UWA DATA ANALYSIS BOOTCAMP
PROJECT 2 REPORT

Repository: https://github.com/jonowood/Project_2

Jono Wood[1] , **Johan Snyman**[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 from .CSV format to PANADA'S data frames is built around dictionaries with index setting to create unique keys. The connectivity and readability of the POSTGRESS SQL database have been tested and the data read back correctly

Keywords #Python #PANDA #POSTGRESS #ERD

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 a region and quarterly centric focus. This vision then inspired the creation of the analysis table have regional change analysis capability.

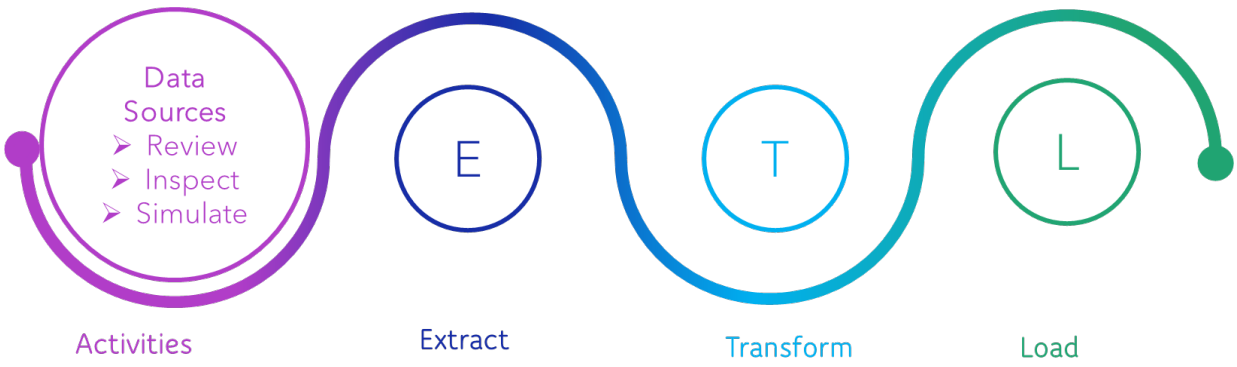


Figure 1 :Project Data development model

[1] Wood, Jono (✉)
e-mail: minsky@iinet.net.au

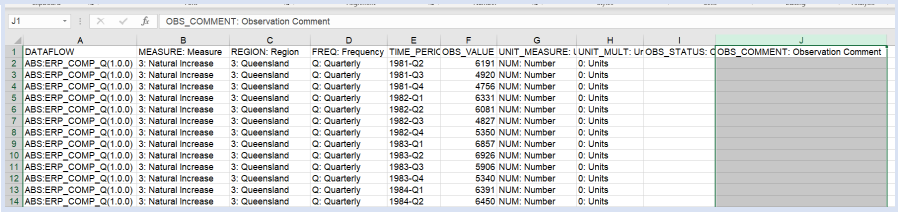
[2] Snyman, Johan (✉)
e-mail: johannes.snyman@southernports.com.au

