CENSUS PROJECT REPORT

MSc in Artificial Intelligence and Data Science

OSAGIE ELLIOT AIBANGBEE

STUDENT NUMBER: 202115576

ABSTRACT

This is a report on the cleaning and analyses carried out on the census data of a specific town of 7,581 people, to help make an informed decision on investing some government resources (a vacant plot of land, and some available public funds).

The first chapter dwells on the approach and steps taken to clean the dataset, and setting it up for robust analyses. Steps include fixing missing data, imputing blank text, casting features to more appropriate data types, getting rid of redundant categories, and implementing consistency in the data format.

The following chapters cover the analyses done to unravel the insights used to develop the basis for the given recommendations. This

OBJECTIVES

- Discover relevant insights through data analysis.
- Reach a data-driven decision based on the derived insights.
- Recommend to the local government the most beneficial way to invest the available public resources (vacant plot of land, and funds).

CHAPTER 1

DATA MUNGING

The original census dataset is a CSV (comma separated values) file containing 7,581 records and 11 features. Each of its feature had a string datatype. Also, missing values, blank text, inconsistent values, and multiplicity of the same categorical values were noticed in some features. Steps taken to clean the data set (feature-by-feature) can be found in a Jupyter Notebook file solely dedicated to data cleaning.

Use of a customized module

In cleaning the dataset, I created a module "census_methods.py" consisting of a class of methods to automate some repetitive coding tasks. This module was imported in the Notebook file where its methods were used. For instance, one of the methods, "check_for_empty_str", which accepts only one parameter (a pandas DataFrame) helped me to get the number of rows per column containing blank text. Please refer to module documentation for more details.

Missing values and blank text

Features where a nan value was intentionally used for a specific category had these changed to another placeholder with a clearer meaning. For instance, the religion and marital status of minors (aged below 18), were changed from nan to 'Ineligible'.

Imputing unintentional nan values and blank text, factors considered include:

a. Using another feature: where it revealed what ought to have been there. For example, first name nan values were easily discovered from their corresponding surnames. While house number nan values were discovered by looking at house numbers of residents having combinations of surnames and street that is the same as those having a nan house number.
b. Checking the living situation: was done after merging the house number and street features into an address feature, and then looking at residents sharing the same address(es). Considering the nature of census dataset, this new feature proved to be invaluable in a number of ways. For instance, a blank surname was easily rectified after looking at the surname of members of that household. Kevin Jones-Lewis' living situation revealed that he is an 18-year-old living with his grandparents and cousin, and thus his marital status was changed from nan to 'Single'.
c. Replacing with a more meaningful placeholder instead of nan or blanks: this was mostly done to keep my interference with the data as minimal as possible. For example, adults (aged above 18 years) with a nan religion, and residents with blank infirmities were categorized as 'Unknown' and 'Undisclosed' respectively.

Multiplicity of categories

Deafness and blindness are forms of physical disability (carehome.co.uk, 2020), and were changed to the latter. The major religions of the world found in the dataset include Christianity, Islam, Judaism, Bahaism, Sikhism, and Agnosticism (The Best Schools, 2021). And Catholic, Baptist, and Methodist are Christian denominations (BBC, 2021) and were categorized as Christians under Major Religions. Jedi was declared by just one resident out of the total population, and was then categorized (together with Private, Undecided, and None) as Unknown.

Data inconsistency

Gender values were changed to either F (for females) or M (for males). And duplicate marital status categories values were changed to a single new category.

Reformatting data types

House number and age features were eventually changed from string to an integer data type. This was after converting values from spellings to figures, and rounding up decimal figures to whole numbers.

Imputing based on research facts

In over 70% of relationships, the males are older than their female partners by an average of 3 to 5 years (OnAverage). Thus, age of 131 years was changed to be 5 years below her husband's. And age -1 was changed to 1 after considering the living situation.

