ETL PROJECT REPORT

**Income and Expenditure**

December 19, 2022



# Emily Coxon Khansa Rubab Malik Brianna O’Connor

# Introduction

Expenditures are a significant but often unnoticed part of household balance sheets. In measuring family financial security, considerable attention is normally paid to income, but much less to whether those resources are appropriate to covering expenditures. For further understanding of this gap, we use the Australian Bureau of Statistics (ABS) data to perform Extract, Transform and Load (ETL) processes.

Seasonally adjusted data for Western Australia from the Average Weekly Earnings1, Labour Force2, and Retail Turnover3 datasets were extracted from the ABS website. These datasets were transformed and filtered according to our requirements and loaded into pgAdmin as an SQL database table format which further analysis could then be performed.

# Extraction

***Data Source:***

From their respective ABS websites, the three datasets contained CSV, JSON files and pulled an API data feed with a vast amount of historical data, however only the CSV files were extracted for this project (Figure 1). The Labour Force2 and Retail Turnover3 datasets contained information from February 1978 and April 1982 respectively, to October 2022. This data was eventually filtered to align with the Average Weekly Earnings1 dataset timeline which contained information between May 2012 to May 2022.

|  |
| --- |
|  |

***Figure 1:***  Data extraction from a CSV file and read initiated in Jupyter notebook performed on the Average Weekly Earnings dataset.

An Entity Relationship Diagram (ERD) (Figure 2.) was created to assist in data sorting and provide insight into merging our tables for the database. A database was then created a based on the ERD aiming to look at the three tables created from the transformed ABS datasets.

|  |
| --- |
|  |

***Figure 2:*** The Entity Relationship Diagram for Average Weekly Earnings, Labour Force and Retail Turnover.

After creating the ERD, three tables containing the table names and column headings for the Average Weekly Earnings, Labour Force, and Retail Turnover datasets were created. These tables were created to load in the data for each respective dataset once data cleaning and transformation had been complete, ensuring the table schema matched the data frame title and column headings.

# Transformation

The Data was cleaned in Jupyter Notebook using Python and the Pandas library.

**Average Weekly Earnings Data**

This dataset was a time series workbook which showed the average weekly earnings for adults in Western Australia. The data was collected in May and November of each year between May 2012 and May 2022 and contained 21 date observations total. Each row was for a reference week specified by the ABS (businesses provide earnings information for that week only). The data was categorized in Males, Females and Persons (assumed to be all) columns, and the unit of measurement for the data was in Australian Dollars ($). Each of which was subdivided into Full Time Ordinary Time Earnings, Full Time Total Earnings, and Total Earnings resulting in nine columns of demographics. The following steps are taken further in transformation process.

* Unnecessary rows (0 to 8) were deleted, these rows contained metadata as present in Figure 1.
* Seven columns required were retained and the three irrelevant columns removed (Ordinary Time Earnings for Males, Females, and Persons).
* Column number 0 header was renamed from “Unnamed: 0” to “Date”.
* Remaining six column headings were shortened e.g., “Earnings; Males; Full Time; Adult; Total earnings ; Western Australia ;” was renamed “Male\_Earnings\_Full\_Time\_$” (Figure 3), which matched the SQL table schema created.

|  |
| --- |
| Table  Description automatically generated |

***Figure 3:*** The final Average Weekly Earnings data frame after completion of data filtering and transformation.

**Labour Force Data**

This dataset contained a time series workbook with monthly statistics for 84 aspects of the labour force in Western Australia. The series dates ranged between February 1978 to October 2022 and contained 546 observations. This dataset was filtered to align with the Average Weekly Earnings dataset observations. The necessary data for this project was the number of full time and total employed males, females, and persons. The unit of measurement for this dataset was Australian Dollars ($) in ‘thousands (‘000). The transformation processes for the Labour Force dataset are as follows:

* Unnecessary rows (0 to 8) were deleted, these rows contained metadata
* Seven columns were retained as representatives of the Australian labour force and remaining 78 columns were removed.
* Column number 0 header was renamed from “Unnamed: 0” to “Date”.
* Remaining six column headings were shortened e.g., “Employed total ; Persons ;.1” was renamed “Employed\_Total\_Persons\_thousands” (Figure 4), which matched the SQL table schema created.

|  |
| --- |
|  |

***Figure 4:*** Final data frame for Labour Force after data cleaning and transformation.

**Retail Turnover Data**

This dataset was a time series workbook that showed monthly retail turnover in Australia. The series dates ranged from April 1982 to October 2022 and contained 487 observations. The data was collected monthly, each business was required to provide their turnover for the calendar month. The total turnover for each State and Territory was an estimate based on the sample data collected. The data was grouped into three series types – Original, Seasonally Adjusted, and Trend. The series types were subdivided into the eight States and Territories of Australia, and total turnover for Australia was also recorded, resulting in 27 columns of data. The data used was from months May and November for dates between May 2012 and May 2022, aligning with the Average Weekly Earnings dataset. The unit of measurement for this dataset was in millions of Australian Dollars ($’000’000)

* Unnecessary rows (0 to 8) were deleted, these rows contained metadata
* Column index 0 was kept as this held date information along with the Total Industry Retail Turnover for Western Australia, remaining columns were removed from the data frame
* Column 0 header “Unnamed: 0” was renamed “Date” and the column “Turnover ; Western Australia ; Total (Industry) ;.1” renamed “Retail\_Turnover\_WA\_millions” (Figure 5)

|  |
| --- |
|  |

***Figure 5:*** Final data frame for Retail Turnover (WA) after data cleaning and transformation.

# Loading to the Database

After the transformation, we proceeded to load the data frames into a Postgres cloud database through PG admin. With the schema file loaded into the Postgres database and ridership. Once we ran the queries with the relevant information, we reconnected to the database and generated additional tables for the data frames.

# Final Database and Future Improvements

# References

[1] Australian Bureau of Statistics (May 2022), [*Average Weekly Earnings, Australia*](https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/average-weekly-earnings-australia/may-2022), ABS Website, accessed 19 December 2022.

[2] Australian Bureau of Statistics (October 2022), [*Labour Force, Australia*](https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia/oct-2022), ABS Website, accessed 19 December 2022.

[3] Australian Bureau of Statistics (October 2022), [*Retail Trade, Australia*](https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/oct-2022), ABS Website, accessed 19 December 2022.