Fundamentals of Data Science Census Project Report

(771766_C23_T3A)

UNIVERSITY of HULL

Made By: Daniyal Kaleem

202411814

Introduction

In this report, we model the examination of a fictitious town's census dataset, which is structured like the UK census of 1881 but with recent modifications. Helping a local government team select priority investment regions based on demographic trends and the best use for an empty block of land are the objectives.

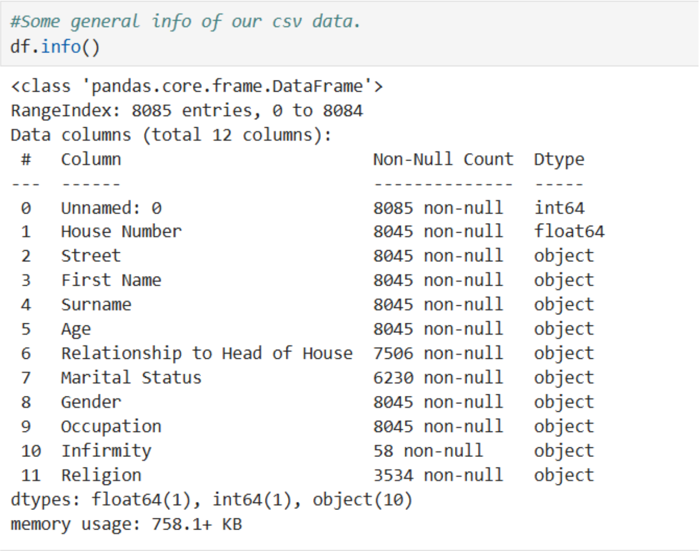
Numerous fields, including street number, street name, names of occupants, ages, relationships, marital statuses, genders, occupations, infirmities, and religions, are included in the data from the fake census. In order to ensure data integrity for further analysis, the dataset first needed to be thoroughly cleaned in order to address missing values and fix incorrect entries.

In order to address these two main questions, the dataset was graphically examined to answer the following questions:

1. What should be built on an unoccupied plot of land?
2. Which area should be prioritized for investment?

This report attempts to provide evidence-based suggestions to direct the local government's growth and investment decisions, ultimately improving the town's infrastructure and citizens' quality of life. It does this by carefully cleaning and analyzing the data from our census.

Chapter 1: Data Cleaning



```
#Some general info of our csv data.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8085 entries, 0 to 8084
Data columns (total 12 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Unnamed: 0                    8085 non-null   int64  
 1   House Number                  8045 non-null   float64
 2   Street                        8045 non-null   object 
 3   First Name                    8045 non-null   object 
 4   Surname                       8045 non-null   object 
 5   Age                           8045 non-null   object 
 6   Relationship to Head of House 7506 non-null   object 
 7   Marital Status                6230 non-null   object 
 8   Gender                        8045 non-null   object 
 9   Occupation                    8045 non-null   object 
 10  Infirmity                     58 non-null     object 
 11  Religion                      3534 non-null   object 
dtypes: float64(1), int64(1), object(10)
memory usage: 758.1+ KB
```

Figure 1: General Information of our census data

1.1 Cleaning House Number

To address the missing house numbers, I created a dictionary using the combination of street name and surname as keys. This approach assumed that individuals with the same surname and living on the same street likely belong to the same household. Missing house numbers were filled using this mapping. For entries without relevant surname or house number matches, the previous house number was incremented by 1 since house numbers are in ascending order within each street in our census data.

1.2 Cleaning Street Names

For missing street names, if the house number of the current entry matched or was one less than the next entry, it was assumed to be the same street. Otherwise, it was assumed to be the next street in the dataset.

1.3 Cleaning First Names

To fill missing first names, I grouped first names by common surnames and calculated the mode (most frequently occurring first name) for each surname group. That mode was then used to fill in the missing first names for that given surname.

1.4 Cleaning Surnames

For missing surnames, a mapping of street and house number was created, similar to the approach used for house numbers. If no relevant matches were found, mode surname for the whole dataset was assigned.

