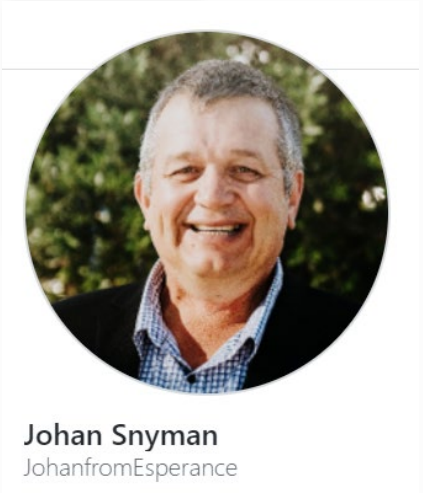
UWA DATA ANALYSIS BOOTCAMP - PROJECT 2 ETL

TEAM 6 PROPOSAL

Our Team



Johan Snyman

Jon Wood

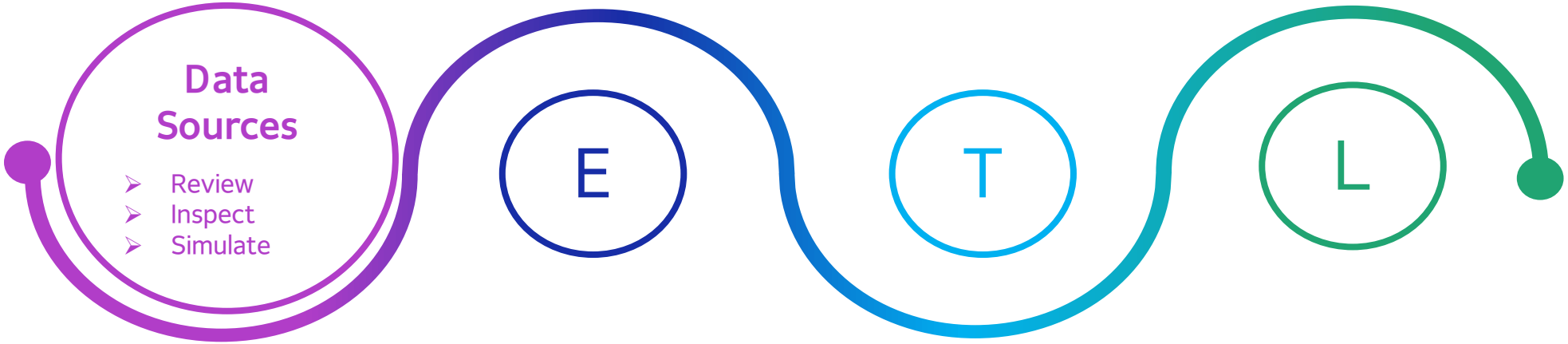
Our Git Repository

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

PROJECT 2-ETL : FINANCE/SOCIAL SERVICES

HOW DOES AVERAGE WEEKLY EARNINGS INFLUENCE ESTIMATED RESIDENCE POPULATION

THE AIM OF OUR PROJECT IS TO UNCOVER PATTERNS IN POPULATION MOVEMENTS AND AVERGE INCOME CHANGES . WE'LL EXAMINE RELATIONSHIPS BETWEEN POPULATION CHANGE AND AVERAGE INCOME, PERCENTAGE CHANGE AND OTHER RELATED RELATIONSHIPS DERIVED FROM THE DATA.



Activities

- Locate and download data
- ABS – AWE.CSV 27K datapoints
- ABS – ERP.CSV 7,5K datapoints
- Prework, structure, common keys, duplicates and NAN
- Data formats

Extract

- Design ERD flow for desired table elements
- Create POSTGRES Tables
- Load .CSV to Python and convert
- Check CSV loaded correctly

Transform

- Wrangling steps
- Cleaning – NAN
- Align Time/reasons
- Rename headers
- Remove columns
- Join

Load

- Connect to PostGres
- Load Data
- Check and test data

EXTRACT

Data Sets

Datasets Sourced From - https://explore.data.abs.gov.au/ (Licence CC01)

File A - **ABS – AWE.CSV (27K datapoints) - Population Movement Data for Australia**

File B - **ABS – ERP.CSV 7,5K (datapoints) – Average Income Data for Australia**

Initial Findings

Both data sets will require filtering of data based off multiple columns. The data is in a similar format across both sets.

Data is incomplete and will require the filling of NaN values. Some column values will require extra characters removed prior to transforming.

For loading this data to Postgres in a relational database, we will need to explode the data into multiple tables and link with date and location ID's.