UWA DATA ANALYSIS BOOTCAMP
PROJECT 2 REPORT

ETL (.CSV) PRIMARY DATA REVIEW

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

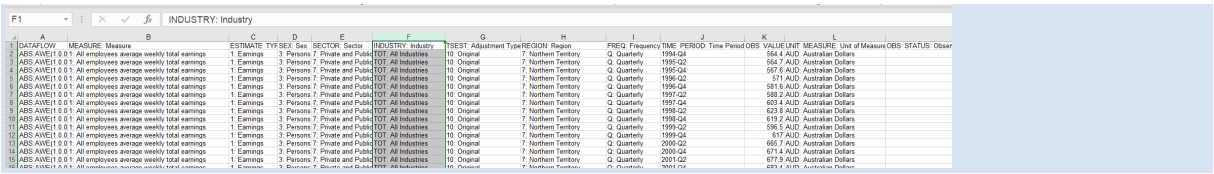
STEP 1: LOCATE AND DOWNLOAD ABS DATA



	A	B	C	D	E	F	G	H	I	J
1	DATAFLOW	MEASURE: Measure	REGION: Region	FREQ: Frequency	TIME_PERIC	OBS_VALUE	UNIT_MEASURE:	UNIT_MULT: Ur	OBS_STATUS: C	OBS_COMMENT: Observation Comment
2	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1981-Q2	6191	NUM: Number	0: Units		
3	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1981-Q3	4920	NUM: Number	0: Units		
4	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1981-Q4	4756	NUM: Number	0: Units		
5	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1982-Q1	6331	NUM: Number	0: Units		
6	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1982-Q2	6081	NUM: Number	0: Units		
7	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1982-Q3	4827	NUM: Number	0: Units		
8	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1982-Q4	5350	NUM: Number	0: Units		
9	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1983-Q1	6857	NUM: Number	0: Units		
10	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1983-Q2	6926	NUM: Number	0: Units		
11	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1983-Q3	5906	NUM: Number	0: Units		
12	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1983-Q4	5340	NUM: Number	0: Units		
13	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1984-Q1	6391	NUM: Number	0: Units		
14	ABS:ERP_COMP_Q(1.0.0)	3: Natural Increase	3: Queensland	Q: Quarterly	1984-Q2	6450	NUM: Number	0: Units		

Figure 2 :ABS_ERP_COMP.csv – Native format downloaded

- ABS – ERP.CSV 7,5K datapoints



	A	B	C	D	E	F	G	H	I	J	K	L
1	DATAFLOW	MEASURE: Measure	ESTIMATE_TYP	SEX: Sex	SECTOR: Sector	INDUSTRY: Industry	TSEST: Adjustment Type	REGION: Region	FREQ: Frequency	TIME_PERIOD: Time Period	OBS_VALUE	UNIT_MEASURE: Unit of Measure	OBS_STATUS: Obser
2	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1994-Q4	564.4	AUD: Australian Dollars	
3	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1995-Q2	564.7	AUD: Australian Dollars	
4	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1995-Q4	567.6	AUD: Australian Dollars	
5	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1996-Q2	571	AUD: Australian Dollars	
6	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1996-Q4	581.6	AUD: Australian Dollars	
7	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1997-Q2	588.2	AUD: Australian Dollars	
8	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1997-Q4	603.4	AUD: Australian Dollars	
9	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1998-Q2	623.8	AUD: Australian Dollars	
10	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1998-Q4	619.2	AUD: Australian Dollars	
11	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1999-Q2	596.5	AUD: Australian Dollars	
12	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	1999-Q4	617	AUD: Australian Dollars	
13	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	2000-Q2	666.7	AUD: Australian Dollars	
14	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	2000-Q4	671.4	AUD: Australian Dollars	
15	ABS:AWE(1.0.0)	1: All employees average weekly total earnings	1: Earnings	3: Persons	7: Private and Public	TOT: All Industries	10: Original	7: Northern Territory	Q: Quarterly	2001-Q2	677.9	AUD: Australian Dollars	

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)

DATA TRANSFORM STEPS

1. **Data Filtering**
   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’
2. **Data Mapping**
   Changed names of states and measures, removing numbers and symbols
3. **Data Deduplication**
   After creating analysis table, duplicate redundant columns were removed
4. **Derived Variables**
   Column created with % change of total average earnings
5. **Splitting data**

- Data from this set has been exploded to match the SQL data load to Postgres

FILE B - ABS – ERP.CSV 7,5K (DATAPOINTS)

DATA TRANSFORM STEPS

1. **Data Filtering**
   Filter column B to only include '4: Internal Arrivals', '5: Internal Departures', '6: Net Internal Migration', '13: Change Over Previous Quarter'
2. **Data Mapping**
   Changed names of states and measures, removing numbers and symbols
3. **Data Deduplication**
   After creating analysis table, duplicate redundant columns were removed
4. **Derived Variables**
   Column created with % change of total estimated population
5. **Splitting data**
   Data from this set has been exploded into tables to match the SQL data load to Postgres

