UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

BSD2343 DATA WAREHOUSING SEMESTER II 2023/2024

GROUP NAME: QUERY

TITLE:

UNVEILING GENDER DISPARITIES: A DATA-DRIVEN EXPLORATION OF COLLEGE MAJORS AND STEM FIELDS (SDG 5: GENDER EQUALITY)

PREPARED FOR: DR AZUANA BINTI RAMLI

Student ID:	Student Name:	Section:
SD22013	AQILAH MAISARAH BINTI AZIZI	01G
SD22005	NURKHAIRUL IZZATI BINTI MOHD SALLEHAN	02G
SD22055	AMIRAH AISYAH BINTI ABDULLAH	01G
SD22066	SHAHIRA BINTI MOHAIDEEN MEERA	02G
SD22057	MUHAMMAD FIRAS BIN IRWAN	01G

Table Of Contents

1.0 Background

1.1 Project Description

In this era, gender equality is as an important issue to talk about in Science, Technology, Engineering, and Mathematics (STEM) fields. Even though there is a positive change in a recent year, women are still left behind in this field and facing many challenges to enter this area compared to men, a lot of factors contributing to this inequality such as social stereotypes, lack of role models, gender bias. For example, in Europe, only 22% of natural science positions and 17.9% of engineering and technology positions are filled by women, which is not even half of the popularity. By highlighting the gender gap, unbiased society is one of the important issues. Women can show their skills and creativity in this field, to prove that they can succeed in these areas of study.

This project follows the Sustainable Development Goal 5 (SDG 5) to get deeper understanding of this topic by analysing data based on the selection of the college majors and the course they graduate. It also aims to understand the reasons behind the lack of women availability in STEM fields. We analysed the educational path chosen by women in STEM, we can identify trends and gap in it, such as percentage of women pursue STEM field and their career success compared to men. In addition, it is important to have strategies and programs that can support gender equality in STEM. As an example, by doing a mentorship program that will help in promoting and creating work environment that inspire and empower women to pursue their career in STEM fields. We need to overcome this issue, since it is important for an equality and make a various surrounding that will lead to innovative workforce that is not only focus on men anymore.

In conclusion, this gender equality issue still be a huge issue nowadays. It is important for every institution and individual to work together in creating a gender equality environment in STEM to get more aware towards talents and contribution of all people. By addressing this issue, we can actively break the gender inequality in this STEM field and not looking down on women representative anymore.

1.2 Problem to Be Solved

The gender gap in STEM fields is still a big problem, it affects the potential to build an equal society. Even though the number of educated women and the openness of people's minds have increased, women are still less likely to choose a career in STEM fields than men. Therefore, the lack of interest is not only a personal issue for every woman that may prevent them from reaching their full potential, but it is also a global issue that prevents the development of society and innovation.

If that is the case, what factors have contributed to overcome this gender gap? The most highlighted factor of all, is the social stereotype, which tells girls from a young age that STEM field is only meant for men. This means that women cannot be in STEM courses nor do women take on paid or volunteer jobs in the STEM industry. It can be clearly seen, when young girls and women cannot see the example of other women before them in that field, it confirms the stereotype that women do not deserve a place in that field. Other than that, the gender bias that make this as a huge problem in hiring and promoting women in this industry.

This gender gaps are not only affecting individual, but it affects wider social and economic issues. In other words, when most women work on research and other STEM-related activities, the nation ignore or undervalue a lot of talents and voices. Gender diversity promotes creativity and contribute to search for solutions of global challenges and problems. Thus, the non-significance and role of women in STEM work is another factor that does not only support criminal justice but also limits and slows down international development.

To summarize, this gender inequality in STEM field, need to be analyse and find for the solution. It is because we need more younger generation like woman in this field so that they can join with men to produce something big and valued to be in STEM field. Societies should prioritize to address the root cause of the disparities to promote innovative and sustainable development of an inclusive and diverse workforce.

1.3 Objectives

The objectives of the project are:

1. To examine the ratio female to male graduates across college major

2. To investigate the ratio female to male students in STEM fields

3. To determine whether the median salary is one of the factors leading to fewer women choosing STEM fields.

4. To investigate correlation between employment outcomes and genders in STEM field.

1.4 Data Schema

A database schema is a design or structure that describes how data is arranged, stored, and retrieved in a database management system (DBMS). It provides information on the logical and physical structure of the database, such as tables, columns, relationships, constraints, and indexes. Our dataset consists of four tables which is allage, recentlygrads, gradstudent and womensstem as shown below:

No	**Table Name**	**Column Name**	**Data Type**	**Description**
1.	allage	major_code	numeric	The code associated with the major
		major_name	String	The specific major of the field of study
		major_course	String	The category of the major
		total_students	numeric	The total number of students in the major
		employed_grad	numeric	The number of employed graduates from the major
		employed_full_time_year_round	numeric	The number of employed graduates from the major who are employed full-time year-round
		unemployed_grad	numeric	The number of unemployed graduates from the major
		unemployment_rate	numeric	The unemployment rate of graduates from the major
		median_salary	numeric	The median salary of graduates from the major
		P25th_salary	numeric	The 25th percentile salary of graduates from the major
		P75th_salary	numeric	The 75th percentile salary of graduates from the major

