|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| ***Data Management Plan for Equalhelper Website Development Project***  **By Team 925ers** |
| |  |  |  | | --- | --- | --- | |  | **Last Updated at 17/05/2023** |  | |

Contents

[1. Project Overview 2](#_Toc135222667)

[2. Data Source 3](#_Toc135222668)

[2.1 Open Data Source 1: Australia Bureau Statistics 3](#_Toc135222669)

[2.2 Data Extraction 5](#_Toc135222670)

[2.2.1 Data’s Website: 5](#_Toc135222671)

[2.2.2 Data Extracted: 5](#_Toc135222672)

[3.Data Usage: 10](#_Toc135222673)

[In Iteration 1: 10](#_Toc135222674)

[In Iteration 2: 10](#_Toc135222675)

[In Iteration 3: 10](#_Toc135222676)

[4.Data Processing: 11](#_Toc135222677)

[4.1Data Transformation: 11](#_Toc135222678)

[4.2 Related Tools/Code: 11](#_Toc135222679)

[4.3 Data Transformation 11](#_Toc135222680)

[4.3.1 Iteration 1: 11](#_Toc135222681)

[4.3.2 Iteration 2: 12](#_Toc135222682)

[4.3.2 Iteration 3: 12](#_Toc135222683)

[5. Database Design: 13](#_Toc135222684)

[5.1 ERD Diagram 13](#_Toc135222685)

[5.2 Data Lake 15](#_Toc135222686)

[6. Data Analytics: 16](#_Toc135222687)

[Introduction: 16](#_Toc135222688)

[Data Blindsight: 16](#_Toc135222689)

[Data Hindsight: 16](#_Toc135222690)

[Data Insight: 16](#_Toc135222691)

[Data Foresight: 17](#_Toc135222692)

[7. Open Data Source Pipeline: 18](#_Toc135222693)

# Project Overview

925ers are developing a modern website that aims to popularise the knowledge of gender inequality in society, provide guidance, help and encourages dialogue around the inequality. In order to let victims, understand the real situation now, all the data on the website comes from the official open data of the Australian government’s organisation. The data we use includes but is not limited to the income difference between men and women in various suburbs in Australia, the time women participate in family affairs, the proportion of women in leadership, etc. In order to meet the needs of data analysis and data visualization, the extracted raw data is converted into a machine friendly format.

Furthermore. This report also includes entity-relationship diagrams and logical data modelling diagram. The data warehouse and data lake are built to optimize database performance based on the specified goals for visualisation and analysis. The data management plan is created to ensure that visitors are not misled by false information. The data must be up-to date, secure and reliable.

# Data Source

Proposed project will involve data collection from third parties and the project team will review and strictly adhere to their copywrite policies to avoid any violations. The project team will only collect data related to Australia and gender inequalities.

The data is then used for analysis and graphing. We will focus on driving insight into various gender gaps such as: gender pays homosexuality and gender-based violence.

## 2.1 Open Data Source 1: Australia Bureau Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Iteration1** | | | | | |
| **Data Name** | **Link** | **Physical Access Used** | **Frequency of Iteration** | **Granularity** | **Copyright details** |
| 1. employees paid at the adult rate, average weekly total cash earnings - industry by sex.xlsx 2. Employee jobs and employee income.xlsx | [Data Source Link](https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/employee-earnings-and-hours-australia/latest-release) | EXCEL | Yearly | Salary per hour; weekly hours worked, etc. | [ABS Copyright Link](https://www.abs.gov.au/website-privacy-copyright-and-disclaimer#copyright-and-creative-commons) |

**Description:**

The data is collected from Australia Bureau Statistics and used for iteration 1. In gender pay gap statistics page, data is used for showing the pay gap across all industry in Australia.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Iteration2** | | | | | |
| **Data Name** | **Link** | **Physical Access Used** | **Frequency of Iteration** | **Granularity** | **Copyright details** |
| 6524055002\_DO004.xlsx | [Data Source Link](https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/employee-earnings-and-hours-australia/latest-release) | EXCEL | Yearly | Salary per Age; weekly hours worked, etc. | [ABS Copyright Link](https://www.abs.gov.au/website-privacy-copyright-and-disclaimer#copyright-and-creative-commons) |