DATABASE SCHEMA DESIGN & ERD

ERD MODEL – RELATIONAL DATABASE

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

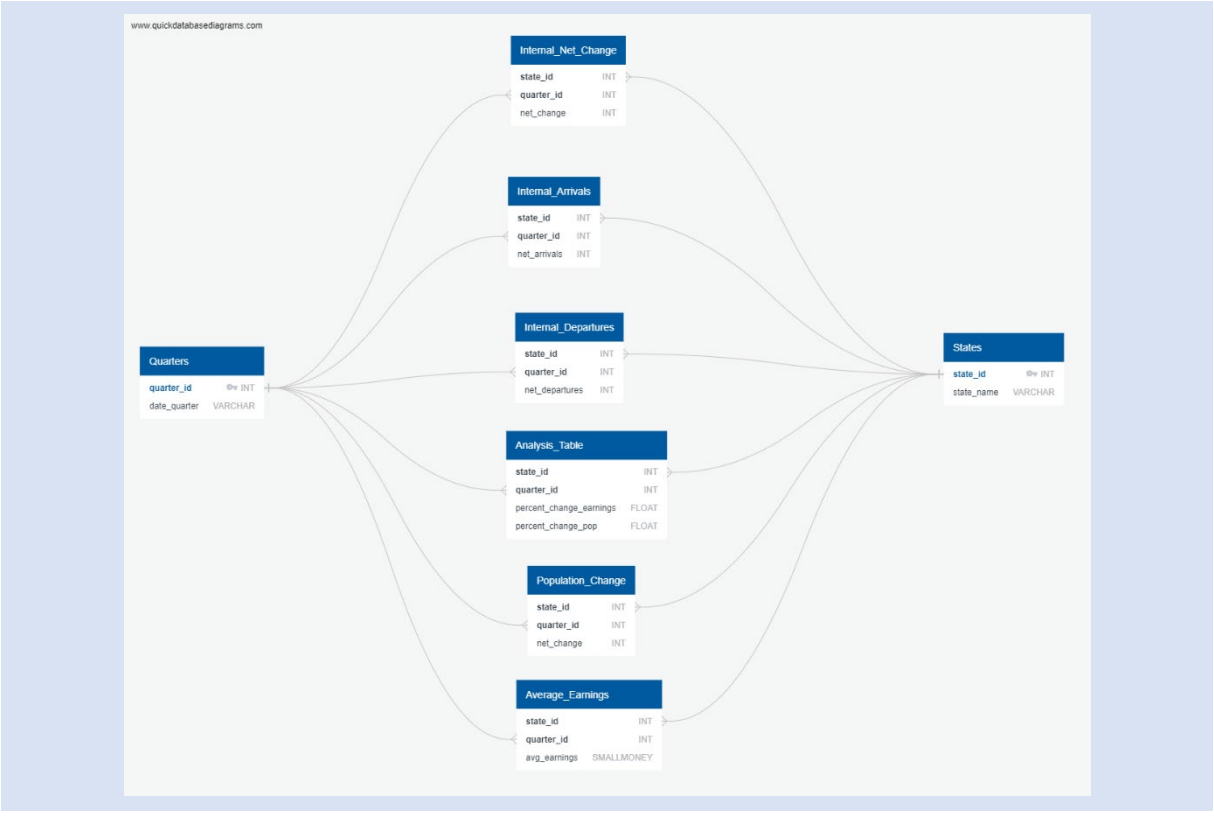


Figure4 :: ERD Model

UWA DATA ANALYSIS BOOTCAMP
PROJECT 2 REPORT

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.

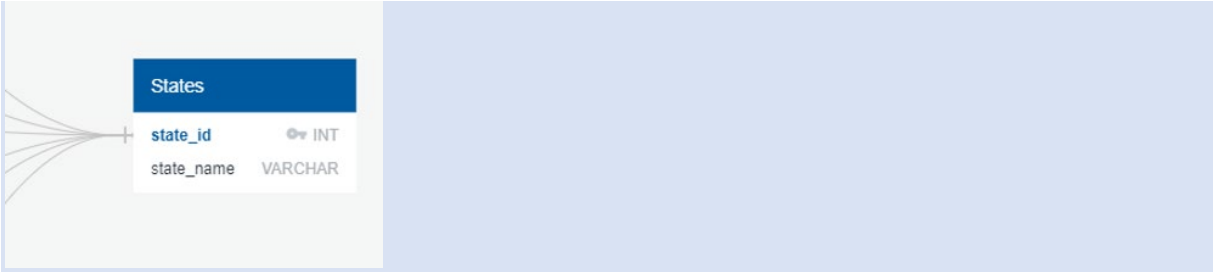


Figure5 :States

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.

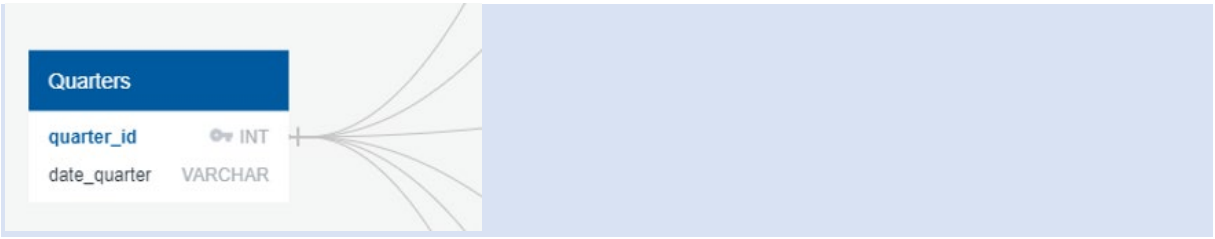


Figure6 :Quarters

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.