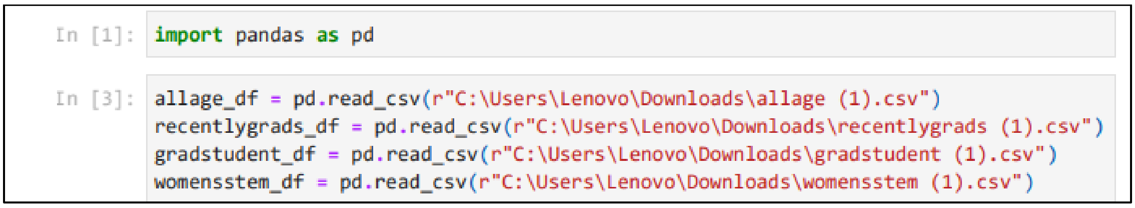
2.	recentlygrads	popularity_rank	numeric	The rank of the major in terms of popularity
		major_code	numeric	The code associated with the major.
		major_name	String	The specific major of the field of study
		major_course	String	The category of the major
		total_students	numeric	The total number of students in the major
		sample_size	numeric	The sample size of the major
		men	numeric	The number of male students in the major
		women	numeric	The number of female students in the major
		sharewomen	numeric	The percentage of female students in the major
		employed_grad	numeric	The number of employed graduates from the major
		full_time	numeric	The number of full-time employed graduates from the major
		part_time	numeric	The number of part-time employed graduates from the major
		full_time_year_round	numeric	The number of full-time year-round employed graduates from the major
		unemployed_grad	numeric	The number of unemployed graduates from the major
		unemployment_rate	numeric	The unemployment rate of graduates from the major

		median_salary	numeric	The median salary of graduates from the major
		P25th_salary	numeric	The 25th percentile salary of graduates from the major
		P75th_salary	numeric	The 75th percentile salary of graduates from the major
		college_jobs	numeric	The number of college jobs held by graduates from the major
		non_college_jobs	numeric	The number of non-college jobs held by graduates from the major
		low_wage_jobs	numeric	The number of low-wage jobs held by graduates from the major
3.	gradstudent	major_code	numeric	The broader category of the field of study
		major_name	String	The specific major of the field of study
		major_course	String	The category of the major
		grad_total	numeric	The total number of graduates from the major
		grad_sample_size	numeric	The sample size of graduates from the major
		grad_employed	numeric	The number of graduates employed
		grad_full_time_year_round	numeric	The number of graduates employed full-time year-round
		grad_unemployed	numeric	The number of graduates
		grad_unemployment_rate	numeric	The unemployment rate of graduates
		grad_median_salary	numeric	The median salary of graduates

		grad_P25th_salary	numeric	The 25th percentile salary of graduates
		grad_P7th_salary	numeric	The 75th percentile salary of graduates
		nongrad_total	numeric	The total number of non-graduates from the major
		nongrad_employed	numeric	The number of non-graduates employed
		nongrad_full_time_year_round	numeric	The number of non-graduates employed full-time year-round
		nongrad_unemployed	numeric	The number of non-graduates unemployed
		nongrad_unemployment_rate	numeric	The unemployment rate of non-graduates
		nongrad_median_salary	numeric	The median salary of non-graduates
		nongrad_P25th_salary	Integer	The 25th percentile salary of non-graduates
		nongrad_P75th_salary	numeric	The 75th percentile salary of non-graduates
		grad_share	numeric	The 75th percentile salary of non-graduates
		diff_salary	numeric	The difference between the median salary of graduates and non-graduates
4.	womensstem	popularity_rank	numeric	The rank of the major in terms of popularity
		major_code	numeric	The code associated with the major
		major_name	String	The specific major of the field of study
		major_course	String	The category of the major

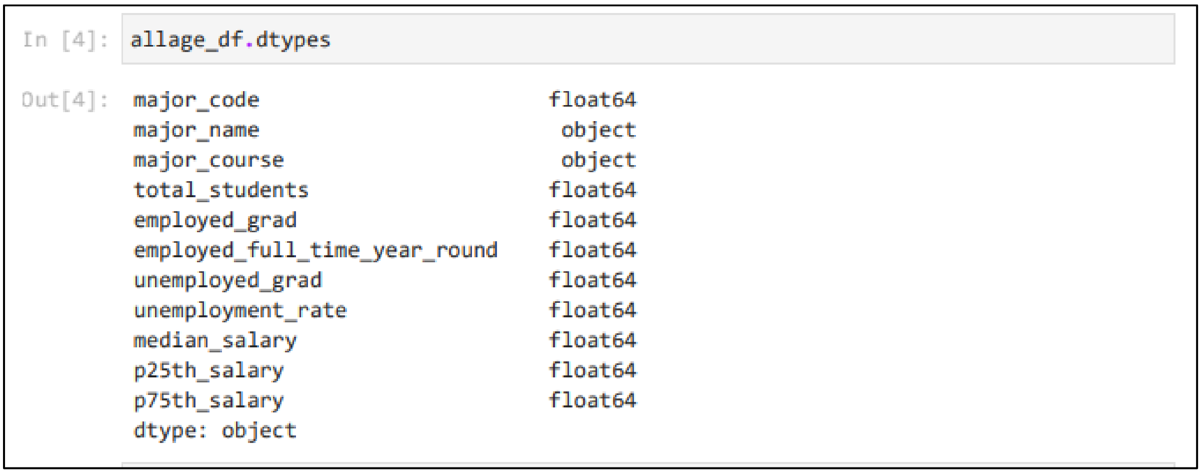
		total_students	numeric	The total number of students in the major
		men	numeric	The number of male students in the major
		women	numeric	The number of female students in the major
		sharewomen	numeric	The percentage of female students in the major
		median_salary	numeric	The median salary of graduates from the major

