TABLE OF CONTENT

1. INTRODUCTION 4

1.1 Background Information 4

2. DATA CLEANING METHADOLOGY 4

3. STATISTICAL ANALYSIS 8

3.1 Population Pyramid 8

3.2 Employment Trend 9

3.3 Religious Affiliation 10

3.4 Marital Status 11

..... 12

3.5 Occupancy Level & High-Density Housing Streets 13

3.6 Commuters 14

3.7 Affluent Population 15

..... 15

3.8 Infirmity 15

3.9 Age Distribution 16

4. DISCUSION 17

4.1 What should be built on an unoccupied plot of land that the local government wishes to develop? 17

4.1.1 High-density housing due to significant population expansion 17

4.1.2 Low-density housing demand for affluent population 17

4.1.3 Train Station for Commuters 17

4.1.4 Need for Religious Building 18

4.1.5 Need for Emergency Medical Building 18

4.2 Which one of the following options should be invested in? 19

4.2.1 Employment and training 19

4.2.2 Old age care 19
4.2.3 Increase spending for schooling 19
4.2.4 General Infrastructure 19
5. DECISION 20
5.1 Decision for Task A 20
5.2 Decision for Task B 20
6. CONCLUSION 20
7. REFERENCES 21

1. INTRODUCTION

1.1 Background Information

For this project, an example of a citizens' census of an imaginary town is included. Census data cleaning and analysis are required in this task to make judgement on how an unused plot of the land can be utilized as well as assessing possible areas of investment. The work will present findings to aid the decision-making process with regards to future strategies of the town.

2. DATA CLEANING METHADOLOGY

As stated earlier, dataset consist of 11 columns with all columns consist of some irregularities. Each column is analyzed in-depth and different approaches were used to handle errors. Potential errors found in data are:

- Presence of NaN values
- Presence of blank entries
- Conflict of data types in single column
- Presence of logically not possible entries
- Inconsistent representation

Approaches implemented to handle each column is described below:

1. **House Numbers:**

There were NaN entries in this column which was handled by looking entries with similar street and surname to NaN entry and replacing NaN with their house number as having same surname and street indicate that the entries belong to same family. If match of similar street and surname is not found then assign the next house number of previous entry as all entries having same street are listed together in dataset.

2. **Street:**

This column included NaN entries; errors were handled in the same way as those in House Number. NaN entries were replaced with street names for those with the same surname and home number. If no matches are found, use the street name from the previous item.

3. **First Name:**

Because there is no way to accurately predict someone's first name, NaN entries in this field were handled using mode of data.

4. **Surname:**

NaN elements in this column were fixed using the same manner as those in House Number and Street. NaN entries were replaced with those with the same house number and street because they could be from the same family and so have the same surname. If no matches are found, use the surname mode, as there is no way to guess someone's surname, just like their first name.

5. **Age:**

First, all categorical entries in the data were converted into integer values to guarantee that all entries were numerical. One record with an obviously erroneous age was addressed by evaluating other information in the record, such as the individual's relationship with the head of the house and occupation. In addition, all age entries that were recorded as floats were recoded to integers to avoid data type incompatibilities. Finally, the data comprised NaN items were corrected by using median age.

6. Relationship to Head of House:

In cases of NaN entries in the Relationship to Head of House column, where the gender was male and the age was less than 18 years, these entries were labeled “Son”. Where the gender was female and age less than 18 years, entries were labeled “Daughter”. Where the age was more than 18 yr where gender was not specified, entries were labeled “Head”

7. Marital Status:

Regarding NaN entries in the marital-status column, the following two primary decisions were made. The entries were defined as “Married” if the relation to head of the household was “Husband” or “Wife. ” Remaining NaN entries were classified as “Single.” It is still more significant to notice some entries with column “Married” entered to the people under 18 years old, which is legally Impossible. To these, the marital status was changed to “Single.”

8. Gender:

First, the gender data entries that incorporated diverse labels were normalized; for instance, “Male,” “Female,”,” male”, “female”, “F,” or “m” were categorized as either “Male” or “Female.” To guess the gender, the relationship of the individuals to the head of the household was used if the entry was left blank. In case of NaN values, a mapping dictionary was formed to assign gender based on the relation to the head of the household. For remaining NaN entries their gender can be predicted by looking at their first names, this method is not efficient if there are large number of NaN entries left so rest NaN entries were labelled as “Undeclared”.

9. Occupation:

Individuals of 0-4 were defined as ‘Child’ similarly, individuals of 5-18 were defined as ‘Student’ likewise, people of more than 68 years were classified as ‘Retired.’ For remaining NaN entries predicting the occupation is a tricky affair. Instead of employing complex methods such as a Random Forest prediction model, NaN entries were filled randomly.

10 Infirmity:

All instances of ‘NaN’ or blanks in the ‘Infirmity’ column were changed to ‘None,’ meaning no infirmity on this record.

11 Religion:

Entries with NaN and Nope conveyed the essence of no religion, hence they were substituted with atheist. There were some absurd entries that were replaced with the most popular religion.

3. STATISTICAL ANALYSIS

3.1 Population Pyramid

Representing the youngest age groups (0-10), the base of the pyramid is noticeably narrower. This suggests that the population is either growing more slowly or is shrinking because of the low birth rate. The pyramid's widest region, which forms a prominent bulge in the intermediate age ranges. This means that a sizable section of the populace is of working age. A linear drop in population size with age is evident in the upper part of the pyramid. As a result of natural mortality, most populations experience a steady decrease in population size as people age. With more women than men in the elder age groups, there is a clear gender disparity as people get older. This is a common demographic trend, reflecting higher life expectancy for women compared to men.