STATISTICAL OVERVIEW OF THE DATASET

DESCRIPTIVE STATISTICS

After cleaning the dataset, there are now 7,581 rows in each column with some additional features (Address, Is Retired, Is Student, Is Unemployed, Major Religions) all in their appropriate data types as shown below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7581 entries, 0 to 7580
Data columns (total 16 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   House Number                  7581 non-null   int64
 1   Street                        7581 non-null   object
 2   First Name                    7581 non-null   object
 3   Surname                       7581 non-null   object
 4   Age                           7581 non-null   int64
 5   Relationship to Head of House 7581 non-null   object
 6   Marital Status                7581 non-null   object
 7   Gender                        7581 non-null   object
 8   Occupation                    7581 non-null   object
 9   Infirmity                     7581 non-null   object
 10  Religion                      5864 non-null   object
 11  Address                       7581 non-null   object
 12  Is Retired                    7581 non-null   int64
 13  Is Unemployed                 7581 non-null   int64
 14  Is Student                    7581 non-null   int64
 15  Major Religions               7581 non-null   object
dtypes: int64(5), object(11)
memory usage: 947.8+ KB
```

	count	unique	top	freq
Street	7581	105	Wharf Wells	314
First Name	7581	369	James	36
Surname	7581	623	Smith	204
Relationship to Head of House	7581	20	Head	2879
Marital Status	7581	5	Single	2596
Gender	7581	2	F	3946
Occupation	7581	1099	Student	1382
Infirmity	7581	6	None	7540
Religion	5864	14	None	2623
Address	7581	2878	27, Brightwater Drive	22
Major Religions	7581	8	Christian	3015

	count	mean	std	min	25%	50%	75%	max