1.5 Cleaning Ages

To address issues with ages, I first converted ages written in words (e.g., "eight," "nine") to numbers. All age values were then converted from strings to numbers. Missing age values were filled by calculating three modes: one for individuals known to be over 18 (e.g., heads of households, husbands, wives), one for those under 18 (e.g., students), and a general mode for all other cases. I also capped all my ages at 100, as it was statistically unlikely to surpass that threshold.

1.6 Cleaning Relationship to Head of House

For missing relationship values, individuals under 18 were assigned "Son" or "Daughter" based on their gender. Individuals over 18 were assigned "Head" due to the lack of a clear method to differentiate between plethora of unique relationship values.

1.7 Cleaning Marital Status

For missing marital statuses, entries for individuals under 18 were set to "Single" as it is legally not allowed to marry under this age. The surname, house number, and street were used to infer marital status by matching similar entries. If no match was found, the most repeated marital status was used.

1.8 Cleaning Gender

For inconsistent gender entries, all variations of male (e.g., "m," "M") were standardized to "Male" and similarly for female to "Female." A mapping dictionary was created using the relationship to the head of house to assign genders to missing values. For entries still missing gender, a random generator assigned either "Male" or "Female."

1.9 Cleaning Infirmity

For infirmity values, null values were replaced with "Healthy." Blank entries were changed to "Undeclared."

1.10 Cleaning Religion

For religion entries, those marked "None" were changed to "Atheist." Null entries for individuals under 18 were set to "Undeclared" as it is inappropriate to ask minors about their religion. Entries marked as "Private" were changed to "Undeclared."

1.11 Cleaning Occupation

For missing occupation values, a Random Forest classifier model was used, utilizing 'Age,' 'Marital Status,' 'Gender,' 'Infirmity,' and 'Religion' as features. A label encoder converted categorical data into numerical form for the model.

Chapter 2: Data Analysis

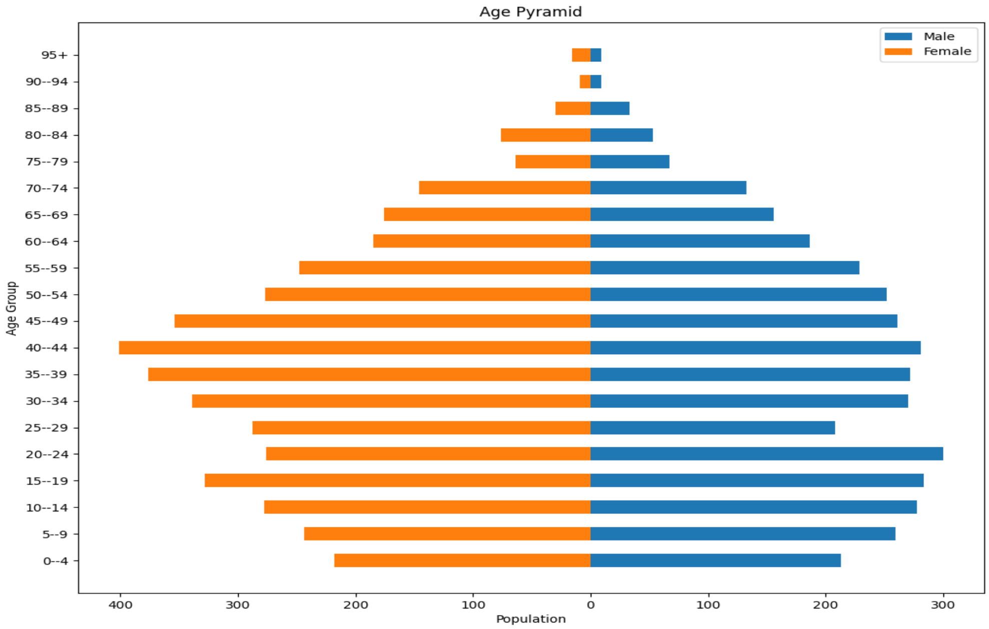


Figure 2: The population pyramid of the town's census data.

The number of individuals between the ages of 40 and 59 has a discernible bulge, suggesting a sizable percentage of middle-aged people. The pyramid also exhibits a sizable proportion of people between the ages of 20 and 29, which represents a young adult population, which may include young professionals or college students who travel to neighboring cities for employment or study.