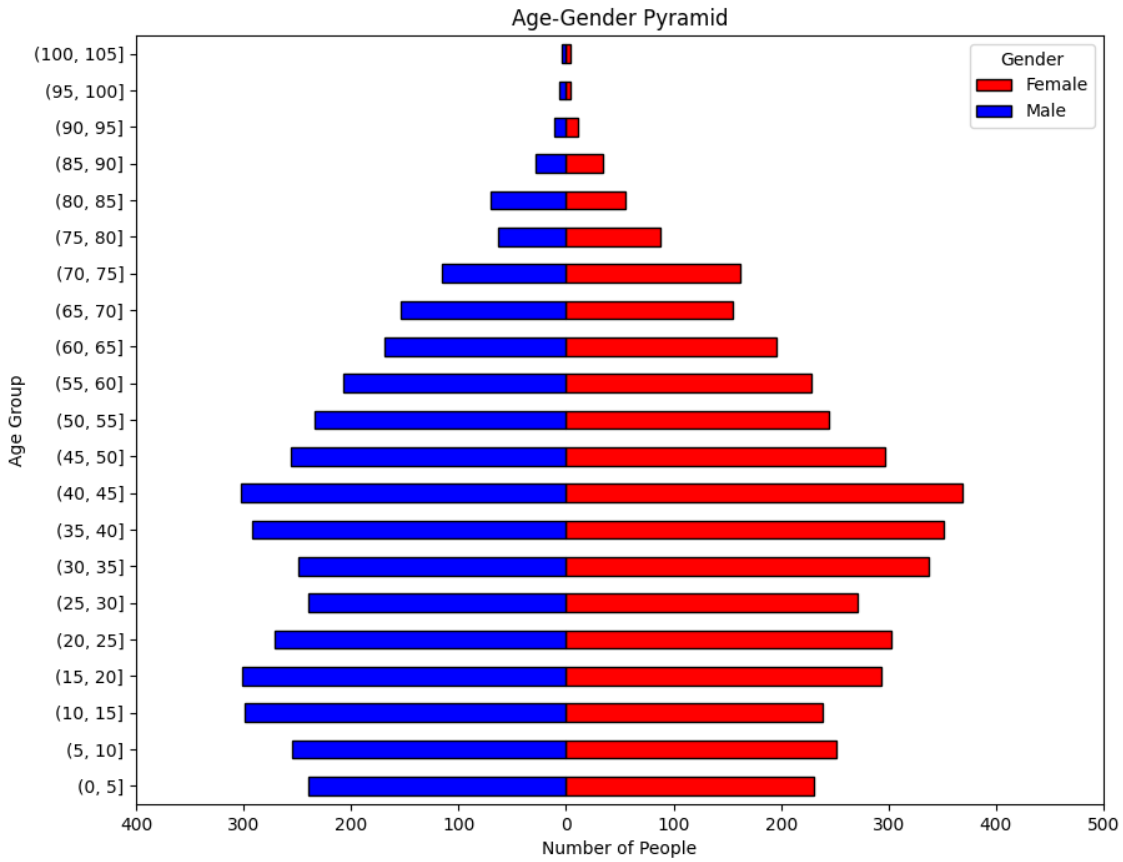

Figure 1 Age Pyramid Graph

3.2 Employment Trend

"Employed" people make up slightly over half of the population (52.6%), which is the largest sector of the population. Strong labor force participation is a prerequisite for both economic growth and stability, and this is indicated by this. The fact that only 6.10% of people are unemployed suggests that unemployment is low.

The unemployment in age group shows that young adults between the ages of 26 and 45 account for the largest proportion of jobless people, with 130 people (1.6%) and 124 people (1.5%) in the 26–35 age group and the 36–45 age group, respectively. Of the elder generation, 46–55, 117 people (1.5%) are unemployed, suggesting that there may be problems with mismatched skills or changes in the economy that impact conventional industries.

