FUNDAMENTALS OF DATA SCIENCE

CENSUS PROJECT REPORT

By-

Prakhar Banka

202117900

Context

This Project report is based on the analysis of the census data of a moderately sized town which has been used to make recommendations for investment in the future services and use cases for development on an unused plot of land in the council area.

We have been provided a mock census data to an imaginary modest town. To carry out the task of providing recommendations to local authorities this dataset needs to be cleaned and analysed. The first section will deal with the details of the data cleaning process, all the methods undertaken to remove anomalies in the dataset.

The subsequent section will highlight the key analyses undertaken aimed at to support the recommendations provided. This includes an overview of town population's demographics then we will be looking into detail analysis of town's population growth, commuters, healthy population and financial trends in households, employment rates and occupancy rates.

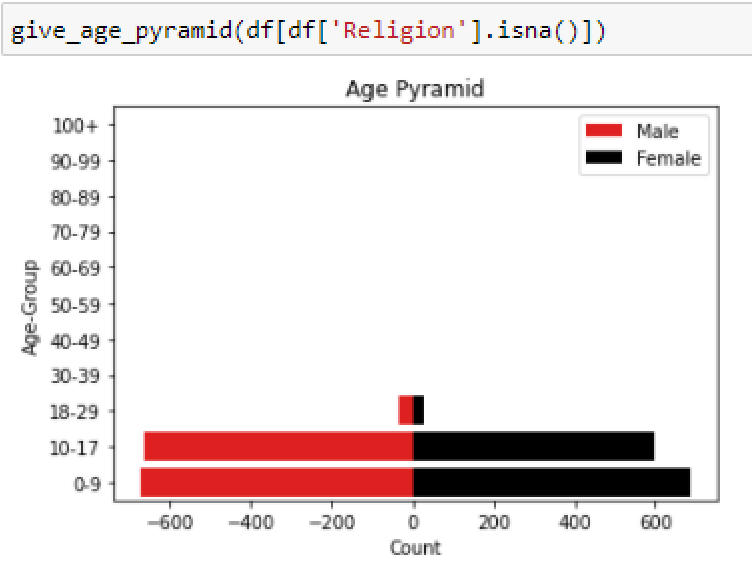
Data Cleaning

At this stage the census data was cleaned to remove any anomalies and inconsistencies in the dataset. The complete log of the data cleaning process undertaken can be viewed in the corresponding jupyter notebook.

Blank Data were imputed with the values by inferring them to the related records. Such as missing surname was imputed by looking at the household and changing it accordingly. Missing ages were imputed by looking at the occupation status which in my case were mostly students, so the mean age of students was calculated, and the blank space was removed by that value, also looking at the household and age of the parents they are staying with which also gave an idea of the possible ages of the individual.

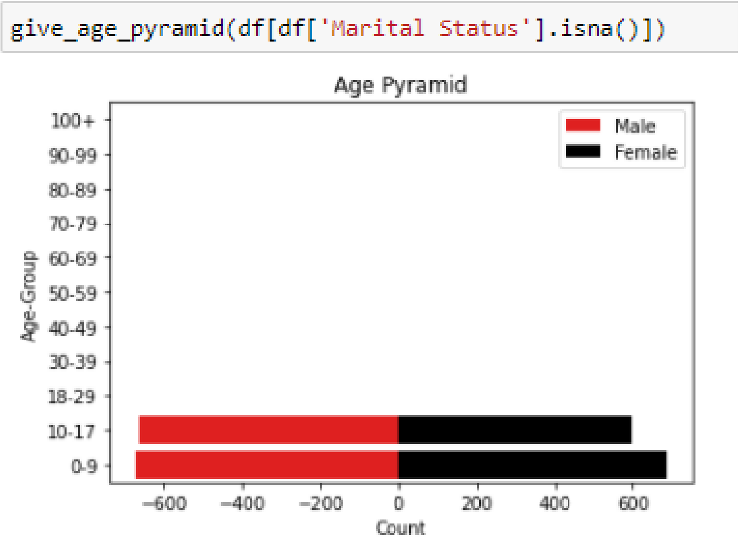
The Ages, which were in float, were converted to the whole number format and the entire age column was then converted to Int64 type for manipulation and analysis of data based on age.

Religions which were converted to None were – Sith, Undecided, Nope. Sith was a deliberate mistake or misleading, Sith cult is considered as a group of people believing in dark forces which is not considered as religion in UK. Nope was added to None as both mean the same.



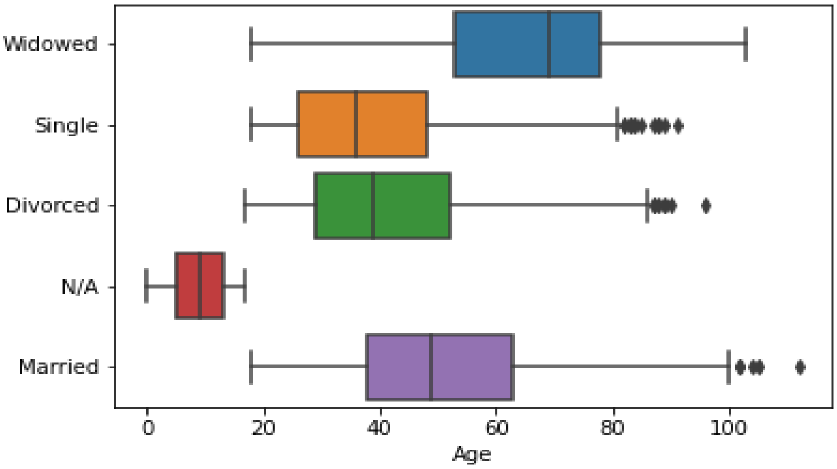
```
give_age_pyramid(df[df['Religion'].isna()])
```

