2021

CENSUS PROJECT

FUNDAMENTALS OF DATA SCIENCE

RUTH ONI

Introduction

The ideal state of any society is one where resources are adequate for the society's population. The population composition in comparison with the available resources determines the quality of life of residents as well as the level of development of the community (National Infrastructural Commission, 2021).

This project aims to generate and analyze statistical data to make decisions about infrastructural development and allocate investment funding.

Specifics include allocating use for a plot of land to one of the following:

- Housing - High-density/ Low-density
- Transportation - Train Station.
- Religious building
- Emergency Medical building

And to allocate funding for one of the following:

- Employment and Training
- Old age care
- Primary and secondary education - School funding

Objectives:

i. Determine population expansion
ii. Determine the demand for large family housing
iii. Determine the population of commuters
iv. analyze religious progeny
v. Project future births/ pregnancies and accident-prone population
vi. Determine unemployment rate
vii. Project retired/aged population
viii. project school children population

Methodology

Data – Census data of UK town. combination of categorical and discrete data

Data features include the following respondents' data: - *House number, Street, First Name, Surname, Age, Relationship to Head of House, Marital status, Gender, Occupation, Infirmity, and Religion*.

Features were expanded to include *Age group*, *employment status*, and *House address* for further analysis and *Housing type*, *commute*, and *key* to enable visualizations. Issues like blank entries and age entry errors were handled in the data primarily by inferring from individual records and households. Examples are in Instances of missing names, house numbers, or street names and age. Blank entries in the occupation and Religion columns and were imputed as 'None', and 'single' in the Marital status column.

All children (younger than 18) were assigned a 'N/A' (Not applicable) Marital Status. The same applied to the Religion column for those younger than 16. In both columns, children accounted for most nan values. Religions like 'housekeeper' and 'Private' were also converted to 'None'.

Three respondents with ages less than 18 had a marital status. Upon further investigation, they all appeared to be entry errors, and the ages were thus adjusted. One respondent record was deleted- Age was presented as 275 which was too far from the mean. The maximum age value for the population is considered 122. Further details of the data cleaning process are presented in the Jupyter notebook.

Software

Python.3.10 (Python Software Foundation, n.d.)

Pandas Python Data Analysis Library. V.1.3.4 (McKinney, n.d.) was used for data cleaning and pre-processing. Also, to generate Statistical profiling of data.

Sklearn for predictive analysis

CatBoostClassifier for predictive analysis. CatBoostClassifier works efficiently for categorical data as we have in this project. It features Inbuilt encoding of variables and management of overfitting.

Seaborn 0.11.2, Matplotlib (Waskom, n.d.) for data visualization

Descriptive statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can either be a representation of the entire population or a sample of it (Hayes, 2021). Measures of central tendency used in this study include the Mean, Median, Mode while measures of variability include the range, interquartile range, and standard deviation.