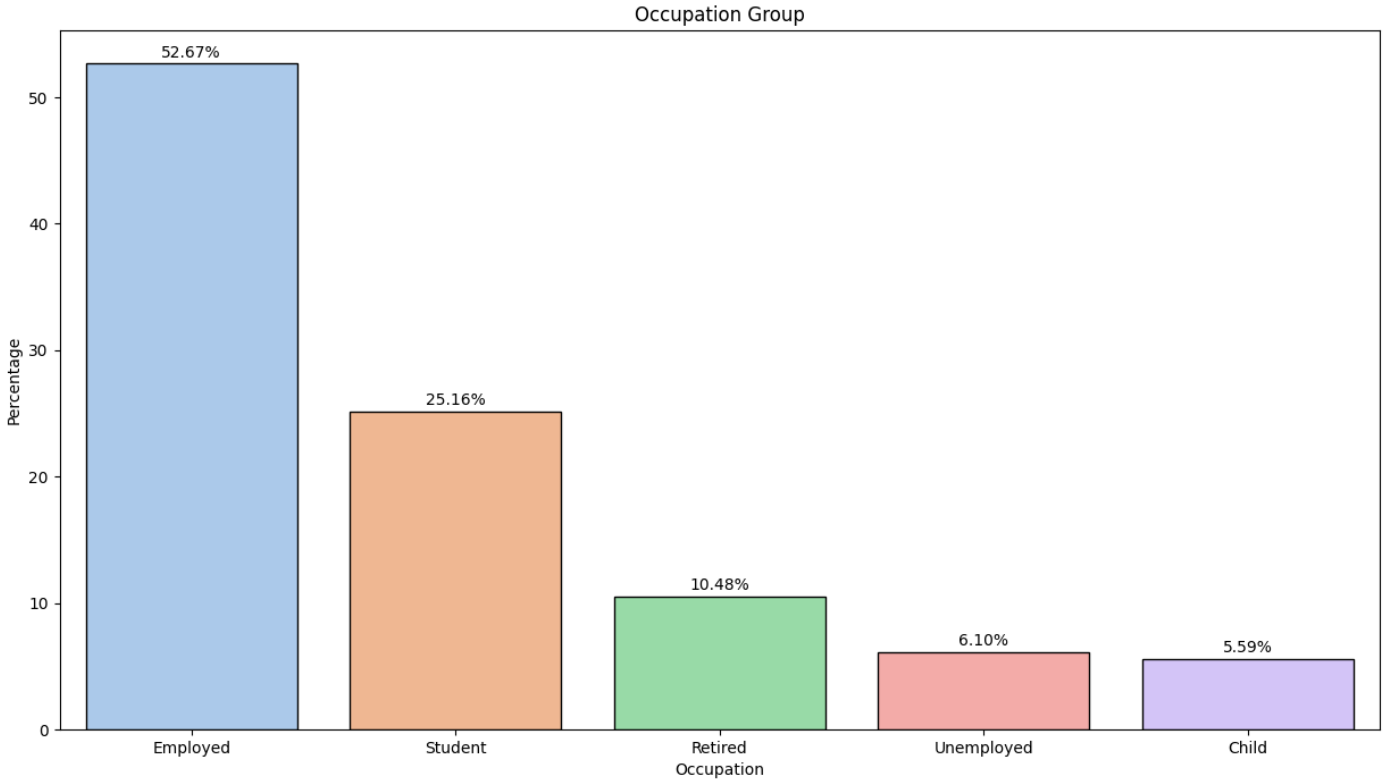

Figure 2 Occupation Group

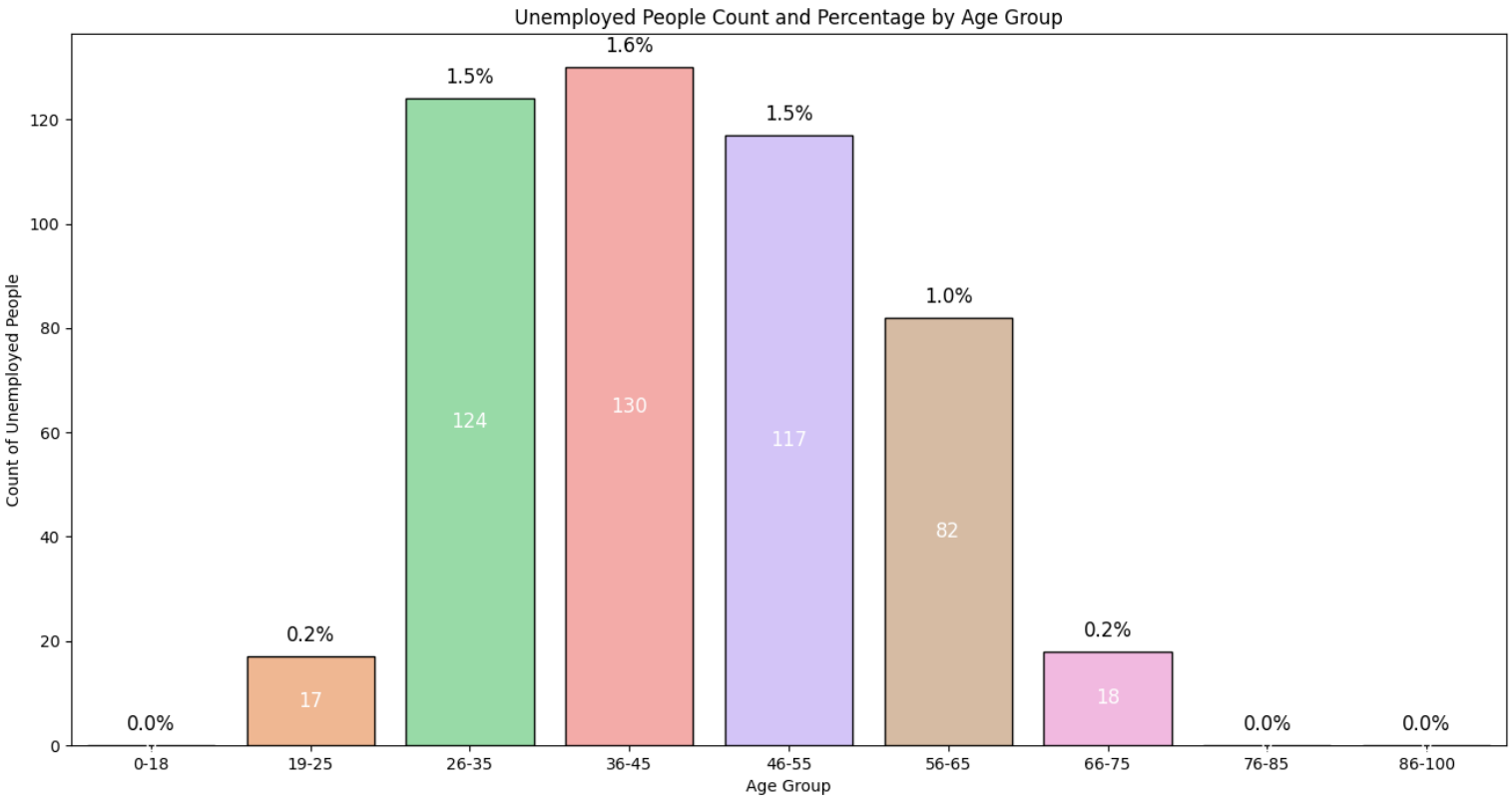

Figure 3Unemployed based on Age group

3.3 Religious Affiliation

The graph reveals that atheism is the predominant belief system in the community, accounting for 38.7% of the population. Following atheism, Christianity comprises 32.6% of the population, with Catholicism at 15.6% and Methodism at 9.7%. Given that atheists do not require places of worship, the focus should shift towards addressing the needs of the Christian population.

