Data Management Plan for Equalhelper Website Development Project

By Team 925ers

Last Updated at 17/05/2023

Contents

1. 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.

2. 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	EXCEL	Yearly	Salary per hour; weekly hours worked, etc.	ABS Copyright Link

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	EXCEL	Yearly	Salary per Age; weekly hours worked, etc.	ABS Copyright Link

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	EXCEL	Yearly	Gender pay gap per year, median salary per suburb.	ABS Copyright Link

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

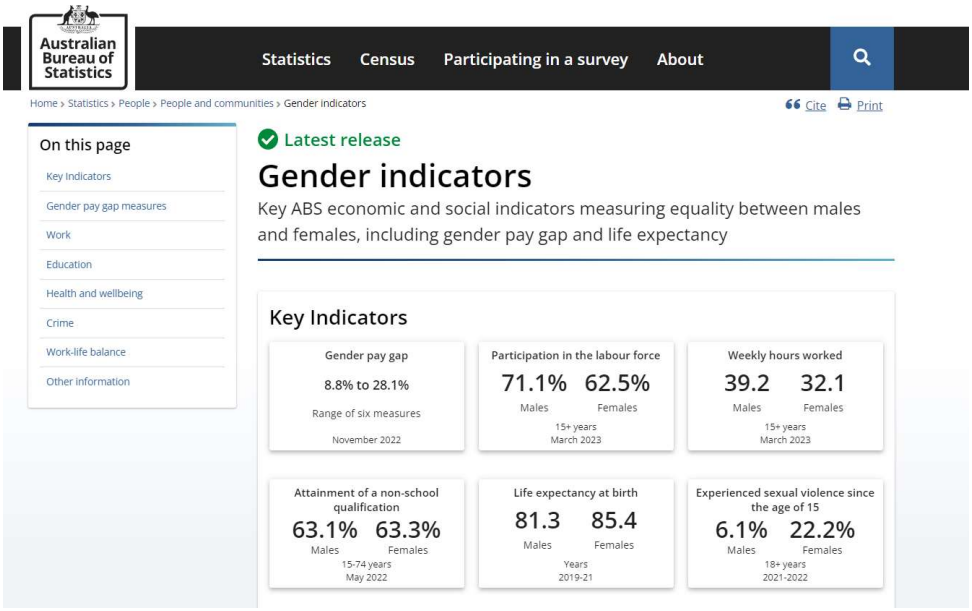


Figure 2.2.1. gender-indicators. Retrieved 25/04/2023 from

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
    a. Scroll down and find **Gender Pay Gap Measures.** Then click on median weekly cash earnings.

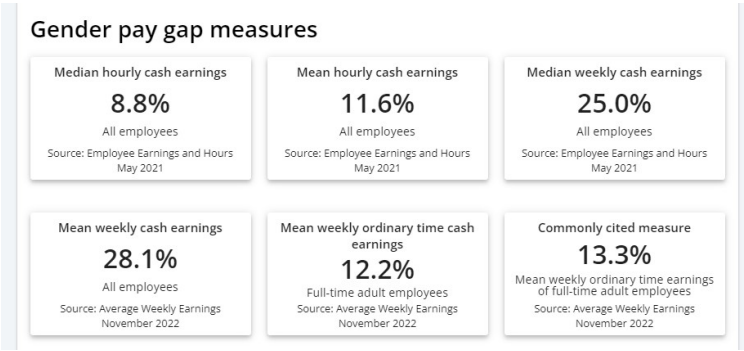


Figure 2.2.2 Gender Pay Gap Measures, Retrieved 25/04/2023 from

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

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

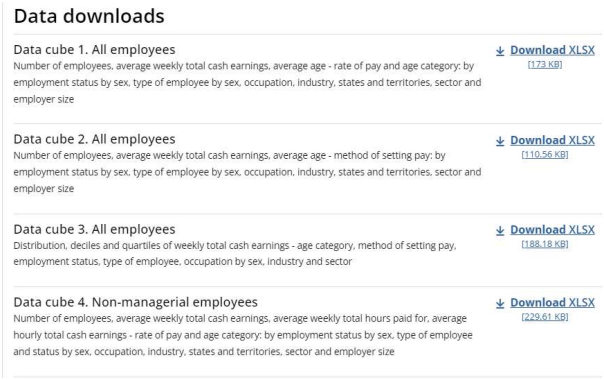


Figure 2.2.3 Gender Pay Gap Measures Data1, Retrieved 25/04/2023 from

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

c. Data snippet:

		MEDIAN EMPLOYEE INCOME PER JOB ($)												
		MALES					FEMALES					PERSONS		
GCCSA	GCCSA NAME	2015-16	2016-17	2017-18	2018-19	2019-20	2015-16	2016-17	2017-18	2018-19	2019-20	2015-16	2016-17	2017-18
Australia (a)		34,649	34,180	35,159	36,699	39,515	22,406	22,369	23,352	23,962	26,834	27,494	27,324	28,312
New South Wales		34,970	34,708	36,000	37,541	40,674	23,651	23,254	24,231	24,908	28,204	28,496	28,163	29,251
1GSYD	Greater Sydney	36,560	36,340	37,588	39,225	42,664	25,158	24,665	25,511	26,253	29,910	30,175	29,838	30,795
1RNSW	Rest of NSW	31,627	31,237	32,595	34,422	36,939	20,904	20,675	21,753	22,210	25,108	25,342	25,119	26,229
Victoria		34,853	34,435	35,232	36,372	39,240	21,561	21,675	22,611	23,069	25,939	26,903	26,648	27,758
2GMEL	Greater Melbourne	35,790	35,372	36,334	37,190	40,340	22,664	22,721	23,614	24,056	26,978	28,070	27,956	28,860
2RVIC	Rest of Vic.	31,076	30,631	31,665	33,279	35,493	18,402	18,604	19,707	20,076	22,790	23,328	23,380	24,451
Queensland		32,128	31,675	32,637	34,552	37,327	21,451	21,640	22,661	23,490	26,132	25,975	25,981	26,922
3GBRI	Greater Brisbane	34,875	34,750	35,702	37,426	40,373	23,812	24,179	25,080	26,000	28,969	28,458	28,700	29,649
3RQLD	Rest of Qld	29,876	29,016	29,973	31,803	34,615	19,569	19,550	20,500	21,226	23,679	23,810	23,567	24,615
South Australia		35,456	34,918	35,089	36,931	38,800	23,393	23,584	24,040	24,949	27,085	28,224	28,137	28,569
4GADE	Greater Adelaide	37,631	37,125	37,430	38,990	40,653	24,907	25,019	25,481	26,325	28,469	30,051	29,977	30,360
4RSAU	Rest of SA	27,733	27,029	27,470	29,697	32,123	18,513	18,784	19,264	20,150	22,280	22,163	22,188	22,791
Western Australia		37,931	36,603	37,503	39,131	41,657	21,898	21,796	22,950	23,213	25,725	28,571	28,000	29,096

Figure 2.2.4 Retrieved 06/04/2023.

d. 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. employees paid at the adult rate, average weekly total cash earnings - industry by sex.xlsx.

a. Same webpages find this plot and click download to download the data

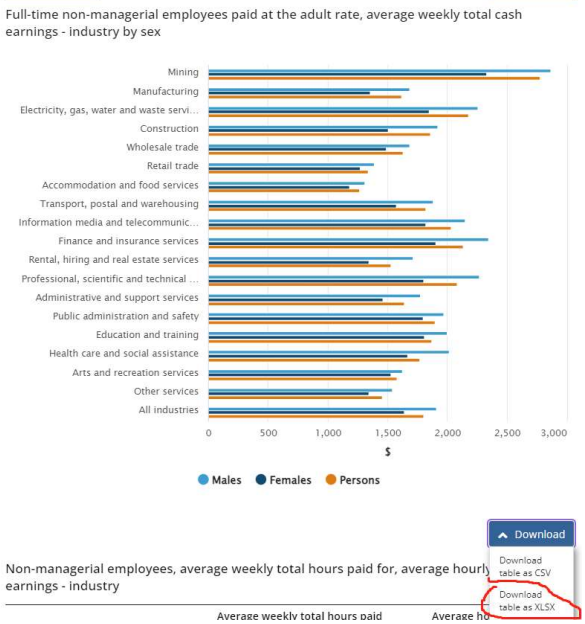


Figure 2.2.5 Gender Pay Gap Measures Data Table, Retrieved 25/04/2023 from

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

b. Data snippet:

INDUSTRY	Males ($)	Females ($)	Persons ($)
Mining	2,863.30	2,326.50	2,774.60
Manufacturing	1,684.70	1,356.20	1,616.70
Electricity, gas, water and waste services	2,257.80	1,846.60	2,176.40
Construction	1,917.30	1,505.30	1,860.40
Wholesale trade	1,686.00	1,486.30	1,626.10
Retail trade	1,386.60	1,271.50	1,337.60
Accommodation and food services	1,311.30	1,179.20	1,266.00
Transport, postal and warehousing	1,880.00	1,574.90	1,820.40
Information media and telecommunications	2,150.50	1,817.30	2,029.30
Finance and insurance services	2,346.30	1,905.10	2,134.10
Rental, hiring and real estate services	1,712.60	1,340.80	1,525.60
Professional, scientific and technical services	2,266.80	1,803.70	2,084.70
Administrative and support services	1,773.90	1,458.30	1,642.20

Figure 2.2.6 Retrieved 06/04/2023

c. Description: it shows the income across all industries by gender, very useful to understand the situation of various industries from a macro perspective.

Iteration 2:

1. 6524055002_DO004.xlsx
   a. Data snippet:

				Median ($)				
STATE	STATE NAME	Age Range	Sex	2015-16	2016-17	2017-18	2018-19	2019-20
0	Australia	24 and Under	Males	25,885	26,070	26,948	27,820	28,620
0	Australia	24 and Under	Females	22,092	22,268	22,750	23,507	24,448
0	Australia	24 and Under	Persons	23,910	24,080	24,756	25,543	26,380
0	Australia	25 to 34	Males	58,652	59,122	61,025	62,486	63,475
0	Australia	25 to 34	Females	45,688	46,216	47,768	49,300	50,414
0	Australia	25 to 34	Persons	52,155	52,655	54,300	55,661	56,677
0	Australia	35 to 44	Males	77,886	78,647	80,779	82,912	85,160
0	Australia	35 to 44	Females	49,151	50,390	52,411	54,531	56,655
0	Australia	35 to 44	Persons	63,233	64,344	66,449	68,563	70,669
0	Australia	45 to 54	Males	79,000	80,000	82,802	84,972	87,484
0	Australia	45 to 54	Females	51,180	52,530	54,471	56,527	58,960
0	Australia	45 to 54	Persons	63,572	65,000	67,225	69,439	71,911
0	Australia	55 to 64	Males	68,961	70,000	72,334	74,279	76,538
0	Australia	55 to 64	Females	48,272	49,279	50,558	51,956	53,838
0	Australia	55 to 64	Persons	57,799	58,834	60,464	62,128	64,218
0	Australia	65+	Males	38,636	39,038	41,339	42,735	47,379
0	Australia	65+	Females	29,263	29,846	30,869	32,000	35,265

Figure 2.2.7 Retrieved 06/04/2023

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

Iteration 3:

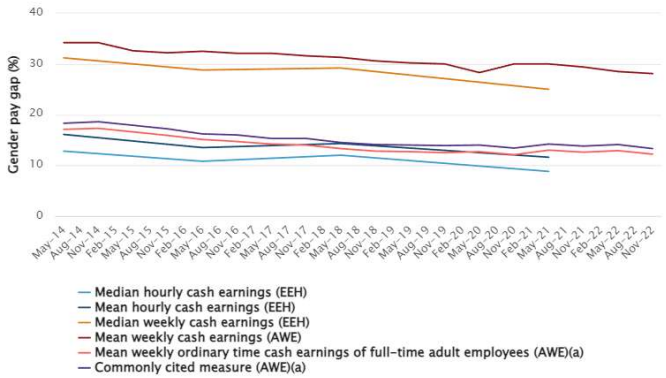
1. Employee jobs and employee income.xlsx
   a. Download steps are as same as iteration1.
   b. Data Snippet1:

		NUMBER OF JOBS ('000)														
		MALES					FEMALES					PERSONS				
SA3	SA3 NAME	2015-16	2016-17	2017-18	2018-19	2019-20	2015-16	2016-17	2017-18	2018-19	2019-20	2015-16	2016-17	2017-18	2018-19	2019-20
Australia (a)		8,462.5	8,750.7	8,941.1	9,123.5	9,038.6	8,077.9	8,410.5	8,578.5	8,915.2	8,819.2	16,540.4	17,161.3	17,519.9	18,038.7	17,857.8
New South Wales		2,661.2	2,775.1	2,825.3	2,878.2	2,828.4	2,532.2	2,672.5	2,730.7	2,825.3	2,752.2	5,193.4	5,447.6	5,556.1	5,703.5	5,580.6
10102	Queanbeyan	22.6	23.4	24.0	24.7	25.2	22.2	23.4	23.5	24.5	24.7	44.8	46.8	47.5	49.2	49.9
10103	Snowy Mountains	7.4	8.0	8.0	8.1	8.3	7.2	7.7	7.8	8.0	8.0	14.6	15.8	15.8	16.1	16.3
10104	South Coast	18.9	19.7	19.9	20.1	20.2	20.6	21.8	21.8	22.7	22.0	39.6	41.5	41.7	42.8	42.2
10105	Goulburn - Mulwaree	12.0	12.5	13.0	13.1	12.8	11.2	11.9	11.9	12.6	12.4	23.2	24.4	24.9	25.7	25.2
10106	Young - Yass	12.5	13.1	13.3	13.5	13.3	11.5	12.0	12.3	12.9	12.7	24.0	25.1	25.6	26.4	26.0
10201	Gosford	55.0	56.9	56.8	56.9	56.2	55.3	58.1	58.6	59.1	58.0	110.3	114.9	115.4	116.0	114.2
10202	Wyong	50.9	52.4	54.4	55.3	54.1	48.7	51.2	52.8	55.2	53.5	99.6	103.6	107.2	110.5	107.6
10301	Bathurst	15.5	15.9	16.3	16.9	16.4	15.0	15.7	16.1	16.6	16.5	30.5	31.6	32.4	33.5	32.9
10302	Lachlan Valley	17.0	18.1	18.4	18.0	17.5	15.2	16.3	16.1	16.7	16.0	32.2	34.4	34.5	34.6	33.6
10303	Lithgow - Mudgee	14.4	15.2	14.8	15.6	15.0	12.4	13.2	13.4	14.4	13.7	26.7	28.4	28.2	30.0	28.7
10304	Orange	19.8	20.7	20.9	21.1	20.9	18.9	20.4	20.8	21.3	20.9	38.8	41.1	41.7	42.4	41.8
10401	Clarence Valley	13.3	14.3	15.1	15.4	14.9	12.8	13.7	13.7	14.2	14.0	26.1	27.9	28.8	29.6	28.9
10402	Coffs Harbour	27.5	28.3	29.2	29.5	29.7	28.3	29.7	30.6	31.5	31.3	55.8	58.0	59.7	61.0	61.0
10501	Bourke - Cobar - Coonamble	7.8	8.5	8.2	7.5	7.1	6.9	7.4	7.2	7.2	6.9	14.7	15.9	15.3	14.7	14.0
10502	Broken Hill and Far West	6.0	6.3	6.3	6.2	5.7	5.8	5.9	6.1	6.2	5.8	11.8	12.2	12.4	12.4	11.5
10503	Dubbo	23.5	24.5	24.8	25.2	25.2	22.8	23.9	24.4	25.1	24.6	46.3	48.4	49.2	50.4	49.8
10601	Lower Hunter	30.0	31.5	32.4	33.1	32.7	25.4	27.3	28.2	29.7	29.7	55.4	58.8	60.7	62.8	62.5
10602	Maitland	26.5	28.3	29.4	30.1	30.2	24.1	26.2	27.4	29.1	29.2	50.7	54.5	56.8	59.2	59.5

Figure 2.2.8 Retrieved 01/05/2023

   c. Description: Employee jobs and employee income by sex, age, business characteristics and geography, 2015-16 to 2019-20
   d. The transform process is as same as iteration1, this data is from table 2.4.
2. Gender pay gap measures.xlsx
   a. Go to Gender Indicator, scroll down and find gender pay gap measures

Gender pay gap measures



— Median hourly cash earnings (EEH)
— Mean hourly cash earnings (EEH)
— Median weekly cash earnings (EEH)
— Mean weekly cash earnings (AWE)
— Mean weekly ordinary time cash earnings of full-time adult employees (AWE)(a)
— Commonly cited measure (AWE)(a)

a. Based on mean weekly ordinary time earnings of full-time adult employees from AWE. These measures exclude part-time employees and overtime earnings. The commonly cited measure also excludes amounts salary sacrificed. For further information, see Gender pay gap guide.

Data Source: Employee Earnings and Hours (published and unpublished) and Average Weekly Earnings.

Graph | Table

Download

Figure 2.2.9 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:

Employee jobs and employee income.xlsx:

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

```
newData = pd.read_excel('state_data.xlsx', sheet_name='Sheet1')
newData = newData.drop(['Unnamed: 1'], axis=1)
df_transposed = newData.T
df_transposed.columns = df_transposed.iloc[0]
df_transposed = df_transposed[1:]
col = df_transposed.columns.tolist()
col[0] = 'Year'
df_transposed.columns = col
df_transposed['Gender'] = Gender
df_transposed['Category'] = Category
df_transposed.reset_index(drop=True)
df_transposed.index = range(len(df_transposed.index))
```

```
NT_DATA = df_selected[NT_row_num:ACT_row_num]

ACT_DATA = df_selected[ACT_row_num:]

df_list = [NSW_data, VIC_data,QLD_data, SA_data, WA_data, TAS_data, NT_DATA, ACT_DATA]

for i in range(len(df_list)):
    df_list[i] = df_list[i].reset_index(drop=True)
    df_list[i].loc[0, 'Unnamed: 1'] = df_list[i].loc[0, df_list[i].columns[0]]
    df_list[i] = df_list[i].drop(df_list[i].columns[0], axis=1)
    df_list[i] = df_list[i].rename(columns={'Unnamed: 1': 'Location'})
    df_list[i]['Unnamed: 21'] = df_list[i]['Unnamed: 21'].apply(custom_conversion)
    df_list[i]['Unnamed: 26'] = df_list[i]['Unnamed: 26'].apply(custom_conversion)
```

Figure 4.3.1.1 Retrieved 06/04/2023

4. Data cleaning code snippet:

```
def replace_negative_and_inf(value):
    if value < 0 or np.isinf(value) or  value == 100:
        return 0
    else:
        return value
```

Figure 4.3.1.2 Retrieved 06/04/2023

5. Output file: Output.xlsx
    a. Data Snippet:

	A	B	C	D	E
1	STATE	STATE NAME	Age Range	Sex	2019-20
2	NSW	New South Wales	24 and Under	Males	28,932
3	NSW	New South Wales	24 and Under	Females	24,997
4	NSW	New South Wales	24 and Under	Persons	26,856
5	NSW	New South Wales	25 to 34	Males	65,199
6	NSW	New South Wales	25 to 34	Females	52,363
7	NSW	New South Wales	25 to 34	Persons	58,587
8	NSW	New South Wales	35 to 44	Males	87,982
9	NSW	New South Wales	35 to 44	Females	59,876
10	NSW	New South Wales	35 to 44	Persons	73,605
11	NSW	New South Wales	45 to 54	Males	88,692
12	NSW	New South Wales	45 to 54	Females	61,133
13	NSW	New South Wales	45 to 54	Persons	73,614
14	NSW	New South Wales	55 to 64	Males	75,838
15	NSW	New South Wales	55 to 64	Females	54,989
16	NSW	New South Wales	55 to 64	Persons	64,409
17	NSW	New South Wales	65+	Males	47,059
18	NSW	New South Wales	65+	Females	36,042
19	NSW	New South Wales	65+	Persons	41,479
20	NSW	New South Wales	Total	Males	67,732
21	NSW	New South Wales	Total	Females	50,266
22	NSW	New South Wales	Total	Persons	58,252
23	VIC	Victoria	24 and Under	Males	26,785
24	VIC	Victoria	24 and Under	Females	23,198
25	VIC	Victoria	24 and Under	Persons	24,862
26	VIC	Victoria	25 to 34	Males	61,691
27	VIC	Victoria	25 to 34	Females	50,685
28	VIC	Victoria	25 to 34	Persons	56,040
29	VIC	Victoria	35 to 44	Males	84,013
30	VIC	Victoria	35 to 44	Females	55,926

Figure 4.3.1.3 Retrieved 06/04/2023

employees paid at the adult rate, average weekly total cash earnings - industry by sex.xlsx.:

1. Data transform code snippet:

```
df_female = df[['INDUSTRY', 'Females ($)']].rename(columns={'Females ($)': 'salary'})
df_female['Gender'] = 'F'
df_male = df[['INDUSTRY', 'Males ($)']].rename(columns={'Males ($)': 'salary'})
df_male['Gender'] = 'M'
df = pd.concat([df_female,df_male], axis =0)
df.to_csv('industry.csv',index=False)
```

Figure 4.3.1.4 Retrieved 06/04/2023

2. After transformation:

	A	B	C
1	INDUSTRY	salary	Gender
2	Mining	2326.5	F
3	Manufacturing	1356.2	F
4	Electricity, gas, water and waste services	1846.6	F
5	Construction	1505.3	F
6	Wholesale trade	1486.3	F
7	Retail trade	1271.5	F

Figure 4.3.1.5 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:

```
VIC_data = df_list[1][['Location', 'Unnamed: 21', 'Unnamed: 26']]
VIC_data['STATE'] = 'VIC'
VIC_data.rename(columns={'Unnamed: 21': 'male_median_salary', 'Unnamed: 26': 'female_median_salary'}, inplace=True
VIC_data = VIC_data.astype({'male_median_salary': int, 'female_median_salary': int})
VIC_data['gender_pay_gap'] = ((VIC_data['male_median_salary'] - VIC_data['female_median_salary']) / VIC_data['male

QLD_data = df_list[2][['Location', 'Unnamed: 21', 'Unnamed: 26']]
QLD_data['STATE'] = 'QLD'
QLD_data.rename(columns={'Unnamed: 21': 'male_median_salary', 'Unnamed: 26': 'female_median_salary'}, inplace=True
QLD_data = QLD_data.astype({'male_median_salary': int, 'female_median_salary': int})
QLD_data['gender_pay_gap'] = ((QLD_data['male_median_salary'] - QLD_data['female_median_salary']) / QLD_data['male

SA_data = df_list[3][['Location', 'Unnamed: 21', 'Unnamed: 26']]
SA_data['STATE'] = 'SA'
SA_data.rename(columns={'Unnamed: 21': 'male_median_salary', 'Unnamed: 26': 'female_median_salary'}, inplace=True)
SA_data = SA_data.astype({'male_median_salary': int, 'female_median_salary': int})
SA_data['gender_pay_gap'] = ((SA_data['male_median_salary'] - SA_data['female_median_salary']) / SA_data['male_med
```

Figure 4.3.2.1 Retrieved 29/04/2023

4. Output file: Java_data.txt:


```
[
    {location: "New South Wales", male_median_salary: 40674, female_median_salary: 28204, state: "NSW", gender_pay_gap: 30.658405861238137},
    {location: "Braidwood", male_median_salary: 36541, female_median_salary: 22951, state: "NSW", gender_pay_gap: 37.19110040776115},
    {location: "Queanbeyan", male_median_salary: 42650, female_median_salary: 34798, state: "NSW", gender_pay_gap: 18.41031652989449},
    {location: "Queanbeyan Region", male_median_salary: 61036, female_median_salary: 41521, state: "NSW", gender_pay_gap: 31.972934006160298},
    {location: "Bombala", male_median_salary: 23400, female_median_salary: 15846, state: "NSW", gender_pay_gap: 32.28205128205128 5},
    {location: "Batemans Bay", male_median_salary: 27166, female_median_salary: 20042, state: "NSW", gender_pay_gap: 26.223956416108 37},
    {location: "Deua - Wadbilliga", male_median_salary: 87201, female_median_salary: 46762, state: "NSW", gender_pay_gap: 46.374468182704 33},
    {location: "Eden", male_median_salary: 27721, female_median_salary: 21740, state: "NSW", gender_pay_gap: 21.575700732296816},
    {location: "Eurobodalla Hinterland", male_median_salary: 30697, female_median_salary: 23406, state: "NSW", gender_pay_gap: 23.751506661888783},
    {location: "Moruya - Tuross Head", male_median_salary: 29473, female_median_salary: 23440, state: "NSW", gender_pay_gap: 20.469582329589795},
    {location: "Goulburn", male_median_salary: 38306, female_median_salary: 27225, state: "NSW", gender_pay_gap: 28.927583146243407},
    {location: "Goulburn Region", male_median_salary: 30805, female_median_salary: 22097, state: "NSW", gender_pay_gap: 28.26813828923876},
    {location: "Yass", male_median_salary: 40984, female_median_salary: 28499, state: "NSW", gender_pay_gap: 30.46310755416748},
    {location: "Yass Region", male_median_salary: 42059, female_median_salary: 30355, state: "NSW", gender_pay_gap: 27.82757554863406},
```


Figure 4.3.2.2 Retrieved 29/04/2023

Industry.csv:

1. Data transformation:

```
df = pd.read_excel('Full-time non-managerial employees paid at the adult rate, average weekly total cash earnings - industry by sex.xlsx')
df_female = df[['INDUSTRY', 'Females ($)']].rename(columns={'Females ($)': 'salary'})
df_female['Gender'] = 'F'
df_male = df[['INDUSTRY', 'Males ($)']].rename(columns={'Males ($)': 'salary'})
df_male['Gender'] = 'M'
df = pd.concat([df_female,df_male], axis =0)
df.to_csv('industry.csv',index=False)
df.head()
```

	INDUSTRY	salary	Gender
0	Mining	2326.5	F
1	Manufacturing	1356.2	F
2	Electricity, gas, water and waste services	1846.6	F
3	Construction	1505.3	F
4	Wholesale trade	1486.3	F

Figure 4.3.2.3 Retrieved 29/04/2023

2. After transformation:

	A	B	C
	INDUSTRY	salary	Gender
	Mining	2326.5	F
	Manufacturing	1356.2	F
	Electricity, gas, water and waste services	1846.6	F
	Construction	1505.3	F
	Wholesale trade	1486.3	F
	Retail trade	1271.5	F
	Accommodation and food services	1179.2	F
	Transport, postal and warehousing	1574.9	F
	Information media and telecommunications	1817.3	F
	Finance and insurance services	1905.1	F
	Rental, hiring and real estate services	1340.8	F
	Professional, scientific and technical services	1803.7	F
	Administrative and support services	1458.3	F

Figure 4.3.2.4 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

```
df = df[['Date', 'Mean weekly cash earnings (AWE) (%)',  'Commonly cited measure (AWE)(a) (%)' ,
         'Mean weekly ordinary time cash earnings of full-time adult employees (AWE)(a) (%)']]
df = df.dropna()
df = df.rename(columns = {'Mean weekly cash earnings (AWE) (%)': 'Mean weekly cash earnings(%)',
                          'Commonly cited measure (AWE)(a) (%)': 'Commonly cited measure(%)',
                          'Mean weekly ordinary time cash earnings of full-time adult employees (AWE)(a) (%)' : 'Mean weekly cash earnings of full-time adult employees(%)'})
df1 = df[['Date', 'Mean weekly cash earnings(%)']].rename(columns = {'Mean weekly cash earnings(%)':'percentage'})
df1['type'] = 'Mean weekly cash earnings(%)'
df2 = df[['Date', 'Commonly cited measure(%)']].rename(columns = {'Commonly cited measure(%)':'percentage'})
df2['type'] = 'Commonly cited measure(%)'
df3 = df[['Date', 'Mean weekly cash earnings of full-time adult employees(%)']].rename(columns = {'Mean weekly cash earnings of full-time adult employees(%)':'percentage'})
df3['type'] = 'Mean weekly cash earnings of full-time adult employees(%)'
```

Figure 4.3.3.1 Retrieved 03/05/2023

4. Output file: pay_gap.xlsx:

	A	B	C	D	E	F
1	Date	percentage	type			
2	May-14	34.2	Mean weekly cash earnings(%)			
3	Nov-14	34.2	Mean weekly cash earnings(%)			
4	May-15	32.6	Mean weekly cash earnings(%)			
5	Nov-15	32.2	Mean weekly cash earnings(%)			
6	May-16	32.5	Mean weekly cash earnings(%)			
7	Nov-16	32.1	Mean weekly cash earnings(%)			
8	May-17	32.1	Mean weekly cash earnings(%)			
9	Nov-17	31.6	Mean weekly cash earnings(%)			
10	May-18	31.3	Mean weekly cash earnings(%)			
11	Nov-18	30.6	Mean weekly cash earnings(%)			
12	May-19	30.2	Mean weekly cash earnings(%)			
13	Nov-19	30	Mean weekly cash earnings(%)			
14	May-20	28.3	Mean weekly cash earnings(%)			
15	Nov-20	30	Mean weekly cash earnings(%)			
16	May-21	30	Mean weekly cash earnings(%)			
17	Nov-21	29.4	Mean weekly cash earnings(%)			
18	May-22	28.5	Mean weekly cash earnings(%)			
19	Nov-22	28.1	Mean weekly cash earnings(%)			

Figure 4.3.3.2 Retrieved 03/05/2023

5. Database Design:

5.1 ERD Diagram

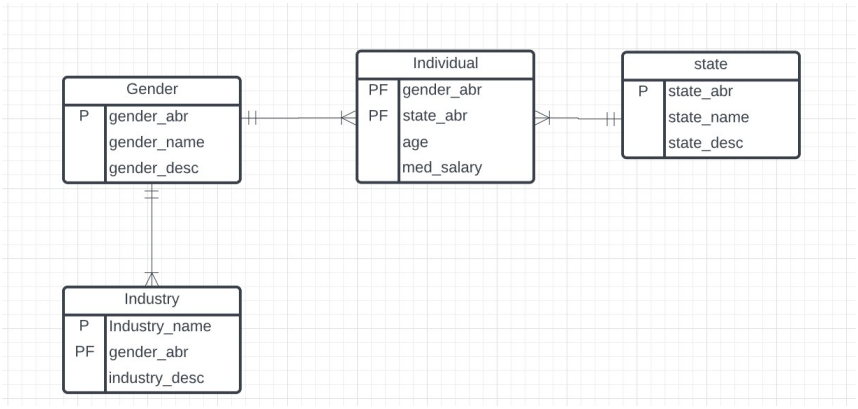


Figure 5.1 Retrieved 28/04/2023

Database creation code snippet:

- The name of the database is tp12_gender_inequality:

```
CREATE DATABASE tp12_gender_inequality;

USE tp12_gender_inequality;

Create Table gender (
    gender varchar(5) primary key,
    gender_desc varchar(100)
);

CREATE Table state(
    state_abr varchar(10) primary key,
    state_name varchar(10) UNIQUE ,
    state_desc varchar(100)
);

CREATE Table individual(
    gender varchar(5),
    state_abr varchar(10),
    age_range varchar(30),
    med_salary int

);

ALTER TABLE individual
ADD PRIMARY KEY (gender,state_abr, age_range);
ALTER TABLE individual
ADD FOREIGN KEY (gender) REFERENCES gender(gender);
ALTER TABLE individual
ADD FOREIGN KEY (state_abr) REFERENCES state(state_abr);
```

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.

```
credentials_path = r'E:\StuDY\FIT5120\db_keys'
sys.path.append(credentials_path)
from db_keys import mysql_host, mysql_user, mysql_password, mysql_database

import pymysql

# Connect to your MySQL database

mydb = mysql.connector.connect(
    host=mysql_host,
    user=mysql_user,
    password=mysql_password,
    database=mysql_database
)

mycursor = mydb.cursor()
```

Figure 5.3 Retrieved 01/05/2023

Insert data into RDS database remotely example:

```
# Insert state data into the `state` table
state_data = [
    ('ACT', 'Australian Capital Territory'),
    ('NSW', 'New South Wales'),
    ('NT', 'Northern Territory'),
    ('QLD', 'Queensland'),
    ('SA', 'South Australia'),
    ('TAS', 'Tasmania'),
    ('VIC', 'Victoria'),
    ('WA', 'Western Australia')
]

sql = "INSERT INTO state (state_abr, state_name) VALUES (%s, %s)"
val = state_data
mycursor.executemany(sql, val)

mydb.commit()
```

Figure 5.4 Retrieved 01/05/2023

Insert pandas Dataframe into database remotely example:

```
try:

    # Insert individual data
    for index, row in df_individual.iterrows():
        sql = f"""
        INSERT INTO individual (state_abr, age_range, gender, med_salary)
        VALUES ('{row['STATE']}', '{row['Age Range']}', '{row['gender']}', {row['med_salary']});
        """
        mycursor.execute(sql)

    mydb.commit()

finally:
    mydb.close()
```

Figure 5.5 Retrieved 01/05/2023

Query RDS database example:

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

```
45
46 •  SELECT * FROM industry;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Co

industry_name	gender	industry_desc	salary
Mining	F	NULL	2327
Manufacturing	F	NULL	1356
Electricity, gas, water and waste services	F	NULL	1847
Construction	F	NULL	1505
Wholesale trade	F	NULL	1486
Retail trade	F	NULL	1272
Accommodation and food services	F	NULL	1179
Transport, postal and warehousing	F	NULL	1575
Information media and telecommunications	F	NULL	1817
Finance and insurance services	F	NULL	1905

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

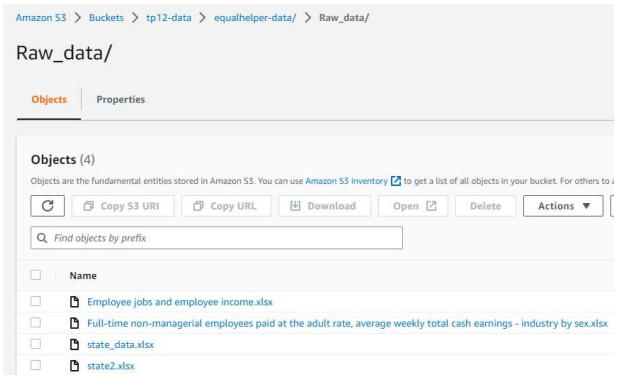


Figure 5.7 Retrieved 02/05/2023

- For processed data

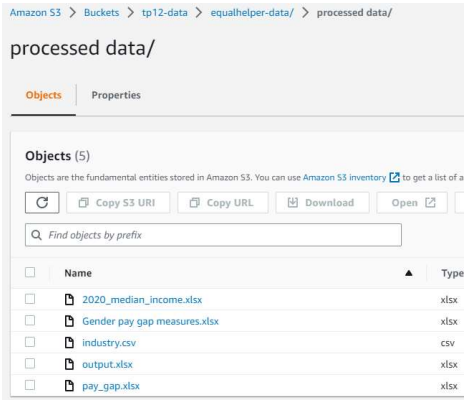


Figure 5.8 Retrieved 02/05/2023

6. 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.

7. Open Data Source Pipeline:

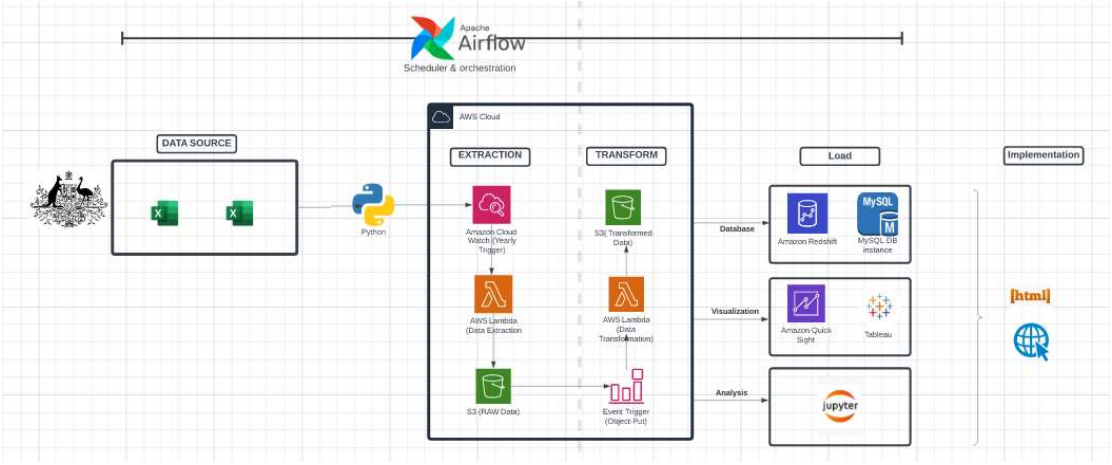


Figure 2.10 Retrieved 06/04/2023

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