UWA DATA ANALYSIS BOOTCAMP - PROJECT REPORT 2 ETL

Jono Wood[[1]](#footnote-1) , Johan Snyman[[2]](#footnote-2)

Abstract

This report will cover the selection, extraction, transformation and loading of ABS (Australian Bureau of Statistics) data and preparing it in a database model to be used by analysts. The data provided a specific challenge (.CSV) with various ABS specific codes as well as empty lines and white space. From the initial data table investigation the schema design built a series of tables to provide the ability to build queries based on specific primary and foreign keys. The transition form .CSV format to PANAD data frames is built around dictionaries with index setting to create unique keys. The connectivity and readability of the PGRESS SQL database have been tested and the data read back correctly

**Keywords #**Python#PANDA#POSTGRESS#ERD

**Repository**: https://github.com/jonowood/Project\_2

# Introduction

The award winning project 2 team was requested to provide a database for analysts to analyze the average weekly earnings (AWE) of Australian income earners with the estimated residence population (ERP) movements. The team reviewed the AWE raw data and the integrity of the ERP data as part of the design review to build a resilient database model with an region and quarterly centric focus. This vision then inspired the creation of the analysis table have regional change analysis capability.

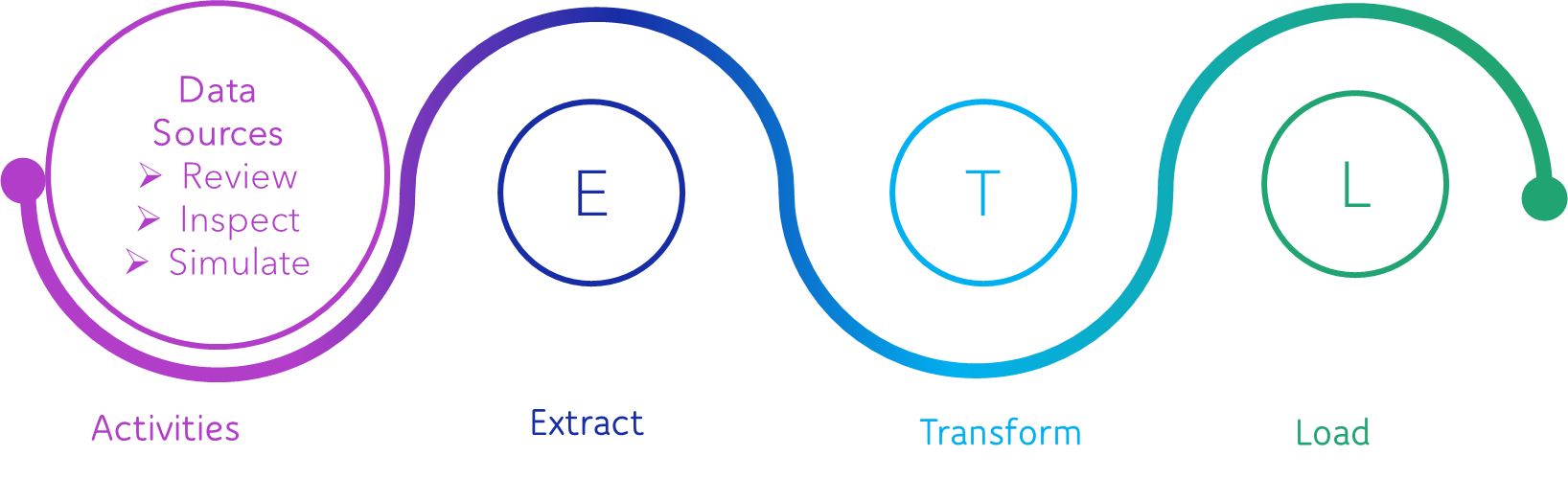


Figure 1 :Project Data development model

# ETL (.CSV) Primary data review

The ABS files required was reviewed to understand the structure, elements and white space.