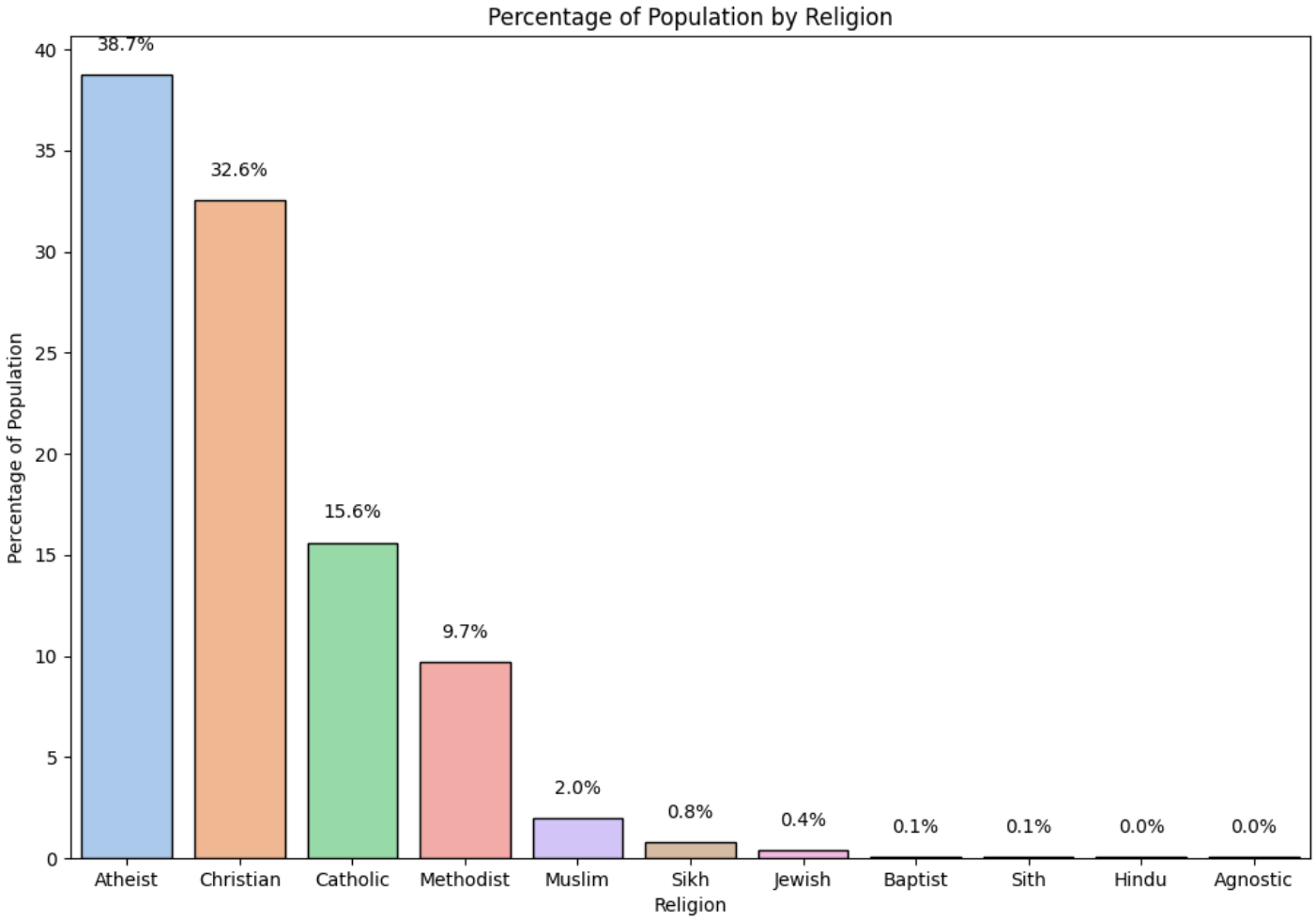

Figure 4 Religion Distribution

3.4 Marital Status

The marital status graph reveals that a substantial majority of the population, 59.2%, is single, while only 25.9% is married. Further analysis of the single population by age group shows that approximately 15% of individuals aged **between** 18-34 are single. This age group is particularly significant as they are likely to marry in the near future, which could influence and stabilize the birth rate.

