**ETL Report - Team Number:** 5

**Team Members:**

1. Joanna Gu
2. Ian Lim
3. Nelson Javier Garcia
4. Babette Blanquet

**The Victorian population and the projections of households until 2036**

1. **Project overview**

In this project, our goal is to build a database to understand the correlation between current population growth and projections of number of households in Victoria.

Our dataset will provide data at the Statistical Area 2[[1]](#footnote-1) (SA2) level which is equivalent to a suburb. For reference, there are 462 suburbs or SA2 in Victoria.

With the new dataset, we will be able to create new queries to answer questions such as:

* What’s the current population growth in each suburb (SA2)?
* Is there a correlation between population growth and population density in each SA2?
* Is there a correlation between household projections and current population growth?
* Is there a correlation between household projections and the type of migration?
* Which suburbs or SA2 will be with the highest population density in 2036?
* What Victoria will look like in terms of population density in 2036?

This analysis will be useful to forecast the needs of schools, aged care facilities, amenities, number and types of dwellings to service each area at Sa2 level but also SA3 and SA4 which are greater areas.

1. **Extract the data sources**

* Dataset 1:

Regional Population Growth 2018-2019

Source: Australia Bureau of Statistics

This dataset provides the data outlined below for Statistical Areas from Level 2 to Level 4 (SA2, SA3, SA4), Greater Capital City Statistical Areas (GCCSA) and Victoria:

* + The estimated population in 2018
  + The estimated population in 2019
  + The components of population change
    - The natural migration
    - The internal migration
    - The overseas migration
  + The size of each area (km2)
  + The population density in 2019

This dataset is in Excel format.

**Extract method:** this dataset has been downloaded from its source and the Table 2 saved as CSV.

<https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3218.02018-19?OpenDocument>

* Dataset 2:

Victoria in Future 2019 (VIF2019) Population and Household Projections – Table “Household Types”

Source: Planning Victoria

This dataset provides the number of households by household type for Statistical Areas from Level 2 to Level 4 (SA2, SA3, SA4), Greater Capital City Statistical Areas (GCCSA) and Victoria:

* + The estimated number of household types in 2016, 2021, 2026, 2031, 2036

With the breakdown below:

* + - Couple family with children
    - Couple family without children
    - One-parent family
    - Other family
    - Group household
    - Lone person
    - All Household Types

**Extract method:** this dataset has been downloaded from its source and the relevant table saved as CSV.

<https://www.planning.vic.gov.au/land-use-and-population-research/victoria-in-future>

1. **Transform: the cleaning, checks and Exploratory Data Analysis process**

**Dataset 1: The Regional Population Growth 2018-2019 from ABS**

The transformations have been applied in a Jupyter Notebook named: (“ABS\_population\_growth18-19\_Transformation.ipynb”)

Basic transformations applied:

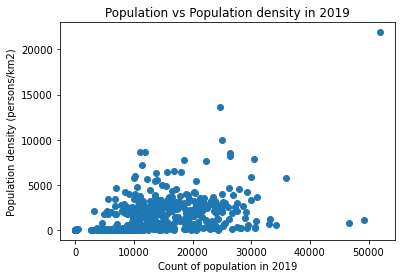
* Setting the row containing the column names as the new header
* Renaming some columns with iloc as 6 columns appeared under the same name ‘no.’
* Removing ‘junk’ rows at the top and bottom of the table using the iloc method
* Selecting the columns needed for our database

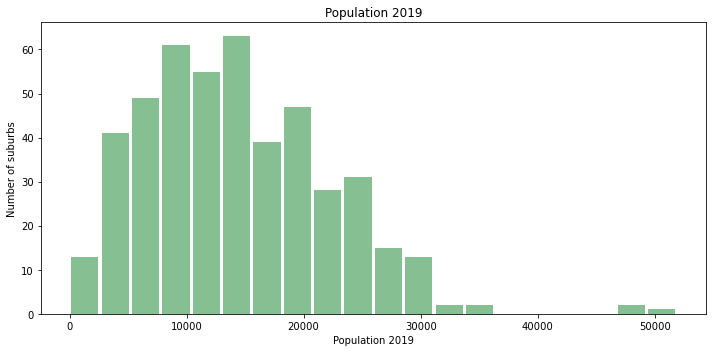
Checking the quality of the data:

* Dropping any NULL value and checking the count of rows
* Checking the type of data
* As they are all string/object, we converted the relevant columns to integer and float

Exploratory Analysis:

* Checking the min and max values with the describe() method
* The min value being at 0 for the population 2018/19 and the population density, we will sort the dataframe to understand the reasons some suburbs might have no population.
* We sorted the dataframe to identify the 10 suburbs with the lowest population in 2019. The suburbs are for example Essendon Airport or Wilsons Promontory. It makes sense people work or walk there but not many live in these suburbs.
* We looked at the correlation between population 2019 and the population density 2019 to ‘map’ all suburbs.



* We can identify a few outliers – so we checked to see if the suburbs with the highest density were relevant. All suburbs with a high population density are the closest to Melbourne CBD and are relevant.
* Lastly, we also plotted the suburbs based on their population in 2019 to visualise the data. Most suburbs have a population between 8000 and 20,000 inhabitants.

**Dataset 2: The Regional Population Growth 2018-2019 from ABS**

The transformations have been applied in a Jupyter Notebook named: (“Household\_types\_Transformation.ipynb”)

Basic transformations applied:

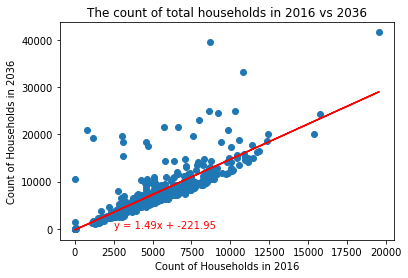
* Setting the row containing the column names as the new header
* Removing ‘junk’ rows at the top and bottom of the table using the iloc method
* A breakdown of the dataset of 45 columns into five smaller datasets - each will contain the households projected at different years: 2016, 2021, 2026, 2031, 2036
* A new column Year/Date has been added to each dataset
* Then a master file has been created by concatenating the five small datasets.
* The master file contains 2310 rows - which includes the count of households in 462 suburbs for five different years.

Checking the quality of the data:

* Checking the type of data: most of the data is in string format although seven columns should be integers to be able to perform aggregations.
* Removing commas from numbers (e.g. 1,675 to 1675) to be able to convert them as integers.
* Dropping Null values
* Dropping eventual duplicates – it happens the dataset had no null value or duplicate

Exploratory Analysis:

* Checking the min and max values with the describe() method
* The min value being at 0 we will group and sort the dataframe to understand the reasons some suburbs might have 0 households.
* There are six suburbs at 0 households, but they are either national parks or airports. Therefore, the data is relevant.
* We compared the number of total households in 2016 versus the projections in 2036 to check if the data is relevant with a linear regression model.



* The linear regression shows most suburbs are expected to grow by 1.5 over 20 years but quite a few suburbs will grow at much higher rate (outliers).
* To verify the quality of the data we wanted to identify these suburbs seen as outliers
* We calculated the growth rate between 2016 and 2036 and sorted the data to identify the suburbs with the highest growth.
* We can identify the main suburbs Planning Victoria is expected to develop intensively over the next 20 years – such as Port Melbourne Industrial Flemington Racecourse and Rockbank- Mount Cottrell.

Conclusion: the quality of the data is good and can be loaded in the database.

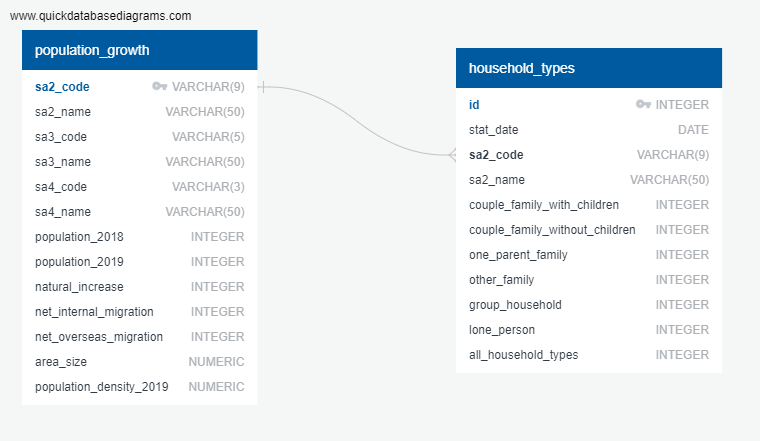
The master file has been saved as csv named “[master\_households\_types.csv](http://localhost:8888/edit/OneDrive/Documents/Data_Analytics/GitHub/ETL_project/2-%20CSV/master_households_types.csv)”

1. **LOAD - Data storage**

We have chosen to load the data and store them in relational database as the data is more structured and a relationship between the datasets can be established.

We have identified SA2 (Statistical Area 2) code as the foreign key to join the datasets. We can then perform queries joining the data from the two tables.

Please see Entity Relationship Diagram (ERD) below depicting the relationship between the two tables from the datasets.



**Queries used after loading to database:**

Query all rows in each table: Select \* from households\_type;

* Query household growth for a sa2 area for each forecast period:
  + Select stat\_date, sa2\_code, sa2\_name, couple\_family\_with\_children from households\_type where sa2\_code = '210051445';

We can also perform query joining information from the 2 tables such as the one below:

SELECT

p.sa3\_code,

p.sa3\_name,

p.sa2\_code,

p.sa2\_name,

p.area\_size,

p.population\_density\_2019,

h.stat\_date,

h.couple\_family\_with\_children,

h.couple\_family\_without\_children

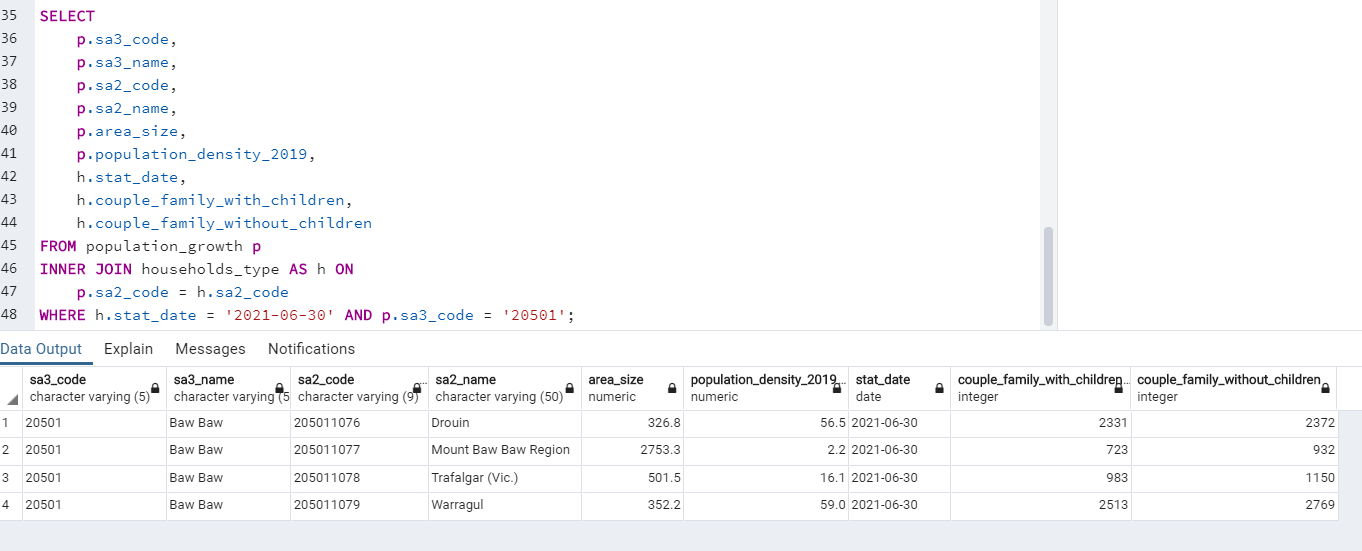
FROM population\_growth p

INNER JOIN households\_type AS h ON

p.sa2\_code = h.sa2\_code

WHERE h.stat\_date = '2021-06-30' AND p.sa3\_code = '20501';

See below print screen of the results.



1. More information on the SA2: The spatial units used in these projections are based on the Australian Statistical Geography Standard (ASGS) where Statistical Areas Level 2 (SA2s) are designed to reflect functional areas that represent a community that interacts together socially and economically. They consider Suburb and Locality boundaries to improve the geographic coding of data to these areas and in major urban areas SA2s often reflect one or more related suburbs. The SA2 is the smallest area for the release of many statistics. Source: <https://www.abs.gov.au/websitedbs/d3310114.nsf/home/australian+statistical+geography+standard+(asgs)> [↑](#footnote-ref-1)