I checked in for the age group of the nan values of the Religion values to know where they lie. The function written to generate the graph can be viewed in the jupyter notebook. This led to a conclusion that most of the nan values were of the people below 18 and on further investigation of the religious plots it was found not even a single record had mentioned the religion for people under 18 this led me to a conclusion that if parents don't provide a religious affiliation under 18 then mostly children are not bound to follow the household religion. But since for young children who might accompany their parents to a religious place till they are mature enough to choose a religion or otherwise should be considered if we think of investing in a religious place because children will also be considered in determining the occupancy size, so for this I changed the religion of young children to the religion of the parent. But for other investigation like size of the religion growing we might need to approach it differently. The rest of the left NaN values were also converted to None.



```
give_age_pyramid(df[df['Marital Status'].isna()])
```

Similarly, the marital status of most of the children has NaN values as shown in the graph. Since except one, none of the entries show a marital status in the census data and even if its allowed for people from 16 to 18 to get married with parental consent I went ahead to change it to N/A for all of the entries below 18 since determining if they are married or not is not possible looking at our dataset so it is better to generalise in this situation.



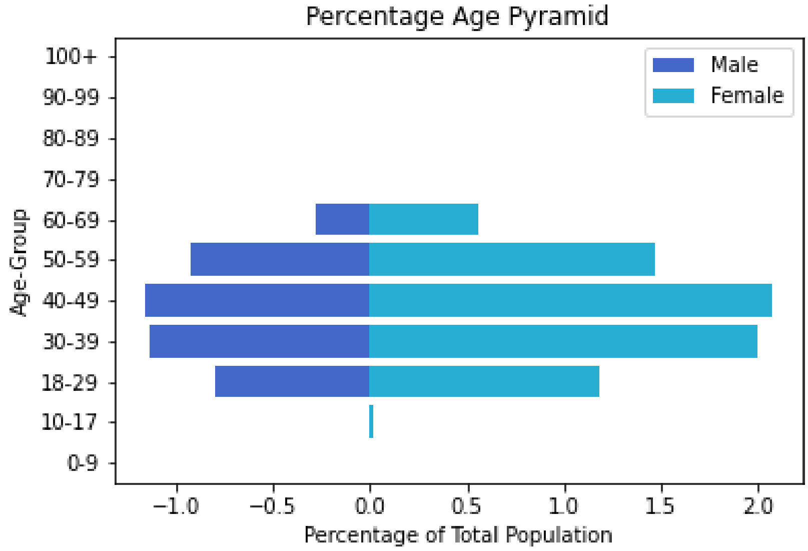
The Outliers in the cases were determined by the boxplot distribution used to visualize the average age of the people corresponding to a particular Marital Status.

On seeing the boxplot, we can infer that the majority divorce rates are high in young couples. And median marriage age is between 40 to 60. This shows less of young married couple as maybe the younger generation move out after marrying.

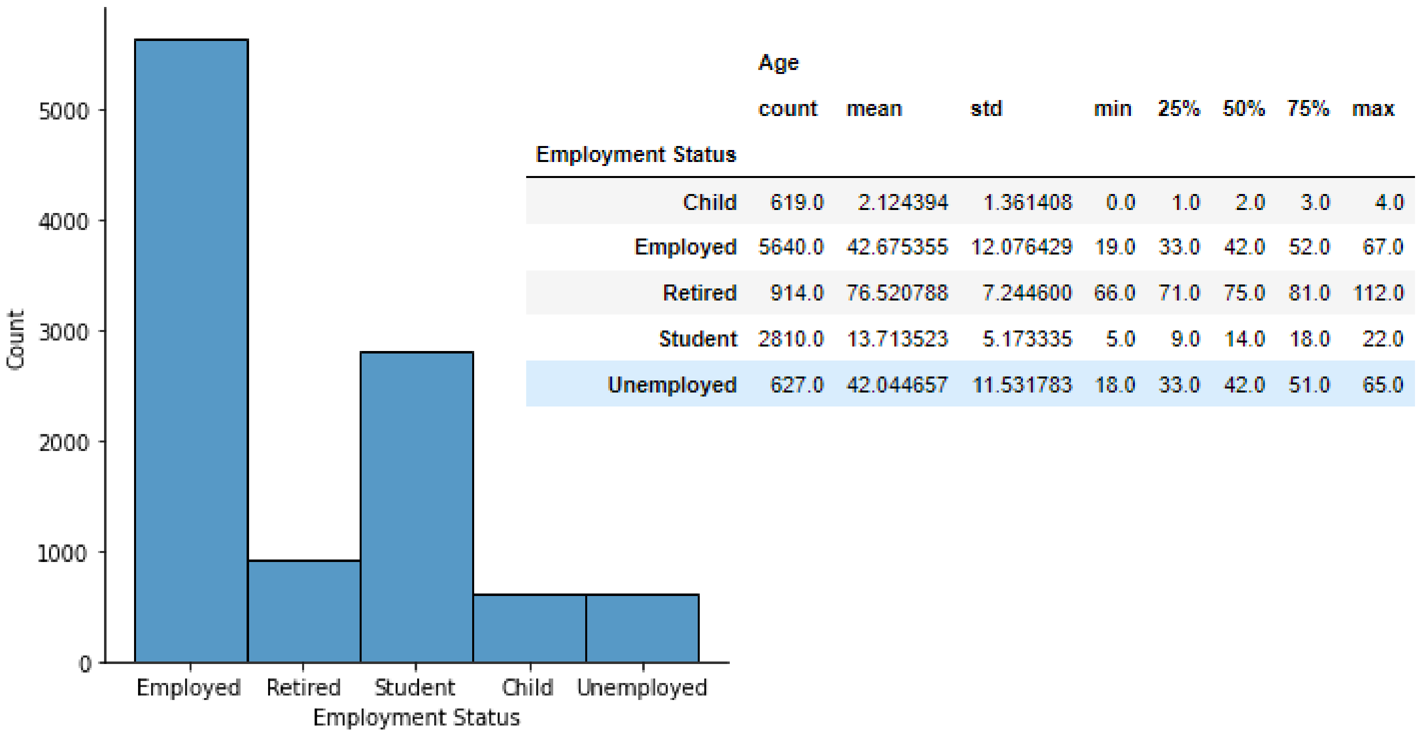
There are very few outliers and aren't absurd looking at our data.