House Number	7581.0	26.807413	25.258199	1.0	8.0	20.0	37.0	150.0
Age	7581.0	37.210790	22.035577	0.0	19.0	36.0	53.0	119.0
Is Retired	7581.0	0.101306	0.301753	0.0	0.0	0.0	0.0	1.0
Is Unemployed	7581.0	0.070571	0.256124	0.0	0.0	0.0	0.0	1.0
Is Student	7581.0	0.248912	0.432411	0.0	0.0	0.0	0.0	1.0

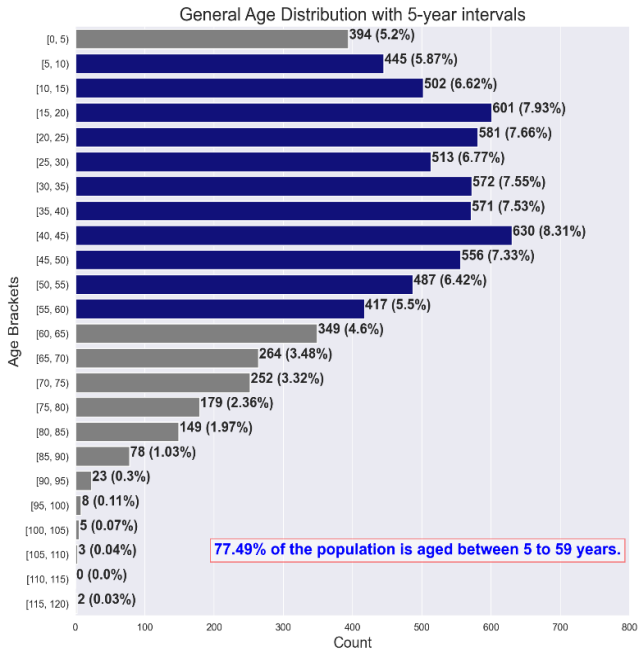
The table above tells us:

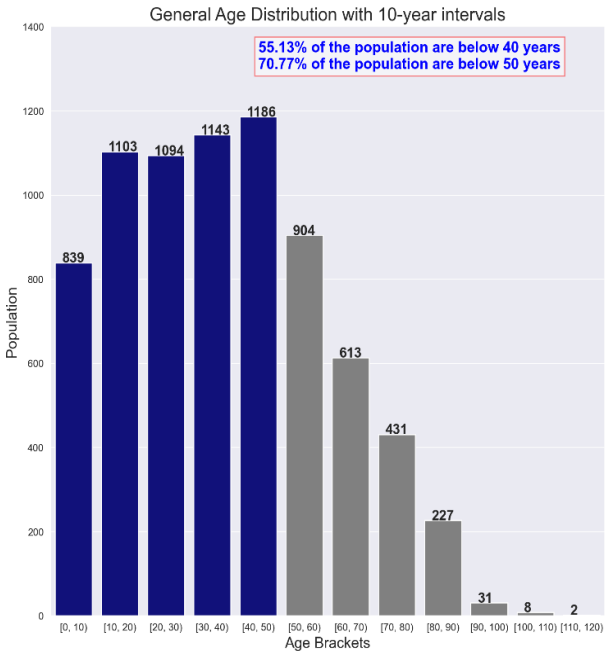
a. Street with the most houses is Wharf Wells (105 houses).
b. House with the highest number of occupants is 27, Brightwater Drive (22 occupants).
c. Single is the most common marital status (2,596).
d. Females (3,946 of them) are more than males.
e. Population is generally healthy (99.5%).
f. Christianity is the major religion, practiced by 3,015 residents (40%).

DETAILED ANALYSIS OF RESULTS

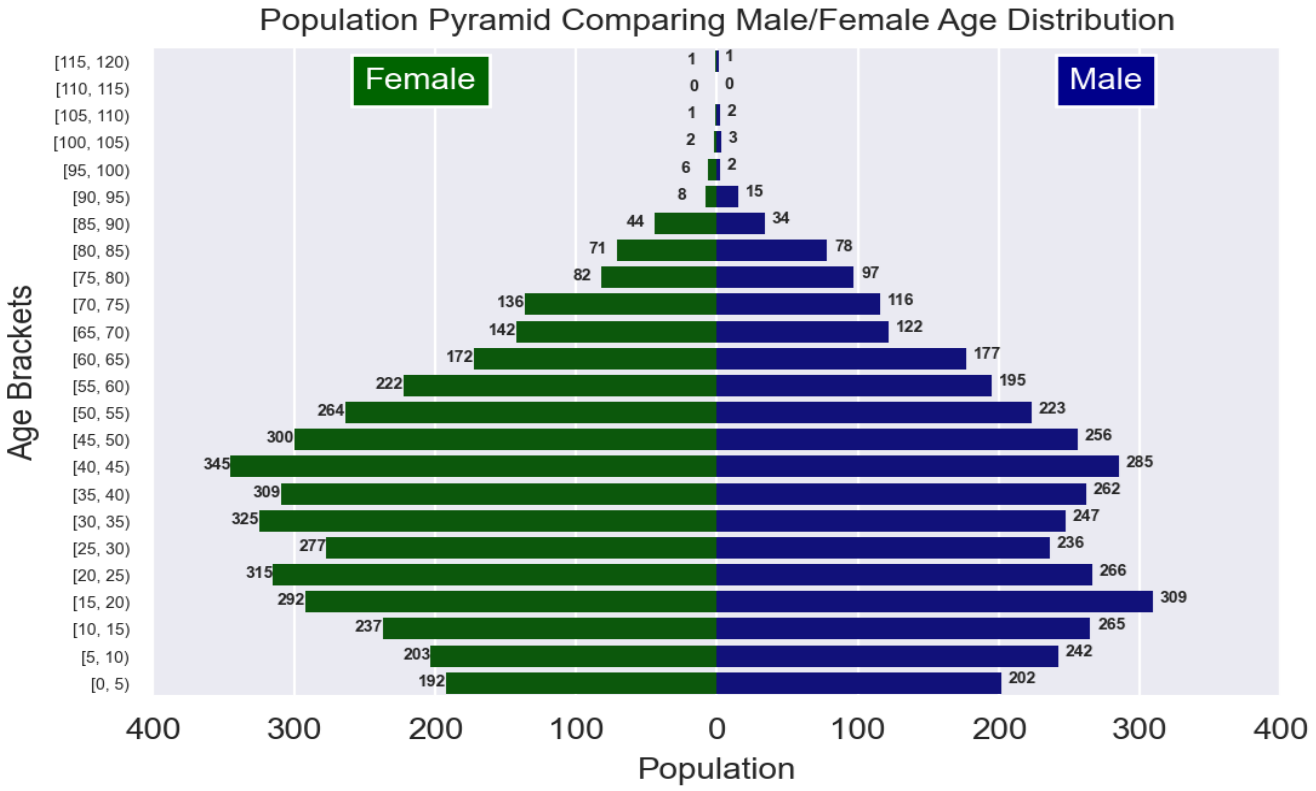
1. AGE DISTRIBUTION

Average age of the population is 37 years (37 years for males, and 38 years for females). The diagram below illustrates that this is a young population with over 70% still under 50 years. Over half the population (55.13%) are under 40 years old, which suggests an active and growing population to contribute economically to the town's development.





The population pyramid below illustrates, on the overall, a higher number of females than males. 15 to 19 years is the age bracket that has the highest male population, and ages 40 to 44 years has the highest female population. It can be said that the male population is younger than the females.

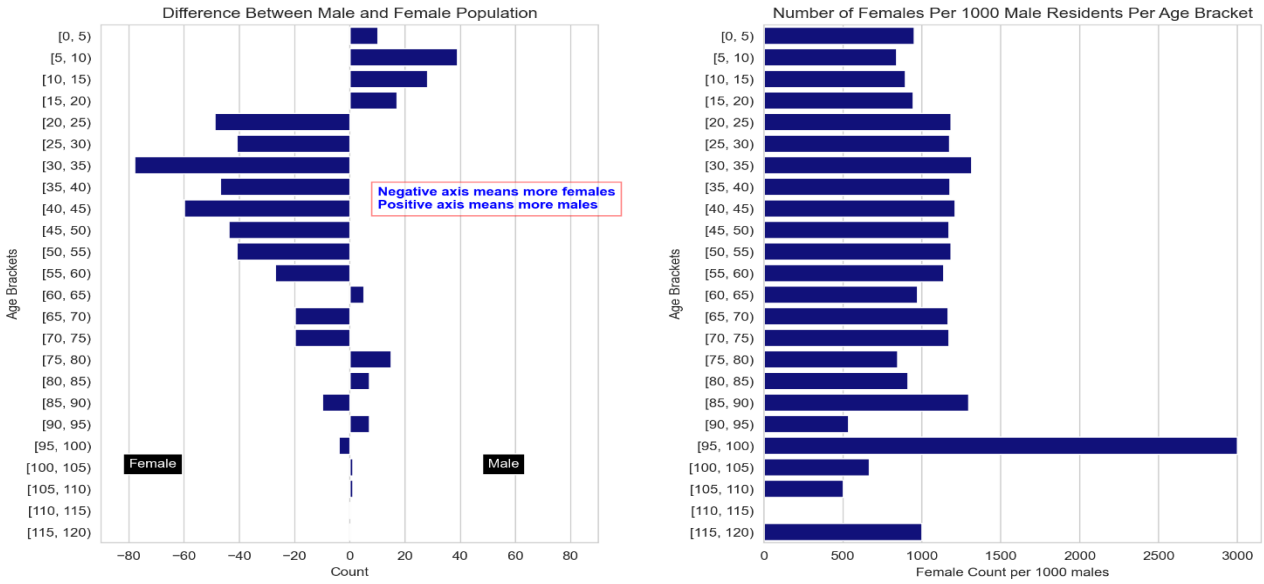


The illustration below reveals that from age 0 to 19 years, males are more than females. While from ages 20 to 59 years, females are visibly more. This could mean that females tend to live longer than males in this town, or that males tend to move to another town (emigration) as they grow older maybe for greener pastures. This should guide us when making gender-based decisions across age groups.