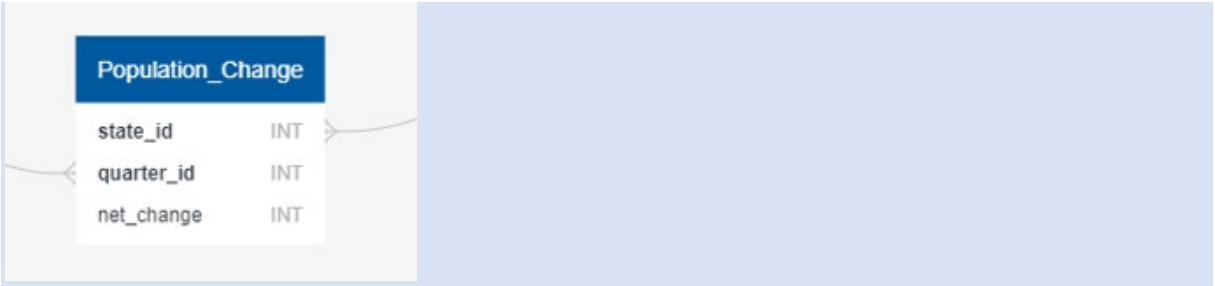


Figure7 :Population Change

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.

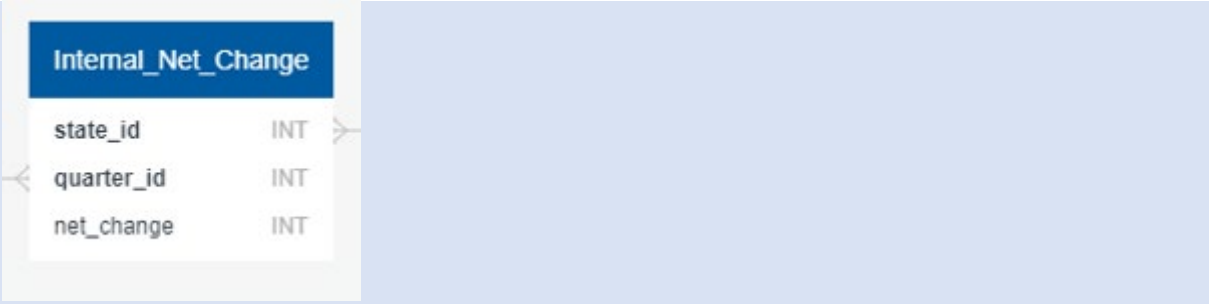


Figure8 :Internal Net Change

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.

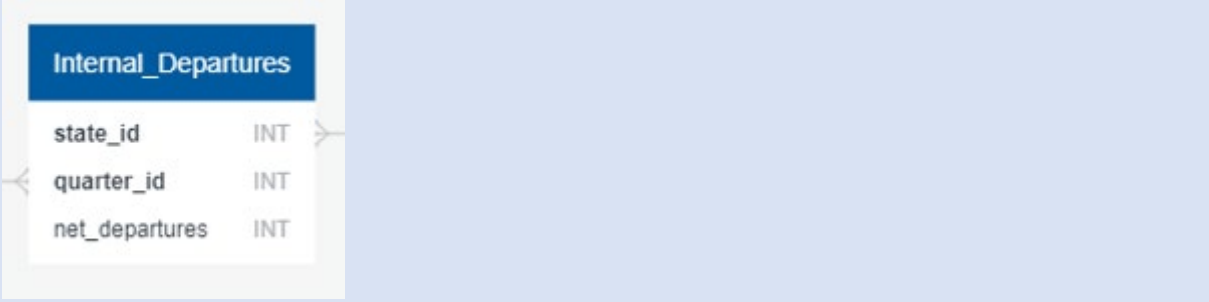


Figure9 :Internal Departures

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.