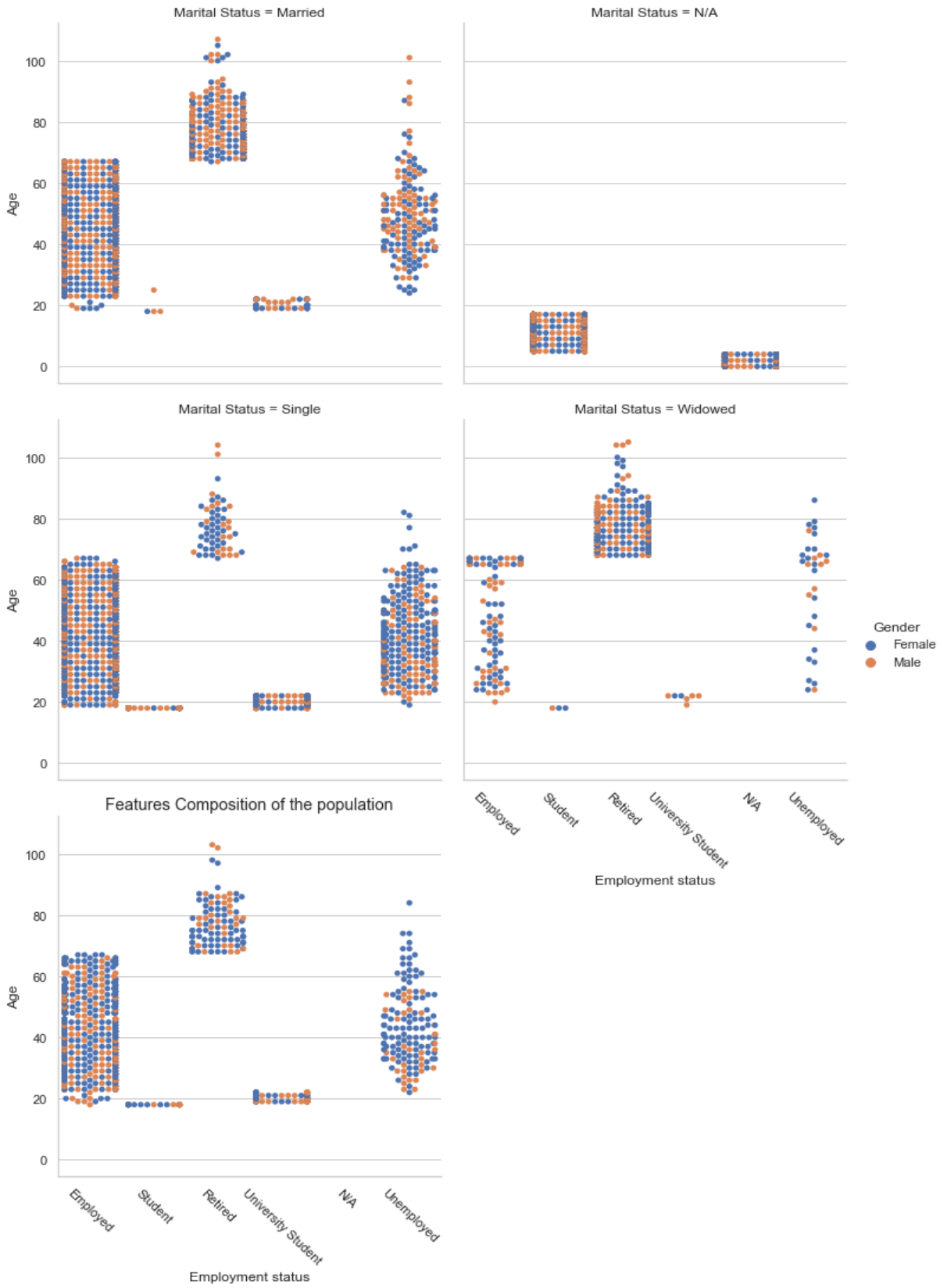


Figure 1. Composition of the population by age, Employment Status, and Marital status.

Demography

Age distribution

The population of the town is over 9900 with the male population making up about 47.8% of that and females comprising the remaining 52.2%

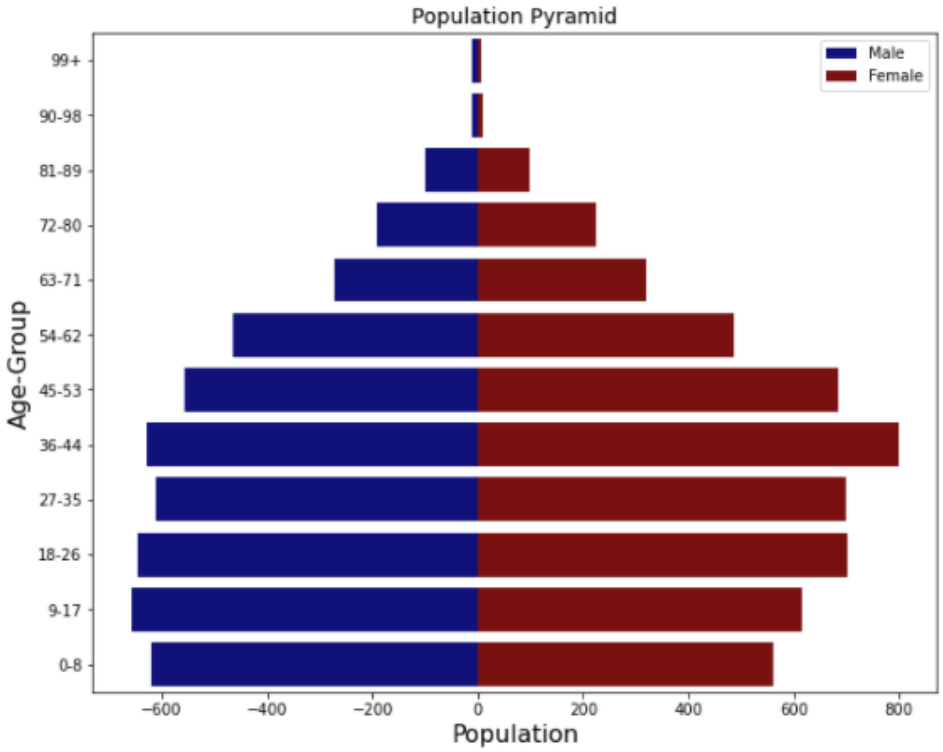


Figure 2. Age- Gender distribution of the population

From the widest sections of the pyramid and the distribution table, it is observable that middle-aged and younger people make up a large proportion of the population. The trend of the pyramid moves from expansive to constrictive and back which illustrates that the population may have been growing but not significantly and is now shrinking (National Geographic Society, n.d.). The average age of the population is 36 and the maximum age value is 107