**Description:**

The data is collected from Australia Bureau Statistics and used for iteration 2. In gender pay gap statistic page, the data is used to show gender pay gap in different age section.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Iteration3** | | | | | |
| **Data Name** | **Link** | **Physical Access Used** | **Frequency of Iteration** | **Granularity** | **Copyright details** |
| 1. Employee jobs and employee income.xlsx 2. Gender pay gap measures.xlsx | [Data Source Link](https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/employee-earnings-and-hours-australia/latest-release) | EXCEL | Yearly | Gender pay gap per year, median salary per suburb. | [ABS Copyright Link](https://www.abs.gov.au/website-privacy-copyright-and-disclaimer#copyright-and-creative-commons) |

**Description:**

Data is used in gender pay gap calculator page to calculate and compare gender pay gap and draw plots.

## 2.2 Data Extraction

### 2.2.1 Data’s Website:

1. Go to ABS’s Gender Indicator

Table

Description automatically generated with medium confidence

Figure 2.2.1. gender-indicators. Retrieved 25/04/2023 from

https://www.abs.gov.au/statistics/people/people-and-communities/gender-indicators

### 2.2.2 Data Extracted:

#### Iteration 1:

1. Employee jobs and employee income.xlsx
   1. Scroll down and find **Gender Pay Gap Measures.** Then click on median weekly cash earnings.

Diagram

Description automatically generated with medium confidence

Figure 2.2.2 Gender Pay Gap Measures, Retrieved 25/04/2023 from

<https://www.abs.gov.au/statistics/people/people-and-communities/gender-indicators>

* 1. Scroll down to the bottom of webpage and click on Download XLSX

Graphical user interface, text, application, email

Description automatically generated

Figure 2.2.3 Gender Pay Gap Measures Data1, Retrieved 25/04/2023 from

https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/employee-earnings-and-hours-australia/latest-release

* 1. Data snippet: Table

     Description automatically generated with medium confidence

Figure 2.2.4 Retrieved 06/04/2023.

* 1. Description: It shows the median employee income per job by sex of each state from 2015 to 2020. It has location, so very useful when creating a geo map. It is a time series data, very useful for visualise the trending.
  2. Data transform code snippet:

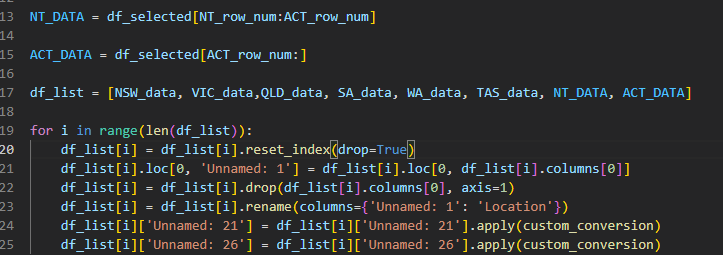


Figure 2.2.5 Retrieved 06/04/2023.

* 1. Data cleaning code snippet:

A screen shot of a computer code

Description automatically generated with low confidence

Figure 2.2.6 Retrieved 06/04/2023.

* 1. After transformation:

A screenshot of a table

Description automatically generated with low confidence

Figure 2.2.7 Retrieved 06/04/2023.

1. employees paid at the adult rate, average weekly total cash earnings - industry by sex.xlsx.
   1. Same webpages find this plot and click download to download the data

Chart

Description automatically generated

Figure 2.2.8 Gender Pay Gap Measures Data Table, Retrieved 25/04/2023 from

https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/employee-earnings-and-hours-australia/latest-release

* 1. Data snippet:

Table

Description automatically generated

Figure 2.2.9 Retrieved 06/04/2023

* 1. Description: it shows the income across all industries by gender, very useful to understand the situation of various industries from a macro perspective.
  2. Data transform code snippet:

A picture containing text, screenshot, font

Description automatically generated

Figure 2.2.10 Retrieved 06/04/2023

* 1. After transformation:

