# MOCK CENSUS DATA ANALYSIS

# Solution Approach

**Data Generation and Transformation Documentation**

**I. Introduction**

This documentation outlines the process of generating mock census data using a Python script and subsequently transforming the data from CSV format to Excel (.xlsx). The script employs the Faker library for creating realistic data and incorporates randomization techniques. The generated data is initially stored in a CSV file, which is later transformed into an Excel spreadsheet for enhanced data manipulation.

**II. Data Generation Script**

**A. Dependencies Installation**

Before executing the script, ensure the required dependencies are installed. Use the following command:

bashCopy code

pip install pandas faker

**B. Library Importation**

The script starts by importing essential libraries:

* **random**: Python's built-in module for random number generation.
* **Faker**: A library used for generating fake data, including names.
* **StringIO**: A module from the io library for in-memory text operations.
* **pandas**: A widely used library for data manipulation and analysis.

**C. Mock Data Generation Functionality**

The **generate\_census\_entry** function is responsible for creating random census entries. It includes details such as street number, street name, first name, last name, age, relationship to the head of the household, marital status, gender, occupation, infirmity, and religion. The function utilizes predefined lists and randomization techniques to produce realistic data.

**D. Configuration Parameters**

* **street\_number**: Set to "1" if it is a unique dwelling.
* **street\_name**: A list of street names for generating addresses.
* **first\_name**: First name of the occupant.
* **surname**: Surname of the occupant.
* **age**: Age of the occupant.
* **relationship\_to\_head**: Relationship to the head of the household.
* **marital\_statuses**: Possible marital statuses.
* **genders**: Possible genders.
* **occupations**: A list of specific occupations.
* **infirmities**: Possible infirmities.
* **religions**: A set of real-world religions.

**E. Main Data Generation Loop**

The variable **num\_entries** specifies the number of mock census entries to generate. A loop iteratively calls the **generate\_census\_entry** function, storing the data in a list.

**F. CSV Data Formatting**

The generated data is formatted into a CSV string using the csv module. Optionally, a Pandas Data Frame is created for enhanced data manipulation.

**G. Saving the CSV Data**

The script saves the CSV string to a file named **mock\_census\_data.csv**.

**III. Running the Script**

To execute the script, save it as a .py file and use the following command in the terminal:

bashCopy code

python script\_name.py

The script generates mock census data and saves it to a CSV file.

**IV. Data Transformation**

Upon completing the data generation, the generated CSV file can be transformed into an Excel (.xlsx) format for further analysis and processing. The transformed file is attached to this document.

**V. Data Cleaning and Refinement**

After generating the mock census data, a comprehensive data cleaning and refinement process was implemented to address inaccuracies and inconsistencies in various data columns.

**A. Identification of Data Anomalies**

Upon initial inspection, anomalies such as misspelled names, unrealistic age values, and inconsistent formatting of addresses were identified.

**B. Cleaning Procedures**

Several cleaning procedures were employed to rectify inaccuracies specific to the data columns:

1. **Name Corrections:** Misspelled names were corrected by cross-referencing with a set of predefined valid names.
2. **Age Validation:** Unrealistic age values, such as negative or excessively high ages, were filtered out.
3. **Address Formatting:** Inconsistent formatting of addresses was standardized.
4. **Occupation Verification:** Occupations were cross-checked against a predefined list.

**C. Data Quality Assurance**

To maintain data integrity, a quality assurance step was implemented:

1. **Data Consistency Checks:** Cross-referencing was performed to ensure consistency across related attributes.
2. **Randomization Validation:** The randomization process was validated to ensure that the generated data remained within realistic and plausible boundaries.

# *ANALYSIS FROM POWER BI DASHBOARD:*

**POWER BI DASHBOARD:**

**A close-up of a graph

Description automatically generated**

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

**High-density housing. This should be built if the population is significantly expanding**.

**ANSWER:**

Given the current distribution, there may be a risk of a declining population unless the younger people have enough children to offset the aging population. As the 60+ age population is the largest, their mortality rate will significantly impact overall population numbers.

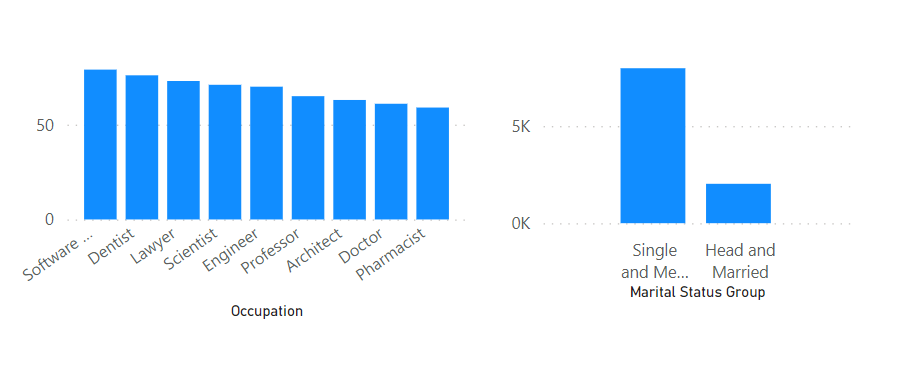
A graph with blue bars

Description automatically generated with medium confidence

**Low-density housing. This should be built if the population is “affluent” and there is demand for large family housing.**