Step 1: Locate and Download ABS data

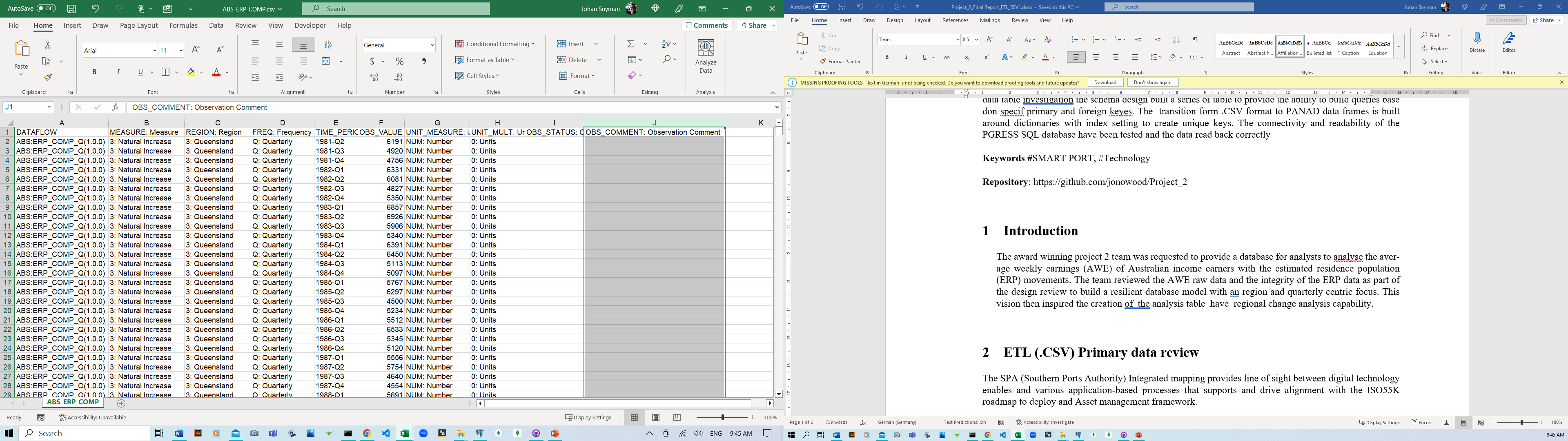


Figure 2 :ABS\_ERP\_COMP.csv – Native format downloaded

* ABS – ERP.CSV 7,5K datapoints

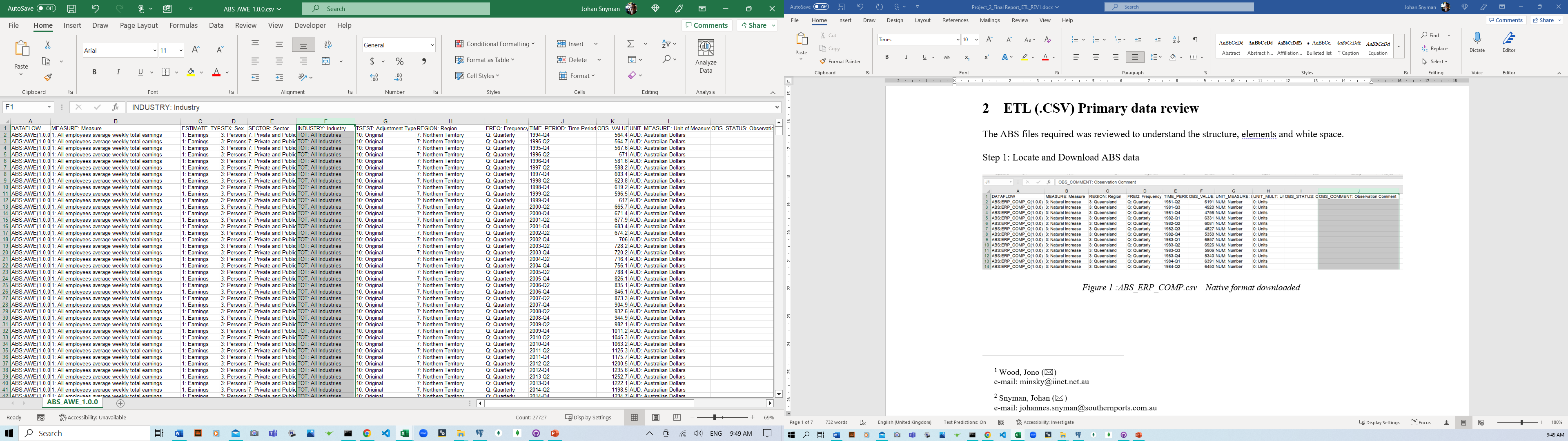


Figure 3 :ABS\_AWE.csv – Native format downloaded

Step 2: Prepare the CSV form for upload.

