Fundamentals of Data Science Census Project Report.

**Introduction**

This project is based on census data of a small town described as a Middleburg because it is sandwiched between two much larger cities that it is connected to by motorways. The main objective of this project is to analyze the census data and based on it decide for the local government on what should be built on an unoccupied plot of land that the local government wishes to develop and the choice of investment in community services.

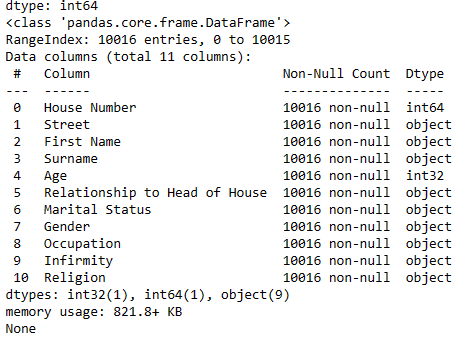
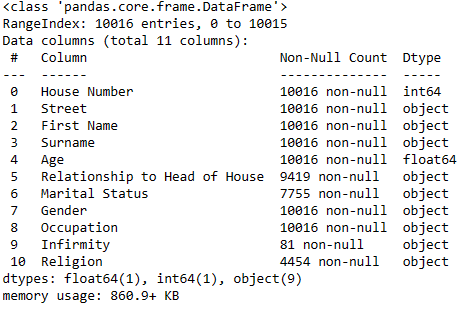
**Data Pre-processing**

Data preprocessing simply deals with the detection, manipulation, and correcting of incomplete data entries in a dataset. This is done by first identifying missing values or incomplete or inaccurate entries in the dataset, using various imputation methods, it is being modified before it is analyzed. The data was a CSV file that was imported into Python with the help of the library called Pandas. The original data contained 10016 rows and 11 columns, with columns labeled as **House Number**, **Street**, **First Name**, **Surname**, **Age**, **Relationship to Head of House**, **Marital Status**, **Gender**, **Occupation**, **Infirmity**, and **Religion.** Columns with missing values include Relationship to Head of House, Marital Status, Infirmity, and Religion. The columns with the missing values were replaced based on the assumption that the respondent did not specify their choice or for some part the respondent was not applicable. These assumptions were held rather than utilizing the mode imputation method because the mode imputation is biased and misinform hence changing the nature of the data and making the integrity of the data questionable. But with these assumptions held, the outcome of this analysis would be the true nature of what the population depicted and therefore, painting the true picture for decision-making.

The details of the replacement process for the missing values are listed below:

1. Relationship to Head of House: This column contained 597 missing entries out of a total of 10016 entries. All these missing entries were replaced with “Not Specified” indicating the respondent did not specify their relationship to the head of the house. Also, Individuals that lived alone and are single were automatically replaced with Head.
2. Marital Status: This column contained 2261 missing entries out of a total of 10016 entries. With marital status, persons below 18 years who had missing values or single status were replaced with “Not Applicable” as they are not legally qualified to be married in the United Kingdom whereas persons above 18 years were replaced with “Not Specified” indicating the respondent did not specify whether married or not.
3. Infirmity: This column also contained 9935 missing entries out of a total of 10016 entries. All these missing entries were replaced with “Not Specified” showing the respondent did not specify whether they had any infirmity or not. This could be the absence of infirmity, or they just failed to specify.
4. Religion: This column also contained 5562 missing entries out of a total of 10016 entries with missing entries replaced with “Not Specified” indicating the respondent did not specify their religion.

The age column did not contain any missing value, but the original data type assigned to the age column was the float data type which was wrongly assigned. The data type was changed to integer data type because we were not considering the fractional age of the respondent just curtate age. **Figure 1** shows a summary of the data before and after data pre-processing.



1. Before Preprocessing b. After Preprocessing

Figure 1: Summary of Data Before and After Preprocessing

**Population Distribution**

The town per description can be described as a Middleburg. From the census data, the town has a population of 10016 people with Females forming the majority part of the population. With gender distribution, the number of males in the population was 4679 forming approximately 47% of the population while females were 5337 of the population forming approximately 53% of the population. The age range of 0 – 18 forms approximately 24% (2426) of the population, the range of 19 – 64 also forms approximately 63% (6355) of the population, and 65+ also forms 12% (1235) of the population. The other part of this section discusses the further demographic characteristics of the population which include birth and death, religion, marriage and divorce, employment and unemployment, and household distribution.