Check for the datatype:

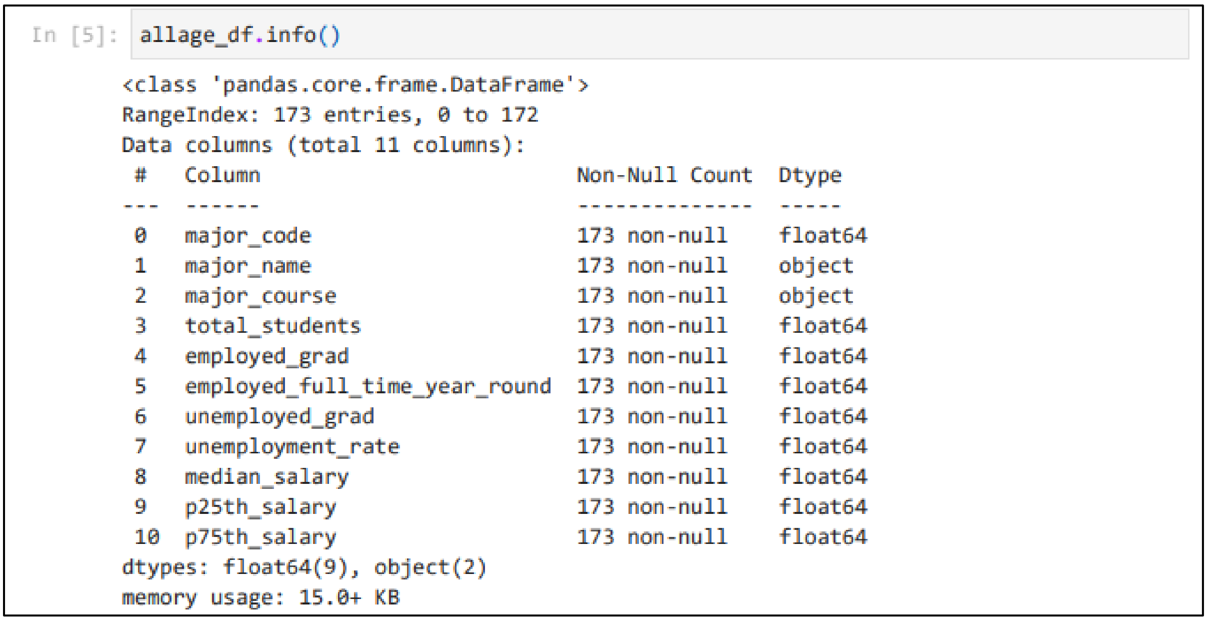


```
In [1]: import pandas as pd

In [3]: allage_df = pd.read_csv(r"C:\Users\Lenovo\Downloads\allage (1).csv")
        recentlygrads_df = pd.read_csv(r"C:\Users\Lenovo\Downloads\recentlygrads (1).csv")
        gradstudent_df = pd.read_csv(r"C:\Users\Lenovo\Downloads\gradstudent (1).csv")
        womensstem_df = pd.read_csv(r"C:\Users\Lenovo\Downloads\womensstem (1).csv")
```