Transformed “text to columns” to remove all the ABS coding in from of the required columns

Removed columns not required

Filter and prepared the files to support the Analysis design of the output database

File A - **ABS – AWE.CSV (27K datapoints)**

* Filter column C values to only include ‘Earnings’
* Filter column B to only include 'All employees average weekly earnings’
* Filter column G to only include ‘Original’
* Copy columns to Data Frame - REGION: Region, TIME\_PERIOD: Time Period, OBS\_VALUE

File B - **ABS – ERP.CSV 7,5K (datapoints)**

* Filter column B to only include ‘4: Internal Arrivals’, ‘5: Internal Departures’, ‘6: Net Internal Migration’, ‘13: Change Over Previous Quarter’
* Copy columns to Data Frame - REGION: Region, TIME\_PERIOD: Time Period, OBS\_VALUE

# Database Schema Design & ERD

## ERD Model

The ERD model was used to design all the tables and their relationship.

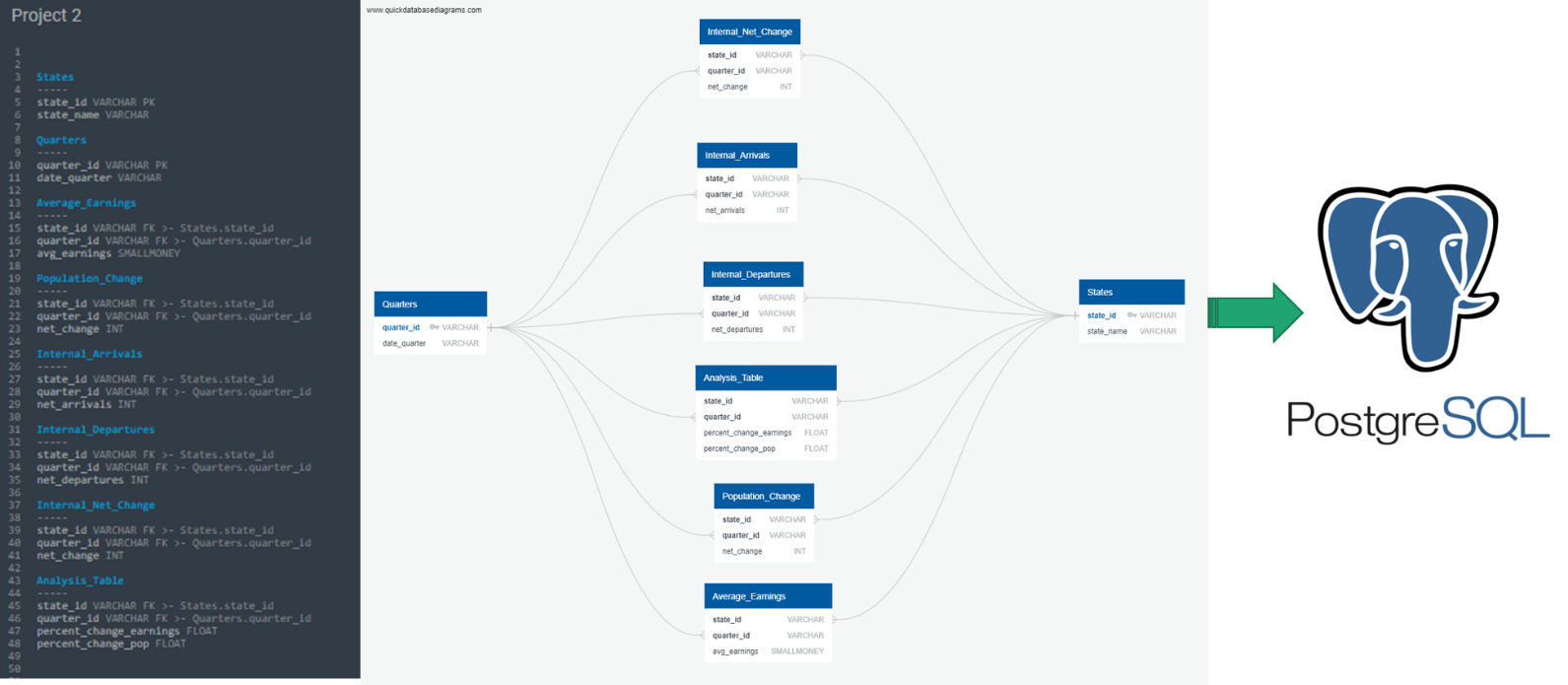


Figure4 :: ERD Model

## STATES Table

The STATES table will use a unique index number and the state name to allow analyst to be able to connect any data to the state.

## Graphical user interface Description automatically generated

Figure5 :States

## QUARTERS Table

The STATES table will use a unique index number and the date quarter to allow analyst to be able to connect any data to the state.

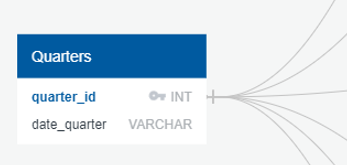


Figure6 :Quarters