On the other hand, there are far fewer people living at the extremities of the age range. There are very few kids in the 0–4 and 5–9 age groups, which may be a sign of a decline in birth rates or younger families leaving town. There are also very few old people, especially those who are 85 years of age or older.

2.1 Birth Rate and Death Rate:

The calculated birth rate for the town is 9.276 per 1,000 people, indicating a relatively moderate level of new births within the community. This birth rate, when compared to the national average, suggests a stable but not rapidly growing young population. In contrast, the latest available data places the UK's crude death rate at approximately 9.4 per 1,000 people (World Bank, 2024).

2.2 Estimated Population Growth rate:

In order to calculate the population growth rate, demographic data is usually compared across time to ascertain changes in the population's size. But for the purposes of this report, we compared the dataset's mean age to the UK Census 2021 mean age (Office for National Statistics, 2024) in order to deduce possible patterns in population increase. We determined if the mean age of our sample differed statistically from the mean age of the entire population by doing a single-sample t-test. In our dataset, a lower average age could be a sign of population expansion due to a higher birth rate or younger migratory patterns. On the other hand, a population that is older on average may indicate slower growth.

Using this logic, I ran a simple t-test Hypothesis testing on our census data and found the population to be young, indicating high birth rate and suggesting a growing population.

2.3 Religious Convictions:

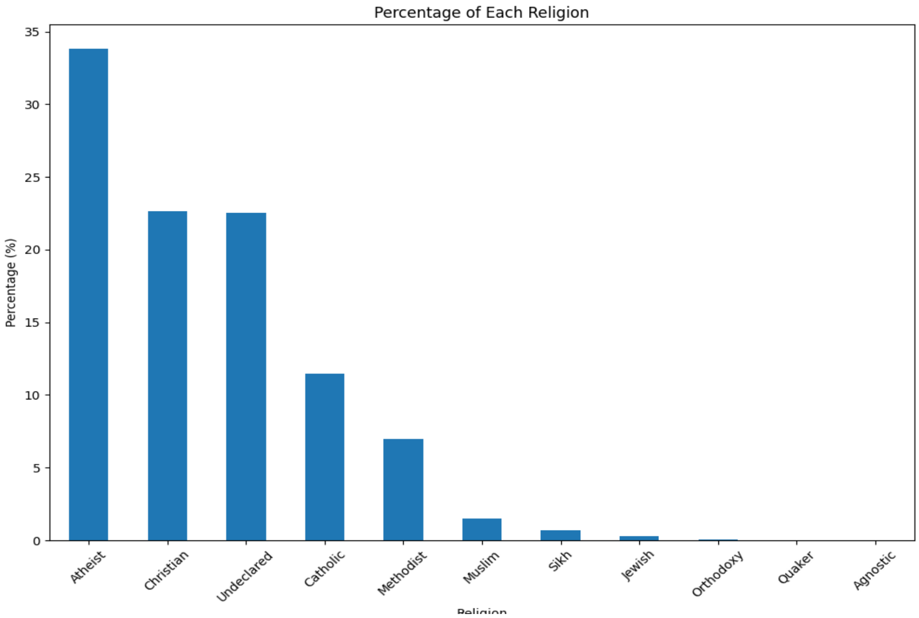


Figure 3: Religious conviction of the populace

From figure 3 we can conclude that Atheists constitute the largest group, comprising approximately 34% of the population. This is followed by Christians and those with undeclared religious affiliations, both making up around 21% and 20% respectively. Catholics represent about 12% of the population, while Methodists account for approximately 7%. Other religious groups, including Muslims, Sikhs, Jews, Orthodox Christians, Quakers, and Agnostics, each constitute a smaller fraction, collectively summing up to around 6%.