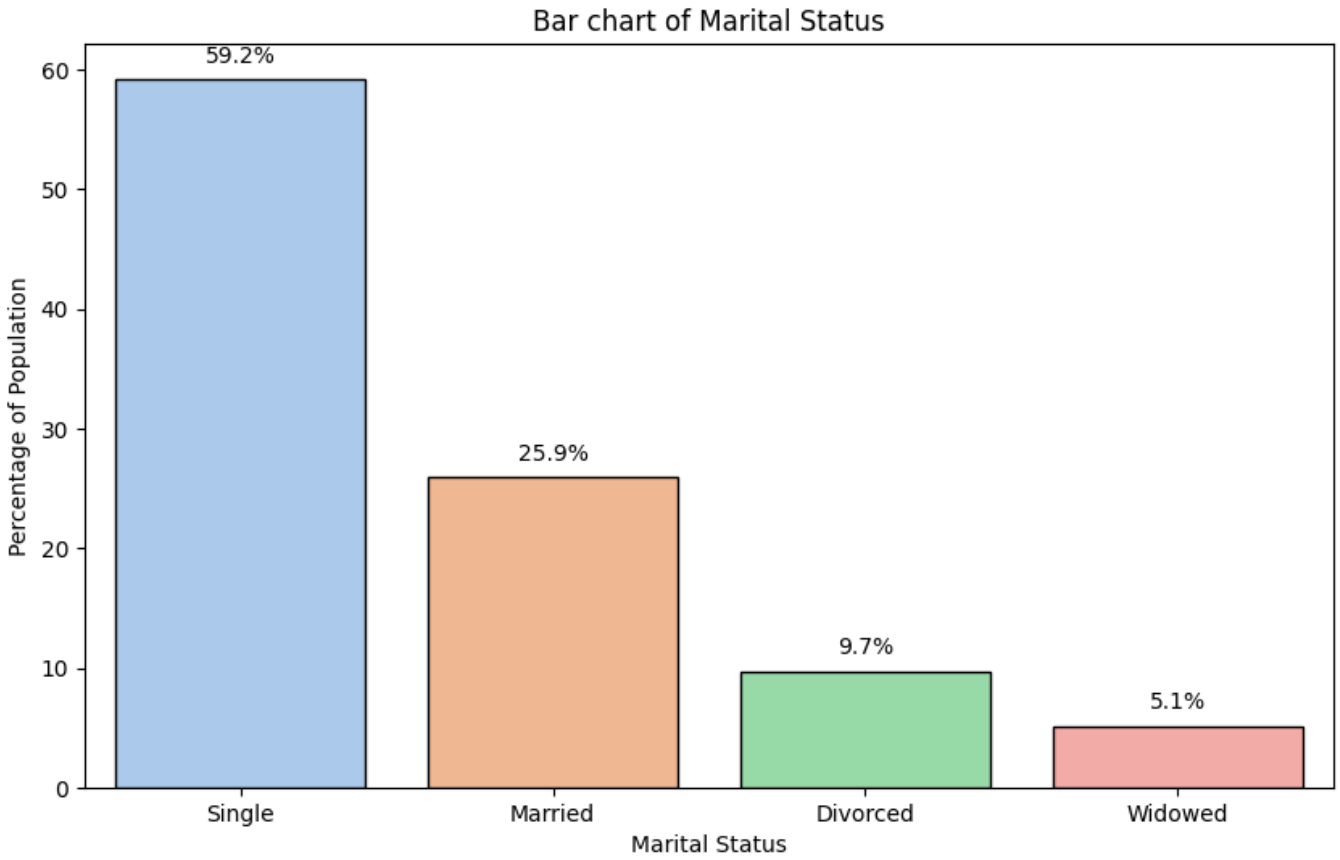

Figure 5 Marital Status Distribution

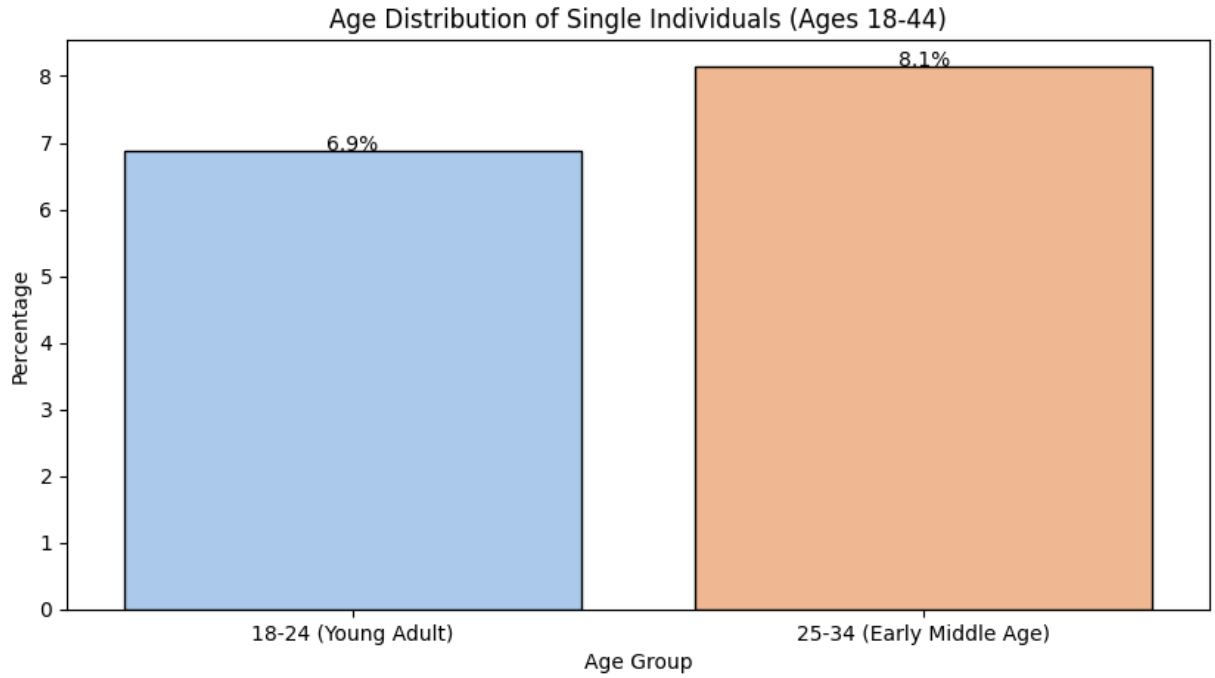

Figure 6 Singles in Aged between 18-44

3.5 Occupancy Level & High-Density Housing Streets

The average household size for the 2022–2023 period was 2.2 people (GOV.UK, 2023). According to an analysis of the occupancy level graph, 1,811 out of 3,084 houses, or 58.7% of all houses, have two or fewer people. This implies that a substantial number of houses are underutilized. In addition, Figure 8 shows the top five streets in terms of housing density. These streets have significance as locations where spending on general infrastructure could be especially advantageous because of their high concentrations of housing.