## Population Change Table

The STATES table will use the state and quarter ID to deliver the net change to allow analyst to be able to connect any data to the state.

A picture containing graphical user interface

Description automatically generated

Figure7 :Population Change

## Internal Net Change Table

The STATES table will use the state and quarter ID to deliver the net change to allow analyst to be able to connect any data to the state.

Graphical user interface

Description automatically generated with medium confidence

Figure7 :Internal Net Change

## Internal Departures Table

The STATES table will use the state and quarter ID to deliver the net departures to allow analyst to be able to connect any data to the state.

Graphical user interface

Description automatically generated with medium confidence

Figure7 :Internal Departures

## Internal Arrivals Table

The STATES table will use the state and quarter ID to deliver the net departures to allow analyst to be able to connect any data to the state.

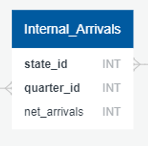


Figure7 :Internal Arrivals

## Average Earnings Table

The STATES table will use the state and quarter ID to deliver the average earnings to allow analyst to be able to connect any data to the state.

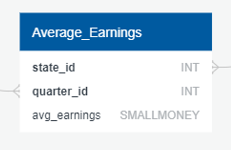


Figure7 :Average Earnings

# Transformation – PANDA

## IMPORT & Clean

The .csv files are read into PYTHON and checked if all the data imported corrected. Various columns are dropped and empty lines removed