Table 1. Age composition of the population

Age group	**population**	**Percentage of total population**	**Population by Gender**	
			Male	Female
Children(0-17)	2454	24.59%	1278	1176
18 - 25	1204	12.07%	586	618
26 - 45	3031	30.38%	1357	1674
46 - 65	2253	22.58 %	1072	1181
Aged(65 and older)	1036	10.38%	480	556

With 78% of the population 16 years and older, the most common marital status in the population is 'single'. Only 28% of the population is married which might indicate a high occurrence of lone occupancy across households.

NB: UK legal minimum age for marriage applied for analysis (UK Parliament, n.d.).

Table 2. Marital composition of the population

Marital Status	**Composition by gender**		**Percentage of total population**
	Female	Male	
Single	1783	1547	33.4%
Married	1413	1396	28.2
Divorced	601	369	9.7%
Widowed	232	183	4.2

Less than 1% of the population have infirmities and except for children who are assumed not to have a religion, 34.4% of the population who do not have a religion, 10 religions are being practiced across the population.

Results

Population expansion and Housing

There are two reasons why population expands:

- Natural increase
- Net Migration.

Natural increase is indicated by a higher birth rate in comparison to the death rate, while net migration is indicated when the difference between immigrants and emigrants is a positive value (Statistical Institute of Jamaica, 2017). The expression of both factors in this study shows no indication of significant expansion in the population size. The *crude birth rate* of the population is calculated to be 9.12 births per 1000 persons and *death rate* at 14.83 deaths per 1000 persons. With death rate greater than the birthrate, there's no indication of expansion due to natural increase. There is also no observed increase due to migration with the population of immigrants during the year put at 16 persons and 373 emigrants resulting in a *net migration rate* of -35.94%. This may rule out the need for high-density housing.

About 67% of the households are well within the house size allowance of 3 occupants which is the *average occupancy* across the 3550 households in the town with only 16% housing exactly 3 occupants. However, the argument for Low-density housing is strong with 28% housing only one occupant and about 25% (882 households) of the remaining 33% households without extra room to-let over-occupied.