TRANSFORM

REQUIRED STEPS TO TRANSFORM DATA;

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 DataFrame - 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 DataFrame - REGION: Region, TIME_PERIOD: Time Period, OBS_VALUE

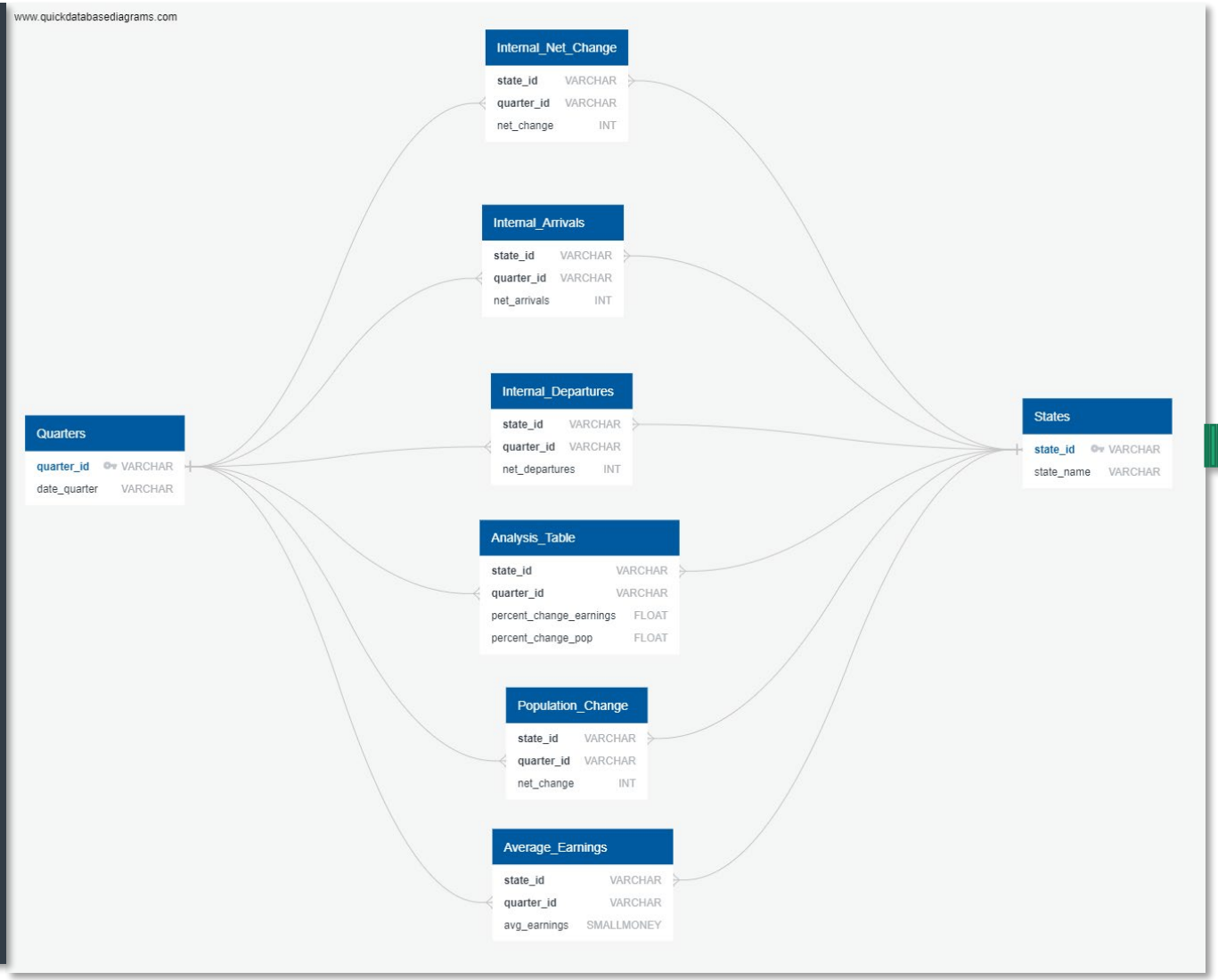
Additional Pandas Transformations – Data Filtering, Data Mapping, Data Deduplication, Derived Variables

LOAD

Relational Database - We will create a Database and Schema in PostgreSQL then Load the data for future Analysis

Project 2

```
States
-----
state_id VARCHAR PK
state_name VARCHAR

Quarters
-----
quarter_id VARCHAR PK
date_quarter VARCHAR

Average_Earnings
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
avg_earnings SMALLMONEY

Population_Change
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
net_change INT

Internal_Arrivals
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
net_arrivals INT

Internal_Departures
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
net_departures INT

Internal_Net_Change
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
net_change INT

Analysis_Table
-----
state_id VARCHAR FK >- States.state_id
quarter_id VARCHAR FK >- Quarters.quarter_id
percent_change_earnings FLOAT
percent_change_pop FLOAT
```



PostgreSQL