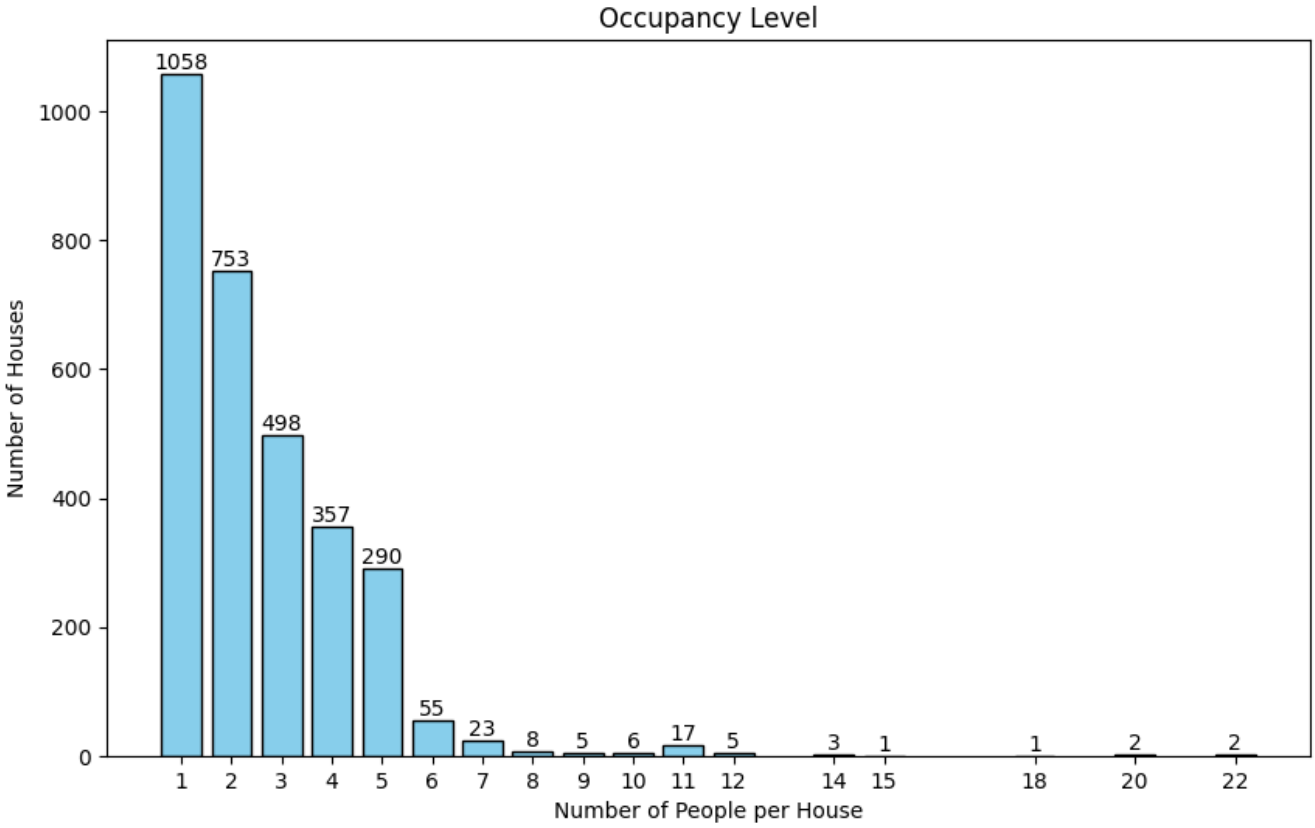

Figure 7 Count Occupancy Level

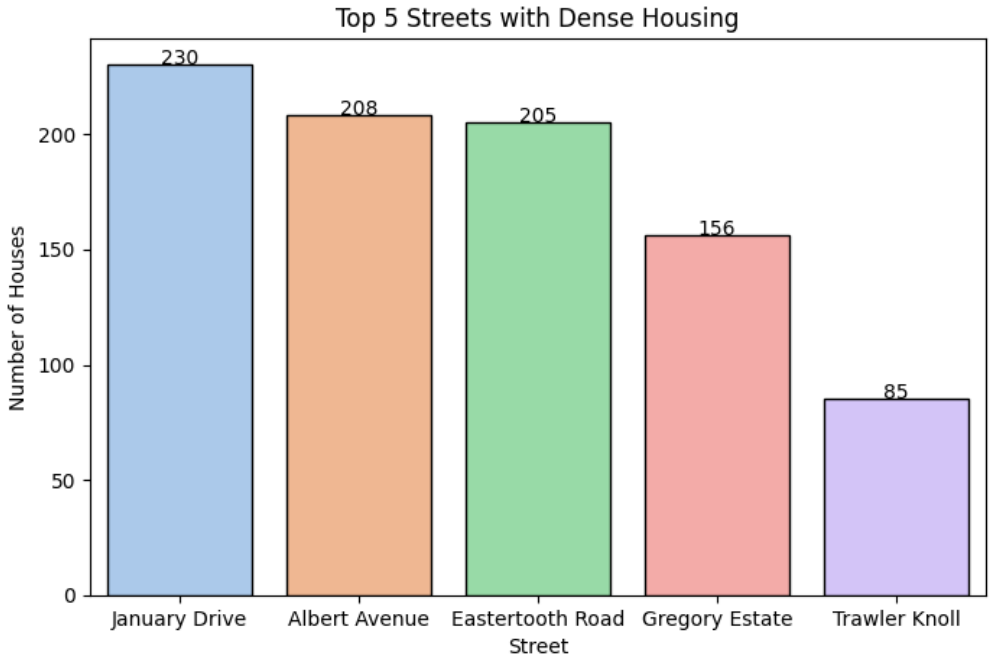

Figure 8 Top 5 Dense housing streets

3.6 Commuters

According to the analysis, commuters make up 33.7% of the population. Notably, students make up 25.2% of all commuters, showing a sizable group of frequent commute users. In the next group of commuters, engineers comprise 4.8% and surveyors, 1.7%, respectively.

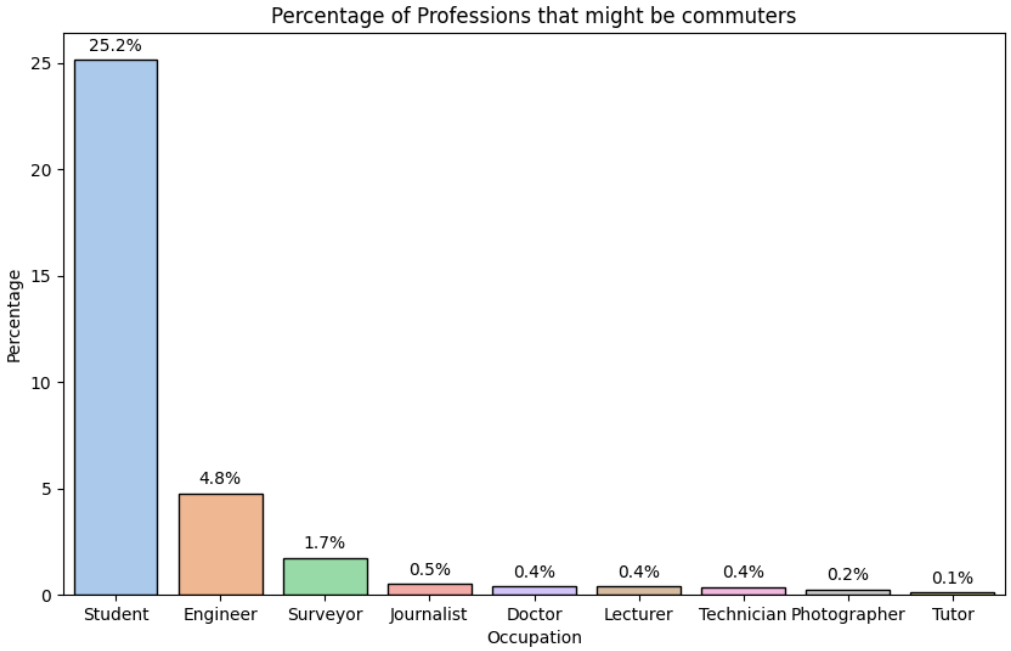

Figure 9 Potential Commuters Percentage

3.7 Affluent Population

The analysis of the affluent population, evaluated from occupation column, indicates that 16.47% of the population falls into the affluent category. Among these, engineers represent the largest segment at 4.8%, followed by managers at 4.6%, and scientists at 2%.