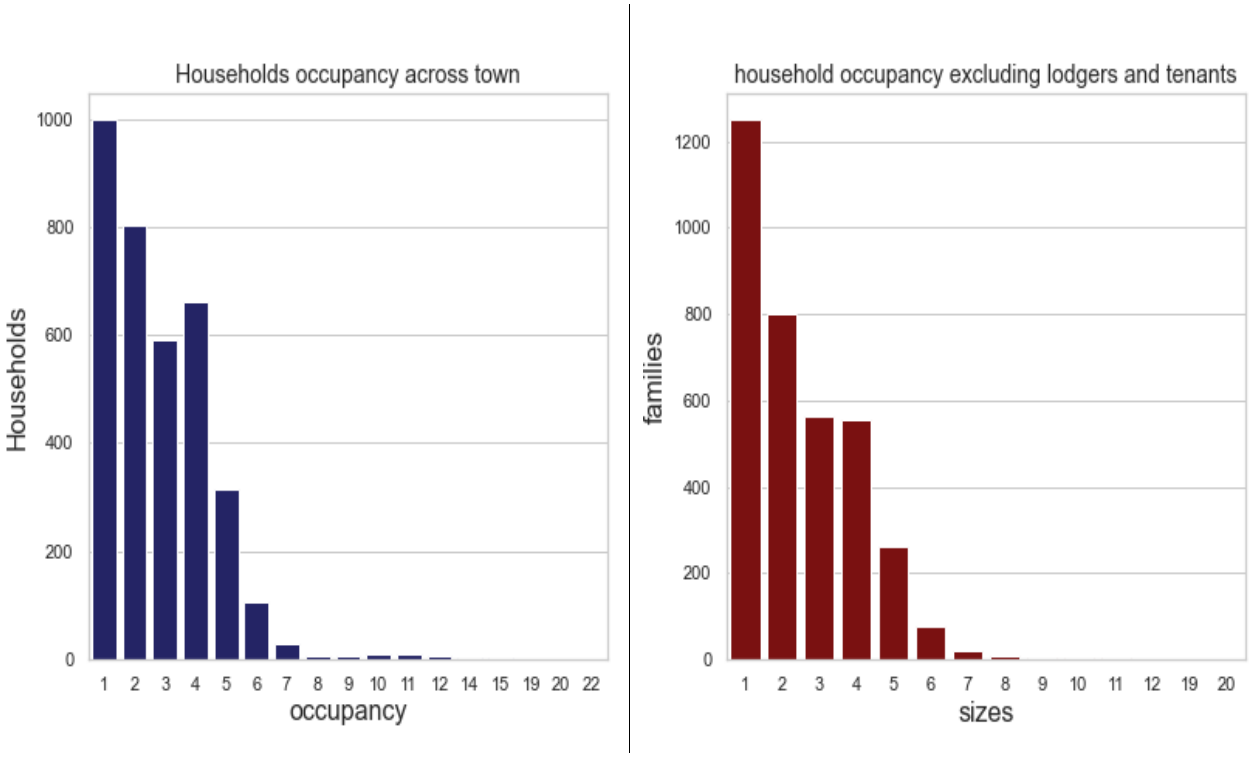


Figure 3 Households and Families

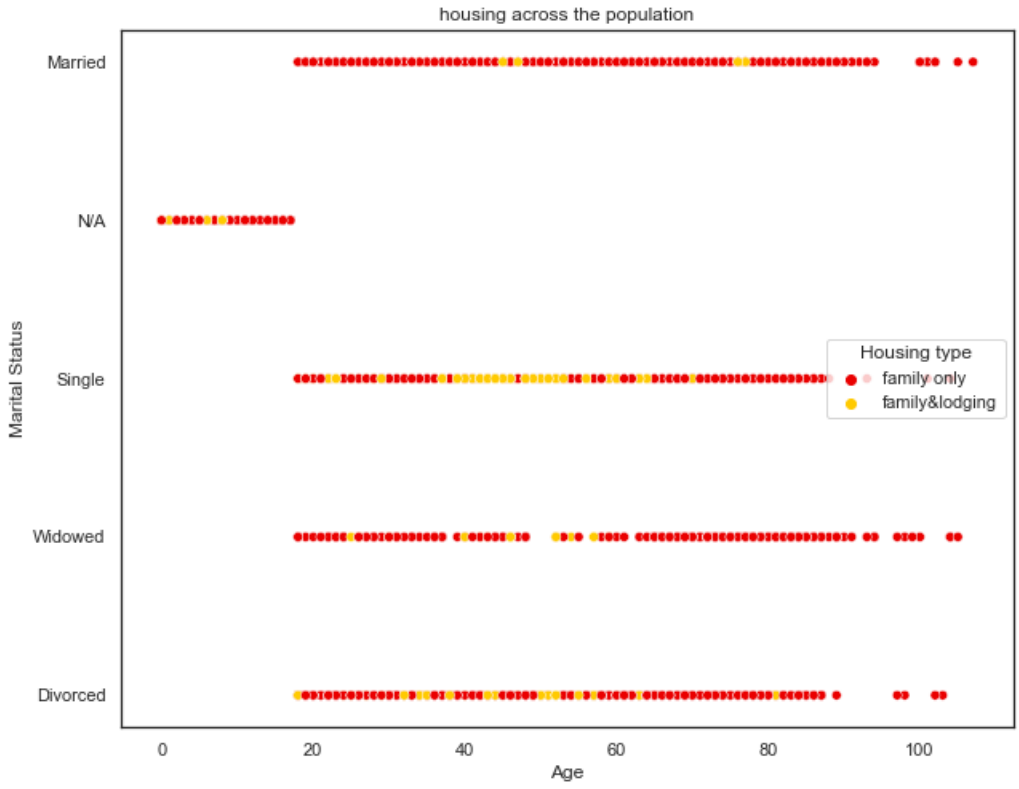
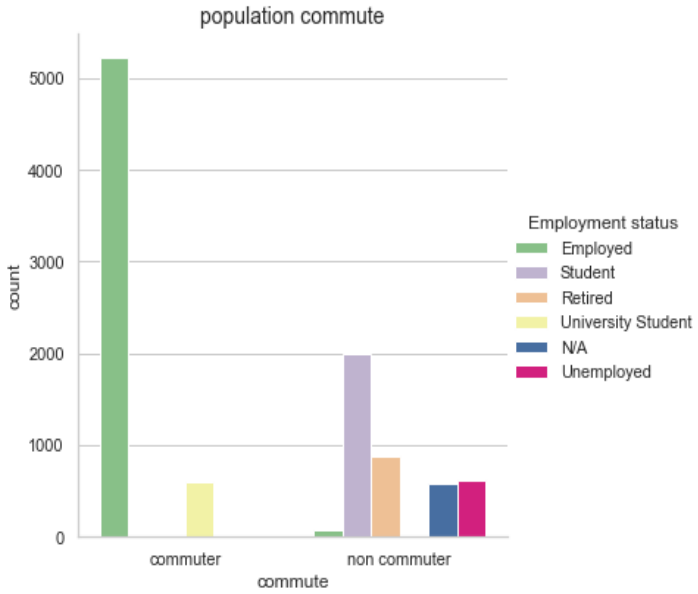