The Unemployed population was as shown by the graphs constitute less than 1% in males in each age group and up to 2% in females. This shows the employment level of the town is quite good.

The older population of age 65 and above stated as employed was converted to 'Retired Unemployed'.



The Occupation data consisted of the occupation details of the population with different job titles, to better analyse the employment status of the population a new column was added which classified the population in the following categories:



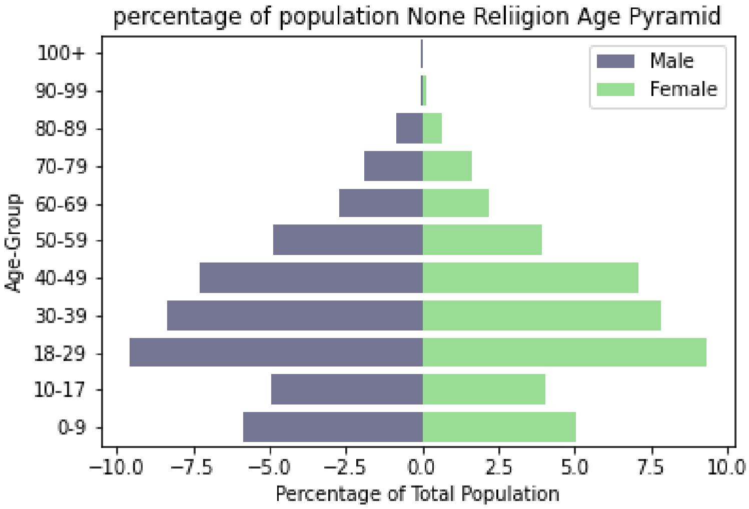
	Age							
	count	mean	std	min	25%	50%	75%	max
Employment Status								
Child	619.0	2.124394	1.361408	0.0	1.0	2.0	3.0	4.0
Employed	5640.0	42.675355	12.076429	19.0	33.0	42.0	52.0	67.0
Retired	914.0	76.520788	7.244600	66.0	71.0	75.0	81.0	112.0
Student	2810.0	13.713523	5.173335	5.0	9.0	14.0	18.0	22.0
Unemployed	627.0	42.044657	11.531783	18.0	33.0	42.0	51.0	65.0

This shows a high employment rate in our town and will also be used further to determine occupancy trends and gender-based analysis.

Data Insights

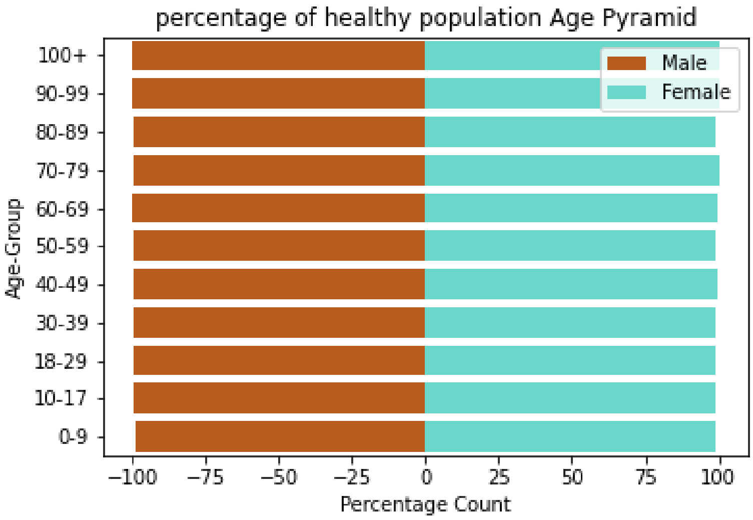
Religion

Large percentage of population in our census identify their religion to be None. This shows the town in consideration is less religious and do not affiliate themselves to a particular religion. It is also possible that people deliberately chose not to specify their religion in the census data. But inferring and imputing such details might disturb the religious distribution since everyone has the right not to be religious, so, we will be going ahead with the inference that majority of the population are not concerned about the religious matters in the census data.



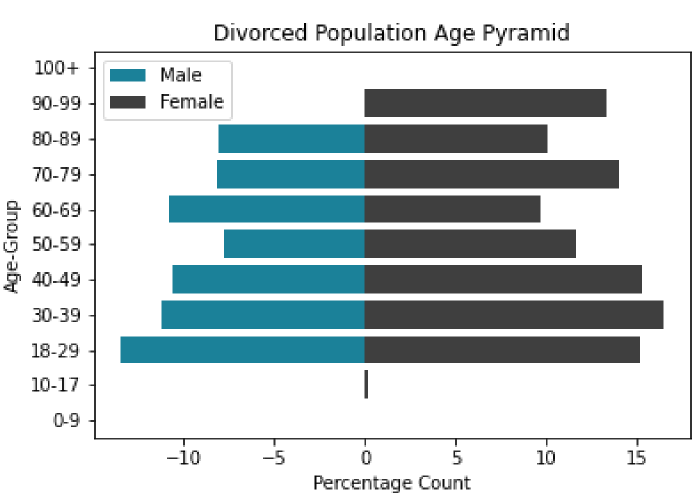
The younger population which are in significant numbers choose not to associate themselves with a particular religion. This shows decreasing popularity of religious beliefs in younger age population.