Ratio of males to females per 1000 residents:

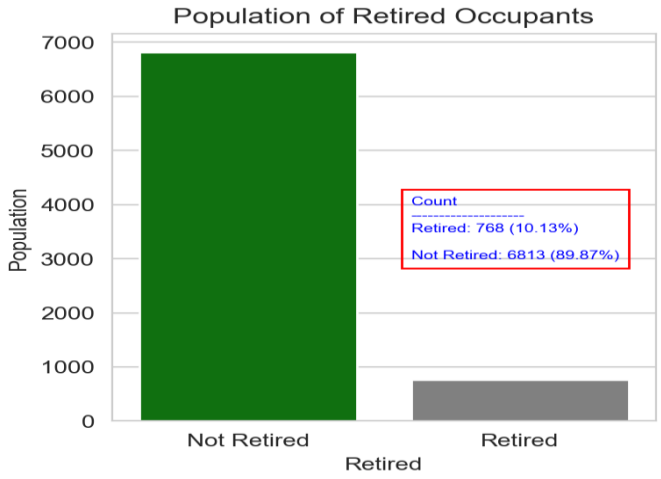
= 1000 * (male population/female population)

For every 1000 females, there are 921 males, showing that females are slightly more than males. The plots below clearly shows the difference in population between males and females by age.

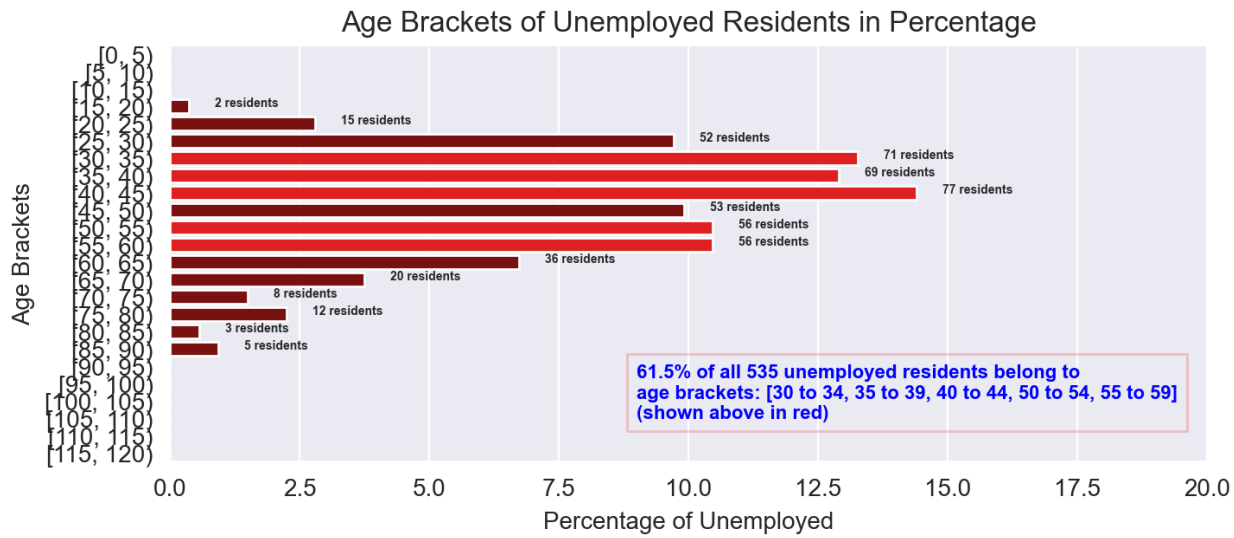


RETIRED

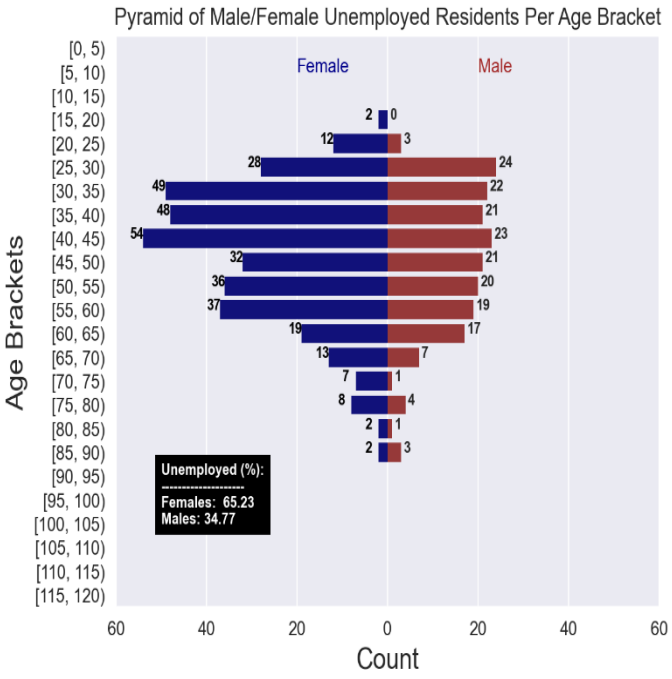
10% of the population (out of 45% who are above 40years old) are already retired, leaving the remaining 35% aging towards retirement. Depending on death rate, attention should be given to healthcare sector to cater for this aging category as they approach retirement.

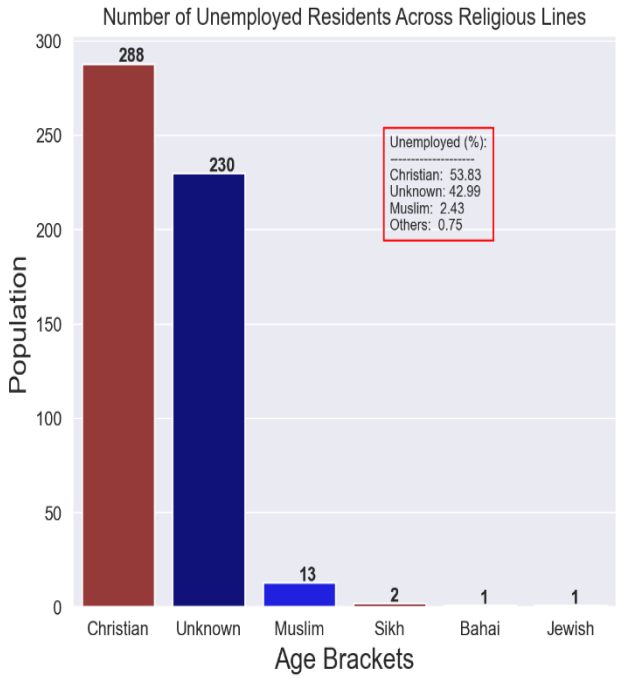


2. UNEMPLOYMENT TREND



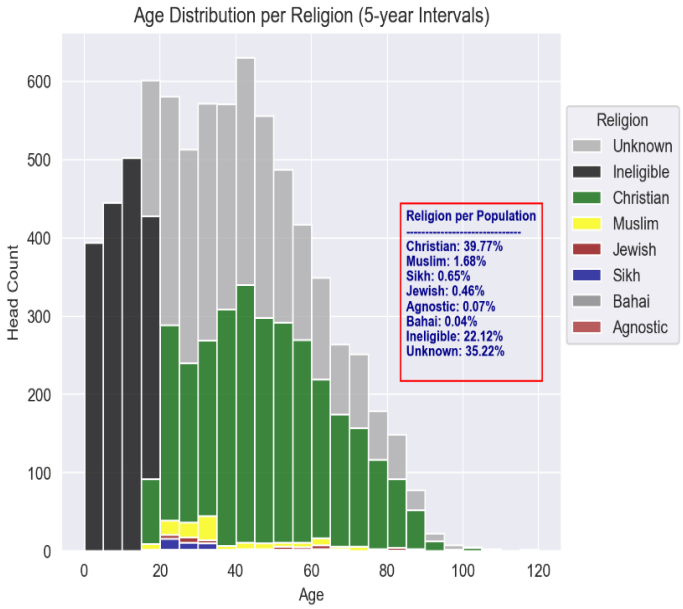
Technically, 535 residents registered as unemployed, but a closer look into their age distribution showed that 41 of them are 65 years or over. Overall, only 7% of the population is unemployed, indicative of an outstanding rate of employment and wealth generation. This also suggests that the economy is booming, and a low level of insecurity. Residents are mostly capable of quality healthcare and diet, which may be one of the reasons for the extremely low level of infirmity. However, it was observed that 53% of unemployed residents are between 25 and 44 years old.

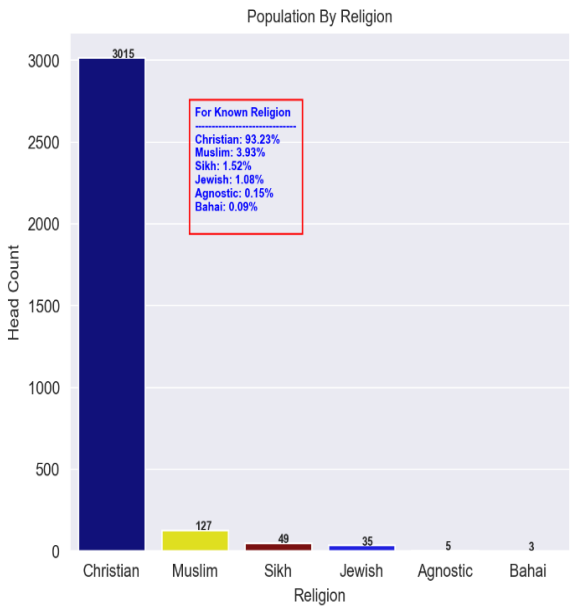




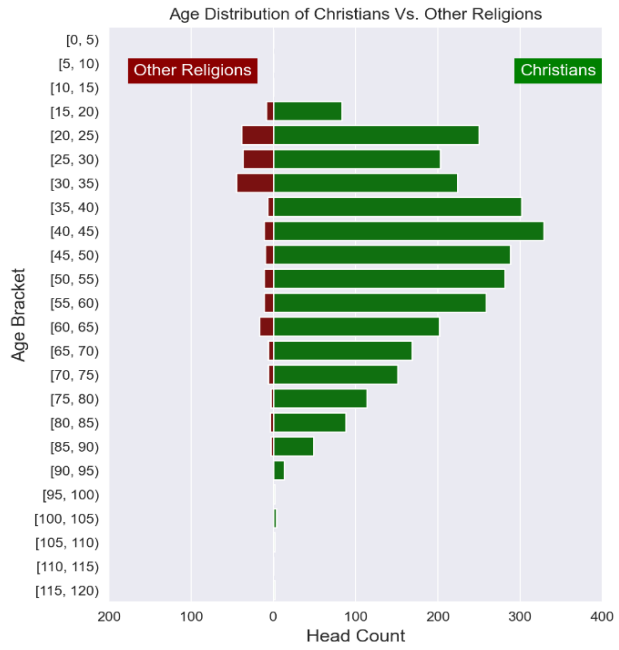
10% of the entire Muslim population are unemployed, which means that religion might be a factor for employment. One of the challenges that goes with being Muslim in a predominantly Christian society. The number of unemployed females is double that of males. This is a reasonable expectation given that a high number of females can be found between ages 20 to 59 years. 30 to 44 year-old females are most likely to be unemployed.