Figure 4. Distribution of Housing types

Commuters and Transport

"Commuting refers to a worker's travel from home to work" (United States Census Bureau, 2021). Commuters within the town were determined by occupation and employment status. Most professions within the population indicate that residents work and commute to the neighboring cities where large corporations, hospitals, and business opportunities better exist in comparison to the town. A sizeable number of university students and lecturers also live and commute to the university from the town.

Analysis reveals that these groups of people make up approximately 85% of the productive population (ages 15-64) of the town. This indicates existing pressure on the roads leading to the cities and makes an argument for a transportation infrastructure like a train station.



Emergencies

"Children under the age of five years and people in later life (those over the age of 65, and particularly those over 75) are most likely to have an accident at home" (Rospa, n.d.). Besides, the projection of future pregnancies and births in the population, for this project, possible emergencies were determined by assessing the growth rate and projected populations of the injury-prone age groups in the population

Growth rate and population projections were respectively calculated as:

PR = ((VPr - Vpa)/VPa) x 100 (PopulationU, 2019)

$P \times (1+R/100)^n$ (PopulationU, 2019)

The percentage growth rate of the population expressed as the percentage change of the population within one year is 0.93%. At this rate, the population of the town is projected to be 10,451 persons in 5 years with an expected 92 pregnancies/ births in one year. This percentage does not indicate a pressing need for a medical center. However, the population of children aged 0-5 is projected to grow from 717 to 1049 in one year and 4810 in five years at a growth rate of 46.3%. There's an expected significant increase in this group of the population. The population of people aged 65 and above is expected to decline in the next five years at a current growth rate of -12.5%. In total, up to 20% of the population will account for emergencies in the town during the next year.

