

# ETL Project – Income and Expenditure

---

## Proposed Use of Data

The project aims to examine the Average of income and expenditure in Western Australia and how it has changed over time. We have income and labour force data for Western Australia and expenditure data for all of the States and Territories of Australia.

## EXTRACT

We decided to use the following dataset from Australian Bureau of Statistics:

- *Table 12e. Average weekly earnings, Western Australia (dollars) - seasonally adjusted.* Format: XLSX  
Link:- <https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/average-weekly-earnings-australia/may-2022#data-download>
- *Table 3. Retail Turnover, by state.* Format: XLSX  
Link:- <https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/oct-2022#data-download>
- *Table 8. Labour force status by Sex, Western Australia - Trend, Seasonally adjusted and Original.* Format: XLSX  
Link:- <https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia/oct-2022#states-and-territories>

## About the data

### Table 12e. Average weekly earnings, Western Australia (dollars) - seasonally adjusted

This dataset is a time series workbook showing the average weekly earnings for adults in Western Australia. The data is collected in May and November each year. Each collection is for a reference week specified by the ABS (Businesses provide earnings information for that week only). The series dates from May 2012 to May 2022 and contains 21 observations.

The data is divided into Males, Females and Persons (assumed to be all). Each of these is subdivided into Full Time Ordinary time earnings, Full Time Total earnings, and Total earnings. We therefore have nine columns of demographics.

### Table 3. Retail turnover, by state

This dataset is a time series workbook showing monthly retail turnover in Australia. The series dates from April 1982 to October 2022 and contains 487 observations. The data is collected monthly, each business is required to provide their turnover for the calendar month. The total turnover for each State and Territory is an estimate based on the sample data collected.

The data is grouped into three series types – Original, Seasonally Adjusted, and Trend. These types are then subdivided into the eight States and Territories of Australia, and total turnover for Australia is also shown, resulting in 27 columns of data.

### Table 8. Labour force status by Sex, Western Australia - Trend, Seasonally adjusted and Original

This dataset is a time series workbook with monthly statistics for 84 aspects of the labour force in Western Australia. The series dates from February 1978 to October 2022. The data we are interested in is the number of employed people.

## TRANSFORM

### Steps

- Limit our dates. We have a lot of data (40 years of retail and labour force, 20 years of earnings). We will examine 1-2 years of data to keep the scope manageable.
- The data for retail turnover will be filtered to what is compatible with the weekly earnings data. Specifically:
  - Select only Seasonally Adjusted data from Western Australia
  - Further reduce this data to the months of May and November
- The data for labour force will be filtered to what is compatible with weekly earnings data. Specifically:
  - Select only Seasonally Adjusted data
  - Select aspects that align with the weekly earnings groupings (Employed full-time, Employed total etc.)
  - Further reduce this data to the months of May and November

## Method

We will use Pandas library to transform our data.

- Cleaning
  - Identify which variables are relevant
  - Confirm data types used and consistent date formatting
  - Eliminate surplus header rows
  - Rename columns
  - Created wanted Dataframe
- Filtering
  - Drop irrelevant/NaN columns
  - Drop rows outside of our date parameters
- Joining
  - Join files based on date

## Intended Result

We should have three tables with a common index (date) that can be used for joining. Retail turnover will have one column (WA Seasonally Adjusted turnover) with the date as index. Weekly earnings will have nine columns with the date as index. Labour force will have approximately six columns with the date as index.

## LOAD

After the transformation of relational databases, we'll proceed to load the data frames into a Postgres local or may be in the cloud database through PG admin. With the schema file loaded into the Postgres database, we'll generate tables to display relevant information.