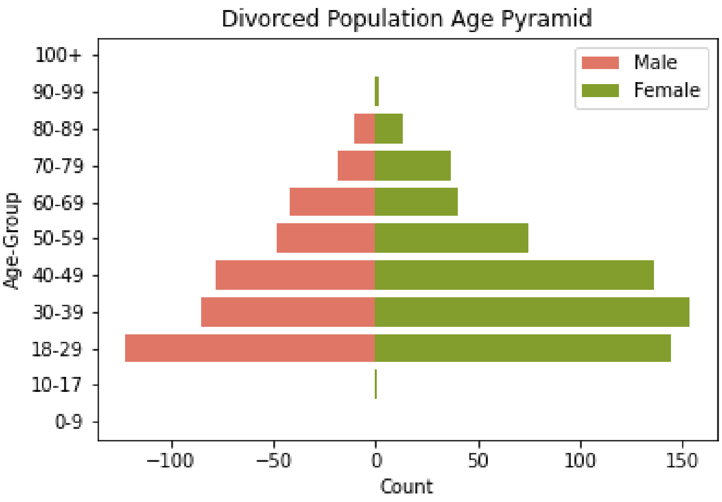
Infirmity



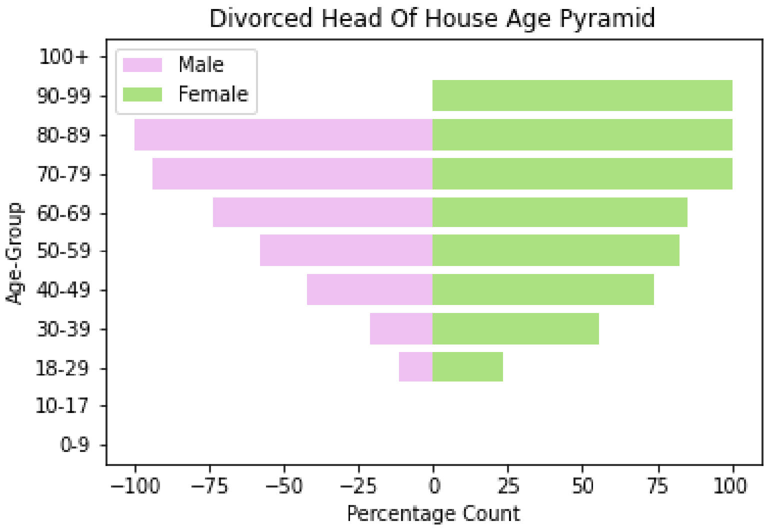
Almost the entire population was identified to be healthy with no infirmities and even the older people were healthy. This shows people in the town might have a larger life expectancy at birth.

Marital Status





As seen by the graphs it is evident that divorces occur from young age until old and going through the marital data of the population a significant trend was observed that the divorce rate is higher in younger generation of population and female divorce rates are higher as compared to male divorce rates.

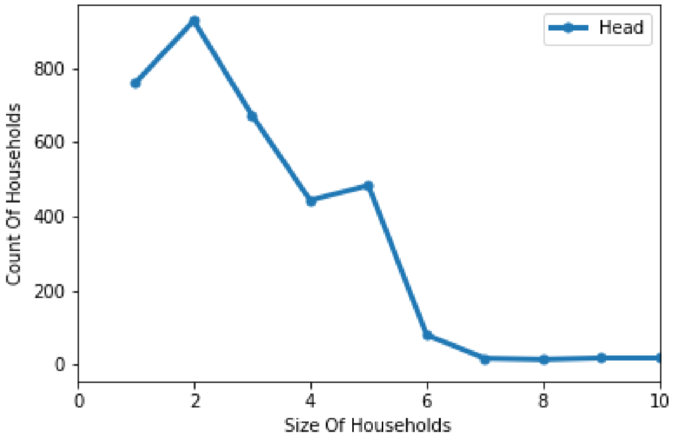


This Graph above shows the percentage of divorced people who are head of a house. This shows the younger divorced people have to move back with parents or are lodgers. This also shows more of the female divorced population are head of a particular house than males even in the younger age groups. We can also infer that maybe after divorce the males move out of the town.

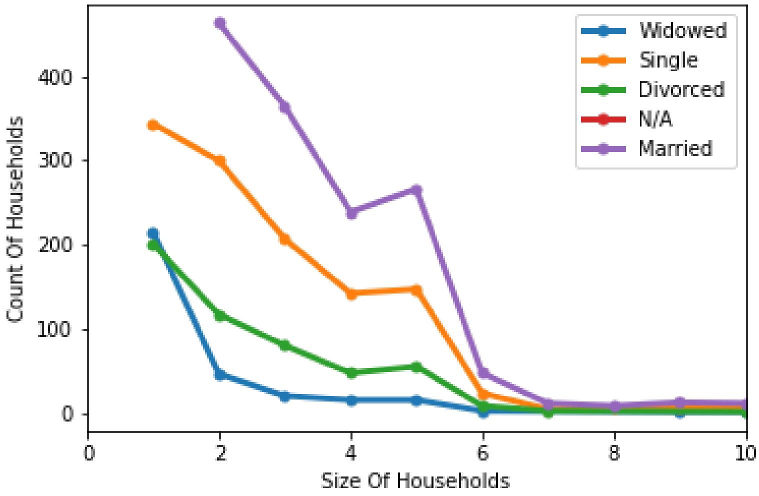
The mathematics behind the above graph is:

$$\frac{\text{Number of Divorced Males or Females Who are head Of House in the specific Age range} \times 100}{\text{Number of Divorced Males or Females in the specific age range}}$$