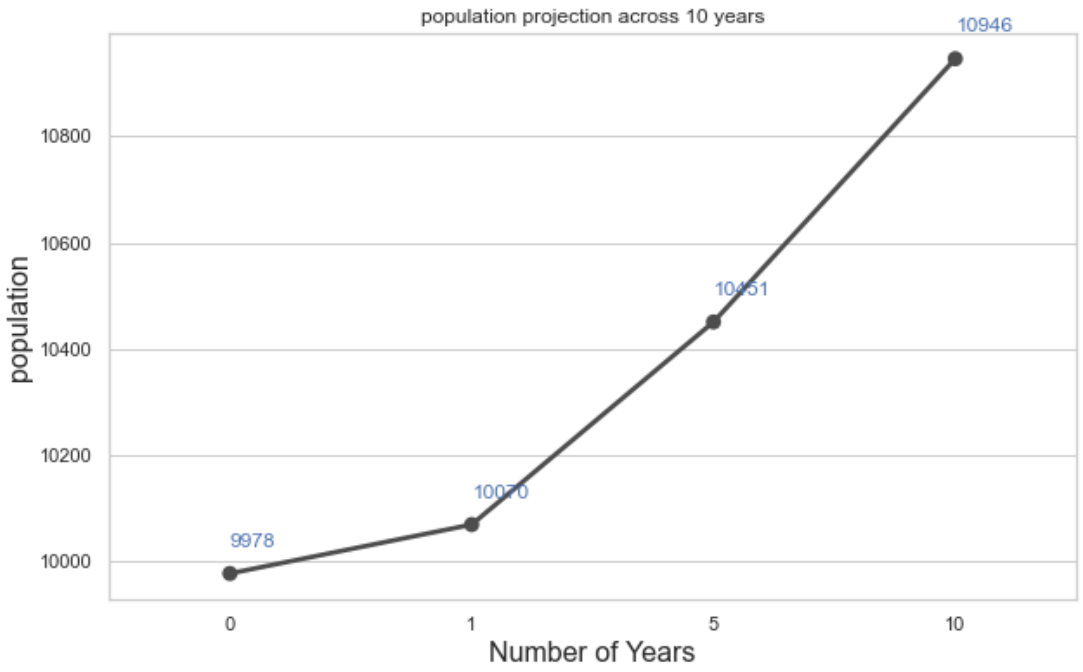


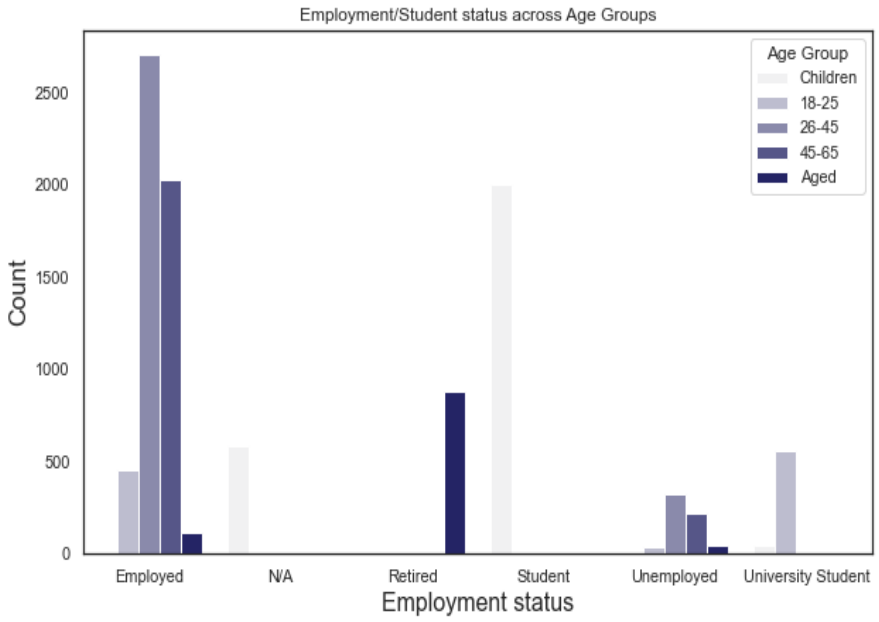
Figure 5: Projected population of the town across 10 years

Employment and Training

The employment status of members of the population was determined by occupation assessment. The official school start age in the UK is 5 years old (GOV.UK, n.d.). Therefore, employment status does not apply to that group of the population. The composition of the population by employment status is represented below

Table 3

Employment Status	**Percentage of total population**
Employed	53.09 %
Unemployed	6.25%
University Student	5.95%
Student	20.05%
Not applicable(children below 5)	5.82%
Retired	8.82%



The headline measure of unemployment for the UK is the unemployment rate for those aged 16 and over. Unemployment rates are calculated, by international guidelines, as the number of unemployed people divided by the economically active population (those in employment plus those who are unemployed) (Office for National Statistics, 2021).

The unemployment rate in the labor force of the population is 10.54% with persons within the age ranges 26 to 65 comprising most of the unemployed population.

Religion

There are currently more people who do not follow a religion than the population of followers of a religion.

Religion	count	mean
Agnostic	1.0	24.000000
Baptist	2.0	36.000000
Catholic	1054.0	42.923150
Christian	2170.0	49.563134
Hindu	1.0	80.000000
Jewish	37.0	40.702703
Methodist	650.0	45.658462
Muslim	106.0	35.698113
N/A	2454.0	8.909535
None	3437.0	41.929881
Orthodoxy	4.0	39.000000
Sikh	62.0	34.080645

Christianity is currently the most dominant religion even with the catholic denomination already having a church. This might indicate the need for a new church building. A prediction of future populations suggests that The Catholic, Christian, and Methodist religions will be dominant in an almost similar order as is currently observable. Other religions may die out with the current generation of adults. However, the median ages of Christianity (51 years old), Catholicism (41 years old), and the Methodist religion (44 years old) do not suggest a large following from the *total population* in the future years.

