**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 was to build a database to understand the correlation between current population growth and projections of number of households in Victoria.

Our datasets provide data at the Statistical Area 2[[1]](#footnote-1) (SA2) level which is equivalent to a suburb. For your reference there are 462 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 (SA2).

1. **Extract the data sources**

* Dataset 1:

Regional Population Growth 2018-2019

Source: Australia Bureau of Statistics

This dataset provides 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.

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

* Dataset 2:

Victoria in Future 2019 (VIF2019) Population and Household Projections

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
    - 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.

<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

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

The transformations have been applied in a Jupyter Notebook untitled: (“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 files gather 2310 rows - which includes the count/projections of households 462 suburbs across five different years.

Checking the quality of the data:

* Checking the type of data: most of the data is string 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 households, but they are either national parks or airport. Therefore, we will keep these suburbs as 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 correct and can be loaded in the database.

The master file has been saved as csv untitled “[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**

The database chosen

An image of the ERD showing how the tables will be joined

An example of queries after the database is built.

The other types of queries we can perform

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)