2.4 Commuters Professions:

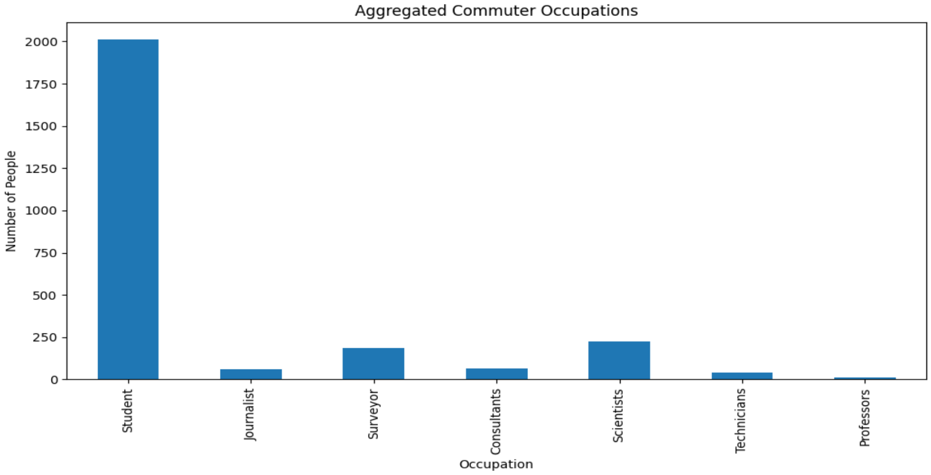


Figure 4: Professions with most commuters

The most prominent group by far is students, with around 2,000 individuals, significantly outnumbering other professions. Scientists follow as the second-largest group, though their numbers are considerably smaller. Surveyors, consultants, journalists, technicians, and professors all have relatively low representation, with each category having fewer than 250 commuters.

2.5 Marital Status

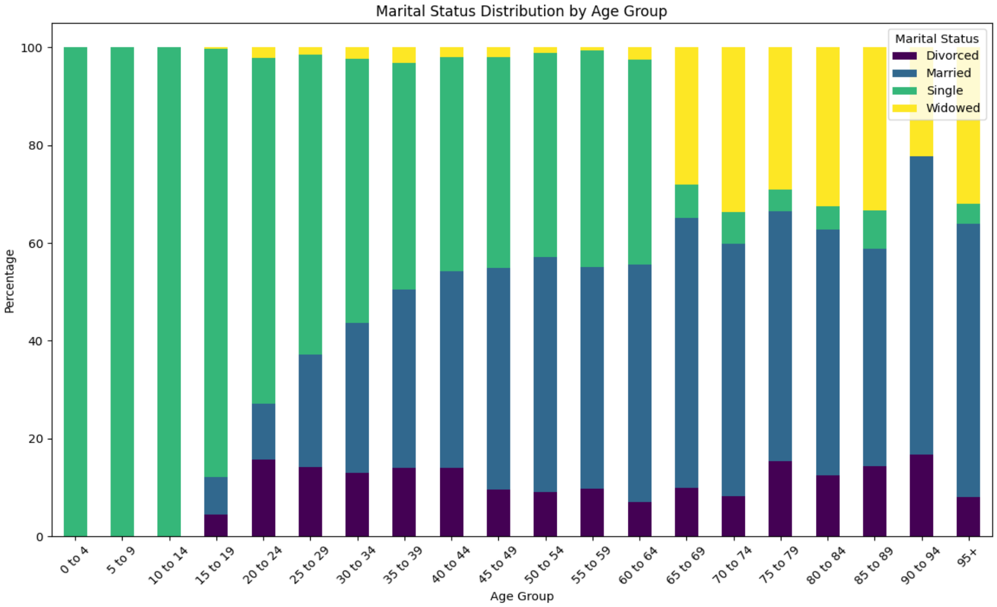


Figure 5: Marital Status of the populace

According to the graphic, the vast majority of people in the earliest age group 0 to 19 are single, which is consistent with predictions for this age range. The percentage of married people rises when people enter early adulthood (20 to 34 years old), peaking between 30 and 34 years old, which is when most people are married. In this age bracket, there is also a minor but discernible proportion of divorced people. As we move into the middle age range (35 to 64 years), marriage continues to be the most prevalent marital status, but the percentage of divorced people rises, indicating that divorces are becoming more frequent at this time.

For older adults (65 and above), the chart shows a gradual decline in the proportion of married individuals and a significant rise in the percentage of widowed individuals, especially in the 75 and above age groups.

