**CENSUS PROJECT REPORT**

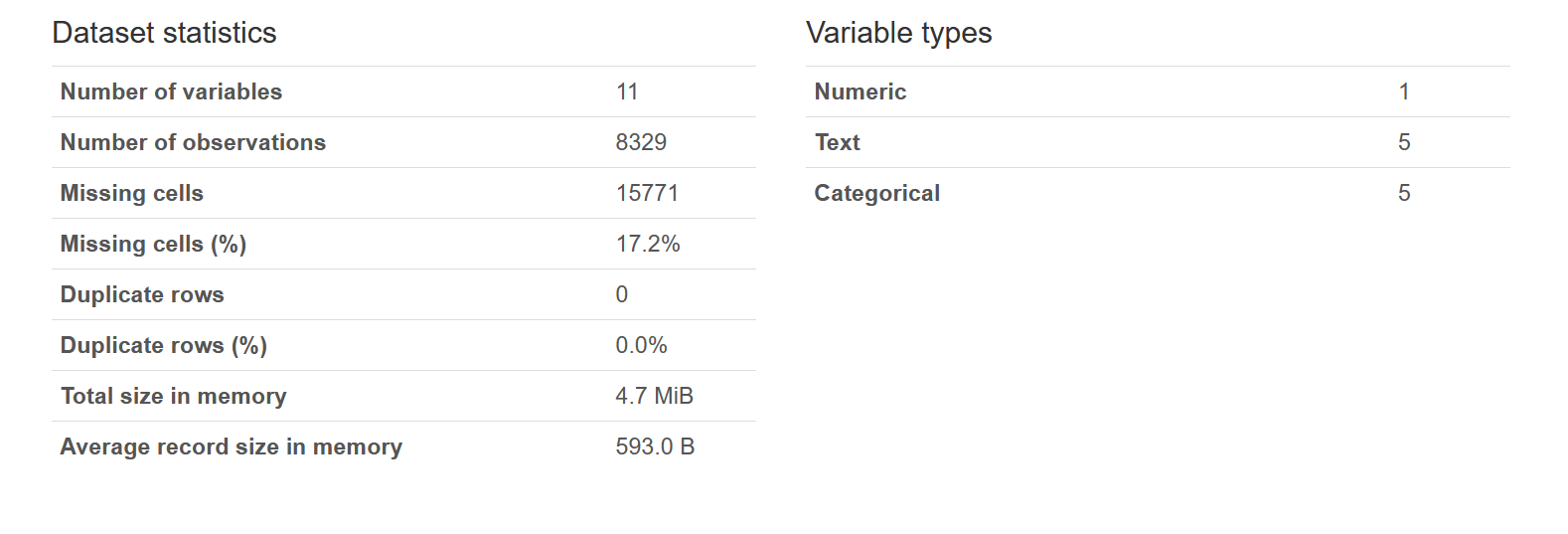
Census data offer a comprehensive snapshot of a town's demographic composition, providing insights into the characteristics, lifestyles, and socioeconomic attributes of its residents. With 8,329 rows and 11 columns, the dataset encompasses diverse information including numeric, text, and categorical variables.

Through meticulous data cleaning and analysis, various aspects are explored, such as age distribution, employment status, religious affiliations, and housing occupancy. This dataset serves as a valuable resource for understanding population dynamics, identifying potential growth areas, and informing strategic decision-making for the town's development, infrastructure, and social services.

**Initial Overview**

The dataset consisted of **8,329 rows** with **11 columns**, including one numeric, five text, and five categorical variables, with 17.2% missing data, no duplicate rows, and a total size of 4.7 MiB in memory.

The columns "Gender" and "Relationship to Head of House" exhibit a strong positive correlation, and both have 6.6% missing values. Additionally, "Marital Status" has 24.7% missing values, "Infirmity" has 99.2% missing values, and "Religion" had 58.9% missing values.