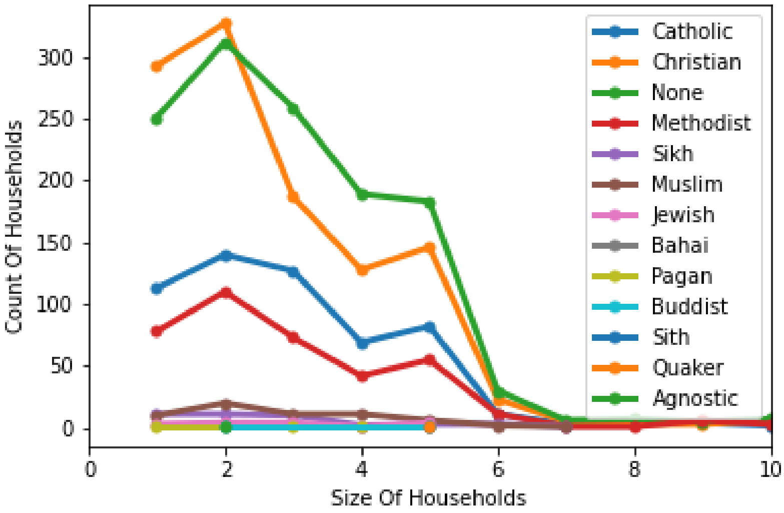
Occupancy



The Median occupancy rate of the town is two. The size of house when compared to other parameters in the dataset gives us some insight into the trends of occupancy level changing.

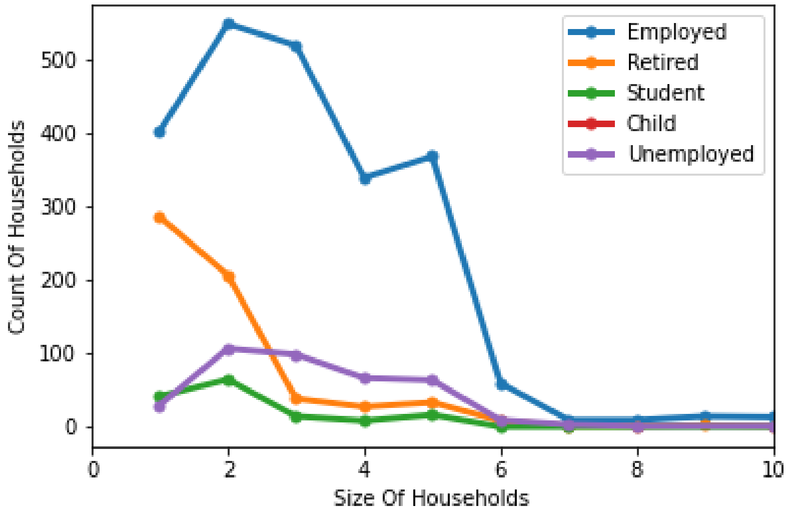


This graph above shows That most of the single, divorced, and widowed people are staying alone with the occupancy of 1 in the household. This also shows the households whose head is single also tend to stay in bigger household of 4 to 5 occupancy showing presence of lodgers in their household.

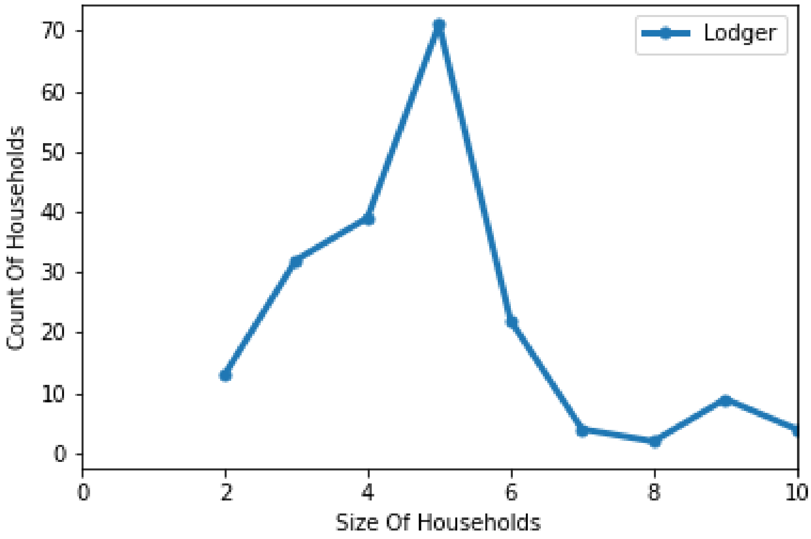


This above graph shows the religious distribution according to Households. This counts the households with Head of the House affiliated to a particular religion. This trend shows a significant number of households have Christian heads followed by Catholic and around 240 households have a head with none as a religion. Christian Heads have large number of bigger households as well.

Most of the Households have a head, who is, employed irrespective of the size of the house.
People with a job tend to stay in larger household size showing growth in family size with employment status. And Retired population which is also the older population in town tend to stay in smaller mostly a 2-3 occupancy size household.