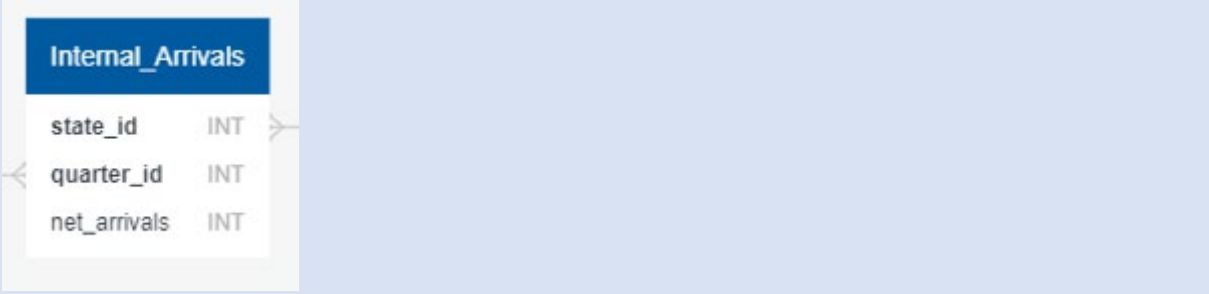
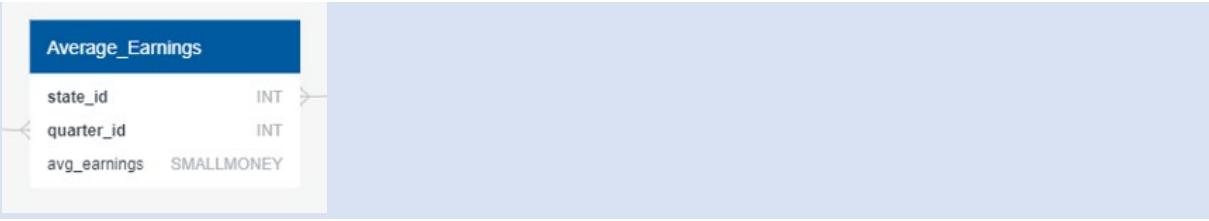


Figure10 :Internal Arrivals

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.