2.6 Occupancy Distribution

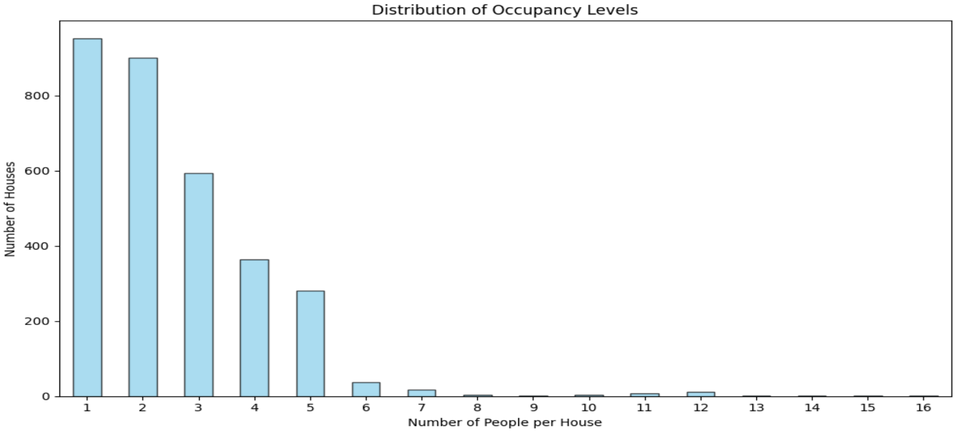


Figure 6: Distribution of Occupancy levels

The distribution of the data is right-skewed, with most households having one to three members. As the number of people living in a house rises, the number of residences drops off quickly, with very few families housing more than ten people. The average housing size of UK is 2.36 people per house (Office for National Statistics, 2024). This implies that the majority of the residents in the town are either single people or have small family groupings.

2.7 Affluent Occupations

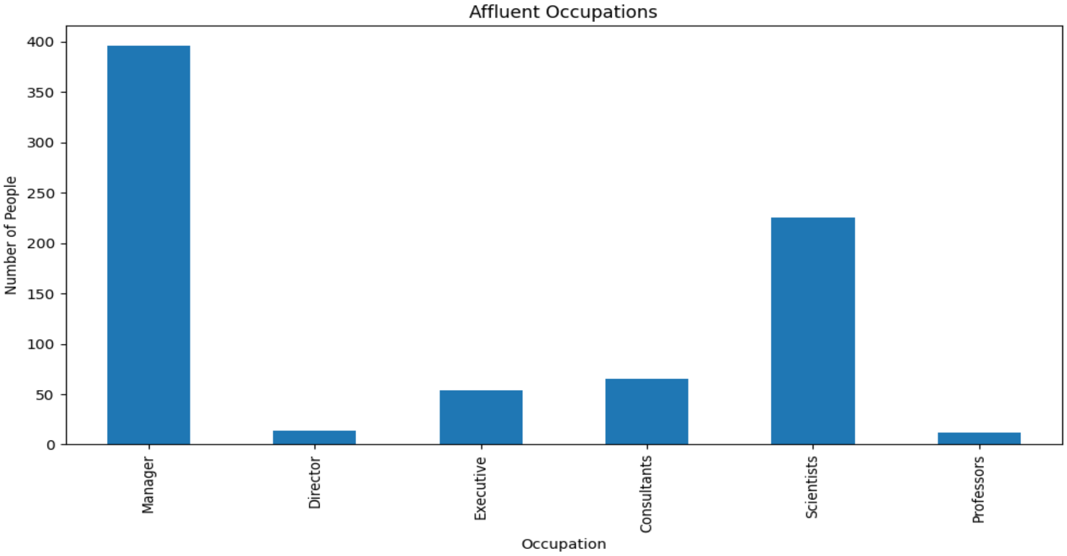


Figure 7: Affluent Occupations

According to the figure, wealthier people are more likely to have management and scientific positions, with management professions being particularly prevalent. This shows that leadership and technical expertise in science are important contributors to wealth. Given our town's population of little over 8000, this graph shows that a sizable proportion of its residents have rich professions.

2.8 Prevalence of Unemployment

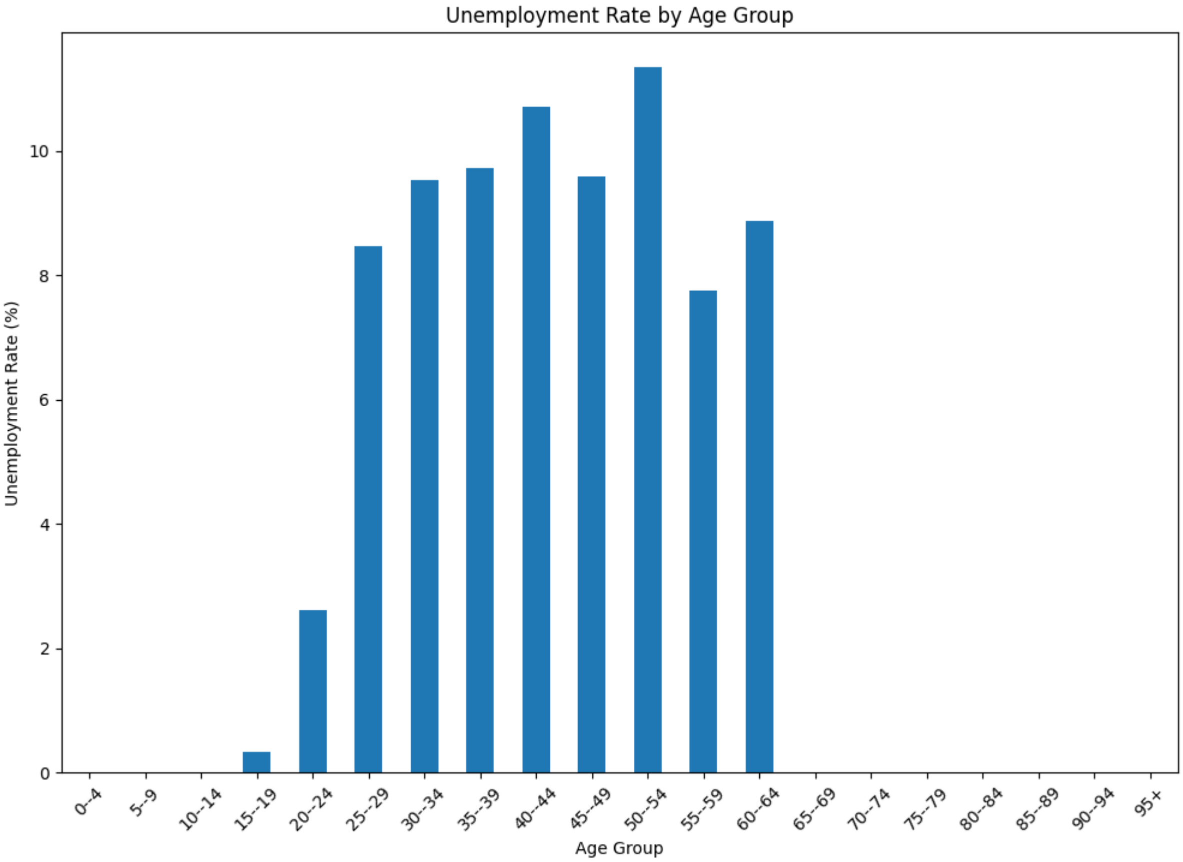


Figure 8: Prevalence of Unemployment in town

This graph highlights that the highest unemployment rates are found among individuals in their prime working years, particularly those aged 35-49. A significant increase in unemployment is observed starting from the 20-24 age group, reaching around 3%.

The highest unemployment rates are recorded in this segment, with the 40-44 and 45-49 age groups experiencing rates slightly above 10%. There is a noticeable decline in the unemployment rate starting from the 55-59 age group, dropping to around 8%. For the oldest age groups, 85-89, 90-94, and 95+, the unemployment rate is very low, likely reflecting retirement.

2.9 School aged and Elderly Population Percentage

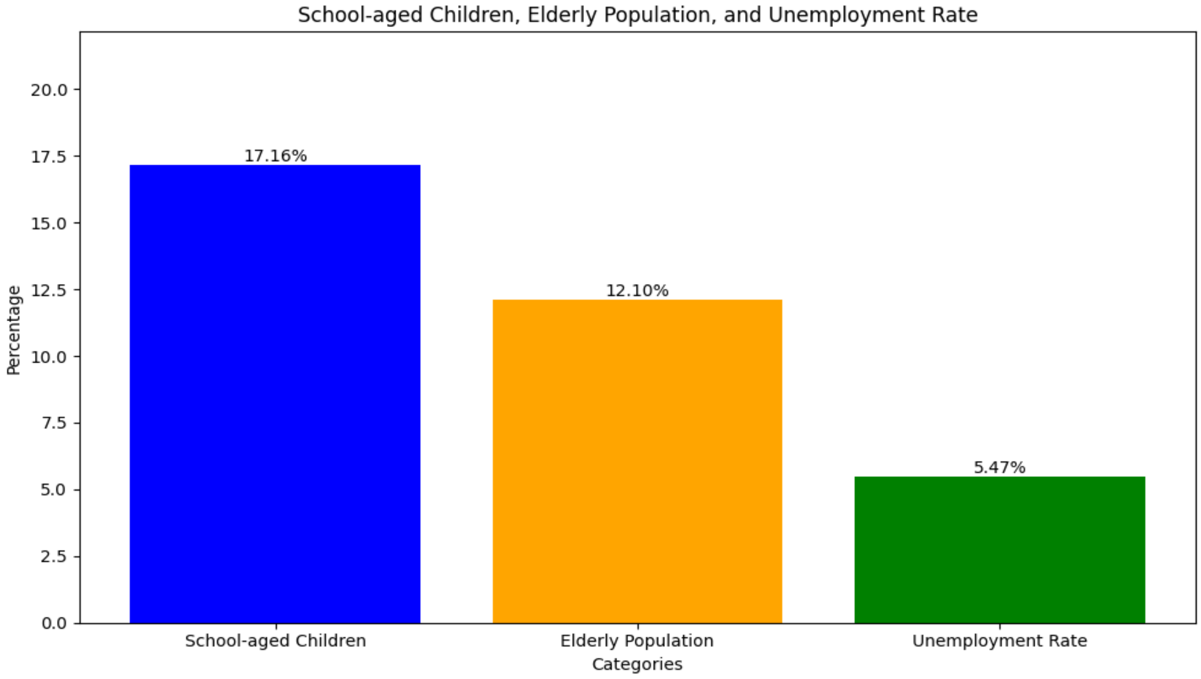


Figure 9: School aged vs elderly population

The graph above helps to put into perspective the unemployment rate of the town. The unemployment rate stands at 5.47%, which is lower than the percentage of both school-aged children and the elderly population. This could suggest that the working-age population is actively engaged in the workforce.

Chapter 4: Recommendations and Insights

Using the census data for the town and the data analysis in Chapter 3 we can justify some recommendations and enlist some insights on what processes need to be undertaken and what infrastructure needs to be deployed to make the best use out of the resources for this town. The table below provides the points to consider in the decision-making process.

Table 1: Criteria for initializing Investment and plot of land

Category	Points to Consider	Decision
Plot of Land	If the population is significantly expanding.	High Density Housing
Plot of Land	If the population is "affluent" and there is demand for large family housing.	Low Density Housing
Plot of Land	There are potentially a lot of commuters in the town and building a train station could take pressure off the roads.	Train Station
Plot of Land	Is there demand for a second Church (if so, which denomination?), or for a different religious building?	Religious Building
Plot of Land	Emergency medical building. Not a full hospital, but a minor injuries center.	Emergency Medical Building
Investment	If there is evidence for a significant amount of unemployment.	Employment and Training
Investment	If there is evidence that there will be an increased number of retired people in future years, the town will need to allocate more funding for end-of-life care.	Old age care
Investment	If there is evidence of a growing population of school-aged children (new births, or families moving into the town), then the schooling spend should increase.	Increase in Spending for school
Investment	If the town is expanding, then services (waste collection; road maintenance, etc.) will require more investment.	General Infrastructure