**Age Distribution**

Every individual not only belongs to a spatial group but is invariably, a member of a temporary class of a generation, progeny, or an age class. Hence, any population necessarily has an age structure. The population in question was divided into 3 age ranges which include 0 – 18, 19 – 64, and 65+. The various age ranges were distributed across the population as follows; 0 – 18 were 2426, 19 – 64 were 6355, and 65+ were 1235 with percentages of 24%, 63%, and 12% respectively. This kind of age distribution corresponds with the progressive age structure where the population is characterized by a high proportion of children and a small number of elderly people. Common characteristics include a high birth rate, high growth rate, and increasing dependency ratio. Figure 2 shows a population pyramid for the age distribution of the population. The shape of the population pyramid shows an expansive population pyramid. There is a wide base indicating a high birth rate and large youth population and also depicts an expanding base in the future. A narrower top indicates lower life expectancy resulting in fewer elderly individuals, and a high dependency ratio due to a large proportion of young dependents. In the future, there will be more school-aged children and more young people. Some merits and demerits of this population structure include the potential for rapid population growth, increased demand for resources and services for the young population, and economic challenges related to providing education and healthcare for a large youth demographic.

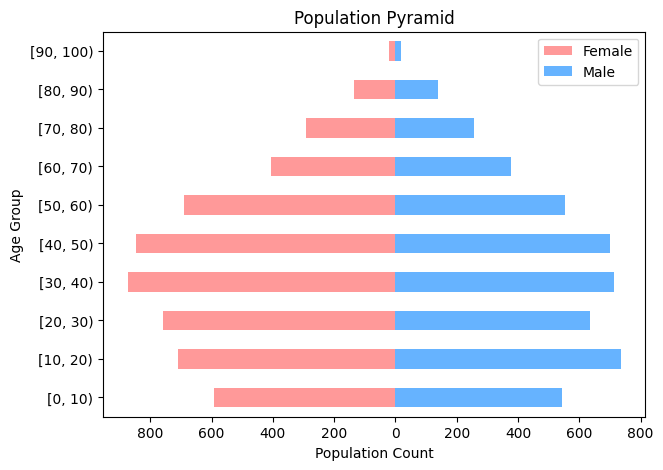


Figure 2 : Population Pyramid of Age Distribution

**Employment and Unemployment**

The town contains many employed people who contribute to the economic growth of the town. The employed individuals were 9370 forming approximating 91% of the working population thus ages between 18 to 65. Common occupations of the population include students, surgeons, textile designers, retail buyers, research officers, geophysicists, and clothing technologists. Figure 3 shows the top common occupation of the population.

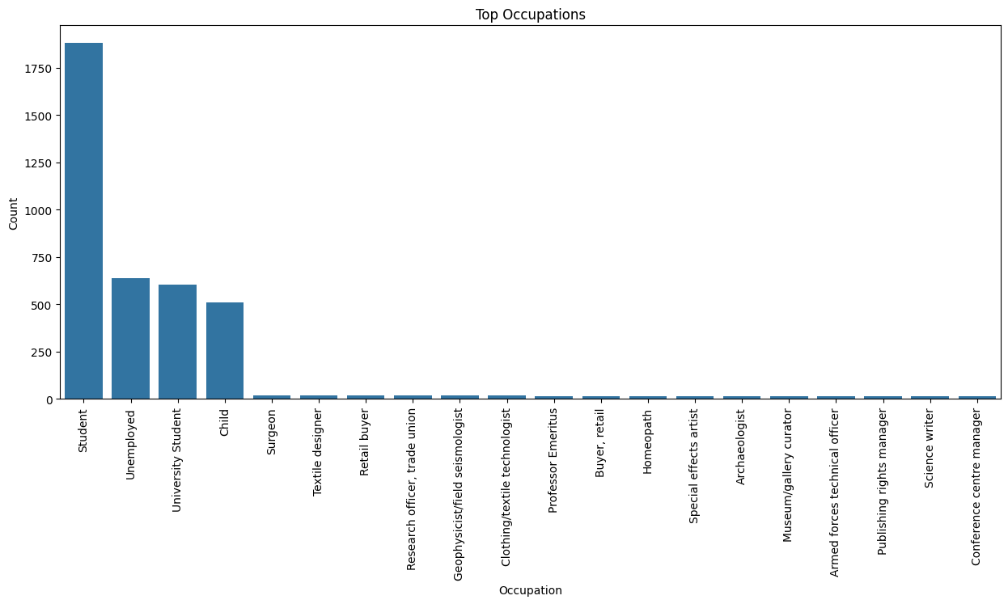


Figure 3: Common Population Occupation

The town's unemployment rate stands at about 9%, with 590 out of the working population of 6,555 being unemployed. This rate is significant, especially when compared to the UK's average which has been between 3.6% and 6.3% in the last 10 years according to the Office for National statistics. A large proportion of the employed are students, who typically do not earn an income, potentially skewing the unemployment figures higher and increasing the dependency ratio. Notably, the majority of the unemployed fall within the 25-54 age range. The accompanying figure illustrates unemployment distribution across various age groups.