UWA DATA ANALYSIS BOOTCAMP
PROJECT 2 REPORT

Figure11 :Average Earnings

TRANSFORMATION – PANDAS

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

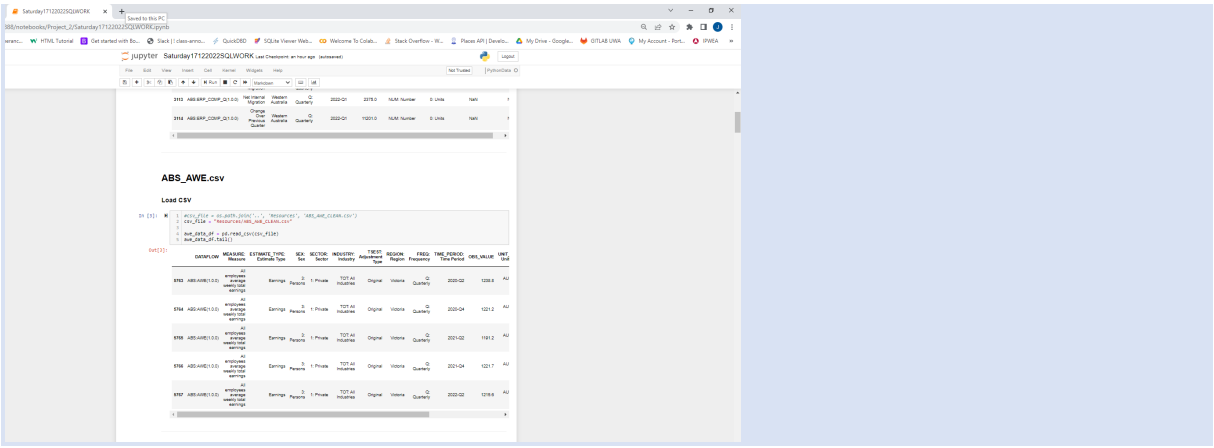


Figure12 :CSV Imported to PYTHON

REFRAME & CHECK

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

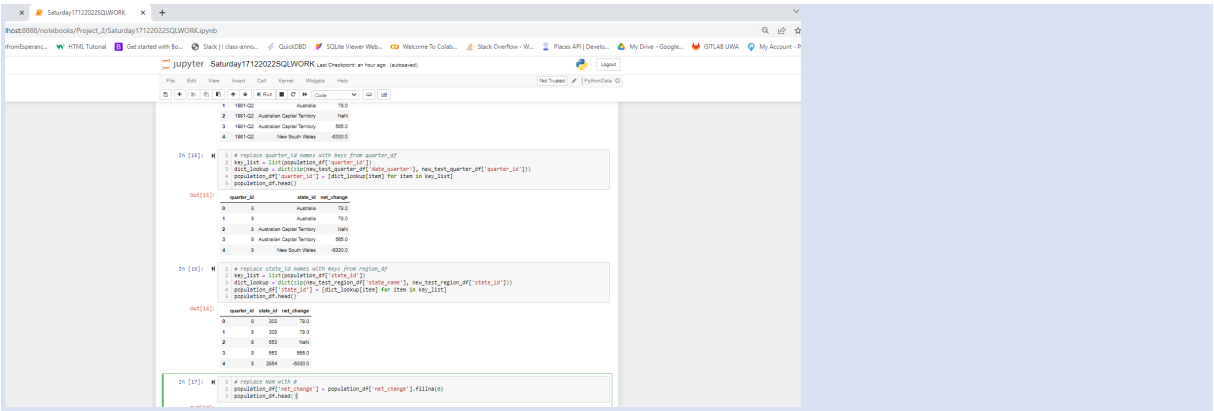


Figure13 :Table creation

CONNECT LOAD & CHECK

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

UWA DATA ANALYSIS BOOTCAMP
PROJECT 2 REPORT

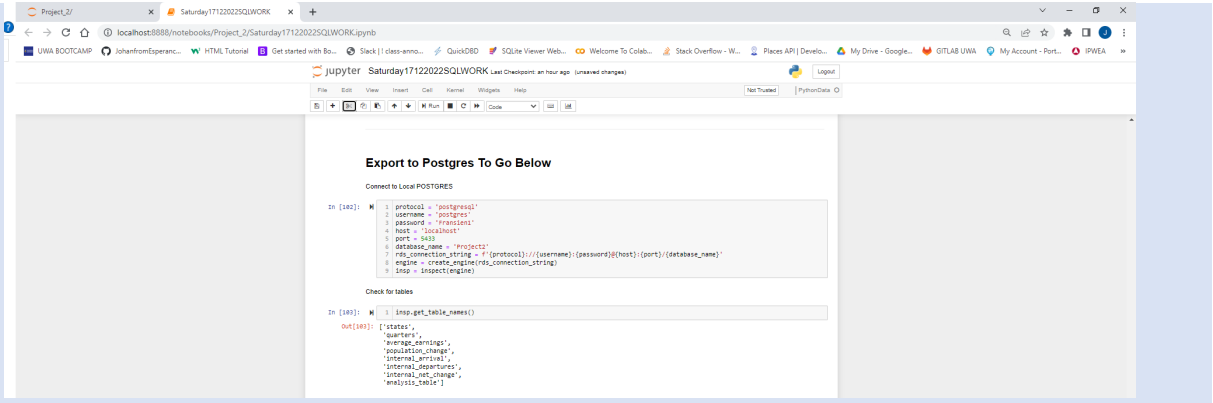


Figure14 :PGRESS Connections

REFRAME & CHECK

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

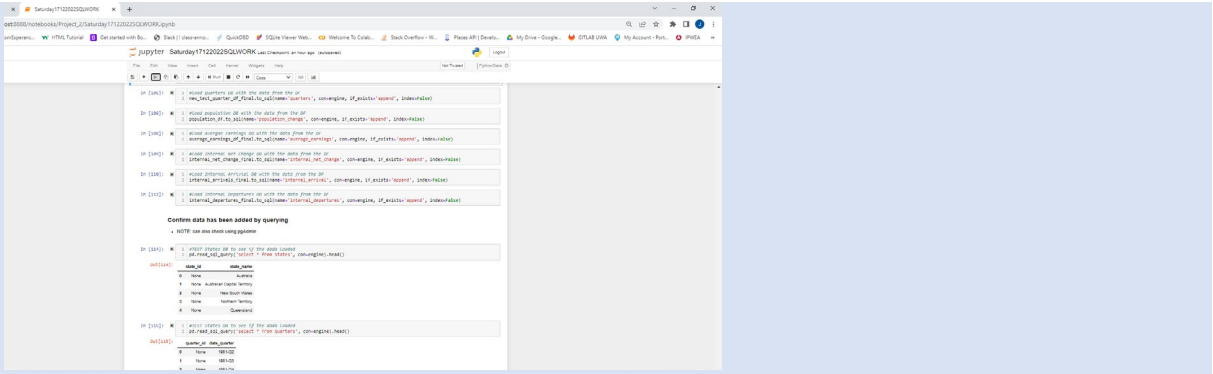


Figure15 :SQL read write to check

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
5. https://app.quickdatabasediagrams.com/#/d/ETsGTa