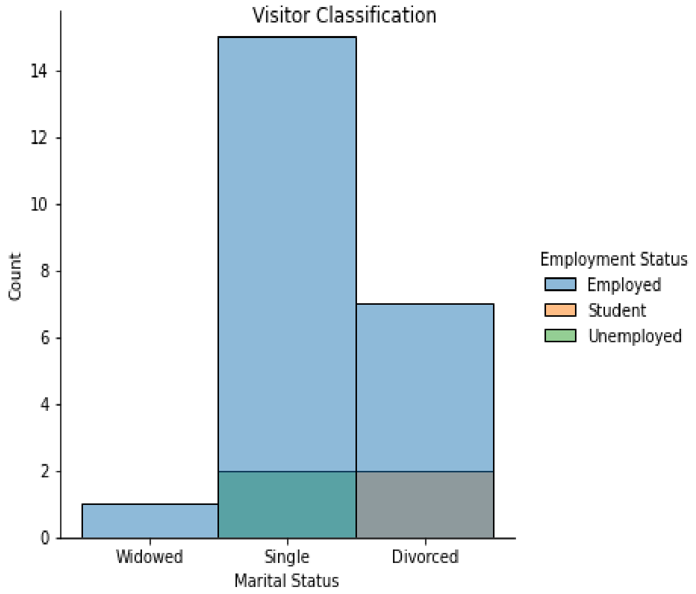
The presence of Lodgers is also significant in the households of size 4 or 5, who are mostly single people or divorced. This shows people instead of downsizing prefer to rent it out to lodgers in the town.

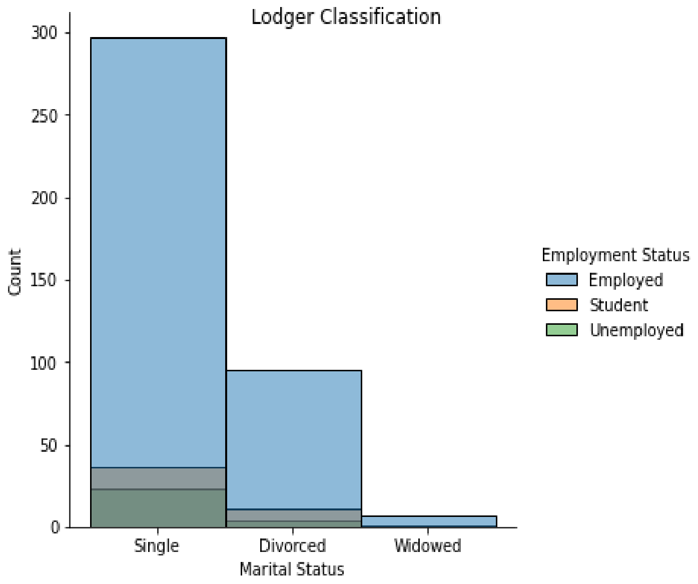


There is not a problem of extra housing in our town since the lodgers are few and also the houses are not much overcrowded and are mostly small in occupancy size. Since the size average size of households in UK is around 2.4.[1]

Immigration and Emigration

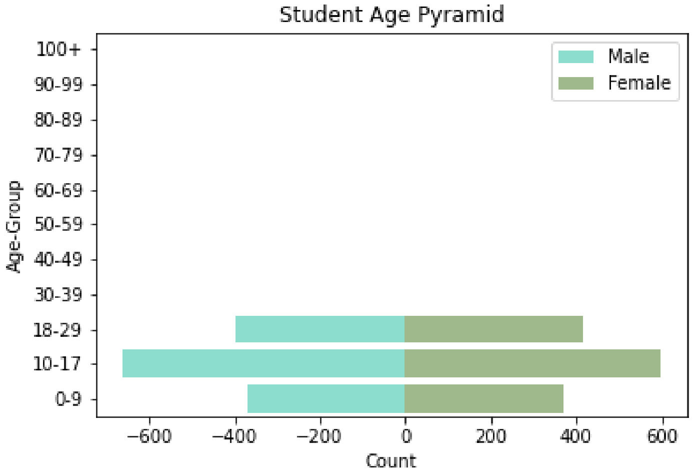
The people coming and settling in the town are mostly determined by the university student who stay as lodgers count of employed lodgers who must have come to town to work and are staying as lodgers. We mostly consider lodgers who are single since there must be lodgers who are divorced and just renting a place after divorce.





The graph above shows the Marital Status of Lodgers and Visitors along with their Employment Status to better classify exactly who can be counted as a immigrant lodger. The above trend doesn’t show a significant number of people who are coming into the town with respect to the population of town.

The people moving out of the town are mostly students who move out to study in the universities as shown by the graph above the count of students who are above 18 when mostly the college and universities begin significantly decrease showing a trend of students moving out of for higher education.



The emigration status can also be determined based on difference of the number of male and female divorces in the town. The difference showing the number of people moved out of town. And also grandparents who have their grandchildren staying with them without the parents living might also show people staying out for work.

Commuters

Commuters in our dataset can be determined using the employment data of our population. People who are employed are quite high in number in our town and significant number of the population are students who are studying in university.

Also including the people who have left children with grandparents must also be determined as commuters since they must be visiting their family to meet their children or the grandparents taking the children to meet their parents.

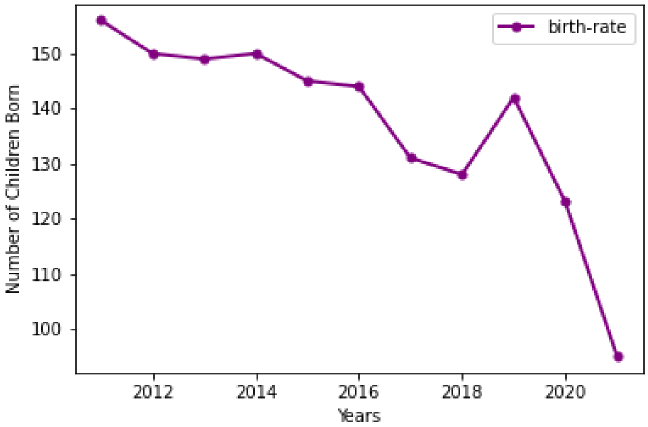
Occupation such as Teachers, retail workers, baristas, social workers are excluded from this list of commuters.