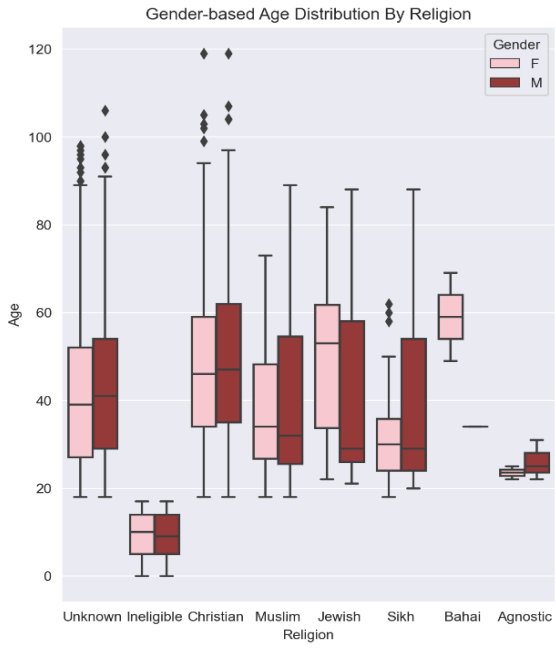
3. RELIGIOUS AFFILIATIONS





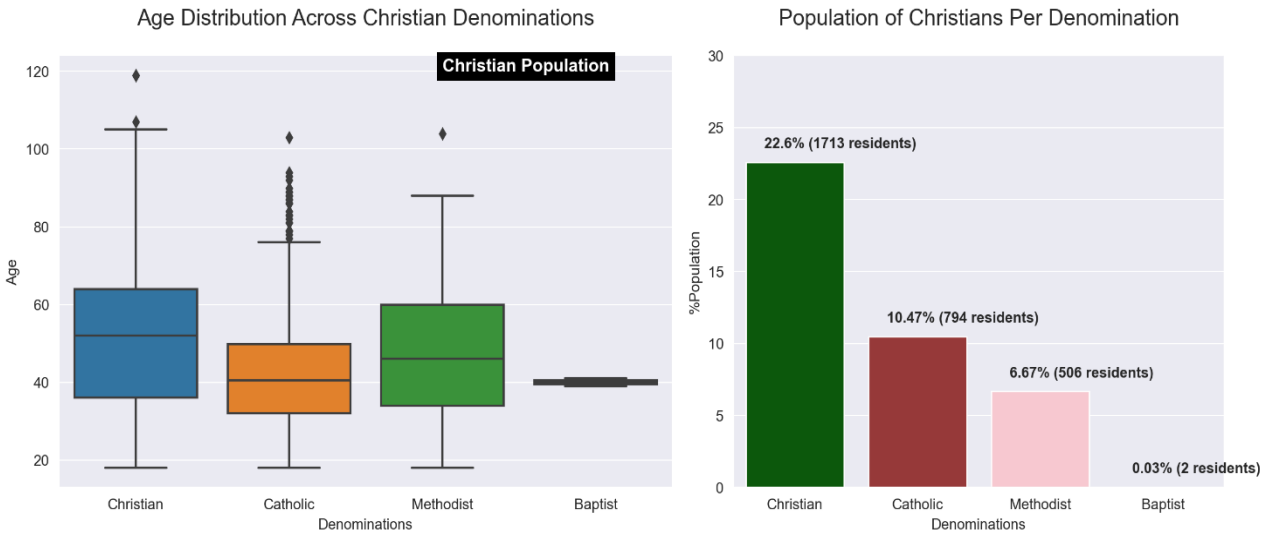
Known Christians actually make up 40% of the entire population, while followers of other religions are only 3% of the population. Excluding minors and residents without a religion, 93% of residents registered as Christians. Christianity is clearly the predominant religion in town. Therefore, it is expected that many of the laws and policies in place would be strongly influenced by Christian teachings and values.





The above diagrams make it obvious that a significant number of followers of other religions are bet ween the ages of 18 and 34 years, implying a growing population. In fact, 68% of followers of other religions are under 45 years, compared to only 46% of Christians being under 45 years. Maybe this i s due to the younger generation being more willing to explore other modes of worship.

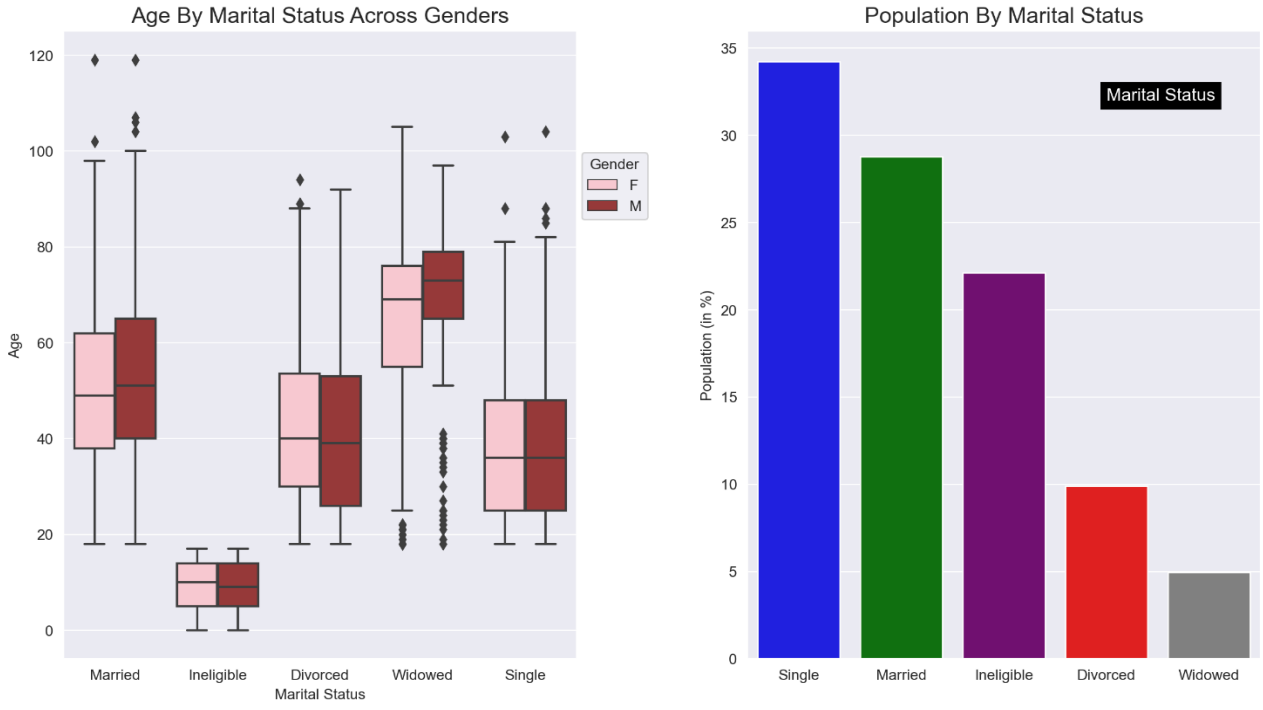
CHRISTIANITY



Among Christians, Catholics are the youngest with some followers above 78 years being outliers. While Methodists are the oldest. Known Catholics (10% of the population) have their own church, but Methodists, Baptists (7% of the population), and other Christian denominations (22% of the population) do not have a church and would have to commute to nearby towns regularly for fellowship. Same goes for all non-Christian worshippers (3% of population). Meaning, 33% of the population have to commute to their places of worship.

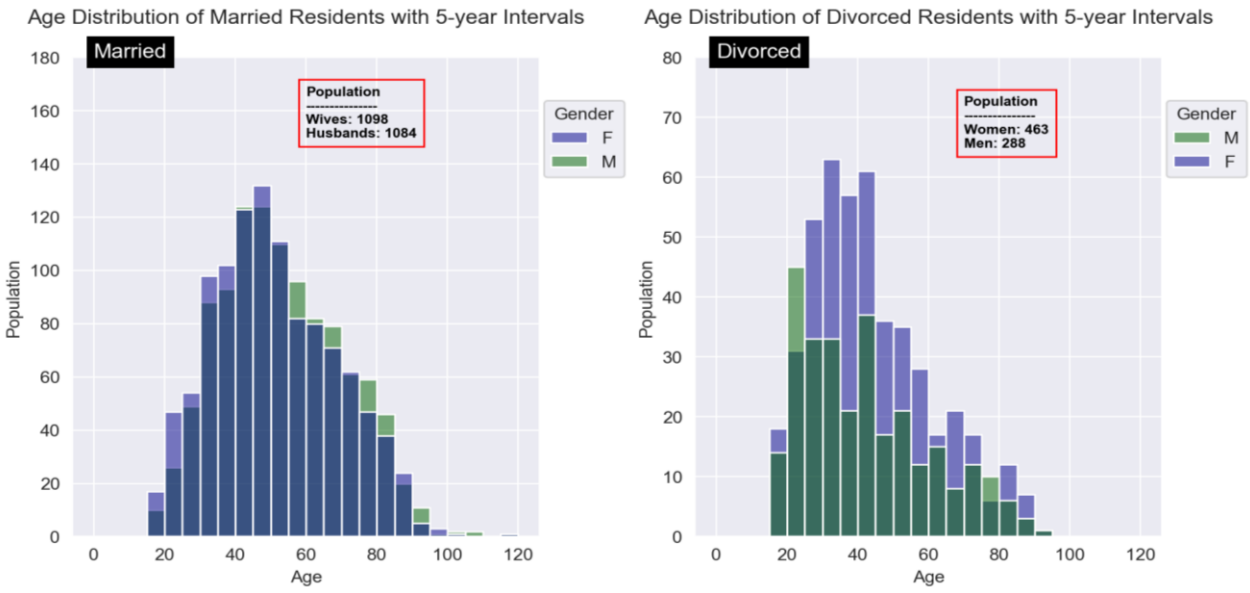
4. MARITAL STATUS

