**ETL Project Proposal**

**Team Number:** 5

**Team Name:**

**Team Members:**

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

**Project overview (200-300 words)**

For each Statistical Area 2 (SA2), we would like to know:

* their population growth
* the types of households growing
* the forecasts of number of households until 2036

This will be useful to forecast the needs of schools, aged care facilities, amenities, number and types of dwellings to service each area at a SA2 level.

**EXTRACT - Proposed data sources (minimum two data sources)**

*Where is the data located?*

*The second dataset has been provided by Planning Victoria.*

* *dataset 1 – the population in 2018/2019 in Victoria*

*The second dataset has been provided by Planning Victoria.*

* *dataset 2 – the household types projections from 2016 to 20136 in Victoria*

*What are the data set formats?*

* *The datasets are both in csv formats*
* *They are downloaded from the ABS website and Planning Victoria*

**TRANSFORM - Proposed clean up and analysis**

* *What are the transformations you will apply to the data?*
* Dataset 1 – Population in Victoria:
  + Keeping only relevant columns – SA2 code, SA2 Name, Population 2018, Population 2019, ERP change No., ERP change %, Population density 2019
  + Removing unnecessary rows tops and bottoms
  + Renaming the columns to load in Database
* Household types:
  + Keeping only relevant columns - SA2 code, SA2 Name, All Household types for 2016,2021,2026,2031,2036
  + Removing rows with SA3 SA4 data
  + Entering the years as a new column ‘Year”
  + With Pandas, creating 5 csv files for each year and then concatenating them to create a master file
  + Renaming columns before loading to database

*How will the data be integrated?*

* Integration with SQL – joins

*How will you apply these transformations?*

* Jupyter notebook Python Pandas

*IMPORTANT → Why did you apply these transformations? How did this enrich your data?*

We applied these transformations to be able to perform:

Group by year, Group by SA2 code, Compare population growth and projections with types of households projections.

**LOAD - Data storage**

*What type of database (relational, document) will you store the data?*

* *Relational with SQL*

*Why did you choose this database over another database?*

*SA2 code will be our PRIMARY KEY and the way we join the tables.*

*Expected tables:*

*Population growth*

*Household types*

**Potential limitations**

* *What are the potential limitations of your above proposed steps?*
* *How can you control these potential issues?*

#########################################################################

**ETL DIAGRAM**

It may be easier to explain the process in an ETL diagram (you don’t have to create a diagram but it may help your group to formulate a plan).

Most data engineers make sure to document the ETL processes for reference. One way to do this is to use ETL diagrams. This is usually drafted at the beginning of the project and finalised at the end of the project.

Here are a two examples of how an ETL diagram would look like:

A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated