non-null   object
 3   Surname                         8370 non-null   object
 4   Age (years)                     8370 non-null   int64
 5   Relationship to Head of House   8370 non-null   object
 6   Marital Status                  6395 non-null   object
 7   Gender                          8370 non-null   object
 8   Occupation                      8370 non-null   object
 9   Infirmity                       8370 non-null   object
 10  Religion                        8108 non-null   object
 11  Age Group                       8370 non-null   object
 12  Occupation Category             8370 non-null   object
 13  Household Occupancy             8370 non-null   int64
 14  Salary (GBP)                    8370 non-null   int64
dtypes: int64(4), object(11)
memory usage: 981.0+ KB
```

**Figure 3.** Cleaned features

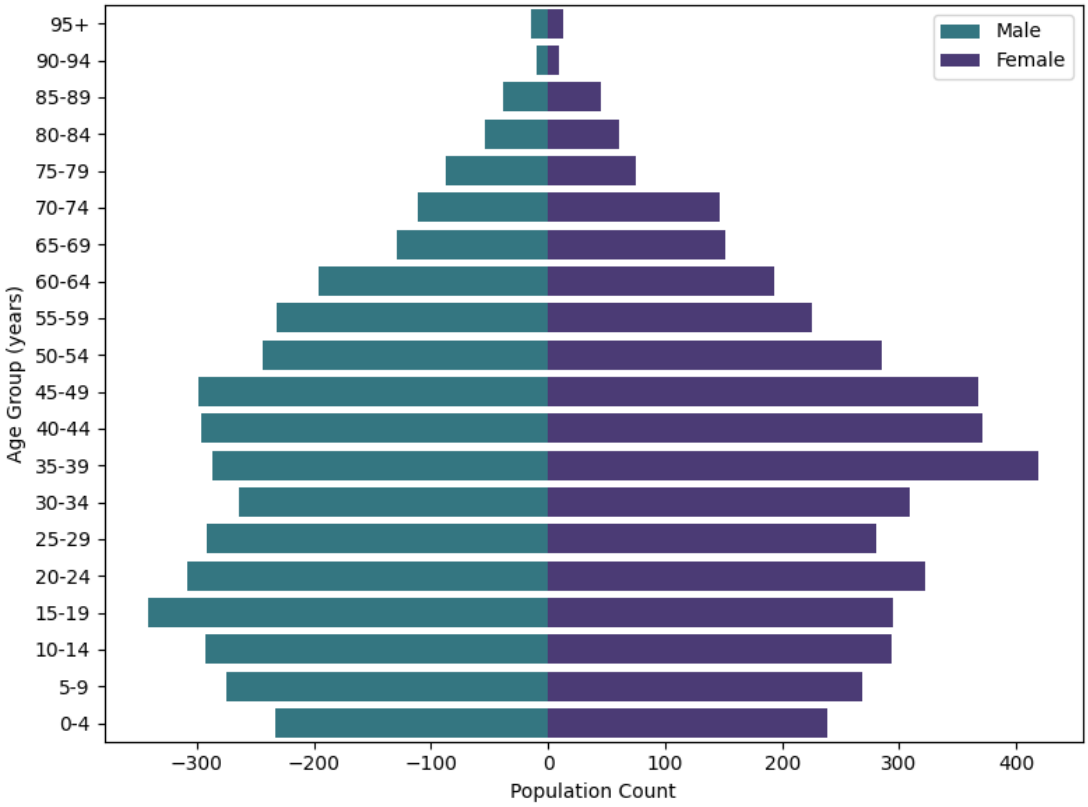