**ANSWER:**

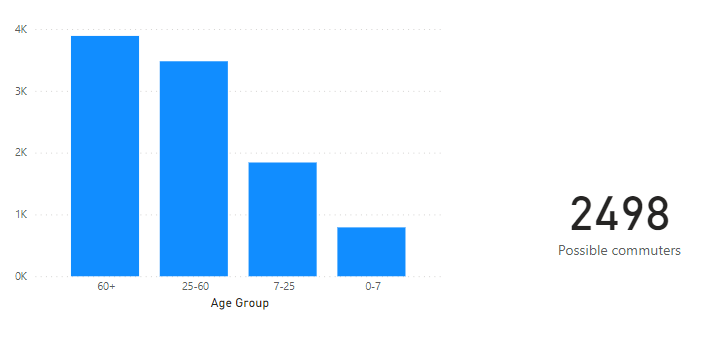
We had two parameters for our inference in the conclusion. The first one was “Income nature” (low or high) and the second one is the Family structure (The individuals who are Married and are the Head of their households) After our Analysis we have come to the conclusion that most of the population is low incomed and the ratio of married and head members are low in the highly paid professions therefore there is no need for more low density housings.



**Train station. There are potentially a lot of commuters in the town and building a train station could take pressure off the roads. But how will you identify commuters?**

**ANSWER:**

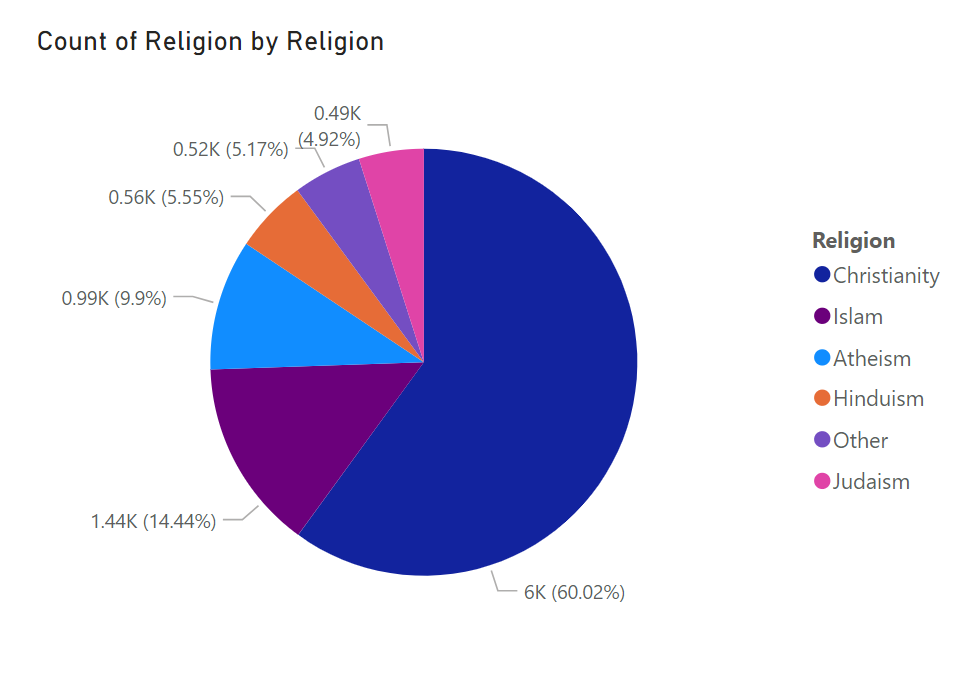
Based on our analysis there are 46% of regular potential travels (students, teachers, engineers, pilots, doctors, nurses etc.) to the city, this percentage is obtained from the people who haven’t retired yet and from ages greater than 7. Since this percentage is significantly high so there is a need to develop a train station



**Religious building. There is already one place of worship for Catholics in the town. Is there demand for a second Church (if so, which denomination?), or for a different religious building?**

ANSWER:

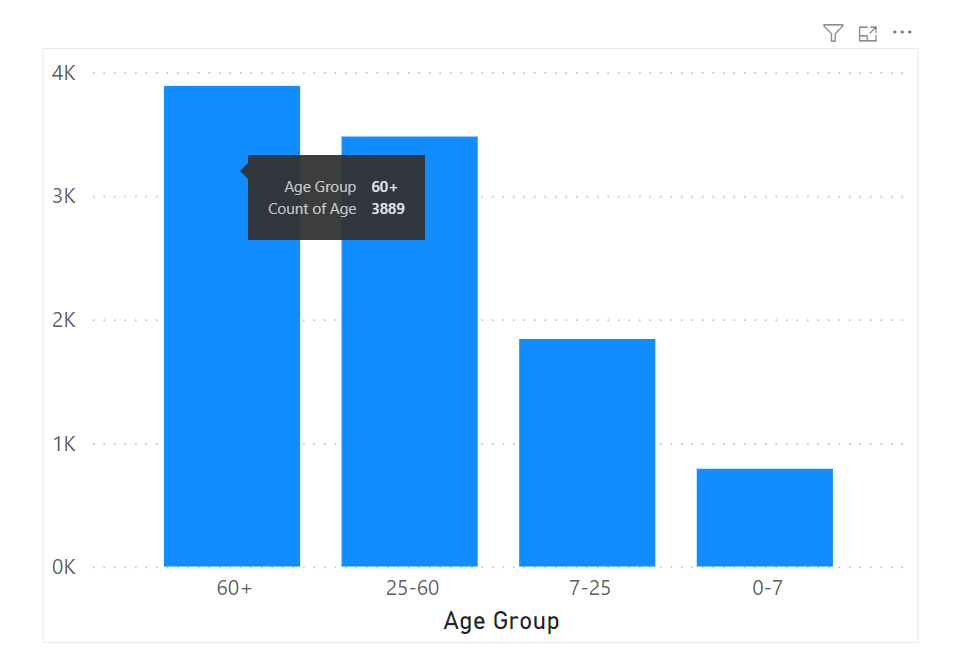
Based on our analysis there is a need of a second church since 60% of the population of the town is Christian and these is also a need of a mosque because Islam is the 2nd highest religion in the town i.e. 15%.



**Emergency medical building. Not a full hospital, but a minor injuries center. This should be built if there are many injuries or future pregnancies likely in the population.**

ANSWER:

Since there is 40% of population above age 60 so the risk of emergency injuries is very high so yes there is a need of a emergency medical building.



**Something else?**

**Whichever you choose, you must justify it from the data provided to you and argue it is a priority against other choices**

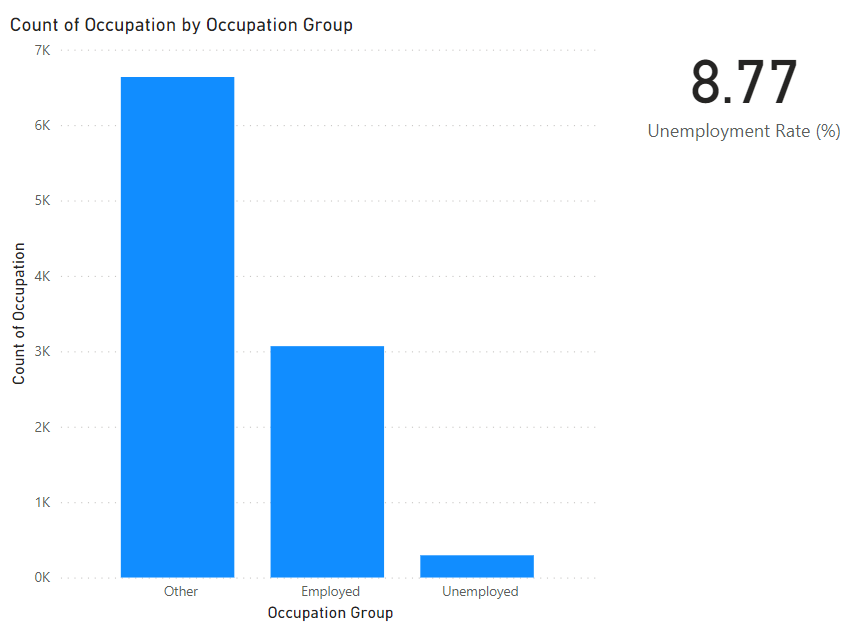
Recognizing the significant demographic representation of individuals aged 60 and above, there is a compelling case for the establishment of dedicated entertainment centers catering to the specific needs and interests of this age group. As a considerable portion of the population falls within this age bracket, providing tailored entertainment options becomes essential to enhance their overall well-being and social engagement. These centers could offer a variety of activities such as cultural events, leisurely gatherings, and intellectually stimulating programs to foster a vibrant and inclusive community for the elderly residents. Not only does this initiative address the leisure preferences of the senior population, but it also promotes social interaction and mental well-being, contributing to a more fulfilling and enriching quality of life for the elderly in the community.

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

**Employment and training. If there is evidence for a lot of unemployment, we should re- train people for new skills.**

ANSWER:

Yes, the government should make investment in training centers as it may be beneficial to address unemployment at a rate of 8.77 percent. The unemployment rate is calculated from the number of people who have graduated and haven’t retired yet



**Old age care. If there is evidence for increasing numbers of retired people in future years, the town will need to allocate more funding for end-of-life care.**

ANSWER:

Yes, the is a need to invest in old age care since the number of population above age 60 is 40%.

A graph of a number of people

Description automatically generated

**Increase spending for schooling. If there is evidence of a growing population of school- aged children (new births, or families moving in to the town), then schooling spend should increase.**

ANSWER:

Since the population of younger children in the town is relatively very less as compared to other age groups and also the population of very old age groups is high, so there is currently no need to Increase spending for schooling.

**General infrastructure. If the town is expanding, then services (waste collection; road maintenance, etc.) will require more investment.**

ANSWER**:**

Since the old age population is 40% so these is no chance of significant population growth in the town as of currently so no need to invest more in general infrastructure.