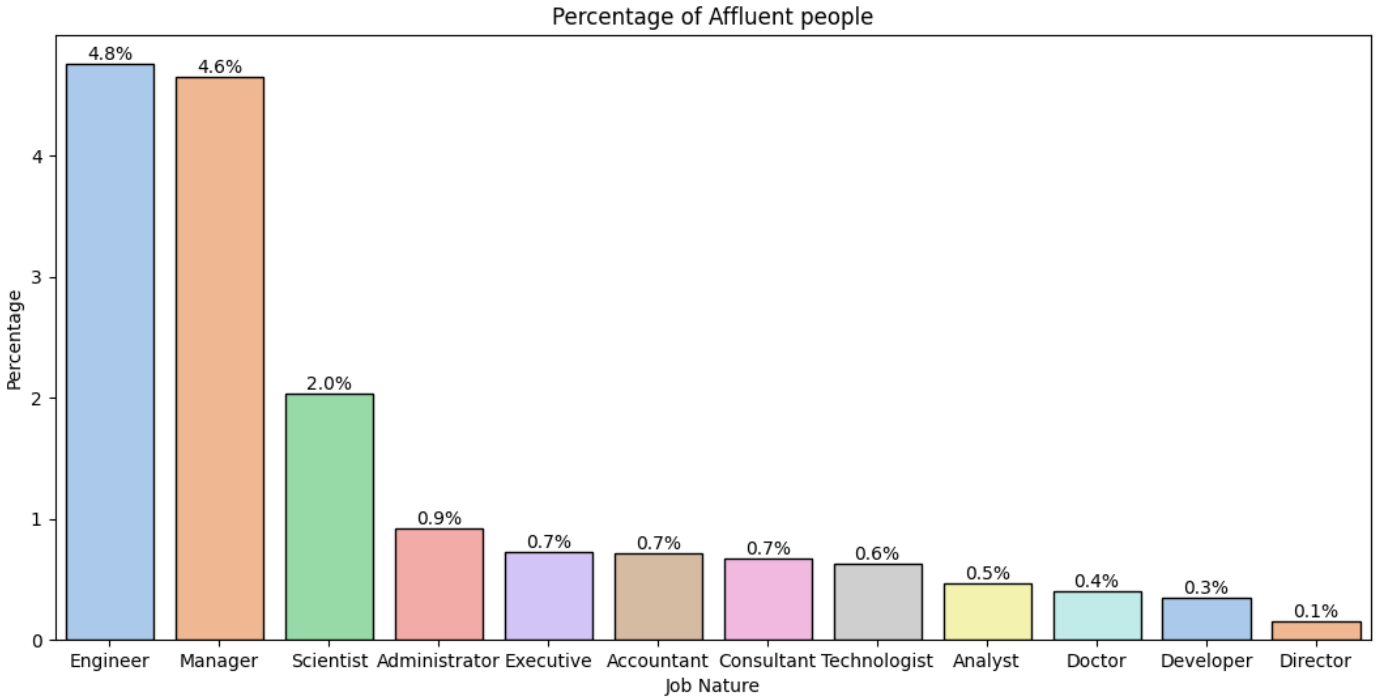

Figure 10 Affluent People Population

3.8 Infirmity

The graph reveals that a substantial 99.3% of the population is in good health, showing no signs of infirmity, this shows that the demand for extensive healthcare resources in not significant. The data indicates that only 0.2% of the population reports having some form of physical disability.

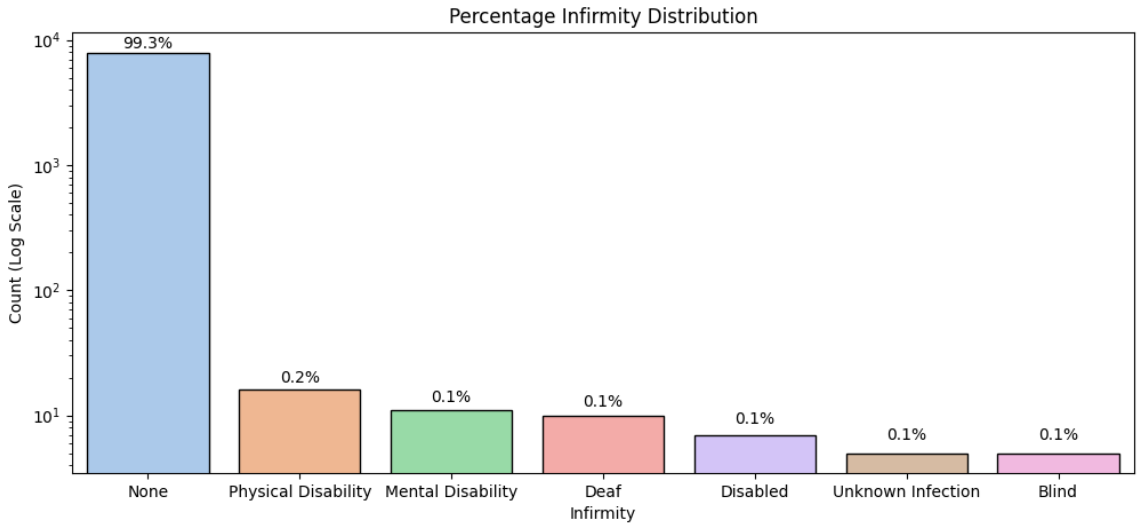

Figure 11 Infirmity Distribution

3.9 Age Distribution

43.3% of the population is composed of people who are between the ages of 18 and 54. This suggests an active working-age population. Given that 16.9% of people in the population are of school age, indicating typically between the ages of 5 and 17, there is a clear need for educational resources. The 12.9% of the population that is 65 years of age or older indicates that a sizeable portion of the population is retired or elderly.

