Provisional Mortality Statistics ETL Report

*By Jack Pan and Heriawan Muhamad*

Extract

A total of nine raw datasets were extracted from the Australian Bureau of Statistics all relating to provision mortality statistics pre and post COVID-19. The data sets are in csv format. Eight of the datasets used had mortality statistics relating to leading causes of death as certified by a doctor:

Cancer

Cerebrovascular disease

Chronic lower respiratory disease

Dementia

Diabetes

Heart disease

Pneumonia

Respiratory disease

The ninth dataset aggregated these mortality statistics into a total count for each period.

These datasets were imported into a PgAdmin4 SQL database with ‘date’ being the primary key to form a relationship between all the datasets. SQL alchemy was then utilised to facilitate data transformation within the database on Jupyter Notebook.

Transform

Our cause for our transformations was to be able to interpret mortality statistics more easily between leading causes of death. As well as providing summary statistics on mortality counts and differences to understand the trends that are happening over time.

Before we could extract the csv datasets into a PgAdmin4 database, we had to go into the csv files to change cell formats to exclude commas. Such that when importing, the values could be interpreted is integers.

Once the database was connected on Jupyter