Figure 1.4.1 shows the libraries that were used to find data schema



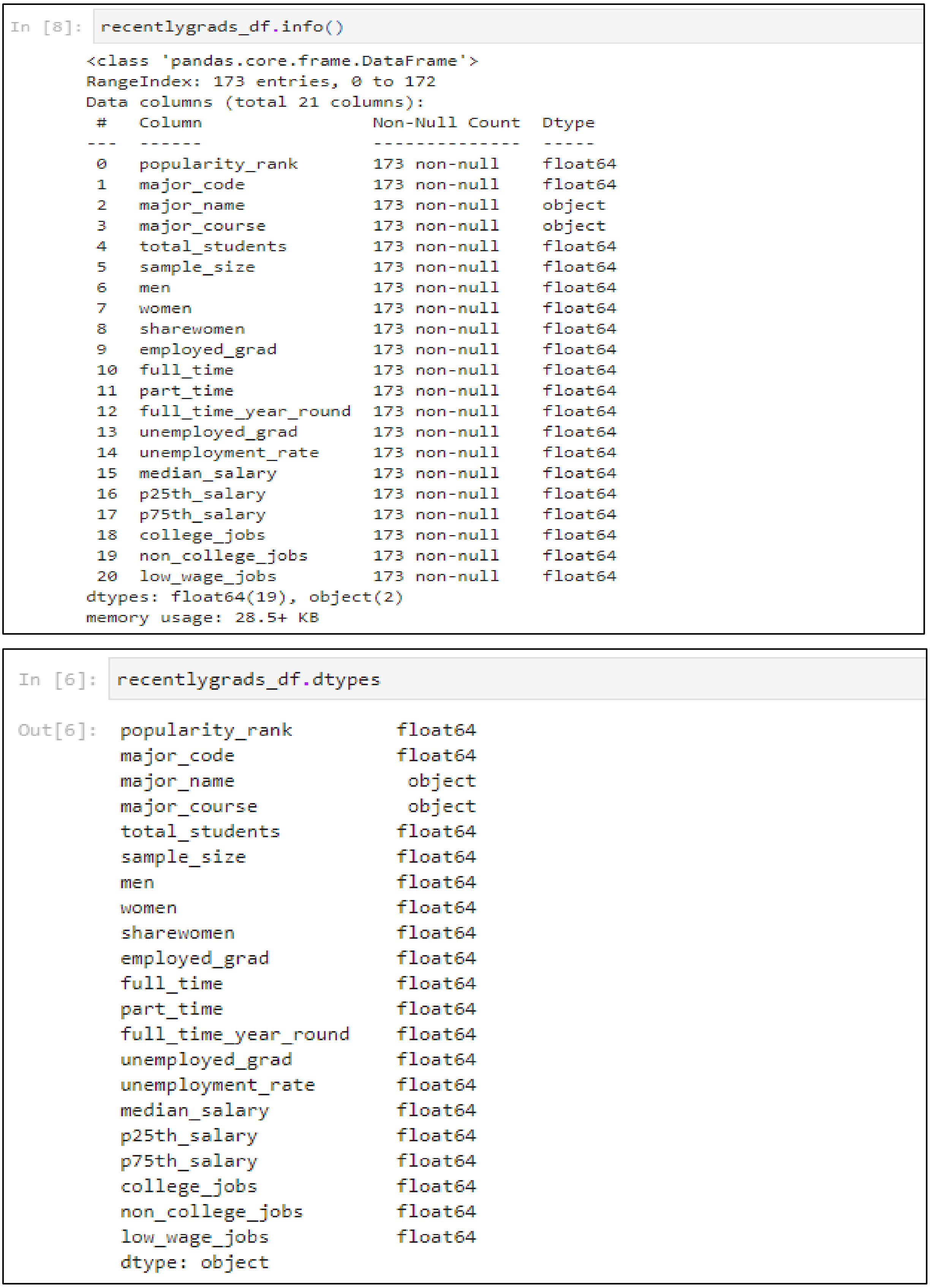
```
In [4]: allage_df.dtypes

Out[4]: major_code                         float64
        major_name                          object
        major_course                        object
        total_students                     float64
        employed_grad                      float64
        employed_full_time_year_round      float64
        unemployed_grad                    float64
        unemployment_rate                  float64
        median_salary                      float64
        p25th_salary                       float64
        p75th_salary                       float64
        dtype: object
```



```
In [5]: allage_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 173 entries, 0 to 172
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   major_code                    173 non-null    float64
 1   major_name                    173 non-null    object
 2   major_course                  173 non-null    object
 3   total_students                173 non-null    float64
 4   employed_grad                 173 non-null    float64
 5   employed_full_time_year_round 173 non-null    float64
 6   unemployed_grad               173 non-null    float64
 7   unemployment_rate             173 non-null    float64
 8   median_salary                 173 non-null    float64
 9   p25th_salary                  173 non-null    float64
 10  p75th_salary                  173 non-null    float64
dtypes: float64(9), object(2)
memory usage: 15.0+ KB
```

Figure 1.4.2 allage Tables

Based on figure 1.4.2 above, the raw dataset for the allage table is basically information about various academic majors, focusing on graduate outcomes and employment statistics. Each row in the table represents a specific major and includes detailed information about that major. There are 11 columns with 2 strings and the rest are numerical data types.

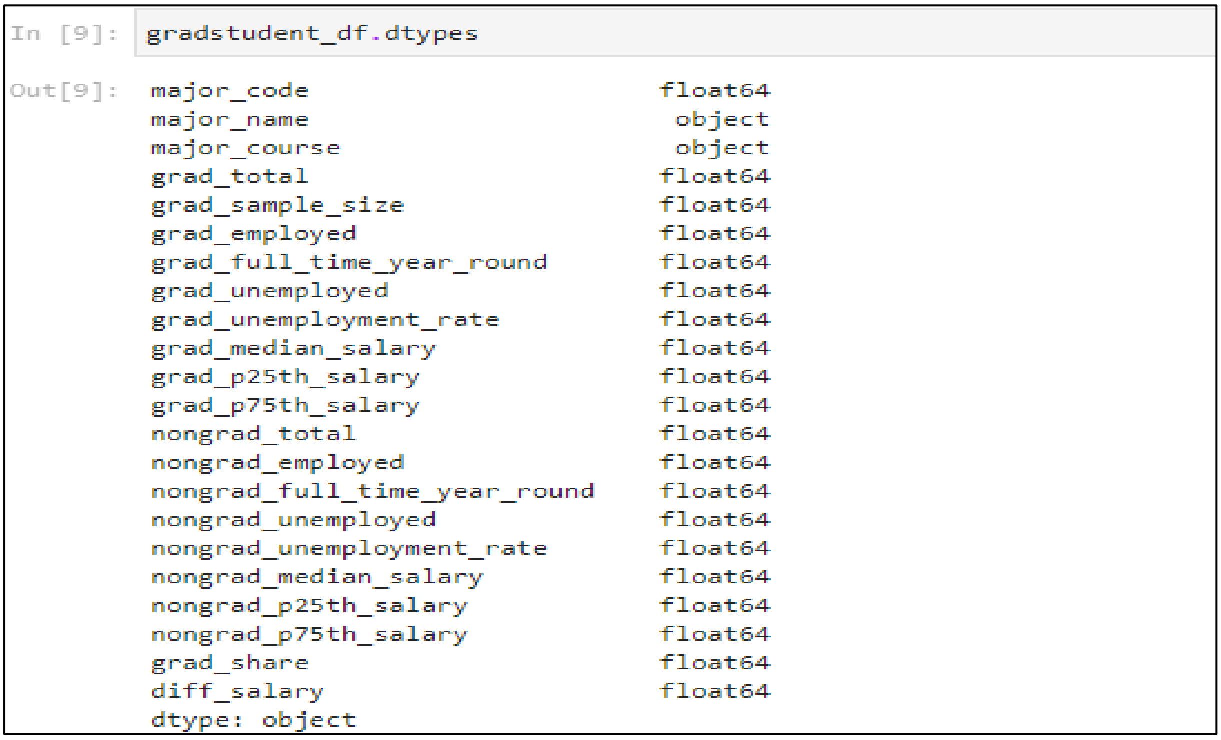


```
In [8]: recentlygrads_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 173 entries, 0 to 172
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   popularity_rank       173 non-null    float64
 1   major_code            173 non-null    float64
 2   major_name            173 non-null    object
 3   major_course          173 non-null    object
 4   total_students        173 non-null    float64
 5   sample_size           173 non-null    float64
 6   men                   173 non-null    float64
 7   women                 173 non-null    float64
 8   sharewomen            173 non-null    float64
 9   employed_grad         173 non-null    float64
 10  full_time             173 non-null    float64
 11  part_time             173 non-null    float64
 12  full_time_year_round  173 non-null    float64
 13  unemployed_grad       173 non-null    float64
 14  unemployment_rate     173 non-null    float64
 15  median_salary         173 non-null    float64
 16  p25th_salary          173 non-null    float64
 17  p75th_salary          173 non-null    float64
 18  college_jobs          173 non-null    float64
 19  non_college_jobs      173 non-null    float64
 20  low_wage_jobs         173 non-null    float64
dtypes: float64(19), object(2)
memory usage: 28.5+ KB
```