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2.2.11 Retrieved 06/04/2023

#### Iteration 2:

1. 6524055002\_DO004.xlsx
   1. Data snippet:

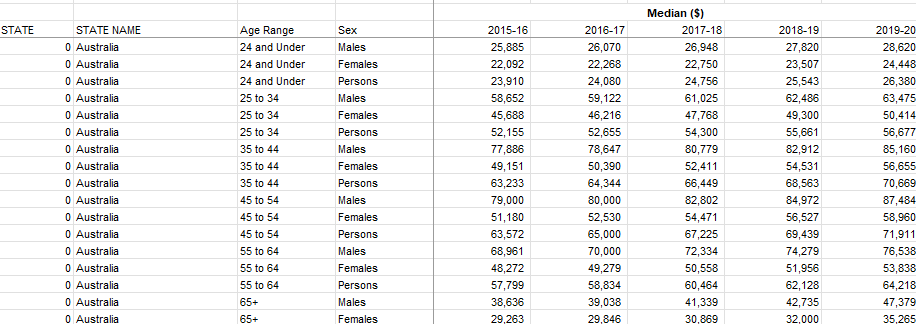


Figure 2.2.12 Retrieved 06/04/2023

* 1. Description: Employee income, earners and summary statistics by age group, sex and state. Can be used to plot a butterfly bar chart.

#### Iteration 3:

1. Employee jobs and employee income.xlsx
   1. Download steps are as same as iteration1.
   2. Data Snippet1:

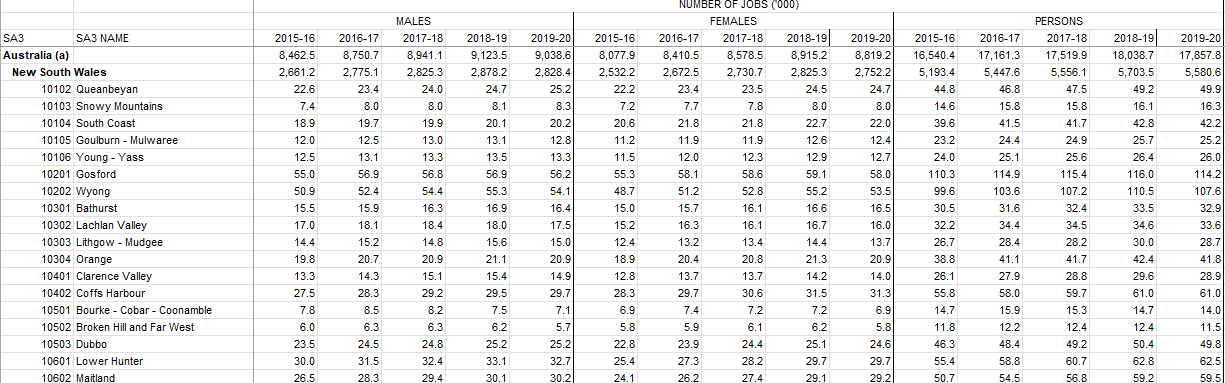


Figure 2.2.13 Retrieved 01/05/2023

* 1. Description: Employee jobs and employee income by sex, age, business characteristics and geography, 2015-16 to 2019-20
  2. The transform process is as same as iteration1, this data is from table 2.4.