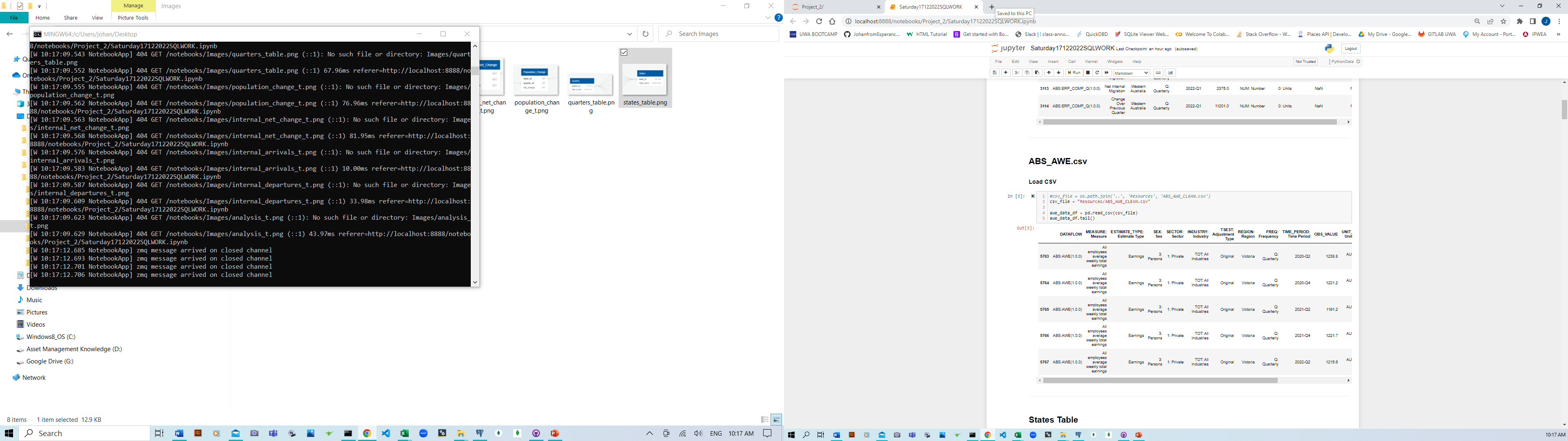


Figure7 :CSV Imported to PYTHON

## Reframe & Check

The new data frame is used to set the new structure of the DF required as the table for the PGRESS solution.

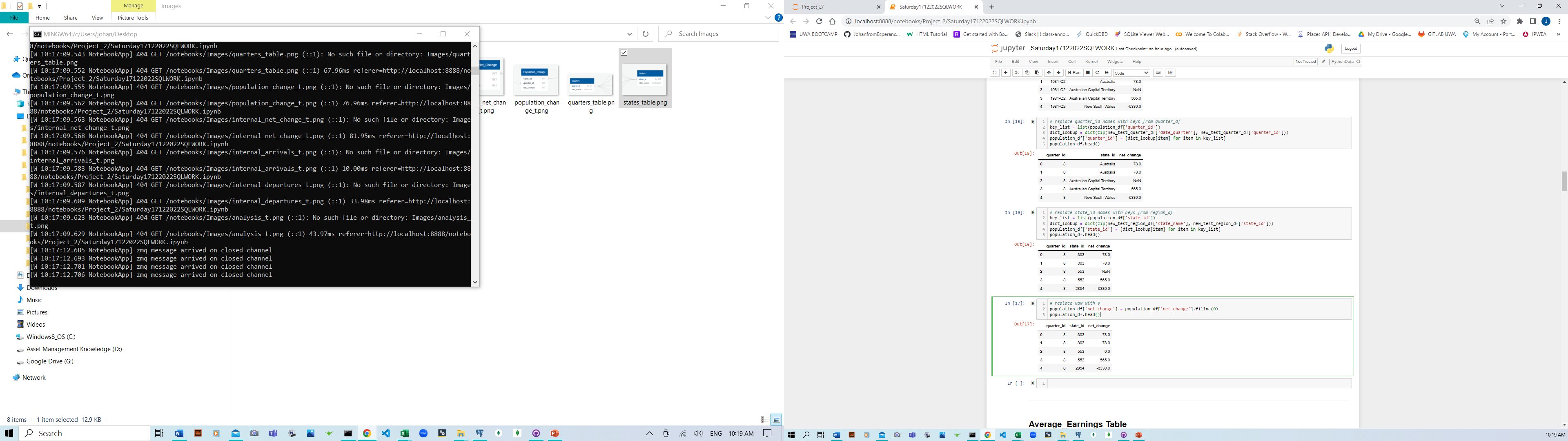


Figure7 :Table creation

## Connect LOAD & Check

Set the connection to the PGRESS and read the table to ensure they were loaded correctly