A screenshot of a computer

Description automatically generated

**Data Cleaning**

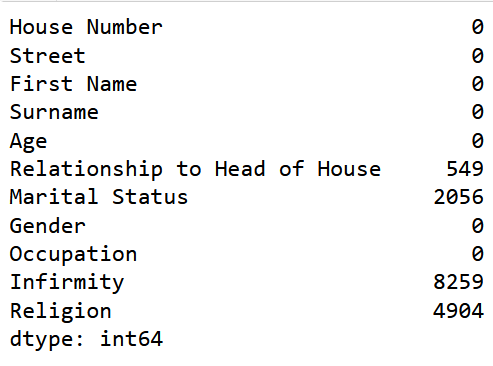
**Columns Datatype**

The initial dataset contains 1 numeric 5 categories and five-string data types. From the observations, the “Age” column needs to be converted to a numeric datatype. Clean the “Age” column by dropping the two rows in the “Age” column with empty strings because they would not significantly impact the outcome of the entire data set.

'Relationship to Head of House,’ ‘Marital Status,’ ‘Gender,’ ‘Infirmity,’ and 'Religion' are converted to categorical data. This conversion to a proper datatype reduces the size of the data.

**Missing Values**

The count of null or missing values in each column of a Data Frame indicates that there are 549 missing entries in the "Relationship to Head of House" column, 2056 in "Marital Status," 8259 in "Infirmity," and 4904 in "Religion," while other columns have no missing values.



**"Relationship to Head of House" missing data**

The "Relationship to Head of House" column has Nan values. Since these Nan values are not known, they are replaced with the value of ‘Unknown.’.

**"Marital Status" missing data**

There were both missing null values and empty strings in marital status. Further, note that an empty value is different from a null value. The United Nations Convention on the Rights of the Child (UNCRC) defines a child as everyone under 18 unless, "under the law applicable to the child, the majority is attained earlier.” A child cannot legally have a marital status. The marital status of a child can be replaced with ‘Child’ while the empty strings can be replaced with ‘Unknown’.

**"Infirmity" missing data**

There were null, empty, and unknown infections as part of the infirmity data values. Infirmity refers to physical or mental weakness, often associated with old age or a lack of vitality therefore ‘Unknown infection' is not an infirmity and can be replaced with 'Unknown' The null and empty values can also be replaced with 'Unknown.'

**"Religion" missing data**

There are both null values and empty strings in the religious dataset, and the Buddhist religion is spelled wrongly and corrected.

**Cleaned Data Overview**

The "Non-Null Count" indicates that all 8327 entries have non-null values for each respective column.

A screenshot of a computer

Description automatically generated

There were no missing or duplicated rows. The total memory is 4.3 MiB, and the average record size is 545.2 bytes. The variable types included two numeric variables, four text variables, and five categorical variables.

A white text box with black lines

Description automatically generated

A screenshot of a computer

Description automatically generated

There is a strong positive correlation between Gender and Relationship to Head of House, and vice versa, denoted by the term "High correlation." Additionally, Infirmity exhibited a high imbalance, with 97.4% of the data falling into one category. Furthermore, Age was associated with 88 instances (1.1%) with zero values.

A screenshot of a computer

Description automatically generated

**Age Distribution Analysis**

To examine the age distribution of the population and determine whether the population is growing or shrinking, or if it will have more retired people in the future, more school-aged children, or more young people, an exploratory data analysis was used to extract these insights and deductions.

An age group was added to the census dataset to group the different ages into a group of ages to aid further analysis. Each age group was grouped into individuals within a 4-year age group.

A graph of a graph

Description automatically generated

N.B.: From the above graph, a **right-skewed** distribution may suggest an older population with a few extremely elderly individuals.

A table with numbers and numbers

Description automatically generated

The age distribution appears to be right skewed, as indicated by the higher mean age **(35.29)** compared to the median age **(35).** Also, the value for the skewness of Age is equal to 0.31. Since the skewness value is positive, it supports the interpretation of a right-skewed distribution.