**Figure 4.** Age population pyramid by gender

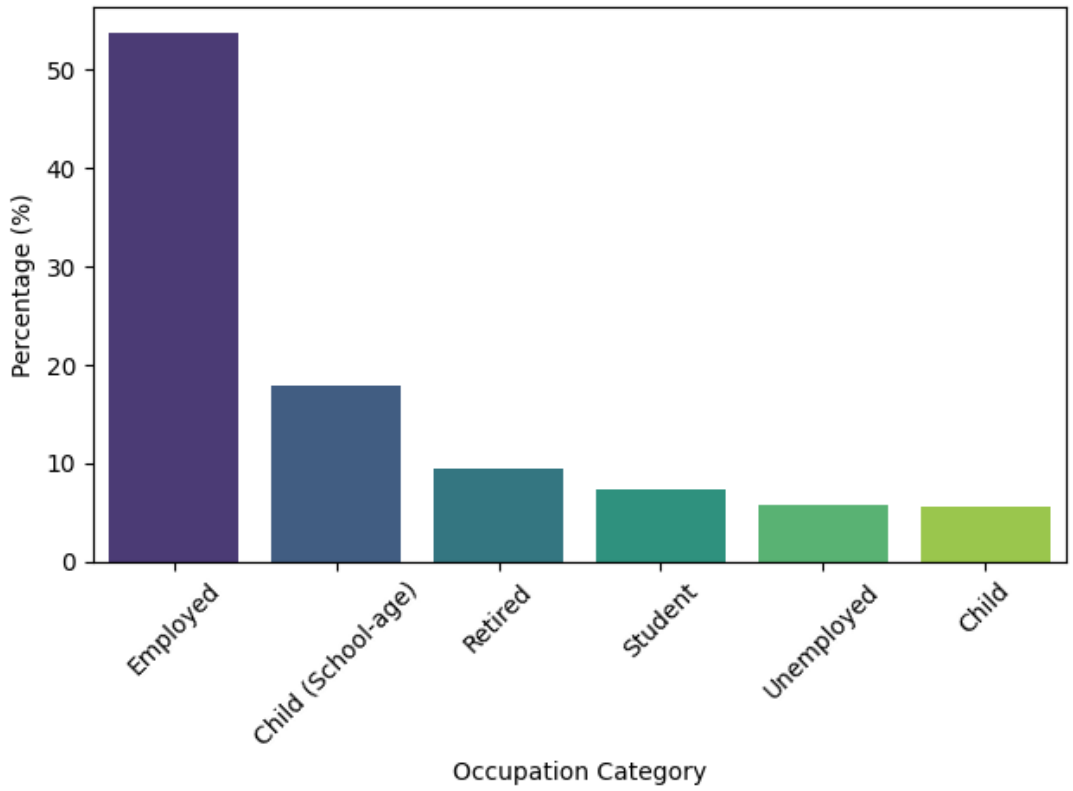


**Figure 5.** Occupation category distribution

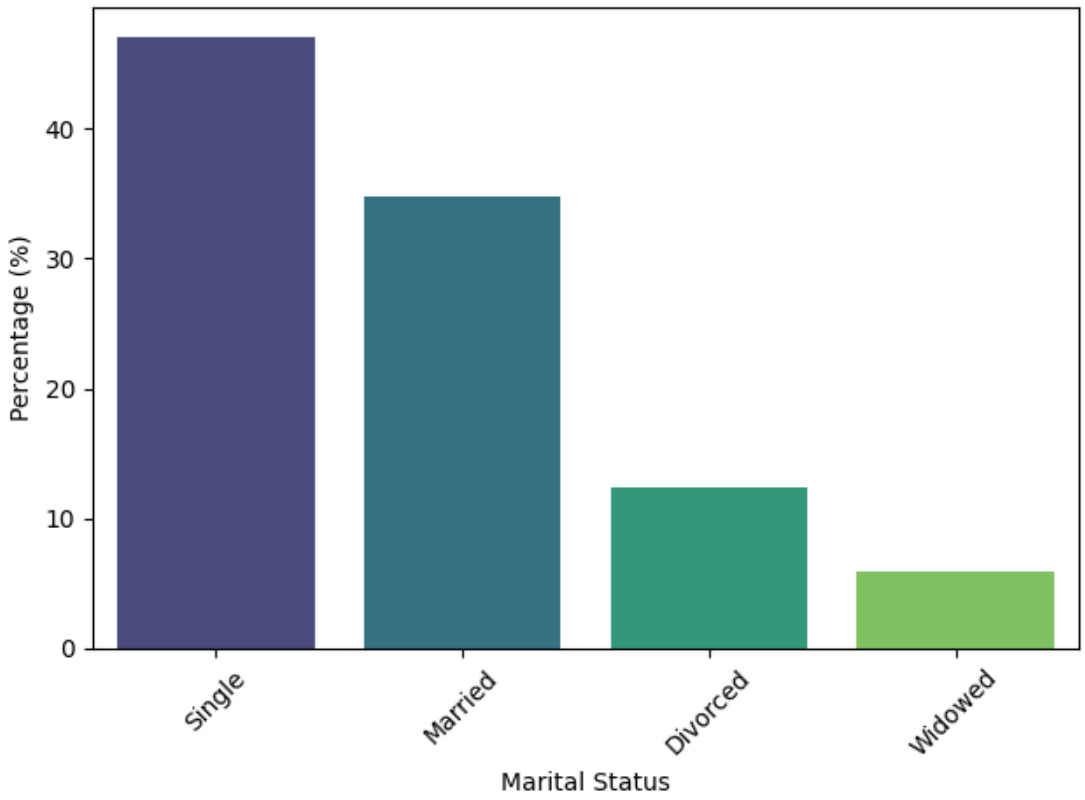


**Figure 6.** Marital status distribution

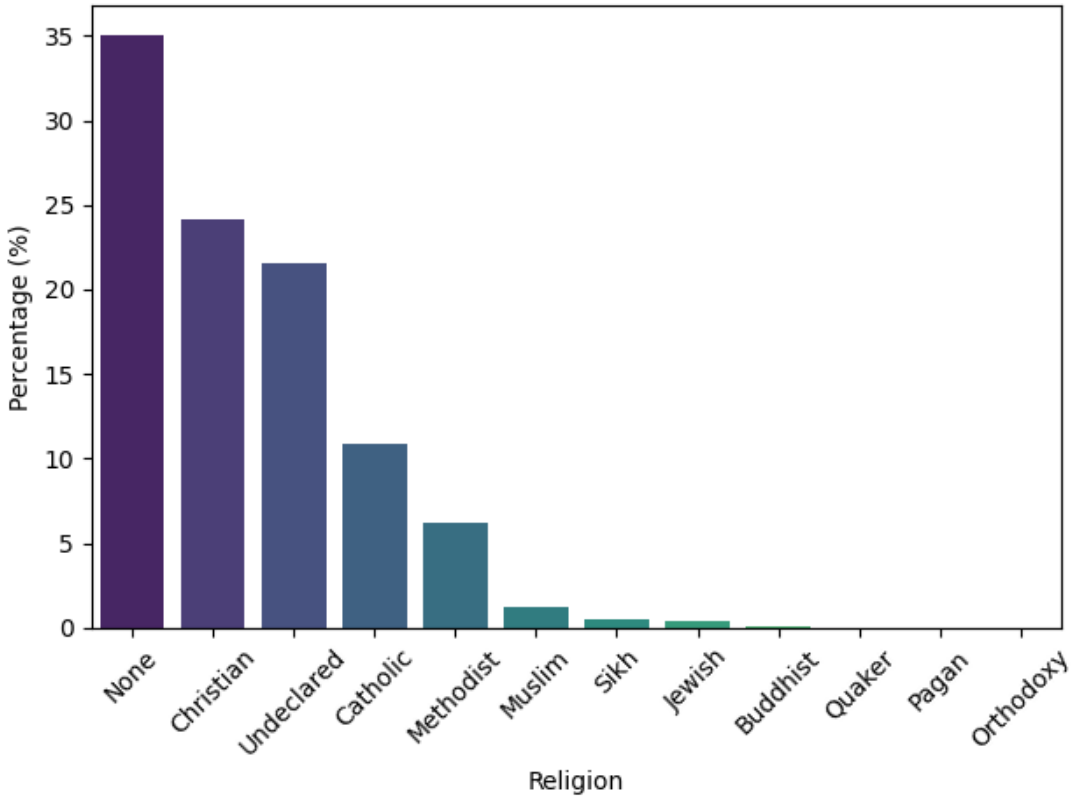


**Figure 7.** Religion distribution

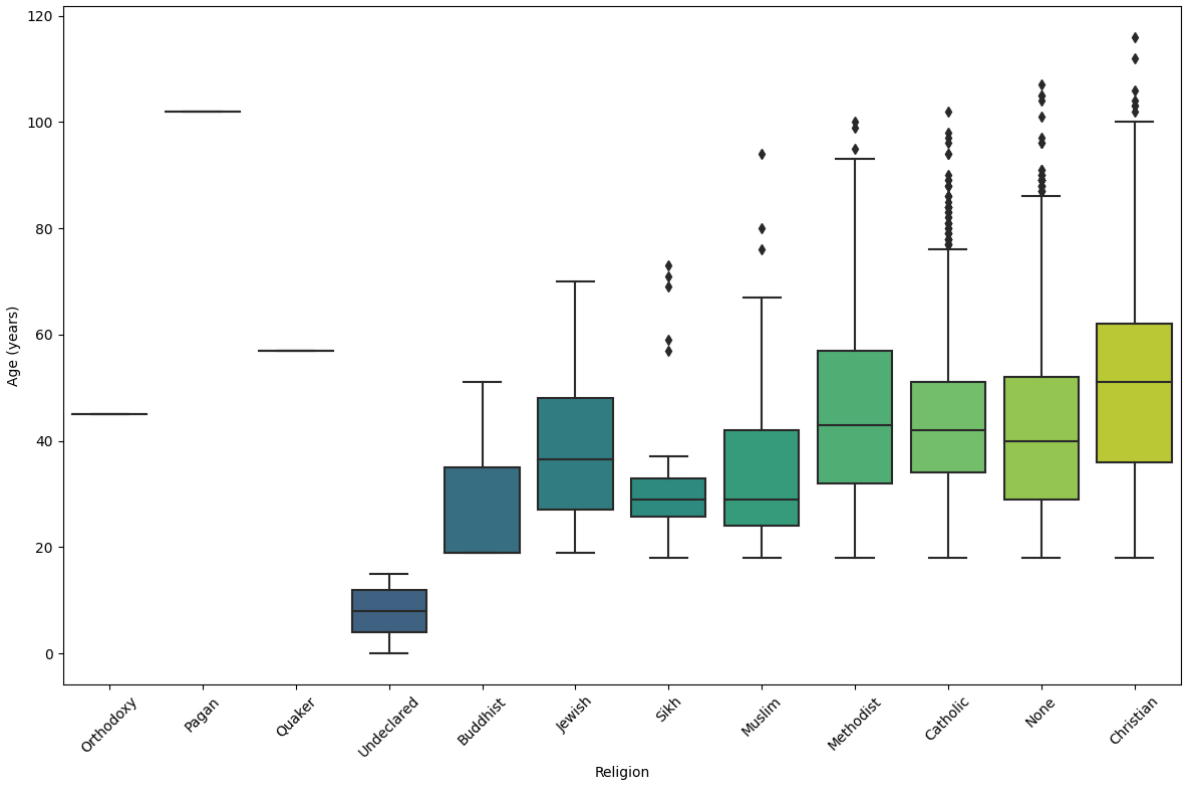


**Figure 8.** Religion by age

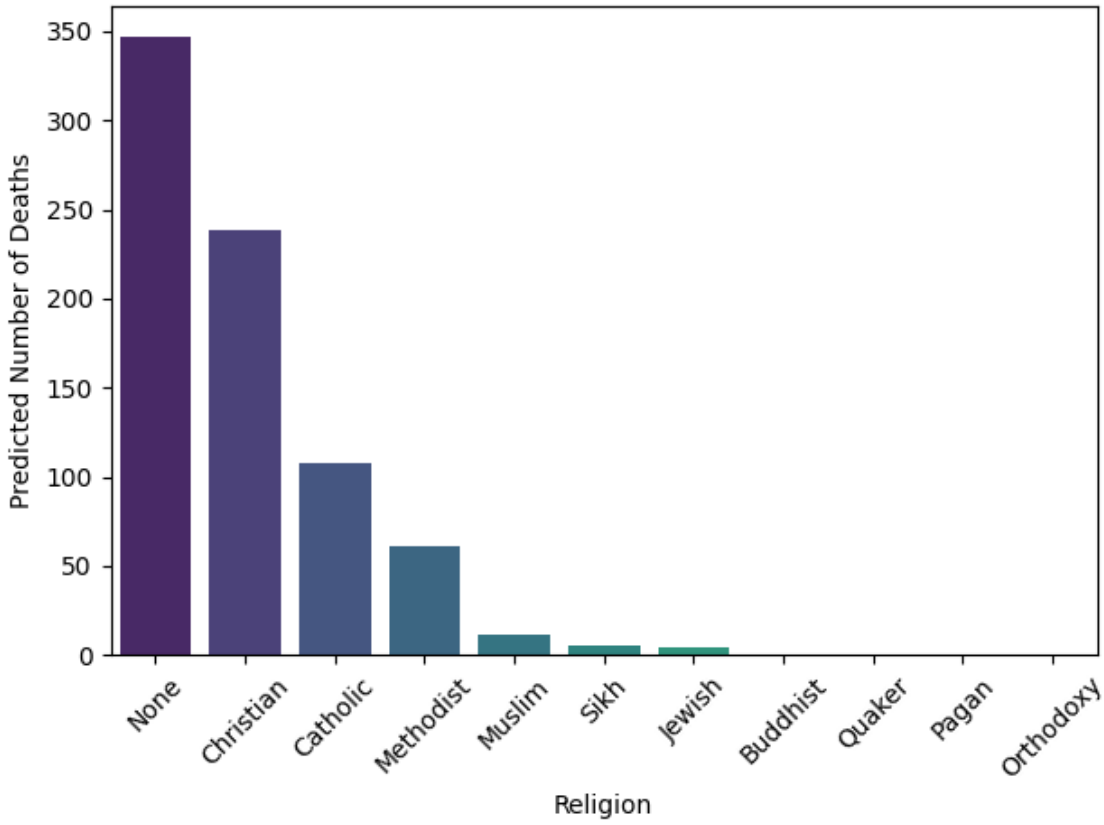


**Figure 9.** Predicted number of deaths by religion

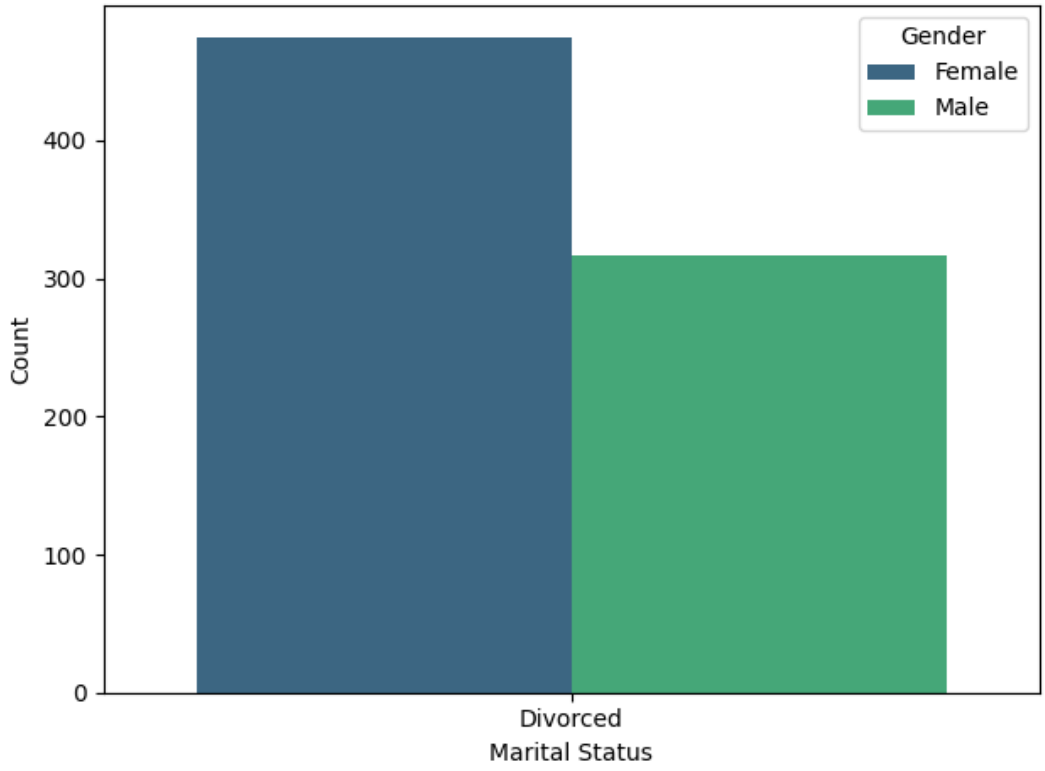


**Figure 10.** Divorced by gender

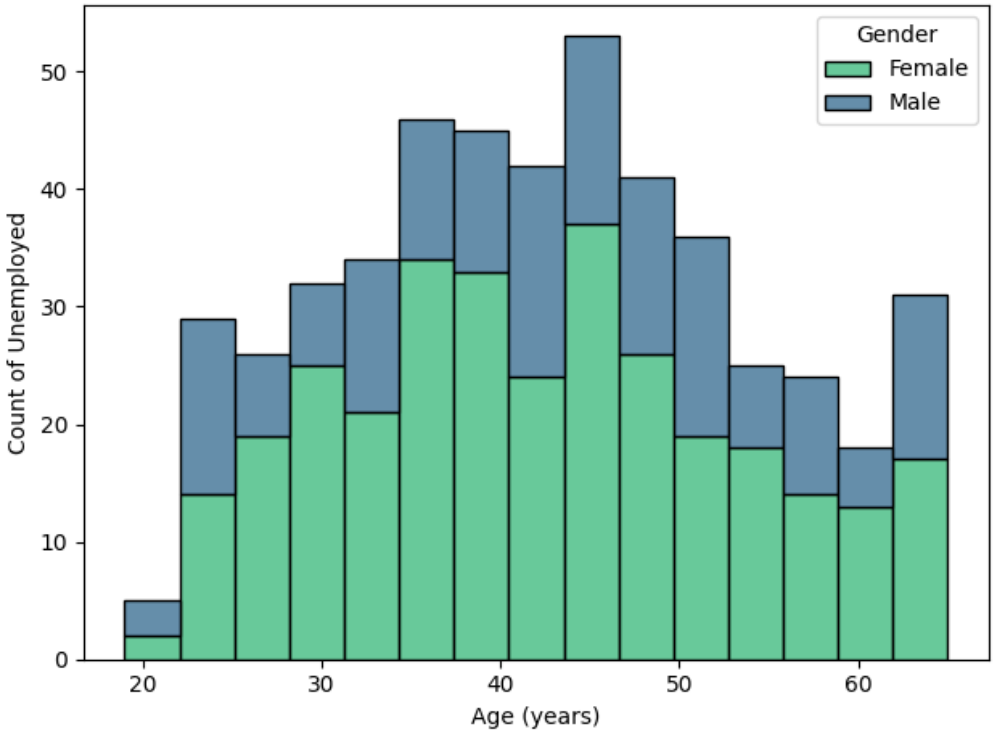


**Figure 11.** Unemployment by age and gender

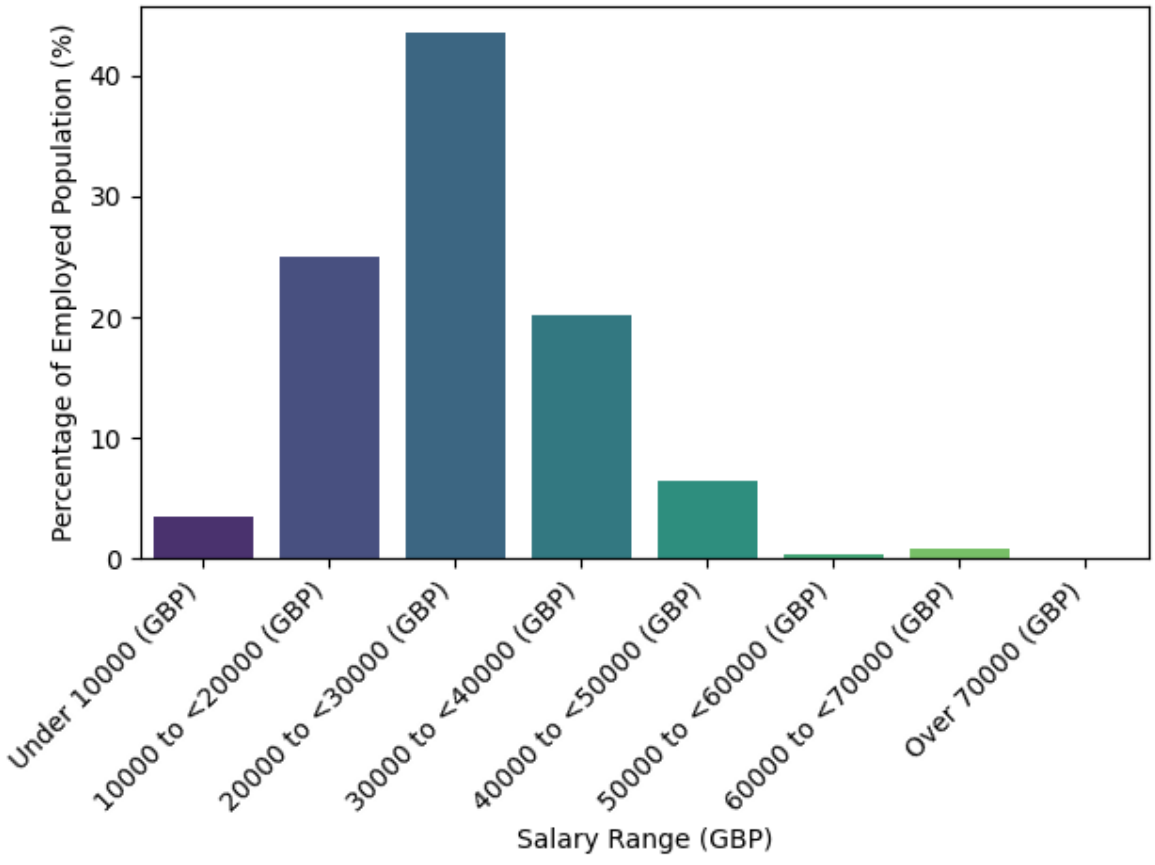


**Figure 12.** Distribution of salary

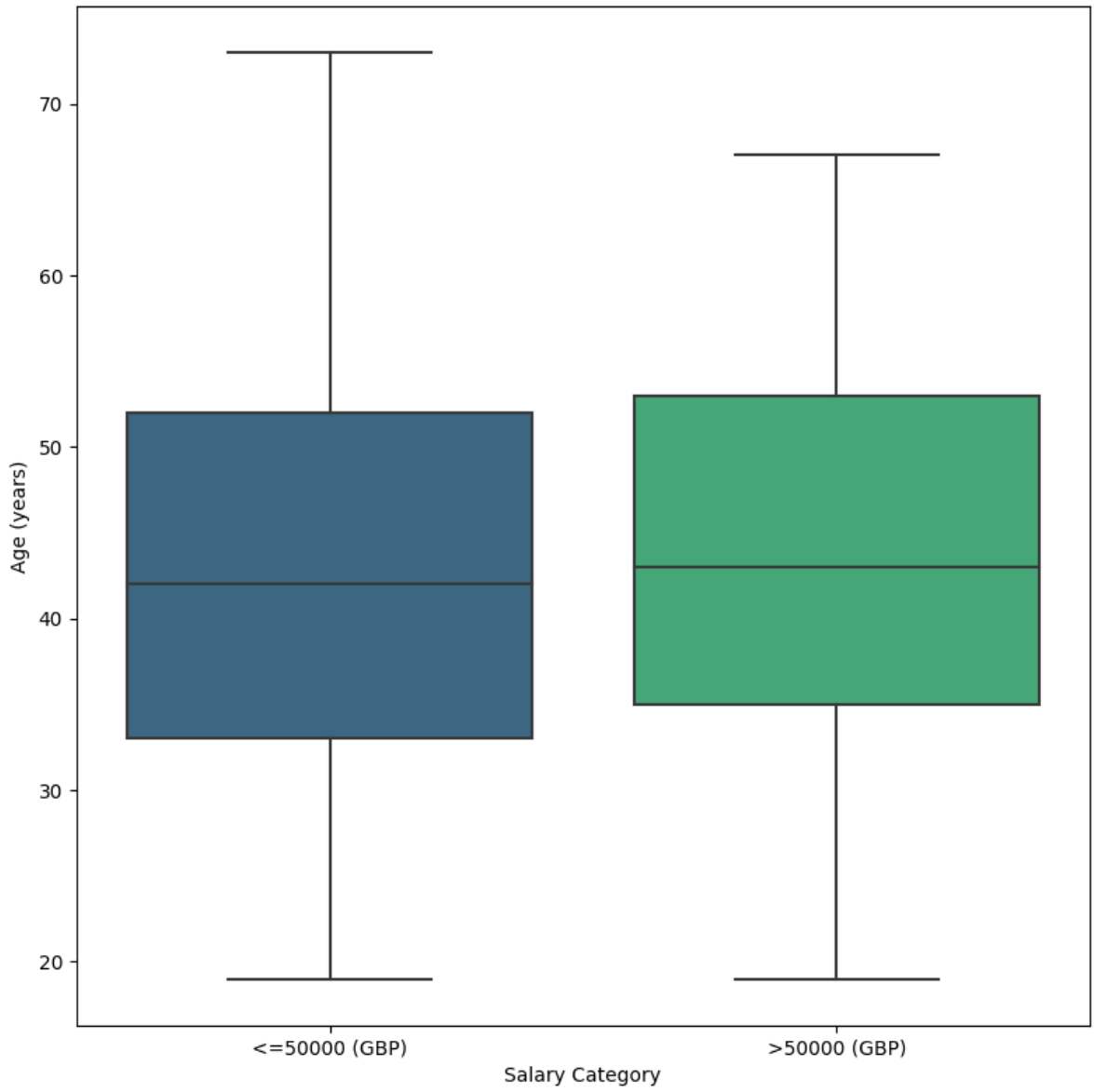


**Figure 13.** Salary by age

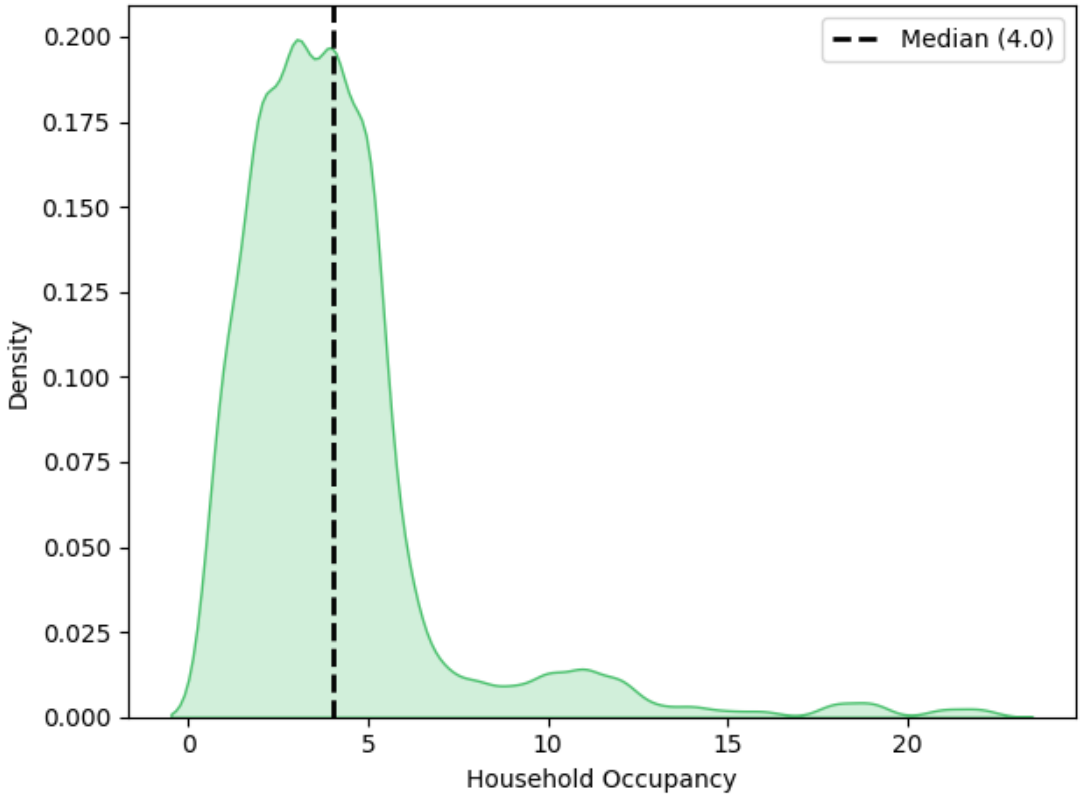


**Figure 14.** Occupancy distribution

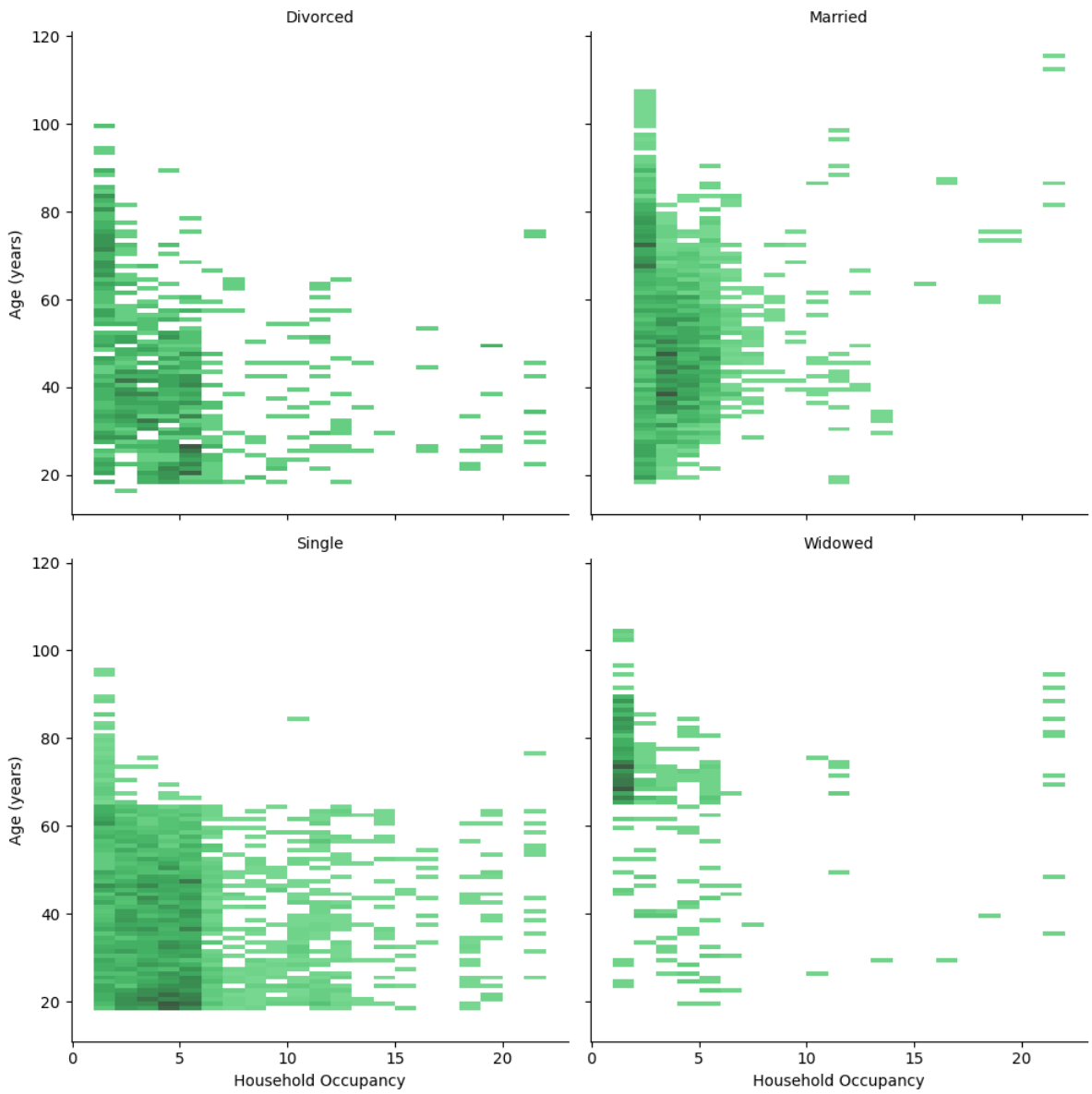


**Figure 15.** Household occupancy by age and marital status

## 6.2 Tables

**Table 1.** Imputation of empty strings

| *Column* | *No. Empty Strings* | *Impute* | *Method* |
|---|---|---|---|
| Age | 1 | Median age | Single-value estimate |
| First Name | 1 | 'Unknown' | |
| Gender | 3 | 'Female' | First name |
| Infirmity | 13 | 'None' | |
| Marital Status | 1 | 'Single' | Adult living with parents |
| Religion | 2 | Mode religion | Single-value estimate |
| Surname | 2 | Household surname | Surname of entire household |

**Table 2.** Marital status for aged 18 and under

| *Column* | *Marital Status* | *No. records* | *Action* |
|---|---|---|---|
| Under 18 | Divorced | 1 | No action |
| | Married | 7 | Deleted (2 households) |
| | Single | 1 | NaN |
| | Widowed | 0 | No action |
| 18 | Widowed | 7 | 'Single' |

**Table 3.** Additional columns

| *Column* | *Description* |
|---|---|
| Age group | Ages banded into 5 year age ranges |
| Occupation Category | Population categorised as Child, Child (school-age), Student, Employed, Retired, Unemployed |
| Household Occupancy | Count of individuals per household |
| Salary (GBP) | ONS median salary (Office for National Statistics, 2023c) |

**Table 4.** Demographics

| *Summary Statistics* | *Age* | *Salary* | *Household Occupancy* |
|---|---|---|---|
| Mean | 36.3 | 34952.3 | 4.1 |
| Median | 36.0 | 34909.0 | 4.0 |
| Min | 0 | 0 | 1.0 |
| Max | 116 | 78599.0 | 22.0 |
| Std Dev | 21.8 | 10270.2 | 3.0 |

**Table 5.** Divorce and marriage rates

| *Statistic* | *Result* | *Description* |
|---|---|---|
| Crude divorce rate | 56.6 | Per 1000 of total population |
| Refined divorce rate | 424.4 | Per 1000 married females |
| Marriage rate | 132.6 | Per 1000 of total population |