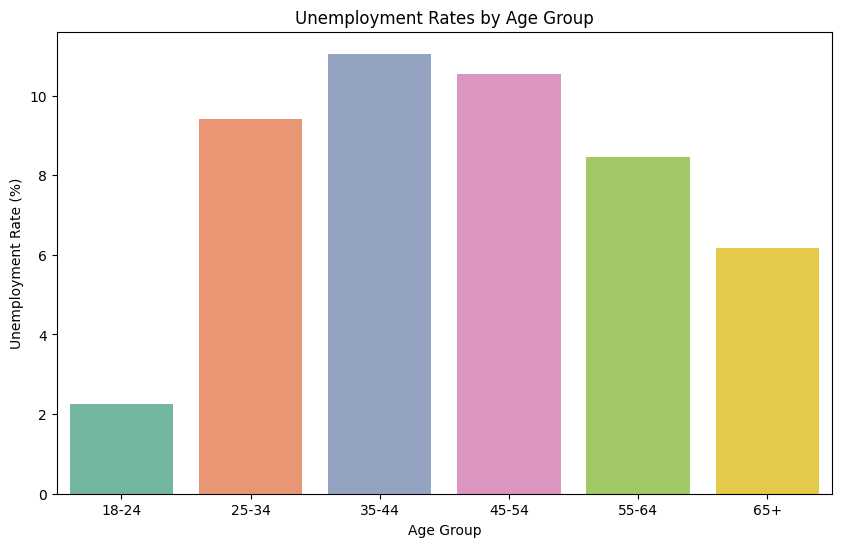


Figure 4: Unemployment Across Different Age Groups

**Religious Affiliation**

The summary of the religious affiliations in the population shows a trend where 55.54% of individuals did not specify their religion, which might indicate a lack of particular religious affiliation or indifference towards religion. Among those who specified, Christians are the most represented group, with 23.22% identifying as Christian. This is followed by other Christian denominations: Catholics at 10.93% and Methodists at 7.32%. Smaller percentages belong to Muslims, Sikhs, Jews, Agnostics, Buddhists, Bahá'ís, Pagans, and others, each contributing less than 2% individually. Additionally, a very minor fraction, 0.02%, reported being undecided about their religious affiliation at the time of the survey. This distribution suggests a diverse religious composition but with a significant proportion of the population not aligned with any particular religion.

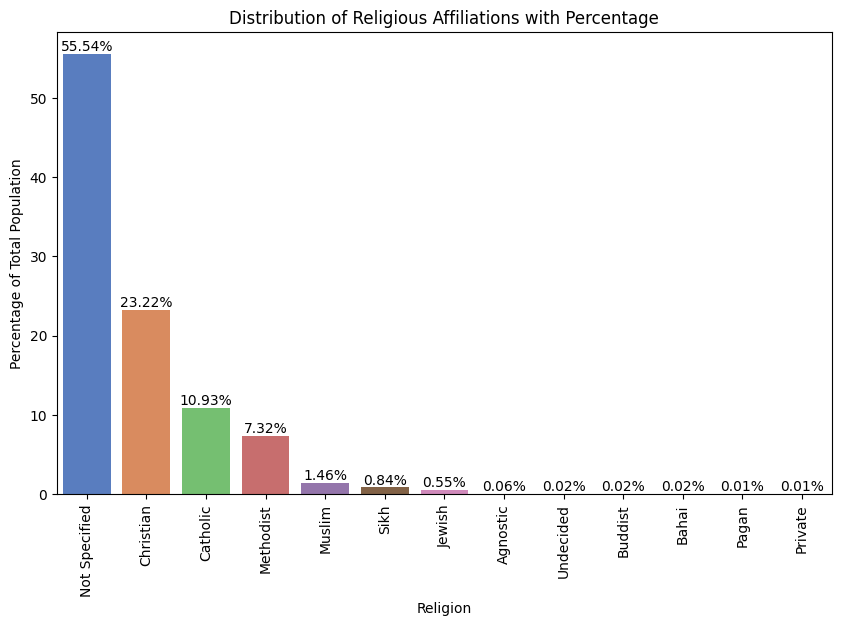


Figure 5: Religious Affiliation of the Population

The population is Christian dominated amongst people associated with religion and has most of its followers around the age of 50. It is safe to say that there is not a potential for growth for Agnostic religion and is the median age is around 70 with the others having potential to grow. Figure 6 shows the median age of the followers of the various religious groups and the number of adherent by religion.