The presence of a few individuals with ages as high as **106** contributes to this right-skewed appearance. This indicates that there might be a tail of older individuals, pulling the mean to the right.

A graph of a person and person

Description automatically generated

The age distribution was diverse, with individuals ranging from 0 to 106 years. The 25th percentile (Q1) was 18, indicating that **25%** of the participants were aged **18 years or younger**. The median age (Q2) was 35 years, suggesting that **50%** of individuals were aged **35 years or younger**. The 75th percentile (Q3) was 50, implying that **75%** of the participants were aged **50 years or older**.

**Employment Status Analysis**

A new column named "Employment Status" in the census Data Frame. It iterates through each row of the data frame and assigns an employment status based on certain conditions related to the "Age" and "Occupation" columns. If the age is **66 or above**, the status is set to **'Retired'**. If the occupation matches patterns indicating unemployment, child, student, or university student, it assigns the corresponding status. Otherwise, it assumes the person is 'Employed'. The patterns are case-insensitive regular expressions, allowing for flexibility in matching different variations of the specified statuses.

A graph of different colored squares

Description automatically generated

The largest category is **"Employed,"** indicating that a significant portion of the population is currently employed.

The combined percentage of individuals categorized as "Student (Child)" and "Retired" was substantial (20.0% + 9.4% = 29.4%). This finding suggests that a significant portion of the population comprises students **(likely children).**

The percentage of individuals categorized as "University Students" is relatively smaller (6.7%). The percentage of individuals categorized as "Child" is also noteworthy **(6.0%).**

The percentage of individuals categorized as "Unemployed" is relatively lower (5.7%), indicating a smaller portion of the population facing unemployment.

A screenshot of a graph

Description automatically generated

**Unemployment Trends Analysis**

A graph of people with different colored bars

Description automatically generated

The age group with the highest unemployment rate is the ages 40-44 therefore between the ages of 40-44 people are more likely to be unemployed. Also, it is important to note that the **average unemployment age is 42**.

A white rectangular object with black text

Description automatically generated

A graph of a number of people

Description automatically generated

N.B.: Statistical analysis of unemployment by gender. The proportion of **unemployed women** was significantly **higher than that of men**.

A graph of a retirement age

Description automatically generated

N.B.: **9.4%** of the individuals in the census data are "Retired.”

**Religious Affiliations Analysis**

A screenshot of a graph

Description automatically generated

The **" unknown "** category was the most prevalent, comprising **58.9%** of the dataset. Most of the **children under 18** do not have any specified religion but will most likely not always have the same religion as one or both parents, and some individuals do not have any religious affiliation, or the information is not available.

The **"Christian**" category is the second largest, representing **21.4%** of the dataset. Other religious affiliations (Catholic, Methodist, Muslim, Sikh, Jewish, Quaker and Buddhist) had smaller percentages in the dataset.

A diagram of a graph

Description automatically generated

The religion of Quaker is almost nonexistent. The most dominant religion is the Christian religion, with an **average age** of worshippers **48** which are middle-aged adults, and most children do not have any religion.

No religion is increasing in number. Quakers seem to be the newer religion since the average age of followers is 21 years of age, but it is not clear if it would experience significant growth.

**Divorce and Marriage Rate Analysis**

A graph of different colored bars

Description automatically generated

A screenshot of a computer

Description automatically generated

N.B.: **35%** of the population are single, **27%** are married and **9%** are divorced.

A graph showing a couple of marriage and divorce rates

Description automatically generated

Further analysis shows that **45-49** had the highest number of married couples (263). **20-24** had the **greatest number of divorced couples** (102). **Younger couples** tend to have **higher divorce** rates than **older couples**.

A screenshot of a graph

Description automatically generated

N.B.: The **unemployed population** appears to have a **high divorce rate**.

**Occupancy Level Analysis**

This analysis begins by inspecting the number of **occupants** residing in a house.

