Data Management Plan For Equalhelper Website Development Project

By Team 925ers

Last Updated at 06/04/2023

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. 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. The data used in this project comes from trusted sources that provide open data. e.g. Australia Government

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

Name	Link	Physical Access Used	Frequency of Iteration	Granularity	Copyright details
Gender indicators	Data Source Link	EXCEL	Yearly	Salary per hour; weekly hours worked, etc.	ABS Copyright Link

Description:

The data is collected from Australia Bureau Statistics, it is an open data source, free and Australia government founded website.

Data Extracted:

1. Employee jobs and employee income.xlsx
    a. 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	42,684	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	26,848	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	21,798	22,950	23,213	25,725	28,571	28,000	29,096

    b. 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. 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

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

3. 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

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

Data 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.

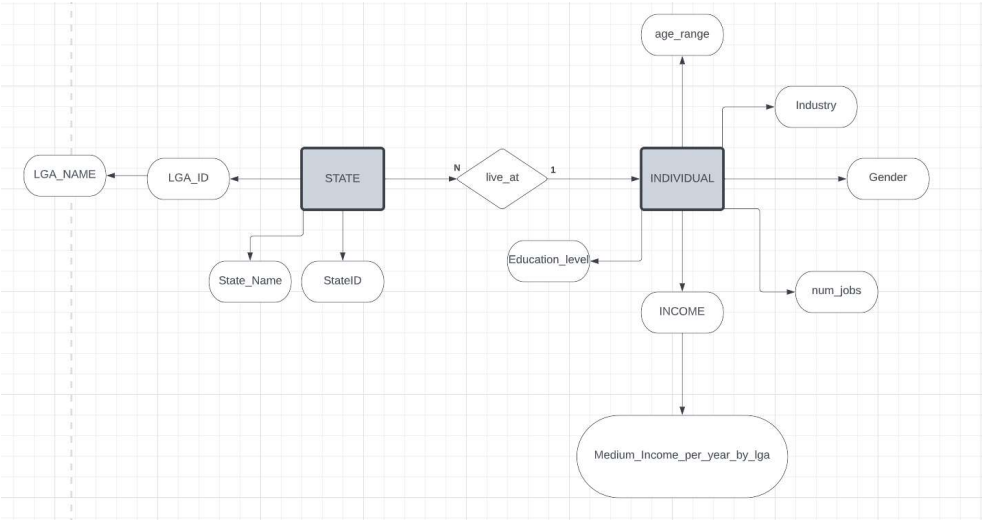
Check data completeness, missing values and duplicate values. Validation Check, make sure the data meet the desired standards, use statistical significance and verify accuracy of calculations.

Related Tools/Code:

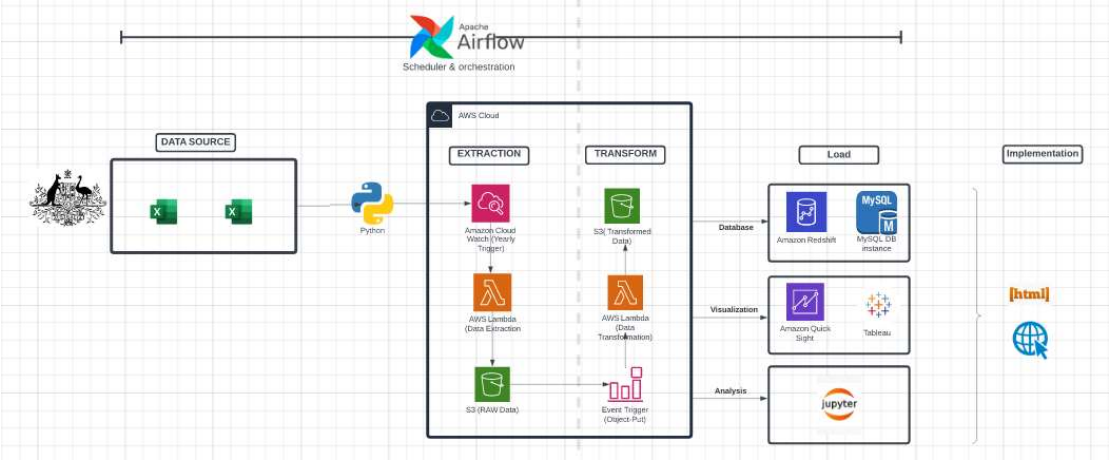
1. Environment: python 3.8.8
2. Tools: Vscode, Jupyter Notebook
3. File: data_transform.ipynb
   Code Snippet:

```
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))
```

ERD Diagram:



2.2 Open Data Source Pipeline:



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 CouldWatch 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

3. Data Generated: To be decided.