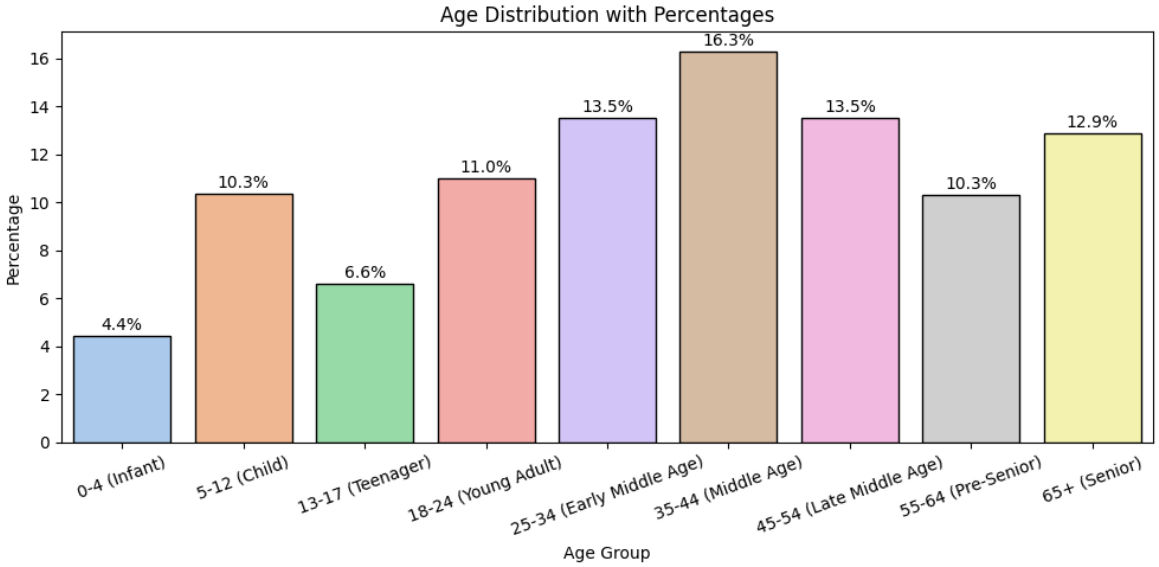

Figure 12 Age Group Distribution

4. DISCUSION

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

4.1.1 High-density housing due to significant population expansion

The choice to invest in high-density housing is mostly influenced by the current housing stock and expected population increase. A low birth rate is currently evident as the age distribution analysis (Figure, 1) shows that the 0–10 age group is quite small. Nonetheless, a noteworthy 15% of unmarried individuals between the ages of 18 and 44 (Figure, 6) are probably going to get married and so raise the birth rate in the future.

Even with the possibility of future population expansion, 58.7% of houses (Figure, 7) have two or less tenants, indicating that many properties are underutilized. There is less need for high-density housing because of this underutilization.

4.1.2 Low-density housing demand for affluent population

The two most important factors that go into the choice to develop low-density housing are the need for large homes and how affluent a given population is. The study found that 16.47% of the people (Figure, 10) are wealthy. While the rich might benefit from low-density housing, they are simply not a large proportion of population to justify an appeal for widespread investment into such kind type of residential at the moment.

4.1.3 Train Station for Commuters

The analysis's findings indicate that a sizable portion (1/3) of the population in this town commutes, as shown in Figure 9, where 33.7% of the participants were determined to be so. Of those, a sizable portion—25.2%—consist of students who are core traveler. A train station would be essential, since one-third of the population commutes.

4.1.4 Need for Religious Building

According to the analysis (Figure, 4), 38.7% of the population identify as atheist, 32.6% as Christian, and 15.6% as Catholic. Because atheists do not typically require places of worship but may benefit from community centers or secular spaces, the major focus should be on meeting the needs of the Christian population.

The 32.6% Christian demography includes denominations other than Catholicism, such as Protestant and Orthodox. As a result, a second church may cater to this considerable segment of the community, possibly covering a broader spectrum of Christian denominations.

4.1.5 Need for Emergency Medical Building

Based on Figure 11, which shows that only 0.2% of the population reports a physical disability and 99.3% have no infirmity, the current demand for emergency medical care due to injuries or health conditions appears to be minimal. Additionally, the age distribution indicates a low birth rate, with a smaller proportion of the population in the early childhood age range, suggesting that future pregnancies are unlikely to significantly increase. Given these factors, the immediate need for a minor injuries center or emergency medical facility appears to be limited.

4.2 Which one of the following options should be invested in?

4.2.1 Employment and training

From Figure 2, it is evident that only 6.10% of the total population is unemployed and majority proportion (52.67%) is employed. Given this relatively low unemployment rate, the need for extensive employment and training programs to address significant unemployment issues is not a pressing concern.

4.2.2 Old age care

Figure 12 reveals that 12.9% of the population is 65 or older. The age pyramid does not indicate that there is a significant number of people who are living longer lives. Therefore, this option will be eliminated.

4.2.3 Increase spending for schooling

From Figure 12, it is observed that 16.9% of the population is of school age, with 10.3% in the child group and 6.6% in the teenager group. This indicates a significant current demand for schooling facilities. Additionally, the presence of 15% of singles aged 18-44 (Figure, 6) suggests a potential increase in births and a future rise in the school-aged population as these individuals may marry and start families. Therefore, increasing spending on schooling is a prudent consideration to accommodate both current needs and anticipated future growth in the student population.

4.2.4 General Infrastructure

The same factors that determine high-density housing should also determine whether to invest in general infrastructure. As previously said, data indicating the town is not currently growing rapidly made the demand for high-density housing seem less urgent. Likewise, general infrastructure expenditure is unnecessary as the town's growth is not apparent.

5. DECISION

5.1 Decision for Task A

The rail station and the religious building have been shortlisted for the first section. Each possibility was evaluated through hypothesis testing. Chi-square and p-value testing were performed on both options. Standard urban planning and community development procedures established a 30% (Berke, 2009) threshold for the rail station and a 15% (Board, 2013) threshold for the church. **When the findings revealed that the church option had the lowest p-value, it was decided that it should be built first.**

5.2 Decision for Task B

For second part, out of four options, three were rejected and only option retained is to **invest on schools.**

6. CONCLUSION

The problem statement for the project was to analyze census data to determine the best use for an empty piece of land and possible investment regions. To remove errors and inconsistencies, extensive data cleaning was required, followed by major statistical analysis to yield actionable insights. Several investment opportunities were considered in light of these disclosures. Hypothesis testing revealed that the most effective investments are **building a new church** and **increase school funding**.

7. REFERENCES

Berke, P. a. G. D., 2009. *Urban Land Use Planning.* 5th ed ed. s.l.:Urbana: University of Illinois Press.

Board, T. R., 2013. *Transit Capacity and Quality of Service Manual..* 3rd ed. ed. Washington, DC: The National Academies Press.

GOV.UK, 2023. *GOV.UK.* [Online]
Available at: https://www.gov.uk/government/statistics/chapters-for-english-housing-survey-2022-to-2023-headline-report/chapter-1-profile-of-households-and-dwellings#:~:text=In%202022%2D23%2C%20the%20mean,had%20the%20smallest%2C%201.8%20persons
[Accessed 08 06 2024].