| Divorce:marriage ratio | 0.43 | No. divorces/ No. marriages |

**Table 6.** Birth and death rates

| *Statistic* | *Result* | *Description* |
|---|---|---|
| Crude birth rate | 10.5 | Per 1000 of total population |
| Fertility rate | 44.1 | Per 1000 childbearing females |
| Evolving birth rate | -10.9 | 25-29 age birth rate compared to 30-34 age birth rate (per 1000 females in age range) |
| Crude death rate | 9.3 | Aged 65 and over per 1000 of total population |

**Table 7.** Immigration and emigration rates

| *Statistic* | *Result* | *Description* |
|---|---|---|
| Immigration rate | 27.4 | Single lodgers/visitors per 1000 of total population |
| Emigration rate | 18.0 | Difference in male and divorced females per 1000 of total population |
| Net migration | 9.4 | Positive if immigration > emigration |

**Table 8.** Occupancy summary statistics

| *Summary Statistics* | *Occupancy* |
|---|---|
| Mean | 4.1 |
| Median | 4.0 |
| Mode | 3.0 |
| Min | 0 |
| Max | 22.0 |
| Standard deviation | 3.0 |
| Skewness | 2.7 |
| Excess kurtosis | 10.0 |
| Variance | 9.2 |

## 6.3 Cleaning log

Cleaning date and time 03/12/2023 at 9.30am.

Used Pandas profiling report to assess data prior to cleaning.

Census Report

Notes from profiling report and initial data exploration

Marital status (1975 entries) and Religion (2012 entries) have null (NaN) values.

- House Number (int data type).
- Street (object) -string data, no inconsistencies, check for empty strings ' '.
- First Name (object) - string data, check everything is hyphenated where it needs to be, 1 empty string ' ' needs dealing with.
- Surname (object) - string data, check everything is hyphenated where it needs to be, 2 empty strings ' ' need dealing with.
- Age(object) - should be integer but some floats, oldest person is 116, there are few ages >100, no negative ages, empty strings ' ' to deal with.
- Relationship to Head of House (object)- There are 'None' inputs in the series.
- Marital Status (object) -There are NaN inputs in the series, 1 empty string ' ' needs dealing with.
- Gender (object)- Male and Female, 3 empty strings ' ' need dealing with.
- Occupation (object) – 1 empty string ' ' needs dealing with, unemployed category, some inputs not helpful e.g. "Make", "sub" and "Best Boy". Formatting not standardised/consistent. There are stacked entries (with comma), forward slash entries, brackets. Different length strings. Retired entries with occupation (helpful). Categorise (new column (category), employed, unemployed, retired, child, student). Need to convert anyone over age of 66 and unemployed to retired, unemployed.
- Infirmity (object)- there is 'None' inputs and empty strings ' '.
- Religion (object)- There are NaN inputs in the series. There are 2 empty strings ' ' and 'NA' inputs (children?), also 'None' category, also 'Private' and 'Housekeeper' inputs, and 'Jedi' too. Buddist spelt incorrectly.

Look out for whitespaces in strings (use .strip())
Any non-ASCII characters- No
Duplicate rows- No

Column headers OK except Age could read Age (years) to include units DONE

**Cleaning performed**

Age (object) DONE

- Should be integer but some floats (currently object with strings)- converted to int (int64) data type
- ' ' empty strings to deal with - imputed the median age as the replacement for any missing values
- Oldest person is 116 (this is legitimate), there are few ages >100, oldest person who ever lived was 122
- No negative ages

- Record 7508 is underage parent, and lives only with child (See marital status below (under 18, head of household, divorced)
- Record 4919, aged 15, have Married status- not legal- see marital status below- deleted entire household due to illegal relationship
- Record 4171, also underage age marriage - deleted entire household due to illegal relationship
- Lost 7 entries in total due to deletion of 2 households (acceptable as not significant to dataset)

Religion (object) DONE

- There are 2 empty strings ' ' - imputed the mode religion as the replacement for any missing values
- For children aged under 16, religion inputted as "Undeclared"
- NaN captures adults that purposefully didn't input a religion
- 'None' represents adults that have inputted that they have no religion
- 'Housekeeper' and 'Jedi' - all assigned to 'None' (after checking household religions)
- ‘Private’ assigned to NaN (not ‘None’ as adult choosing not to disclose religion
- Buddist spelt incorrectly - changed to Buddhist (affects 3 records)
- 8109 non-null values

Marital Status (object) DONE

- There are 6395 non null values, 1 empty string ' ' needs dealing with
    - Converted the 1 instance of empty string to Single
    - The participant is living at home, aged 25, imputed as Single because if put NaN, this would be only instance of this in anyone 18 or over (creating an outlier in itself!)
    - Look at under 18-
    - Widowed
        - no widows under 18
        - Plotted box plot of marital status versus age, few outliers, possible to be widowed at 18 as in old age (but less likely)
        - Anyone 18 and widowed is unlikely and Single was imputed to replace the 7 instances of this
    - Divorced
        - One individual divorced under 18, Chloe Lewis aged 16
        - She has a has a child aged 0- possible underage parent
        - She lives just with her child and aged under 18- keeping record in because it is legal to live alone with child at 16 (but parents still responsible for wellbeing until 18
    - Married
        - Underage marriages
        - For record 4919, Valerie Smith, aged 15 has Married status (19 year old husband, no children)- not legal- deleted entire household (small impact to dataset)
        - Record 4171, Abigail Begum, can't determine whether she got married before 27 February 2023, when it was legal to marry at 16 or 17 with parental consent, not legal so deleted entire household (small impact to dataset)
    - Single
        - One instance and listed as head of household
        - Ok to live in household with over 18s as 17 year old, but because under 18 Single status changed to NaN

Street (object) DONE

- String data, no inconsistencies
    - No empty strings
    - No NaN
    - Removed whitespaces
    - Formatting all the same for Street strings

Relationship to Head of House (object) DONE

- There are 'None' inputs in the series
    - Formatting fine
    - No empty strings
    - No NaN
    - Removed whitespaces

Gender (object) DONE

- Male and Female, 3 empty strings ' ' need dealing with
    - Converted all 3 empty strings to Female based on first name and no NA/None option
    - All entries are either Male or Female
    - Removed whitespaces
    - No NaN

Infirmity (object) DONE

- There is 'None' inputs and empty strings ' '
    - Converted all 13 empty strings to 'None' as have no indicator if they are disabled or not
    - Removed whitespaces
    - No NaN

First Name (object) DONE

- String data, check everything is hyphenated where it needs to be, 1 empty string ' ' needs dealing with
    - 1 empty string replaced with 'Unknown'
    - Removed whitespaces
    - No NaN
    - No hyphenated names in this series and no hyphens need adding

Surname (object) DONE

- String data, check everything is hyphenated where it needs to be, 2 empty strings ' ' need dealing with
    - For 2 empty strings, looked up the households for two children and imputed the surnames of others in household (record 60 = 'Patel' and record 4236 = 'Lloyd')
    - Removed whitespaces
    - No NaN
    - No names missing a hyphen

Occupation (object) DONE

- Legal age for state pension is 66
    - Updated anyone aged 66 or over as 'unemployed' to 'Retired Unemployed'
    - 1 empty string replaced with 'Unknown'

- Removed whitespaces
- Don't need to format the column more because will look for partial string matches (case-insensitive using .lower()) in strings) for salary fuzzy match
- Instances where occupation is not empty but isn't determinable e.g. 'Best boy', will be converted to 'Unknown' and excluded from any salary analysis
- Chloe Lewis (aged 16) was identified previously as divorced and a single underage mother
    - She is the head of the household, living alone with her newborn and has inputted occupation as unemployed
    - She is still living under parental responsibility until 18 even though she lives apart from parents
    - Because she is an exception can class as either unemployed (as head of household) or as child
    - Keep as unemployed

Classifying data (create new columns) DONE

- Age ranges (5 years age bands for age pyramid) DONE
    - Created histograms of age by age groups (5 year ranges)
    - Converted data type from category to object

- Occupation category- employed, unemployed, retired, child, student DONE
    - Include in child category anyone aged under 5
    - Include in child (school-age) category anyone aged >= 5 and under 18
    - Include in student category any string containing 'student' aged 18 or over
    - Include in retired category any string containing 'retired' (will include retired unemployed as retired because it comes first in function)
    - Include in unemployed category any string containing 'unemployed'
    - Everyone else is employed

Additional metrics required (Create new columns) DONE

- Household occupancy (will help with determining under- or over-occupancy etc.) DONE
- Salary (created by fuzzy matching with ONS occupation/salary for UK) DONE