Figure 6. Religion statistics currently

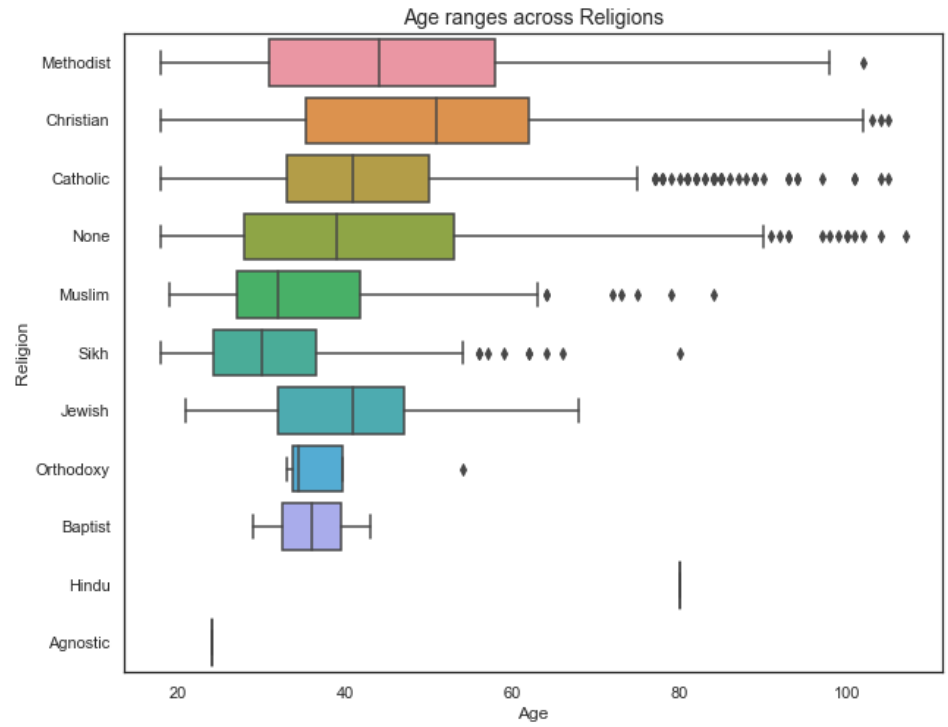


Figure 7 Age ranges across religions

Religion	count	mean	std	min	50%	max
Catholic	217.0	21.705069	3.297814	17.0	22.0	27.0
Christian	426.0	22.596244	3.246194	17.0	23.0	27.0
Methodist	144.0	21.805556	3.082270	17.0	21.0	27.0
None	806.0	21.710918	3.149612	17.0	22.0	27.0

Figure 8. Religion statistics in ten years

Old Age care

Based on a current death rate of 14.8 deaths per 1000 persons, the population of aged people (65years and older) in the town is expected to reduce by 82 deaths over the next five years. However, the population of current 59- to 64-year-olds indicates an increase in the total aged population by 514 people in five years bringing the current population of 1106 to 1537.98 persons by then. This is sufficient evidence of an increasing aged population.

School funding

The population assessed for this area is the total of school-aged children and potential school-aged children (ages 0-5) in the population (Gov.UK, n.d.). The comparison of the current population of children aged 0 to 16 and the population of the same group from the last year reveals that there was no increase in the past year. The result is a negative growth rate value for this group of the population. The current population of school children and potential school children in the town is 2307. This number is projected to be 2046 in five years based on a growth rate of -2.37%.

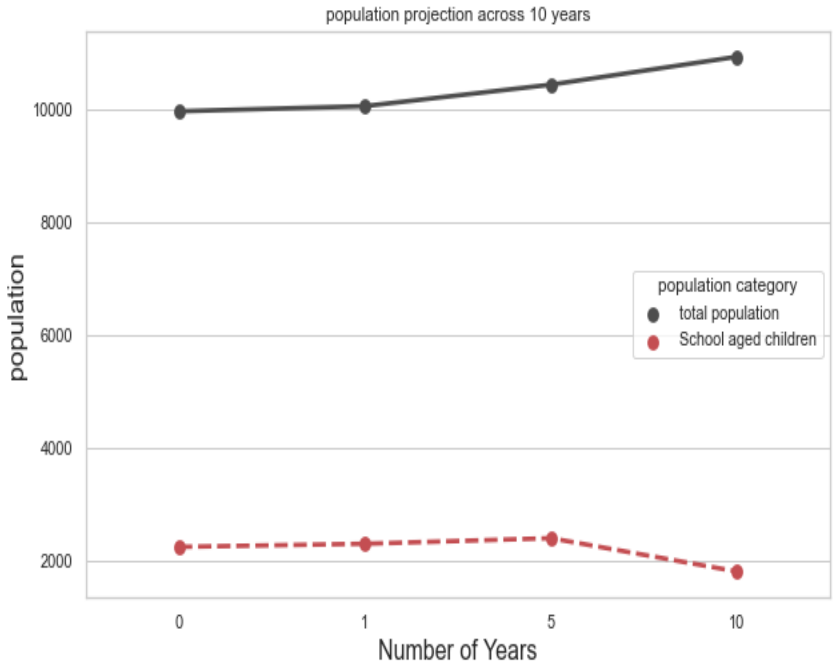


Figure 9. Population projection of School Aged Children and the Total Population

Recommendations

Infrastructure

There is currently no significant evidence of the population expanding significantly as the birth rate is low in comparison with previous years and the current death rate. In addition, more people are moving out of the town than are coming in. This nullifies an expedient need for High-density housing. While there is an almost similar percentage of over-occupied (24%) and under-occupied (28%) homes without letting within the town, downsizing and restructuring of settlements might be more reasonable than building low-density housing.