A screenshot of a number

Description automatically generated

A graph of a number of levels

Description automatically generated with medium confidence

A graph of distribution of people per house

Description automatically generated

From the occupancy-level distribution analysis, **five persons per house** is the **most appropriate** number of **occupants** that should ideally be in a living apartment. However, some houses had more than 5 occupants.

A table with numbers and percentages

Description automatically generated

N.B.: The **average number** of people per household was **3**. The minimum occupancy was 1, and the **maximum occupancy** was **22**.

Further analysis showed that **1868 houses** were **moderately used**, whereas **956** were **overly used**.

A group of blue squares

Description automatically generated with medium confidence

From the graph above, Single and Divorced people have higher occupancy rates across various occupancy levels. It also suggests that most **singles either cohabit or stay with friends**. In addition, **divorced people** stay with either their **children or lodgers**. Most **married people** stay within **occupancy levels of 1 to 5**. The **widows** had the **lowest occupancy levels**, but most remained alone in the house.

**Number of Commuters Analysis**

Possible commuters include a combination of those **employed, university students, and PhD students**.

A graph showing a number of communicators

Description automatically generated

The number of possible **commuters** was **4913**, while the number of possible **non-commuters** was **3414**. Thus, almost half of the population are **possible commuters**.

**Birth Rate and Death Rate Analysis**

The dataset has no birth or death dates to obtain the estimated birth and death rates.

A blue and orange rectangles with black text

Description automatically generated

**Steps to calculate the birth rate and death rate is as follows:**

**Child population:** These are suggested to be individuals with ages between **0 and 14** in the 'census' dataset.

**Adult population:** These are suggested to be individuals with ages between **15 and 64**.

**Elderly population:** These are suggested to be individuals with ages **65 and above**.

**Estimate birth rate and death rate per 1,000 population:**

Childbirth rate: Calculated the birth rate for children by taking the ratio of the child population to the total population, then multiplying by 1000 to get the rate per 1,000 population.

Elderly death rate: Calculated the birth rate for children by taking the ratio of the elderly population to the total population, then multiplying by 1000 to get the rate per 1,000 population.

The **birth rate (children) per 1,000** people is **204**, and the **death rate (elderly)** per 1,000 people is **101**.

**Recommendation**

**(a) Development of Unoccupied Plot of Land**

Based on the analysis of census data, the suggested development for the unoccupied plot of land would be a **Train Station**.

**Justification:**

**Commuter Presence:** The Data analysis indicates a significant portion of the population as possible commuters. The establishment of a train station offers an efficient transportation option, potentially alleviating road congestion.

**Employment and University Students**: With a considerable percentage of the population categorized as employed or university students, the presence of a train station can facilitate commuting for work or educational purposes.

**Population Growth:** While age distribution analysis hints at a diverse population, the presence of commuters suggests potential town growth. A train station aligns with the needs of the growing population.

**(b) Investment decision**

The primary investment priority should be to **increase spending on schooling**.

**Justification:**

**Growing Population of School-Aged Children**: The analysis revealed a diverse age distribution, with a significant portion of the population falling within the school-aged category. This finding suggests a growing population of children.

**Educational Needs:** Investing in schooling and education becomes imperative to accommodate the needs of the younger population and provide adequate facilities for their development.

**Future Demands:** Addressing the educational requirements aligns with the potential growth of the town, ensuring that the infrastructure is in place to support the educational needs of the community.

**In conclusion,** establishing a Train Station on the unoccupied plot of land and prioritizing an increase in spending for schooling are recommended based on the insights derived from the census data analysis. These choices aligned with the needs and trends identified in the dataset.

**Reference**

NSPCC 2023, 'Children and the law', NSPCC Learning, 20 December 2023, <https://learning.nspcc.org.uk/child-protection-system/children-the-law>

Ronish, R. 2023, 'Difference Between Infirmity and Sickness', Medium, 4 October 2023, <https://medium.com/@ronishrh5/difference-between-infirmity-and-sickness-9471b93c6d4d>