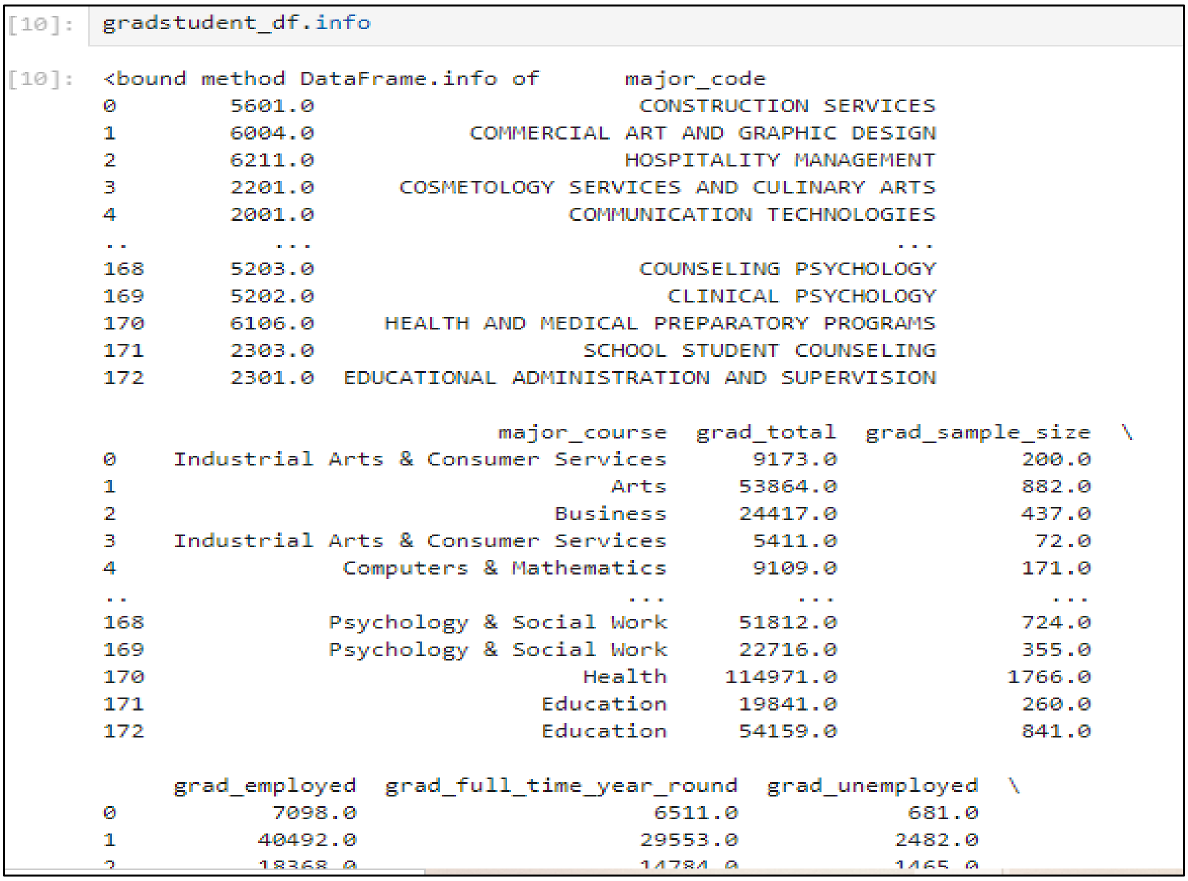
```
In [6]: recentlygrads_df.dtypes

Out[6]: popularity_rank         float64
        major_code              float64
        major_name               object
        major_course             object
        total_students          float64
        sample_size             float64
        men                     float64
        women                   float64
        sharewomen              float64
        employed_grad           float64
        full_time               float64
        part_time               float64
        full_time_year_round    float64
        unemployed_grad         float64
        unemployment_rate       float64
        median_salary           float64
        p25th_salary            float64
        p75th_salary            float64
        college_jobs            float64
        non_college_jobs        float64
        low_wage_jobs           float64
        dtype: object
```