4.1 What should be built on an unoccupied plot of land that the local government wishes to develop?

i. High-Density Housing

The age pyramid shows a balanced distribution between males and females across various age groups, with a relatively large proportion of the population in the 20-44 age range, indicating a substantial working-age population. Additionally, there are significant numbers of individuals in younger age groups (0-19), suggesting a potential for future population growth, but the birth rate for the town shows that there is no specific evidence in the provided data indicating significant population growth. Furthermore, I used a p-hypothesis testing (p-value of 0.72) to prove that there is no evidence of town's population expanding. Therefore, high-density housing may not be the priority.

ii. Low-Density Housing

The figure 7 suggests a considerable number of affluent occupations, such as managers and scientists, indicating an affluent population with a potential demand for larger family homes as compared with UK’s average affluent household percentage of 10% (Office for National Statistics, 2024). Thus, building low-density housing would cater to this demographic.

iii. Train Station

The data shows a significant number of students and commuters, indicating that a large segment of the population travels regularly. Also, all the University and PhD students already travel to the two cities adjacent to the town thus constructing a train station could alleviate pressure on roads and provide efficient transportation.

iv. Religious Building

Given that there is already one place of worship for Catholics and the percentage of Catholics is relatively low compared to other groups, there is no significant demand for a second Catholic Church. Alternatively, a non-religious community center could cater to the large Atheist and Undeclared populations.

v. Emergency Medical Building

While the census data does not provide explicit data on injuries or pregnancies, it mentions various other infirmities. Considering that only 0.25% of the towns populace was either physically or mentally disabled, establishing a minor injuries center would not be a prudent way of utilizing the plot of land.

(b) Which one of the following options should be invested in?

i. Employment and Training

The data mentions the presence of various professional occupations but does not highlight a significant amount of unemployment. The unemployment rate was 5.47% for the town which was less than the number of retired old pupil in the town therefore, while employment and training are important, they may not be the highest priority based on the current data.

ii. Old Age Care

The census data does not provide specific evidence of an increasing number of retired people. Also, using Hypothesis testing (p-value of 1.0) we can conclusively prove that there is no significant difference of elderly people to that of General population. As such we can state that there is no need for an investment in old age care.

iii. Increase Spending for Schooling

There is a substantial number of students, indicating a growing population of school-aged children. Increasing spending on schooling would support the educational needs of these children, ensuring that the town's educational infrastructure can accommodate and provide quality education to a growing student population.

iv. General Infrastructure

The calculated birth rate for the town is 9.276 per 1,000 people, indicating a relatively moderate level of new births within the community. This birth rate, when compared to the national average, suggests a stable but not rapidly growing young population. Therefore, we can state that there is no need for additional spending to improve waste collection and road maintenance

Chapter 5: Conclusion

This report presented a detailed analysis of a mock census dataset for an imaginary town, aimed at assisting local government decision-making regarding the use of an unoccupied plot of land and priority areas for investment. The analysis was carried out in two major phases: data cleaning and data analysis, each providing crucial insights into the town's demographic trends and needs.

In conclusion, this comprehensive analysis provides evidence-based recommendations to guide the local government’s development and investment strategies. By addressing the identified needs, the town can enhance its infrastructure and services, ultimately improving the quality of life for its residents.

Chapter 6: Bibliography

Office for National Statistics. (2024, August 6). Retrieved from Office for National Statistics: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householdincomeinequalityfinancial/financialyearending2022

Office for National Statistics. (2024, August 6). Retrieved from Office for National Statistics website: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2022#:~:text=Households-,There%20were%20an%20estimated%2028.2%20million%20households%20in%20the%20UK,both%202012%20and%20in%202022.

Office for National Statistics. (2024, August 6). Retrieved from Office for National Statistics website: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicgroupbyageandsexenglandandwales/census2021#:~:text=Between%202011%20and%202021%2C%20the,39%20years%20to%2040%20years.

World Bank. (2024, August 6). Retrieved from World Bank Website: https://data.worldbank.org/indicator/SP.DYN.CDRT.IN?locations=GB