1. Gender pay gap measures.xlsx
   1. Go to Gender Indicator, scroll down and find gender pay gap measures

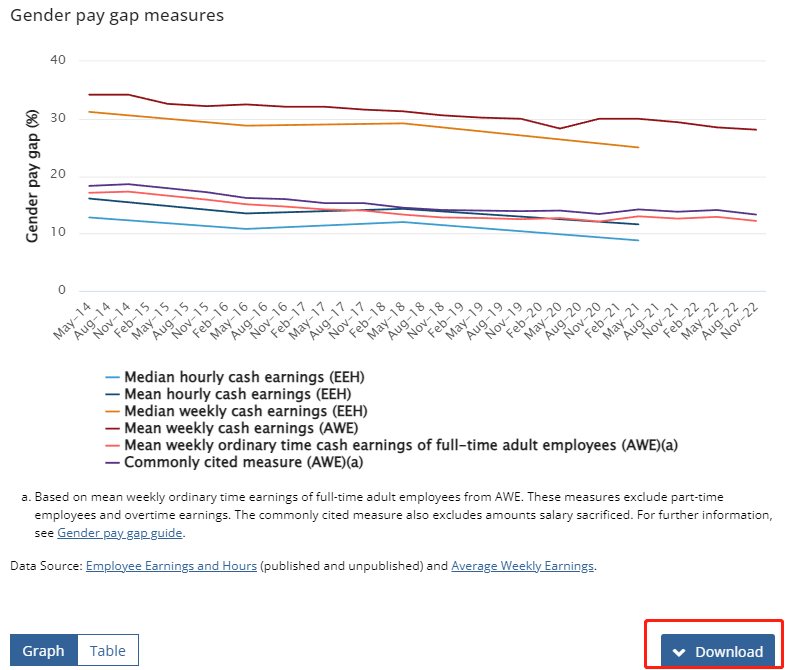


Figure 2.2.14 Retrieved 01/05/2023

* 1. Data Snippet2:

A picture containing text, screenshot, line, number

Description automatically generated

Figure 2.2.15 Retrieved 01/05/2023

* 1. Data transformation:

A picture containing screenshot, text, font

Description automatically generated

Figure 2.2.16 Retrieved 01/05/2023

* 1. After transformation:

A screenshot of a spreadsheet

Description automatically generated with medium confidence

Figure 2.2.17 Retrieved 01/05/2023

# 3.Data Usage:

The Data Governance will be constantly updated during the project development

## In Iteration 1:

* Data in xlsx files will be cleaned and transformed locally, processed data will be stored locally as well.
* Data is stored at frontend for use.

## In Iteration 2:

* Data in xlsx files will be cleaned and transformed locally, processed data will be stored in MySQL database and WordPress database.
* MySQL database is used for dashboard, that is connected to Tableau or PowerBI, WordPress database is used for website.

## In Iteration 3:

* There's no necessity to store the data in relational databases, as the data have no interconnections and aren't suited for segregation into distinct data tables.

# 4.Data Processing:

## 4.1Data Transformation:

Since the format of the data cannot be directly used for plotting, and the formats of different tables are different, it is necessary to transform the data into a unified format.

Check data completeness, missing values and duplicate values. Validation Check, make sure the data meet the desired standards, use statistical significance and verify accuracy of calculations.

## 4.2 Related Tools/Code:

1. Environment: python 3.8.8
2. Tools: VScode, Jupyter Notebook

## 4.3 Data Transformation

### 4.3.1 Iteration 1:

1. File: data\_transform.ipynb
2. Aim: Transfer data into machine friendly for build dashboard
3. Code Snippet:

Text

Description automatically generated

Figure 4.3.1.1 Retrieved 06/04/2023

1. Output file: Output.xlsx
   1. Data Snippet:

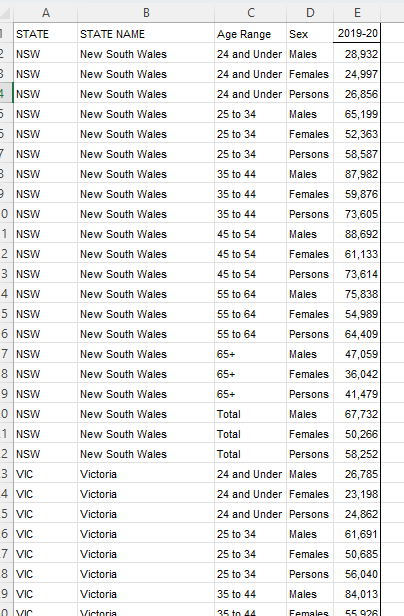


Figure 4.3.1.2 Retrieved 06/04/2023

### 4.3.2 Iteration 2:

1. AU\_PAY\_GAP\_BY\_STATE.ipynb
2. Aim: Transfer data into machine friendly format and output for as JavaScript list format.
3. Code Snippet:

A picture containing graphical user interface

Description automatically generated

Figure 4.3.2.1 Retrieved 29/04/2023

1. Output file: Java\_data.txt:

A picture containing text

Description automatically generated

Figure 4.3.2.1 Retrieved 29/04/2023

### 4.3.2 Iteration 3:

1. gender\_pay\_gap\_measures.ipynb
2. Aim: Transfer data into machine friendly format.
3. Code Snippet

A picture containing screenshot, text, font

Description automatically generated

1. Output file: pay\_gap.xlsx:

A screenshot of a spreadsheet

Description automatically generated with medium confidence

# Database Design:

## 5.1 ERD Diagram

Diagram

Description automatically generated

Figure 5.1 Retrieved 28/04/2023

**Database creation code snippet:**

* The name of the database is tp12\_gender\_inequality:

A screenshot of a computer program

Description automatically generated with medium confidence

Figure 5.2 Retrieved 01/05/2023

**Connect to RDS MySQL database:**

* This procedure will persist consistently across all iterations.
* Keys are stored under E:\StuDY\FIT5120\db\_keys this path, for safety reasons.

A screen shot of a computer

Description automatically generated with medium confidence

Figure 5.3 Retrieved 01/05/2023

**Insert data into RDS database remotely example:**

A picture containing text, screenshot, font

Description automatically generated

Figure 5.4 Retrieved 01/05/2023

**Insert pandas Dataframe into database remotely example:**

A picture containing text, screenshot

Description automatically generated

Figure 5.5 Retrieved 01/05/2023

**Query RDS database example:**

You can use SQL command to query the database, here is an example:

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5.6 Retrieved 01/05/2023

## 5.2 Data Lake

As mentioned above in data usage, not all data is necessary to put in relational database, and to manage outdated data, I will use AWS S3 to store all data. It features cost-optimized storage classes and lifecycle rules, which facilitate the transition of data to more economical storage classes. Older transactional data and pictures can be transferred to S3 Glacier. I can establish policies such as automatically shifting data to Glacier after a period of 30 days.