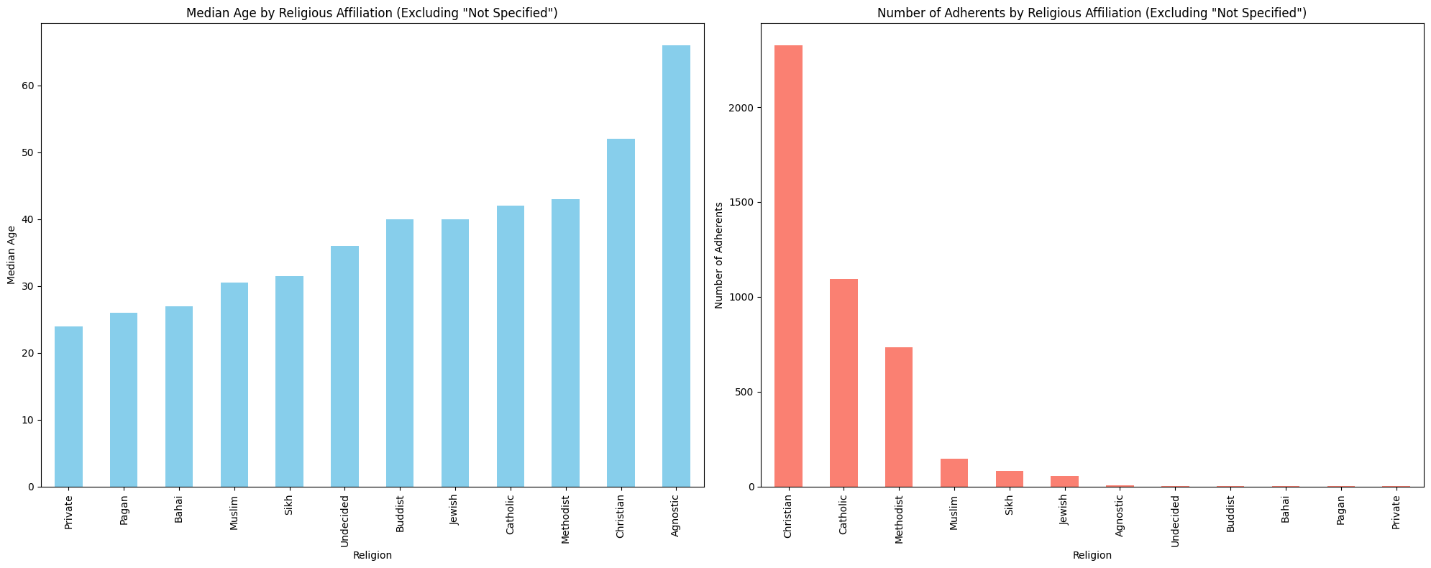


Figure 6: Median Age of Adherent followers of the Various Religious Groups.

**Marital Status**

In the population data, the marital status recorded was married, divorced, single, widowed and Not applicable where individual was below 18 years. It was observed that most of the people married were around the age range of 41 – 70, and most of the single people were around the age range of 18– 40. The Figure below shows the distribution of marital status across different age groups.

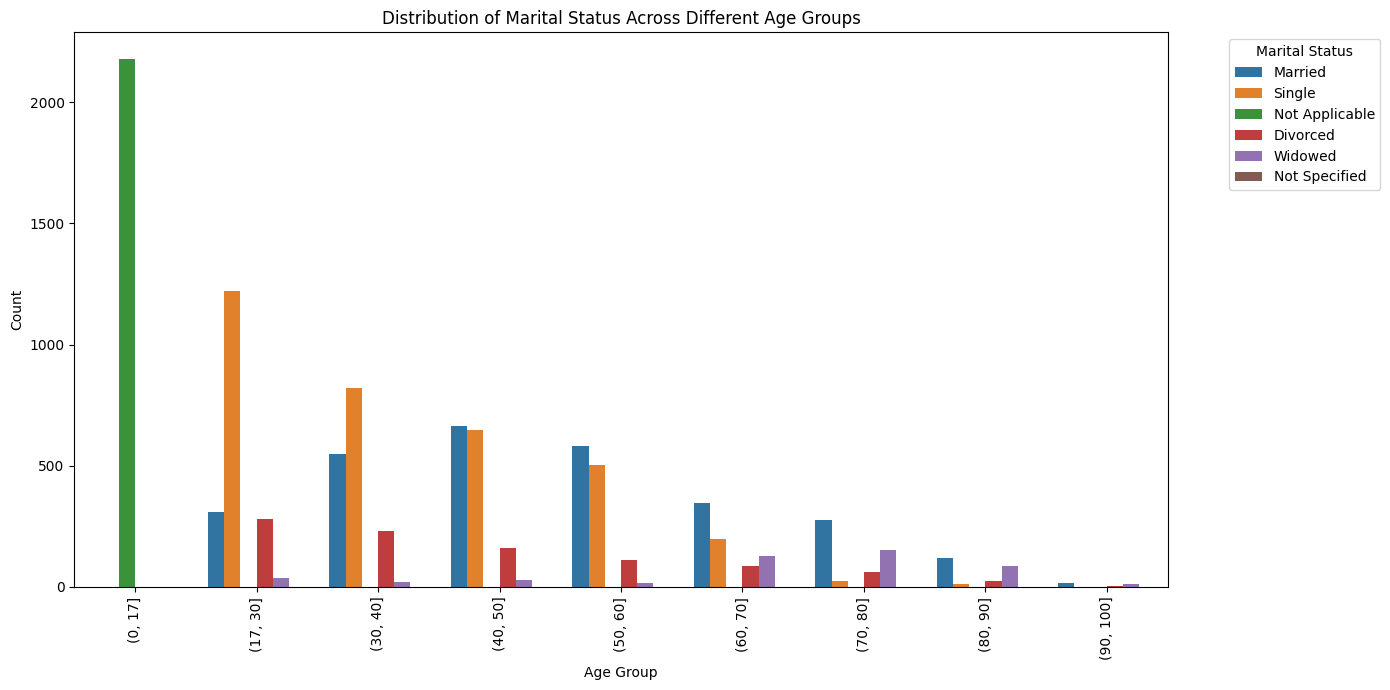


Figure 7: Distribution of Martial Status Across Different Age Groups

**Household Distribution**

The study identified 3,850 unique households in the town by counting distinct house numbers on each street. Although the average household size was calculated at 2.60, the data revealed that the most common household type consists predominantly of single-occupant homes. This trend towards smaller household units may be influenced by several factors including an aging population, higher divorce rates, lifestyle preferences for living alone, or changes in family structures. Demographic analysis suggests that individuals aged 0-17, categorized as Not Applicable, and married individuals living with their families comprise about 51% of the population. In contrast, those who are single, divorced, or widowed—who are more likely to live alone—account for approximately 48% of the population.