Figure 1.4.3 recentlygrads Tables

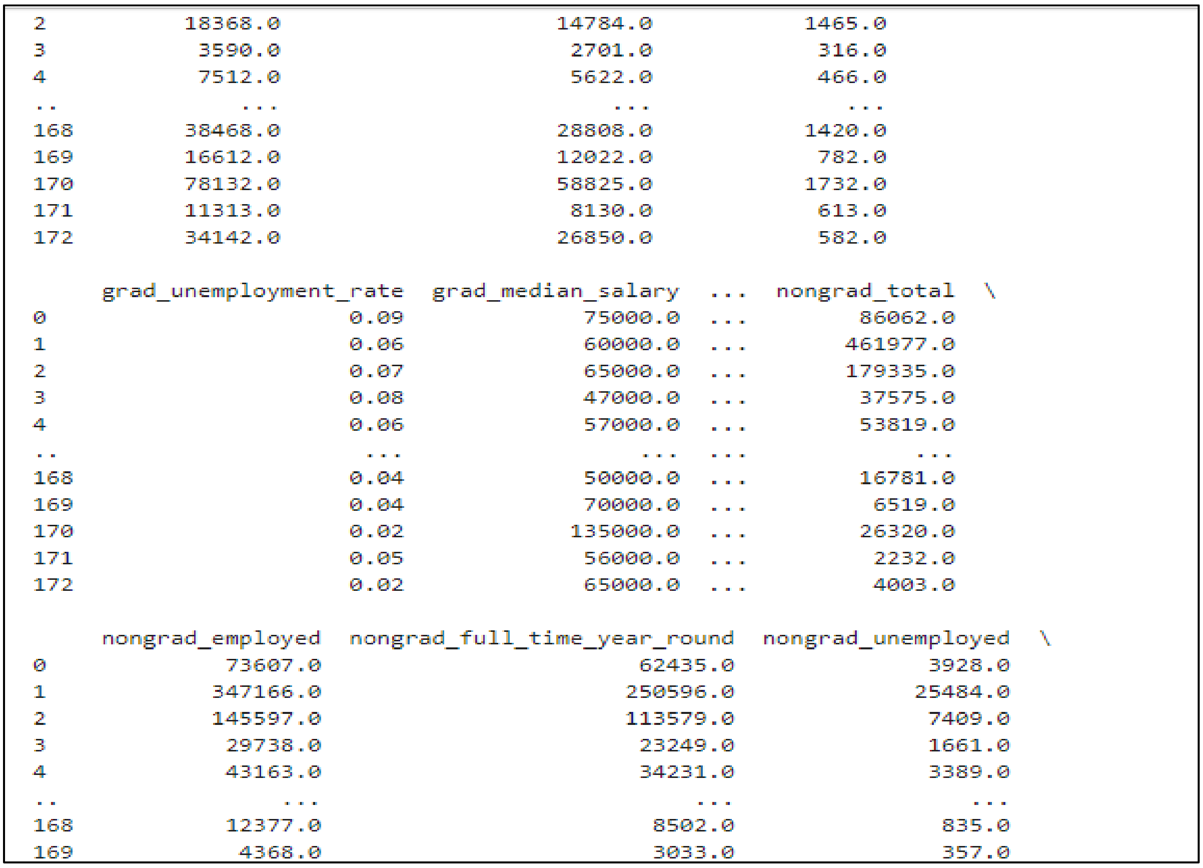
Figure 1.4.3 shows data schemas about table 2. The recentlygrads table is about the demographics and job outcomes of recent graduates from different majors. This dataset consists of 21 columns, where 2 columns are strings, and the other 19 columns are of numerical data types.