Diagrams below illustrate, based on each median values, that married residents are older than divorced and single residents. Widowed residents are clearly the oldest among them all. However, there seems to be some outliers. Several male and female widows are abnormally young. Those widows below 20 years might need further investigation to tackle any attempt to falsify facts for personal gains.

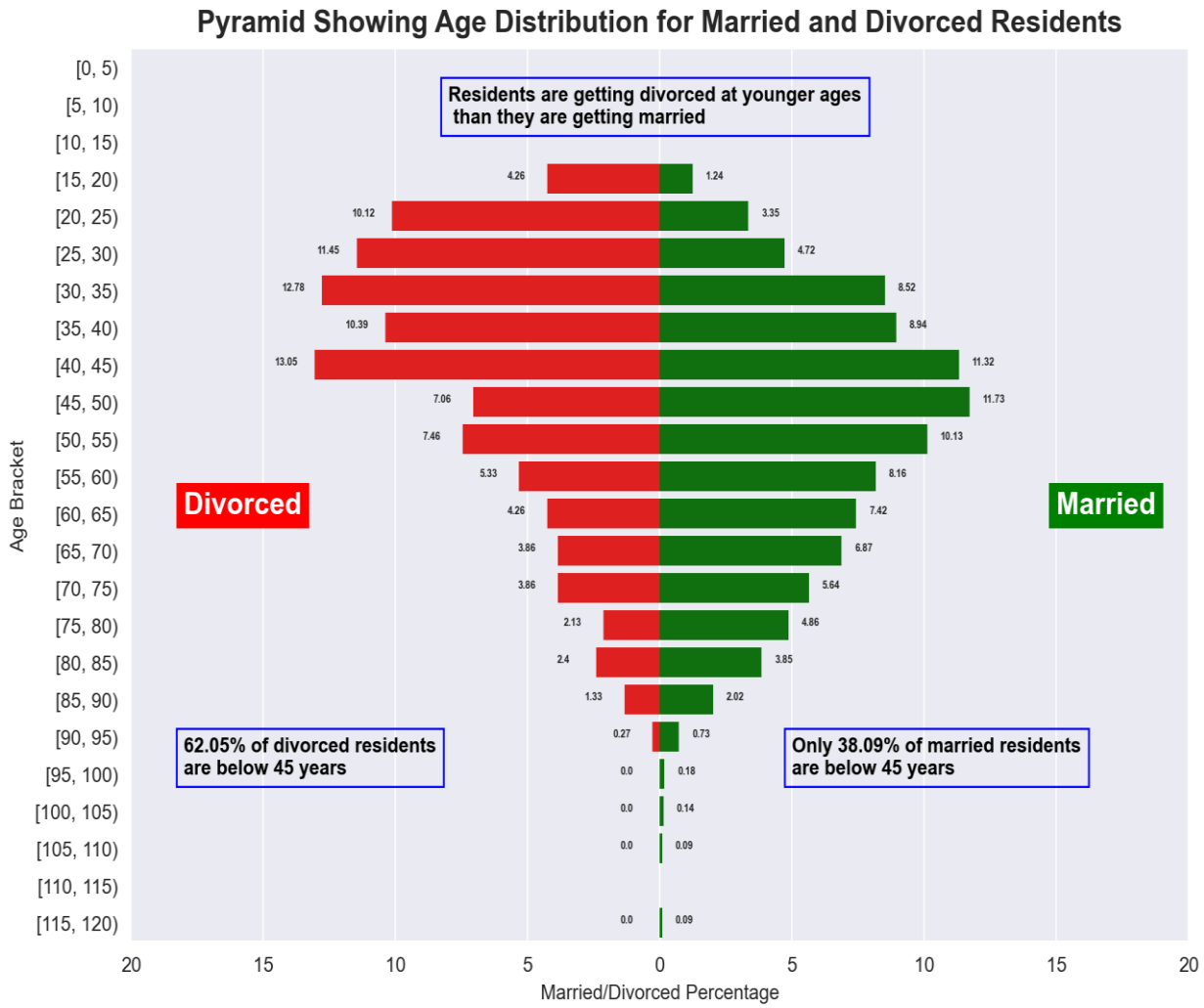


MARRIED VS. DIVORCED

Generally, there is numeric balance between the population of wives and husbands.



Among 18 to 54 year-old residents, wives are more than husbands, but for 55 years and older, there tends to be more husbands than wives. This implies that the tendency to divorce is lower for men and men tend to stay in marriages as a result. Looking at the age distribution of divorced residents, it is evident that there is a significantly higher number of divorced women than men. The number of divorced men who are between 20 to 25 years is higher than the women within that age bracket.



The above pyramid shows that husbands tend to get divorced at earlier ages before 45 years (62%) and wives tend to get divorced at 45 years and above (62%).

Married Rate

Crude Marriage Rate:
CMR = (Number of marriages per year / Total population) * 1000
Number of marriages per 1000 resident

It is assumed that people tend to get married between 25 and 29 years. Thus, the number of marriages this year (newly-weds) would be the average of the total number of wives (54) and husbands (49) between 25 and 29 years. Therefore, 7 marriages tend to occur per 1000 residents annually.

Divorced Rate ¶

Crude Divorced Rate:
CDR = (Number of divorces per year / Total population) * 1000
Number of divorces per 1000 resident

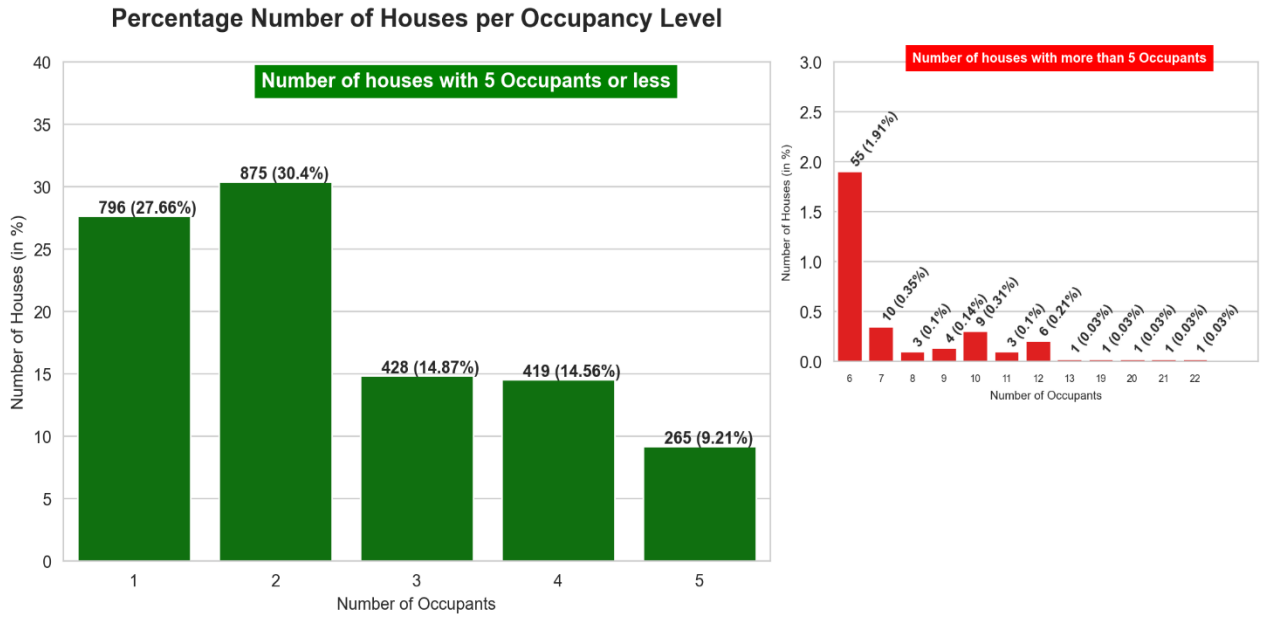
Divorce is most likely to occur between 45 to 49 years for men and women (Crisp & Co Solicitors, 2021). Thus, the number of divorces this year would be the average of the total number of divorced men and women between ages 45 to 49 years. Therefore, 3 divorces tend to occur per 1000 residents annually