* For raw data

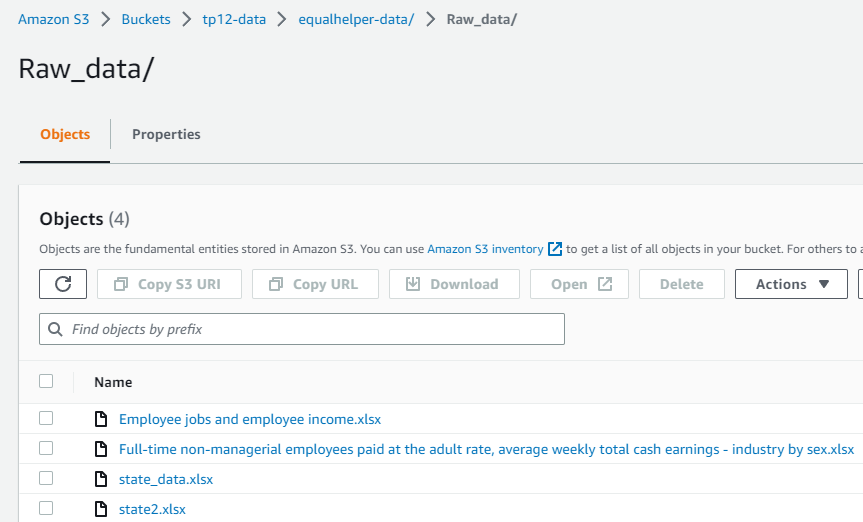


Figure 5.7 Retrieved 02/05/2023

* For processed data

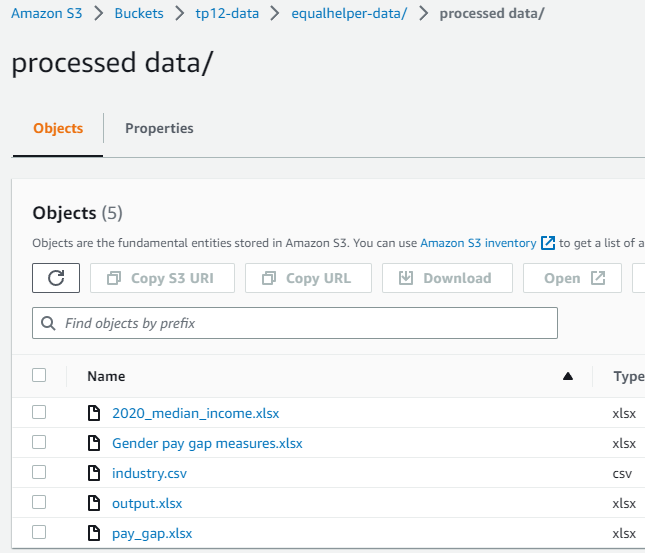


Figure 5.8 Retrieved 02/05/2023

# Data Analytics:

## Introduction:

Like many countries around the world, the gender pay gap has been a topic of concern in Australia for many years. This report aims to analyse the gender pay gap in Australia through data blindness, data hindsight, data insight and data forward look. By understanding historical trends, current conditions, and future projections, we can work towards creating a fairer workplace for all.

## Data Blindsight:

Before starting our analysis, firstly, we need to identify relevant data resources and collect the information necessary to understand the gender pay gap in Australia.

**Data Sources:**

* Australia Bureau of Statistics

**Gender Indicators or Key Metrics:**

* Median male/female salary by suburb/state
* Median male/female salary from 2012 – 2020
* Male/female salary in different industries

## Data Hindsight:

By analysing historical data from ABS, we can observe the trends and changes in the gender pay gap over the past years in Australia.

**Key findings:**

* Australia's gender pay gap has reduced over the past few decades; the current gender pay gap is around 20%.
* The pay gap varies by industry, with some industries, such as financial services and mining, showing larger gaps; However, for female, mining industry has highest salary, which does not mean that mining industry is not a good career choice.
* The gap also varies between occupations, with even wider gaps in managerial and professional roles.

## Data Insight:

By examining current data, we can find out the main drivers behind the gender pay gap and understand the current situation.

**Key drivers:**

* Occupational segregation: Women tend to work in low-paying industries and jobs, creating a pay gap. For example, most of teachers are women, but teacher at primary school, middle school and high school is not a high paying job.
* Part-time and casual jobs: Women are more likely to work part-time or casual jobs, which generally pay lower average salaries.
* Caregiving responsibilities: Women are more likely to take time off or reduce their working hours to care for children or other family members, which affects their career advancement and earning potential.
* Discrimination and bias: Discrimination and unconscious bias in hiring, promotion, and compensation decisions lead to pay gaps. As we can see, across all industry, women get pay lower than men.

## Data Foresight:

By examining the future trends and considering various scenarios, we can predict potential changes in the gender pay gap in Australia and we may identify some strategies to improve it.

**Possible scenarios:**

* If current trends continue, we can use regression model to predict when the gender pay gap will be fixed. Currently, we can see the percentage of gender pay gap is continuing to reduce.
* The pay gap could be reduced faster if we put more effort to popularise gender equality knowledge in the workplace and society.

**Suggestion:**

* Implement policies and practices that support work-life balance, such as flexible work arrangements and paid parental leave.
* Encourage and support women to enter high-paying industries and positions.
* Trains managers and HR professionals to identify and address bias in the workplace.
* Set gender diversity goals at all levels of the organization and regularly report on progress.

# Open Data Source Pipeline:

Diagram

Description automatically generated with medium confidence

Figure 2.10 Retrieved 06/04/2023

**Description:** This is a flow chart of designed automate data pipeline, it can avoid manually code and format data and allowing transformation happen on platform.

The pipeline is scheduled and managed by using Apache Airflow and AWS CloudWatch. Because the data is updated once a year, this is a batch type pipeline, I used CloudWatch to create a trigger which allows the lambda function to run once a year, the lambda function has the code to extract, transform and load the data into database and data lake.

Data is stored in AWS Redshift, which is an OLAP database, the database is connected to Tableau, so the dashboard will be updated automatically.

**Cost:** we are using the free tier and free tools, so the cost for now is 0 dollar