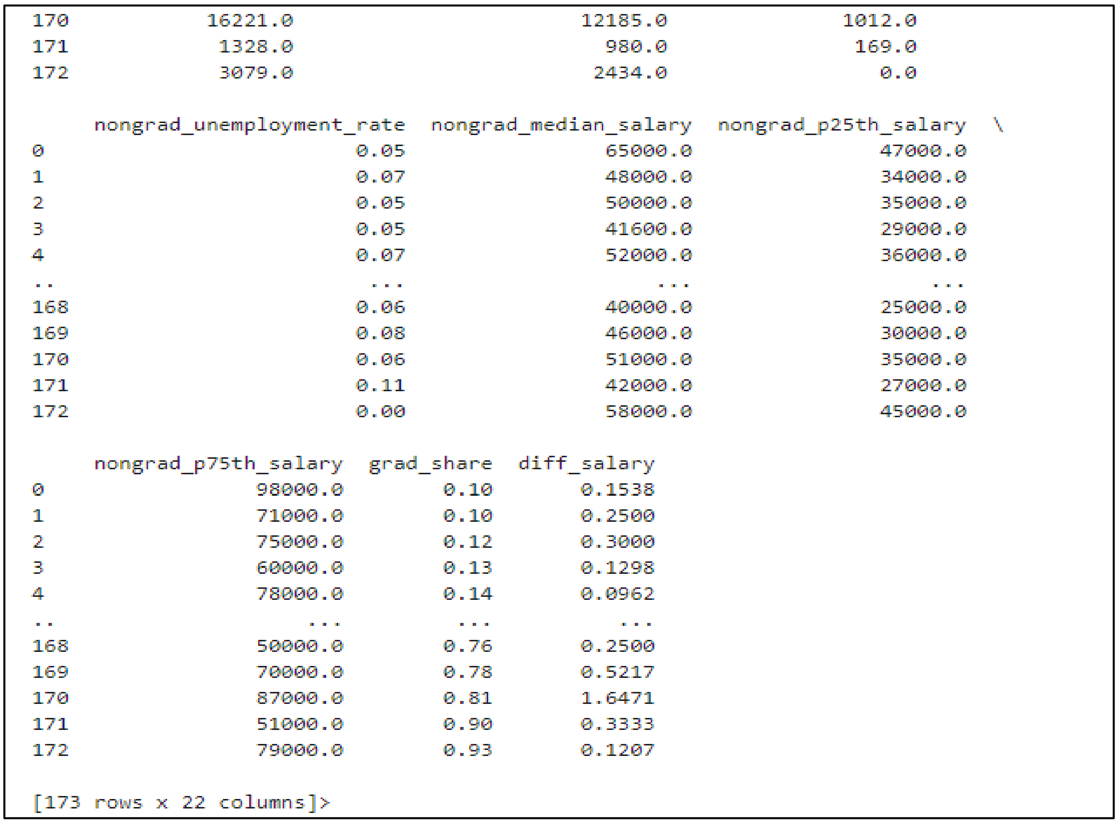
```
In [9]:  gradstudent_df.dtypes

Out[9]:  major_code                      float64
         major_name                       object
         major_course                     object
         grad_total                      float64
         grad_sample_size                float64
         grad_employed                   float64
         grad_full_time_year_round       float64
         grad_unemployed                 float64
         grad_unemployment_rate          float64
         grad_median_salary              float64
         grad_p25th_salary               float64
         grad_p75th_salary               float64
         nongrad_total                   float64
         nongrad_employed                float64
         nongrad_full_time_year_round    float64
         nongrad_unemployed              float64
         nongrad_unemployment_rate       float64
         nongrad_median_salary           float64
         nongrad_p25th_salary            float64
         nongrad_p75th_salary            float64
         grad_share                      float64
         diff_salary                     float64
         dtype: object
```



```
[10]: gradstudent_df.info

[10]: <bound method DataFrame.info of      major_code
      0         5601.0                     CONSTRUCTION SERVICES
      1         6004.0           COMMERCIAL ART AND GRAPHIC DESIGN
      2         6211.0                    HOSPITALITY MANAGEMENT
      3         2201.0       COSMETOLOGY SERVICES AND CULINARY ARTS
      4         2001.0                  COMMUNICATION TECHNOLOGIES
      ..           ...                                        ...
      168       5203.0                     COUNSELING PSYCHOLOGY
      169       5202.0                       CLINICAL PSYCHOLOGY
      170       6106.0      HEALTH AND MEDICAL PREPARATORY PROGRAMS
      171       2303.0                  SCHOOL STUDENT COUNSELING
      172       2301.0   EDUCATIONAL ADMINISTRATION AND SUPERVISION

                              major_course  grad_total  grad_sample_size  \
      0    Industrial Arts & Consumer Services     9173.0             200.0
      1                                   Arts    53864.0             882.0
      2                               Business    24417.0             437.0
      3    Industrial Arts & Consumer Services     5411.0              72.0
      4               Computers & Mathematics     9109.0             171.0
      ..                                   ...        ...               ...
      168              Psychology & Social Work   51812.0             724.0
      169              Psychology & Social Work   22716.0             355.0
      170                                Health  114971.0            1766.0
      171                             Education   19841.0             260.0
      172                             Education   54159.0             841.0

           grad_employed  grad_full_time_year_round  grad_unemployed  \
      0           7098.0                     6511.0            681.0
      1          40492.0                    29553.0           2482.0
```



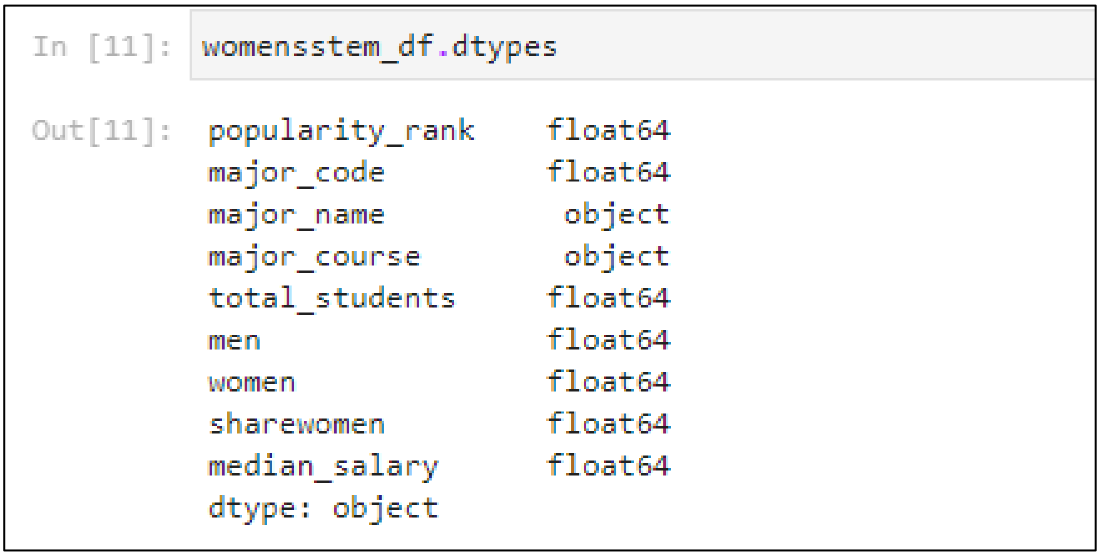
```
2         18368.0                     14784.0           1465.0
3          3590.0                      2701.0            316.0
4          7512.0                      5622.0            466.0
..            ...                         ...              ...
168       38468.0                     28808.0           1420.0
169       16612.0                     12022.0            782.0
170       78132.0                     58825.0           1732.0
171       11313.0                      8130.0            613.0
172       34142.0                     26850.0            582.0

     grad_unemployment_rate  grad_median_salary  ...  nongrad_total  \
0                      0.09             75000.0  ...        86062.0
1                      0.06             60000.0  ...       461977.0
2                      0.07             65000.0  ...       179335.0
3                      0.08             47000.0  ...        37575.0
4                      0.06             57000.0  ...        53819.0
..                      ...                 ...  ...            ...
168                    0.04             50000.0  ...        16781.0
169                    0.04             70000.0  ...         6519.0
170                    0.02            135000.0  ...        26320.0
171                    0.05             56000.0  ...         2232.0
172                    0.02             65000.0  ...         4003.0

     nongrad_employed  nongrad_full_time_year_round  nongrad_unemployed  \
0             73607.0                       62435.0              3928.0
1            347166.0                      250596.0             25484.0
2            145597.0                      113579.0              7409.0
3             29738.0                       23249.0              1661.0
4             43163.0                       34231.0              3389.0
..                ...                           ...                 ...
168           12377.0                        8502.0               835.0
169            4368.0                        3033.0               357.0
```



