**Project Report – ETL JobSeeker Challenges**

**Team:** Diane Yaneza and Mwansa Mwango

**Objective**

Source, prepare and create a database to aid the analysis of records from the Australian Bureau of Statistics (ASBS) for the period ending August 2020, to explore the characteristics and main difficulties faced by job seekers.

The compiled database would be beneficial for the following audiences:

- Jobseekers

- Employment Agencies

- Recruiters

**Extraction (E)**

***Data sources***

The data was selected and downloaded manually from the ABS website in MS Excel format (xls).

The data was sourced from the Australian Bureau of Statistics

* [Characteristics of Jobseekers](https://www.abs.gov.au/statistics/labour/employment-and-unemployment/participation-job-search-and-mobility-australia/feb-2020#seeking-a-job)
* [Qualifications of Successfully Employed](https://www.abs.gov.au/statistics/people/education/qualifications-and-work/2018-19#data-download)
* [Employment Vacancies](https://www.abs.gov.au/statistics/labour/employment-and-unemployment/job-vacancies-australia/nov-2020#data-download)

The Excel files were imported via Pandas Library for manipulation.

***Limitations***:

1. The available data was already aggregated and did not have individual survey records posing a challenge.
2. Additionally, the required data was present across multiple sheets and relevant data had to be imported accordingly.

**Transform (T)**

The following transformation were completed to clean and prepare the data for CSV format before uploading into PostgresSQL Database:

* Aggregated data records had to be split into respective tables by:
  + Age
  + Location
  + Gender
  + Main difficulty finding work – Too many applicants
  + Number of unemployed vs employed person
  + Number of Job Vacancies
* Unused columns and rows were deleted by drop relevant roles come as a result of headings, subheadings and footnotes present in the excel table
* In the Main difficulty table, other reasons were dropped to only keep the main reason given for not finding work – “Too many job applicants”
* Data from previous years was removed to only keep was relevant as at Aug 2021.

Each of the above subset tables were exported to CSV for loading into the database.

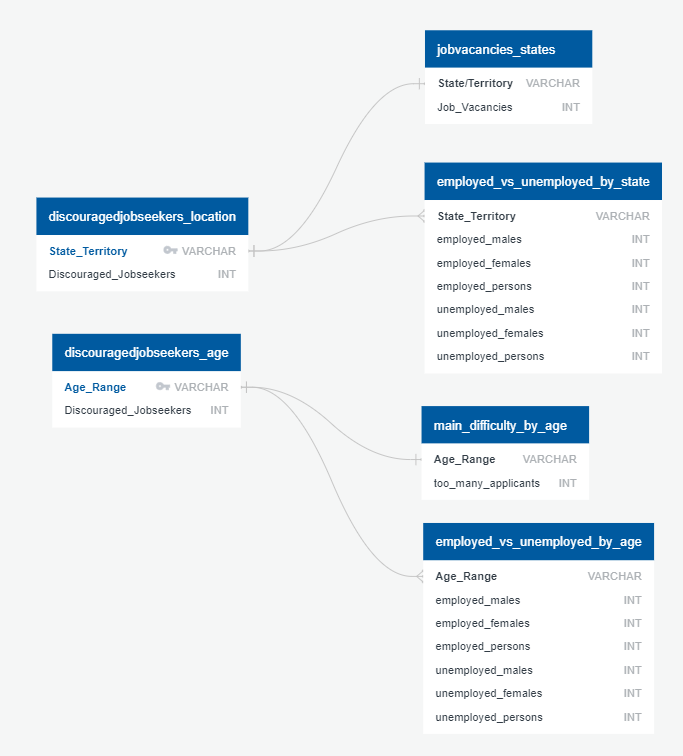
**Load (L)**

After cleaning the data, the following data Entity Relation Diagram was created using QuickDatabase.

The database was created using PostgresSQL pgAdmin.

Relationships between tables are linked via the Discouraged Job Seeker Tables.

<https://app.quickdatabasediagrams.com/#/d/0UDHAz>

****