On calculating this by considering all the above scenarios over 60-70% of the population were determined to be commuting.

Birth and Death Rate

Looking at the dataset for a single year makes it tough to determine the actual birth and death rate of the town, since we must assume that there has been no family migration into or out of the town since a particular year. This makes the birth and death rate too hard to assume.

Looking at the current year the birth rate has been around 8 births per thousand people. Which is quite low compared to previous years assuming the facts as stated earlier.



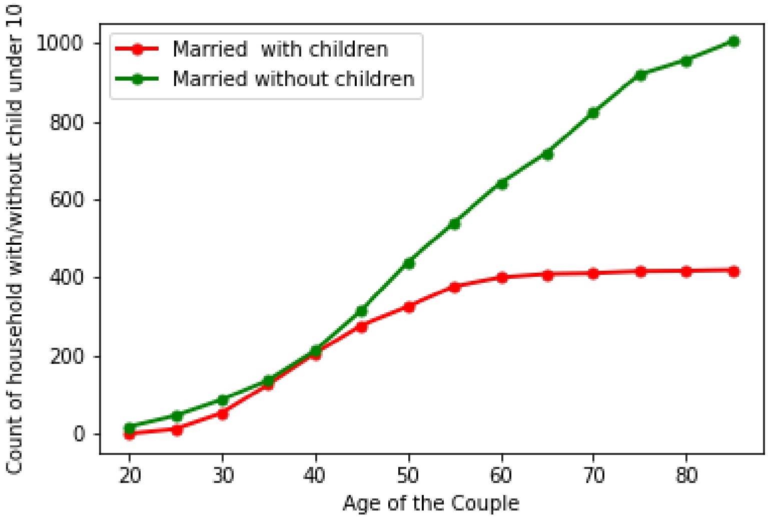
This graph depicts the falling birth-rate in our town which is deduced based on the assumption. This could be showing the true picture of the rate but it’s hard to conclusively say that this is the accurate birth-rate graph.

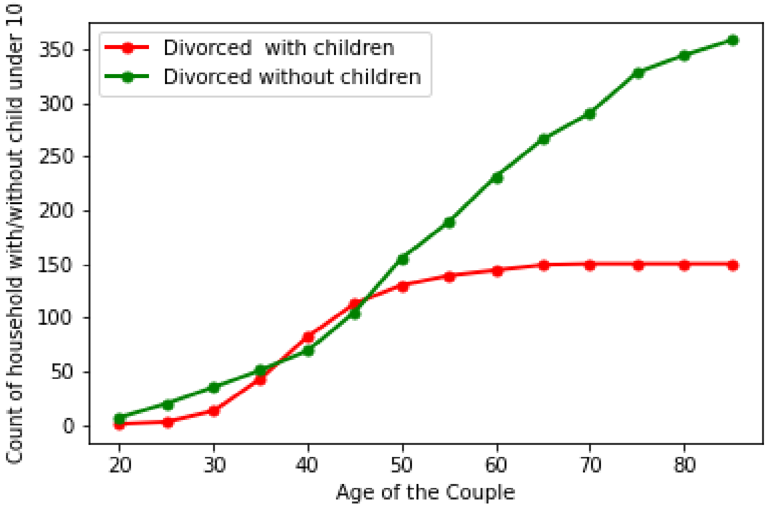
The Death-rate calculation made on the dataset of the current year can be roughly drawn by assuming the age of people above 65 and calculating the decrease in its frequency as the age grows assuming people under 65 do not move out of the town. These Death and Birth rates cannot be roughly determinant of the growth of the town size but might give some idea about the census growth rate.

Growth of Household with children

The growth of family in the coming years can be predicted with the trend in the current data looking at the growth rate of families with respective of the ages of the couple. The major analysis for predicting the future needs of the town with respect to family.

The Graphs below shows the number of households with child aged under 10 that determines born in recent years, staying with the parent who is married or divorced with respective to their parents' age. The trend shows that the average age when couple decide to have a child in the family is around 30 to 45 which shows the exponential growth in the number of children.