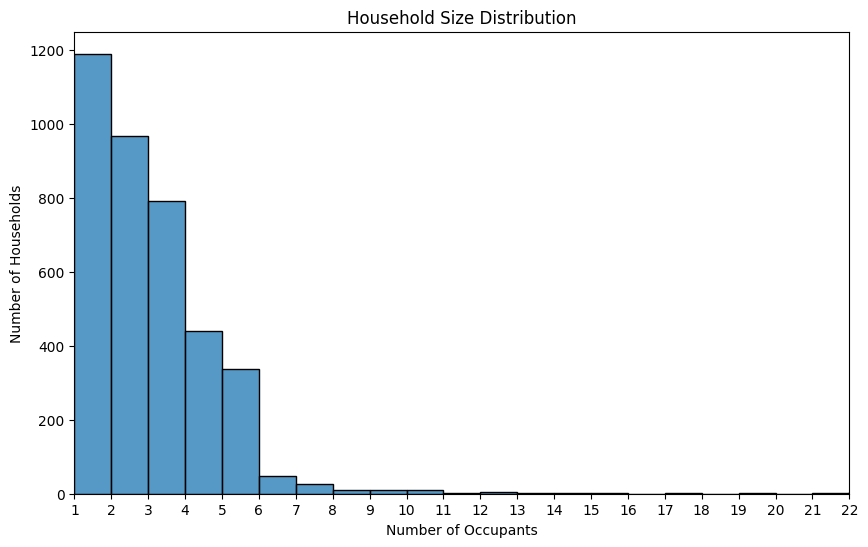


Figure 8: Household Size Distribution by Number of Occupants

**Commuters**

Commuters here are based on the assumption that the town in question does not have a university so all university students are potential commuters and all employed individuals within the working age thus 18 – 65 are also likely possible commuters. Individuals in the age group of 65+, those unemployed, and children in the town are not likely to commute to other cities. Based on these assumptions, the number of potential commuters in the town recorded was 6004, which forms a proportion of about 60% of the population. See the distribution below.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 9: Proportion of commuters and non commuters

**Birth and Death Rate**

**Birth Rate Analysis**: We based on the current data to identify trends across various age groups, utilizing a direct calculation of births per 1000 women. The analysis revealed an initially surprising peak in birth rates among the 44-48 age group at approximately 395 births per 1000 women, indicating a high previous reproductive age. To project into the future, we assumed a steady fertility rate among the current population of younger women. From the graph below, we can tell there is a steady increase in birth rate based on the data. These insights are critical for planning, suggesting an increase in demand for educational and recreational facilities to support a growing younger population.

A graph with blue lines and red dots

Description automatically generated

Figure 10: Birth rate and future projections

A graph of a number of people

Description automatically generated**Death Rate Analysis:** We analyzed the mortality across different age groups. The assumption is that the change in population size between the two age groups is due to deaths. It does not account for other factors such as migration (both in and out) or other demographic changes. There was a general increase in the rate from the age group 66+. The highest observed death rate is in the 86-90 to 91-100 age group, with approximately 113 deaths per 1000 individuals annually, emphasizing a significant mortality rate among the elderly.

Figure 11: Death rate across the elderly age group

**Net Population Growth**

The town's demographic trends suggest net population growth, driven by high birth rates among middle-aged women but tempered by high mortality rates among the elderly. Additionally, analysis of the head of household data revealed that the presence of 268 lodgers, who are likely to leave, could further moderate this growth. These factors collectively highlight the complex dynamics influencing population changes and the need for strategic planning.

Top of Form

# Key Suggestions and Recommendations

## 1. Development on Unoccupied Plot of Land

**Based on the demographic growth and commuting patterns analyzed:**

* Recommended Development: The growing young and working-age population supports the construction of high-density housing. This type of housing will effectively utilize the available space to accommodate more residents as the population expands. Alongside, a train station would benefit the large percentage of commuters by providing an efficient, alternative transport option, reducing road congestion, and fostering economic activity through easier access to nearby cities.

**Counterpoints to Other Developments:**

* Religious Building: Despite the presence of multiple religious affiliations, a significant 55.54% of the population did not specify their religion, suggesting a potential secular or non-religious majority. Building another religious facility might not serve the broad interests of the community, especially when other more universally beneficial facilities are needed.
* Low-density Housing: Given the limited space and the high percentage of young individuals likely to start their own households, low-density housing would not be as effective in maximizing land use or meeting the housing demand as high-density housing.

## 2. Recommendation for Investment in community service

Investing in employment and training programs is strongly recommended due to the town's significantly high unemployment rate of about 9%, compared to the UK's national average which has fluctuated between 3.6% and 6.3% in the last 10 years. This high rate, particularly pronounced in the 25-54 age range, indicates a critical need for initiatives that enhance employability and align workforce skills with current job market demands. Such programs will not only benefit individuals by improving job prospects and increasing earnings but will also positively impact the local economy through enhanced consumer spending, reduced social welfare costs, and increased tax revenue. The diversity of occupations in the town, from students to surgeons and textile designers, suggests that a targeted training approach could foster economic diversification and growth by focusing on underrepresented and emerging sectors.

While investments in old age care and schooling are important, they do not directly address the immediate economic challenges posed by high unemployment. Enhancing employment and training tackles this issue head-on, providing broad economic benefits that extend beyond individual employment. Moreover, improved employment rates can support better care for elderly dependents and justify future infrastructure improvements, ensuring these investments are sustainable. Thus, prioritizing employment and training is not merely a response to current challenges but a strategic move to equip the town for future economic stability and growth, making it the most impactful investment choice given the current demographic and economic data.

REFERENCES

1. Office for National Statistics: [Unemployment rate (aged 16 and over, seasonally adjusted): % - Office for National Statistics (ons.gov.uk)](https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms)