**Extract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc)**

* The databases were sourced from the Australian Bureau of statistics and are in CSV/Excel format.

**Transform: what data cleaning or transformation was required**

Cultural diversity

TAKE OUT ELECTORAL DIVISION FOR PURPOSES OF NORMALISATION

*Diversity*

* Checked for datatypes. As expected, these were either floats (percentages) and one object (division name).
* Checked for duplicates and there are no duplicates in divisions and as such, no transformation required.
* Checked for missing values and no NAN values so no transformation required.
* Renamed columns to ensure consistency
* percentages show % of all australians that is culturally diverse? TBC
* Count rows - 151 rows

*Division*

* Created for purposes of merging with multiplate databases
* For completeness, I performed a check for duplicates of divisions (none) and datatypes (integer and objects, as expected) and as such, no transformation was required.
* I merged the diversity and division dataframes through a pd.merge function (inner join), and joined both dataframes on electoral division name.

*Merged dataframe*

* Checked datatypes (object for division name), float for all others, as expected)
* The column “division\_id” changed from a rounded number to a number with one decimal, but with still the same datatype (float). Based on research, it appears that you cannot convert a floating number to an integer when there are NAN values. As such, I counted the rows before and after the merge. Post merge (inner join), the number of rows went from 158 to 143, which is not consistent with the actual csv files as there are no missing values. To be investigated.

Age and population

*Age*

* Checked the columns headers
* Checked the datatypes and which are all as expected
* Counted the rows (151), dropped NAN values and counted rows again (151 rows). No NAN were present and as such, no transformation as required.

*Division*

* Created for purposes of merging with multiplate databases
* I merged the age and division dataframes through a pd.merge function (inner join), and joined both dataframes on electoral division name.

*Merged dataframe*

* I dropped the “state” column TBC
* I counted the rows post merge (158 rows)
* Division\_id changed from rounded number to number with one decimal, but with still the same datatype. Apparently, you can’t convert a float to an integer when there is NAN values. Did a len(df) check and after merge the number of rows went from 158 to 143! Odd as there are no missing values in the csvs.

*Population*

* I counted the rows (151)
* I merged the merged dataframe (as per above) and population dataframes through a pd.merge function (inner join), and joined both dataframes on electoral division name.
* I renamed age-related columns to make it more readable

*Merged dataframe 2*

* I counted the rows post merge (158 rows)
* I grouped by division and division\_id -> 150 rows.
* Difference in rows to be investigated.

**Load: the final database, tables/collections, and why this was chosen.**

* All of the above dataframes were exported to a csv file.
* These csv files are ready for analysis, for instance:
  + Cultural diversity per division. Are certain areas within Austraia more culturally diverse than others?
  + Cultural diversity database can be merged with federal election results in order to analyse the results / identify trends patterns on election results in electoral divisions with varying cultural diversity. Political parties can target their campaigns based on these results to gain more voters.
  + Population and age csv can be merged with federal election results or Australian marriage postal survey results: identify trends, for instance whether more densed divisions tend to vote for specific political parties or whether they tend to be more progressive towards equal marriage. Political campaigns can be adjusted/targeted based on these results.