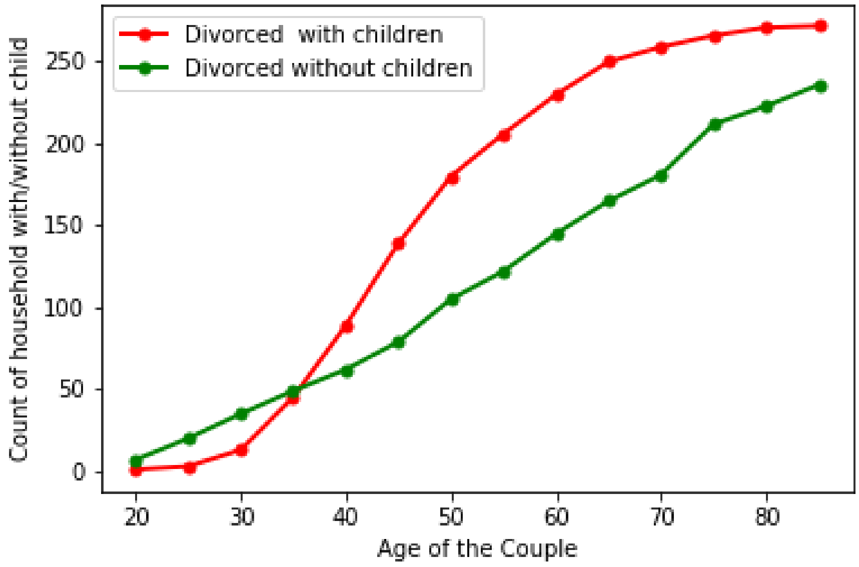


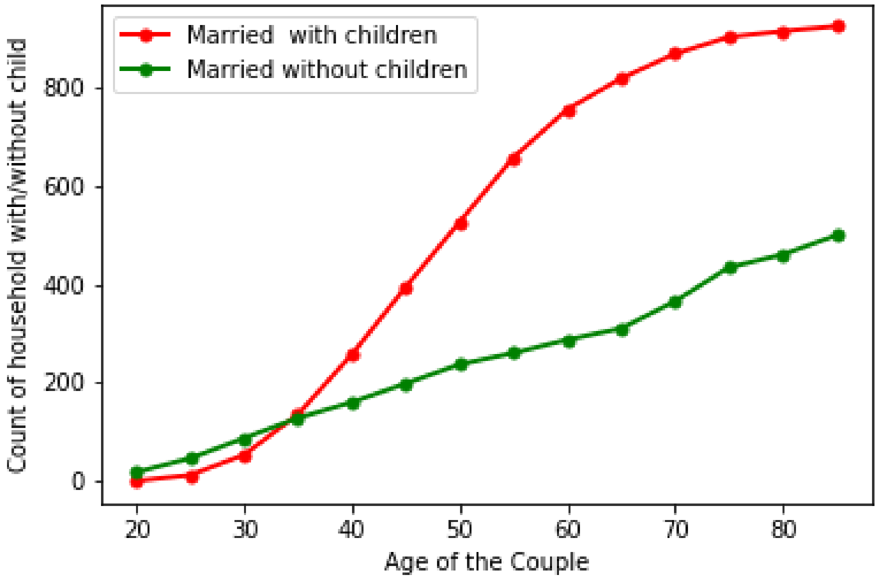
The growth of young children comes to a constant when the ages of married couple increase which is quite logical as older couple are less likely to have young children in the house.

Since most of the population in our census are around 20 to 50 age groups, we can predict several families deciding to have a child in the coming 10 years as the couple with growing ages till 50 are deciding to have children. For instance, the couple in their 40's or 30's who don't have a child in the house that is depicted by green line might move towards the red line in coming years and majority of those couple tend to have a child.

The following graphs below show the number of people in a household living with a married couple as their parents or grandparents. The children are of any age just staying with the family in a household.

The growth rate of children show that even older couples tend to have their children stay with them and this keeps on increasing.

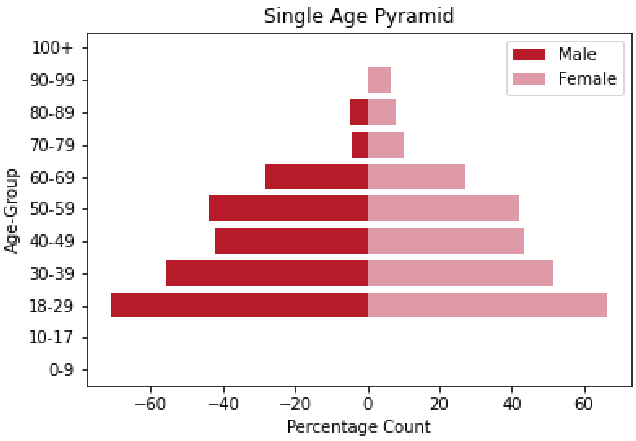


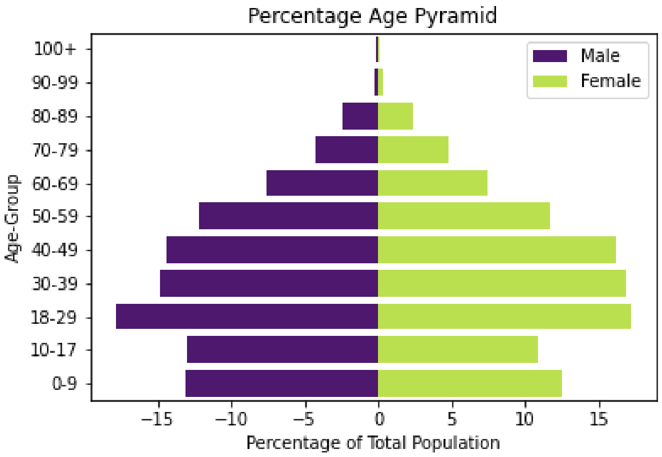


We can infer that, children in this town do not move out of houses to live independently or move out of town and tend to stay with their parents and the people moving out leave their kids with the grandparents this shows they keep visiting back to the town to visit their children or it saves them money to leave the kids with the grandparents while they are working in other cities.

Recommendations

Looking at the analyses of my data I think that there is a need for a Train Station to be build as the population in the town consist of various people who need better commuting services, and also potentially invest in a medical centre with maternity ward this might help the growth of the town as many people seem to travel for work and stay far from families., with train station built people likely to stay in town with majority of the population in 25 to 50 age group might help in increasing the birth-rate with better maternity facilities in place too.





The graphs above show the percentage of single people within that age range. For instance, 70% of population within the age range of 18-29 is single.

Around 12% of the total population are young children born within last 10 years so a better pregnancy ward might also help the future married couple as 13% of total population is in the age group of 18 to 29 and is single and around 6% precent of total population in that age group is married without children.

The Housing needs are not much of a priority in coming years since people tend to stay in smaller households and the growth of population currently is also not very significant.

The town will need to allocate more funding for end of life care since 30% of the total population lies within the age range of 40 to 60 which is potentially ageing population and might be retiring soon.

Other services might be needed in years in distant future once the town starts growing but the current needs of the town might be well served with the recommendations above.

References

1. Families and households in the UK:2020 https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2020