Although there is an expected increase in the total population of persons vulnerable to accidents in the future, only 20% of the total population account for the group in the next year. Ambulance services to the nearby cities may still be sufficient to cater to medical emergencies. Therefore, an Emergency Medical center is not recommended for immediate consideration.

Most religions are also not predicted to grow and those which might, may not do so significantly due to the currently observable age ranges across religions. The large population of commuters in comparison to the total population, however, makes the most compelling argument for a train station as most of the12 population will be benefitting from such infrastructure.

Investment

Given that there is evidence of unemployment in the town, employment training might be considered for future investments to reduce the unemployment rate. However, Old-age care is a more imperative need as the aging population is expected to increase significantly over the next few years. The population of school children is also projected to increase but much less significantly than the aging population.

Acknowledgments

For code, methods, or insights gleaned, I would like to acknowledge the following sources

i. Stack Overflow
   - Count unique values per group with Pandas [duplicate].
   - How to get data labels on a Seaborn pointplot
ii. Dan Friedman
   - cut() Method: Bin Values into Discrete Intervals
iii. Data Carpentry
   - Indexing, Slicing, and Subsetting DataFrames in Python
iv. Edureka
   - What is the easiest way to implement IN and NOT IN in Pandas dataframe
v. United States Census Bureau
   - Calculating Commuter-Adjusted Population Estimates
vi. Statology
   - How to Create a Population Pyramid in Python
vii. Jake VanderPlas
   - Customizing Plot Legends
viii. Delfstack
   - Change Seaborn Plot Size
ix. Pretag
   - How to subtract rows of one pandas data frame from another
x. Analytics Vidhya
   - Simple Methods to deal with Categorical Variables in Predictive Modeling
xi. GitHub
   - TypeError: Cannot convert 'b'Y'' to float
xii. Catboost
   - CatBoostClassifier
xiii. Elite Data Science
   -Seaborn Tutorial
xiv. Effective ML
   - Parameters to tune for Classification
xv. In determining emmigrants. (Mohsenin, 1983)

Bibliography

Python Software Foundation, n.d. *Python.* [Online]
Available at: https://www.python.org/

Gov.UK, n.d. *School Admissions.* [Online]
Available at: https://www.gov.uk/schools-admissions/school-starting-age
[Accessed 06 December 2021].

GOV.UK, n.d. *School Admissions.* [Online]
Available at: https://www.gov.uk/schools-admissions/school-starting-age
[Accessed 8 December 2021].

Hayes, 2021. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/d/descriptive_statistics.asp
[Accessed 30 November 2021].

McKinney, n.d. *Pandas.* [Online]
Available at: https://pandas.pydata.org/
[Accessed 2021].

Mohsenin, I. C., 1983. Note on Age Structure of College Students.. *History of Education Quarterly,* 23(4), p. 491–498.

National Geographic Society, n.d. *Population Pyramid.* [Online]
Available at: https://www.nationalgeographic.org/encyclopedia/population-pyramid/
[Accessed 6 December 2021].

National Infrastructural Commission, 2021. *NIC-Infrastructure-Towns-and-Regeneration-Report.* [Online]
Available at: https://nic.org.uk/studies-reports/infrastructure-towns-and-regeneration/infrastructure-towns-regeneration-final-report/
[Accessed 30 November 2021].

Office for National Statistics, 2021. *A guide to labour market statistics.* [Online]
Available at:
https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/aguidetolabourmarketstatistics
[Accessed 5 December 2021].

PopulationU, 2019. *Population Formula.* [Online]
Available at: https://www.populationu.com/gen/population-formula
[Accessed 30 November 2021].

Rospa, n.d. *Facts and Figures.* [Online]
Available at: https://www.rospa.com/home-safety/advice/general/facts-and-figures
[Accessed 5 December 2021].

Statistical Institute of Jamaica, 2017. *Population and Demography.* [Online]
Available at: https://statinja.gov.jm/demo_socialstats/DemoMethodology.aspx
[Accessed 7 December 2021].

UK Parliament, n.d. *The law of marriage.* [Online]
Available at: https://www.parliament.uk/about/living-heritage/transformingsociety/private-lives/relationships/overview/lawofmarriage-/
[Accessed 30 November 2021].

United States Census Bureau, 2021. *Commuting.* [Online]
Available at: https://www.census.gov/topics/employment/commuting/guidance/commuting.html
[Accessed 30 November 2021].

Waskom, n.d. *Seaborn.* [Online]
Available at: https://seaborn.pydata.org/
[Accessed 2021].