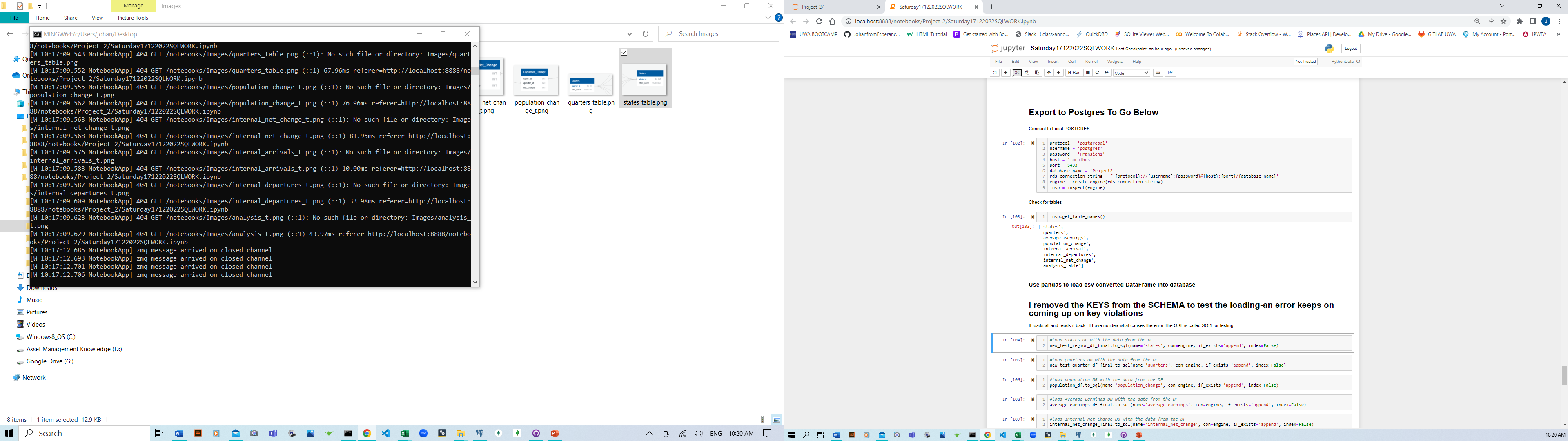


Figure7 :PGRESS Connections

## Reframe & Check

The data is written to the SQL database and then read back to check if it is correct

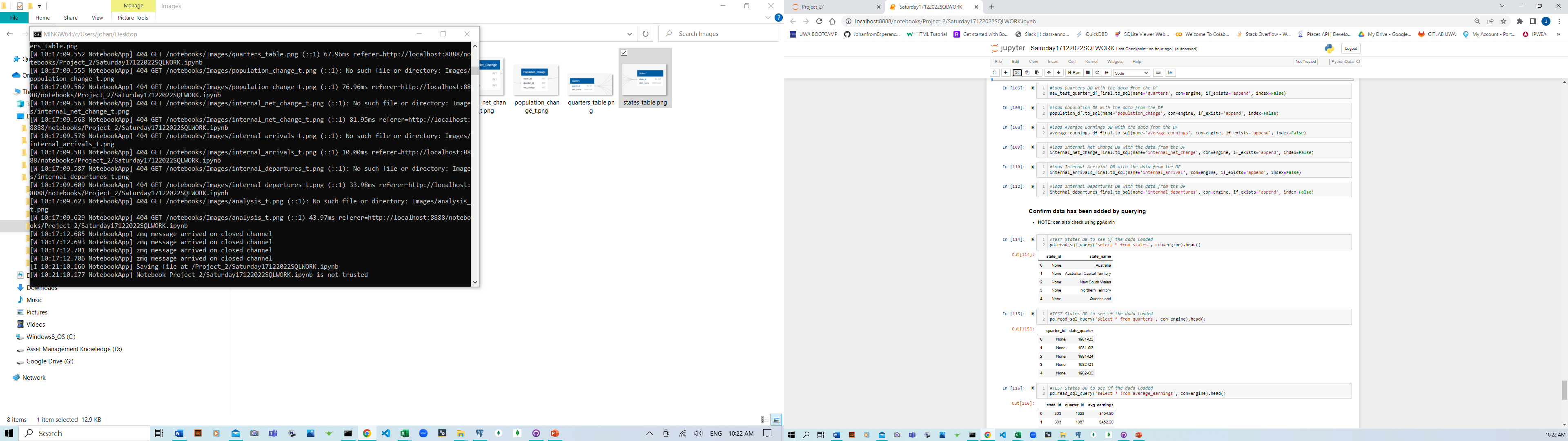


Figure7 :SQL read write to check

# Conclusion

The project team successfully created and loaded a database for use by the data analysts.

References

1. Project Proposal: Project 2-ETL - ERP vs.AWE.pptx
2. UWA Bootcamp Week4,5,6 Lesson Plans and Activities
3. <https://stackoverflow.com/questions/39598238/sql-database-is-giving-me-this-error>
4. <https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release>

1. Wood, Jono (🖂)  
   e-mail: minsky@iinet.net.au [↑](#footnote-ref-1)
2. Snyman, Johan (🖂)  
   e-mail: [johannes.snyman@southernports.com.au](mailto:johannes.snyman@southernports.com.au) [↑](#footnote-ref-2)