5. OCCUPANCY LEVEL

The town has 2,878 houses on 105 streets for a population of 7,581. Houses are occupied at rate of 2.6 (approximately 3) occupants per house, and there are 380 houses available per thousand residents.

The diagram below shows that:

a. 796 houses are reportedly has just 1 occupant each. Meaning, 796 residents (10% of the population) live in 796 houses (28% of available houses).
b. Another 875 houses (30% of available houses) each have 2 occupants (1,750 residents – 23% of population).
c. 428 houses (15% of available houses) each have 3 occupants (1,284 residents – 17% of population).
d. 419 houses (15% of available houses) each have 4 occupants (1,676 residents – 22% of population).
e. 265 houses (9% of available houses) each have 5 occupants (1,325 residents – 17% of population).
f. The remaining 95 houses (3% of available houses) are occupied by 750 residents (10% of population)



The occupancy level of the town implies its residents are mostly wealthy or financially comfortable, which is one of the benefits of its low unemployment rate. Half of the population live in a house having a maximum of 3 occupants, which amounts to 50% of residents using 73% of the available houses (2,099 houses). While the remaining half of the population only have 27% of houses available for them to share. The town seems to be meant primarily for high class living.

6. COMMUTERS

Commuters are university students and non-Catholic worshippers who have to travel out of town regularly for fellowship. The table below shows the number and percentage of the population who have to commute frequently for religious purposes. Non-Catholic Christians alone are about 30% of the population. This is 33% of the population.

```
Christian     1713    22.60
Methodist      506     6.67
Muslim         127     1.68
Sikh            49     0.65
Jewish          35     0.46
Agnostic         5     0.07
Bahai            3     0.04
Baptist          2     0.03
Name: Religion, dtype: int64
```

248 university students (3% of the population) have unknown religions, and there are 30 university employees. Making the total commuters 36% of the population

7. BIRTH RATE VS DEATH RATE

From statistical computation, there are 8.97 births per thousand residents. The previous birth rate was 11.34, implying a drop of 2.37 births per thousand residents. Age specific fertility rate for women between 25 and 29 years was 245.49; 264.62 for women between 30 and 34 years; 281.55 for 35 to 39 years. This proves that there is a drop in birth rate of the population.

Town's death rate is quite low at 15.3 deaths per 1000 resident. This shows the population has a high life expectancy.

CHAPTER 4

RECOMMENDATION

After detailed analysis, I recommend that a multi-purpose worship hall should be built on the vacant plot of land to serve as worship center for Methodists, Muslims, Sikh, Jews, and other non-Catholic followers. Worship days and time for each of the major religions differ. For instance, worship days for Muslims is primarily Friday, for Methodists and Baptists it's Sunday (but they can do so at different times), while Jews primarily worship on Saturdays. Proper attention should be given to management and allocation of this important building, to ensure there is fairness and equity. And representatives from each religious parties should form part of this management. This can be an avenue to promote unity and religious tolerance in the town. This should be also boost the economy as more businesses opportunities will spring up in and around the worship center.

Given the population is relatively young and healthy, 16% of the population are 60 years or older. And 35% are already above 40 years. Thus, I recommend investing in old-age care, like care homes, exercise centers for the ageing and elderly.

REFERENCES

Oxford Learner's Dictionaries (No Date)

Available Online: https://www.oxfordlearnersdictionaries.com/definition/english/census?q=census [Accessed 6/12/2021]

Carehome.co.uk (2020) *Types of Physical Disabilities: August, 2020*

Available Online: https://www.carehome.co.uk/advice/types-of-physical-disabilities [Accessed 7/12/2021]

The Best Schools (2021) *18 Major World Religions – Study Starters: September, 2021*

Available Online: https://thebestschools.org/magazine/world-religions-study-starters/ [Accessed 7/12/2021]

BBC (2021) *God and Authority in Christianity – Branches of Christianity (No Date)*

Available Online: https://www.bbc.co.uk/bitesize/guides/zbj48mn/revision/8 [Accessed 31/11/2021]

OnAverage (No Date) *Average Age Difference Between Couples*

Available Online: https://www.onaverage.co.uk/age-averages/average-age-difference-between-couples

[Accessed 4/12/2021]

Crisp&Co Solicitors (2021) *Divorce Statistics 2021: 2021*

Available Online: https://www.crispandco.com/site/divorce-statistics/#:~:text=Women%20in%20opposite-sex%20marriages%20are%20also%20most%20likely,to%20cost%20the%20taxpayer%20%C2%A348%20billion%20per%20year. [Accessed 7/12/2021]