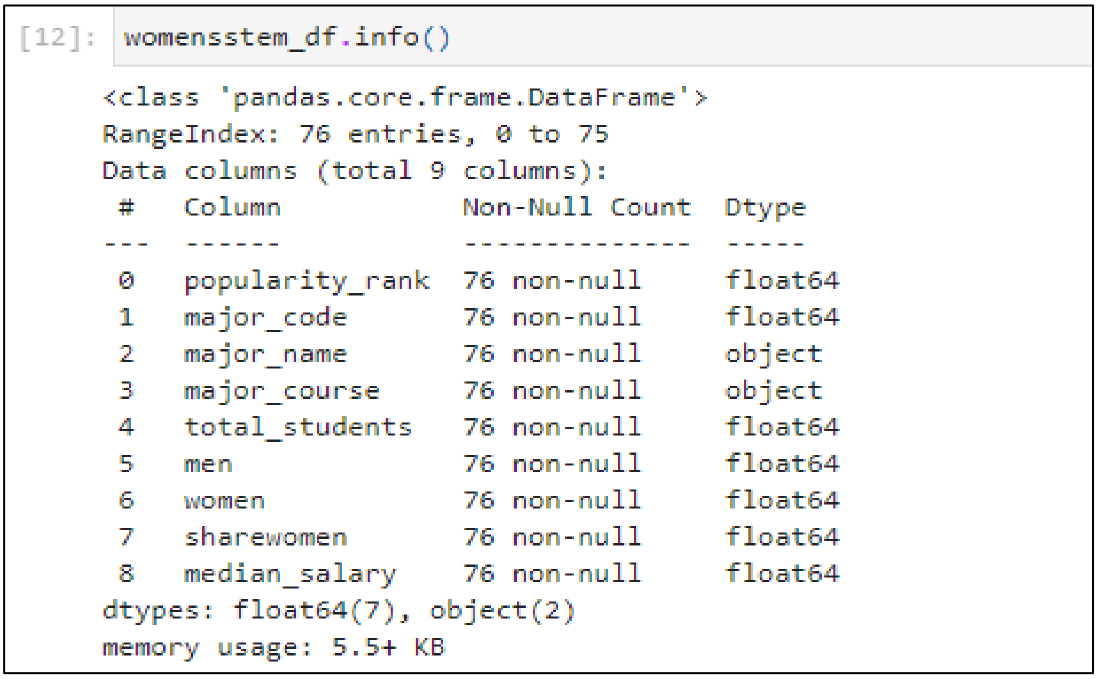
```
170          16221.0                          12185.0                1012.0
171           1328.0                            980.0                 169.0
172           3079.0                           2434.0                   0.0

     nongrad_unemployment_rate  nongrad_median_salary  nongrad_p25th_salary  \
0                         0.05                65000.0               47000.0
1                         0.07                48000.0               34000.0
2                         0.05                50000.0               35000.0
3                         0.05                41600.0               29000.0
4                         0.07                52000.0               36000.0
..                         ...                    ...                   ...
168                       0.06                40000.0               25000.0
169                       0.08                46000.0               30000.0
170                       0.06                51000.0               35000.0
171                       0.11                42000.0               27000.0
172                       0.00                58000.0               45000.0

     nongrad_p75th_salary  grad_share  diff_salary
0                 98000.0        0.10       0.1538
1                 71000.0        0.10       0.2500
2                 75000.0        0.12       0.3000
3                 60000.0        0.13       0.1298
4                 78000.0        0.14       0.0962
..                    ...         ...          ...
168               50000.0        0.76       0.2500
169               70000.0        0.78       0.5217
170               87000.0        0.81       1.6471
171               51000.0        0.90       0.3333
172               79000.0        0.93       0.1207

[173 rows x 22 columns]>
```

Figure 1.4.4 gradstudent Tables

Gradstudent data schemas are shown in Figure 1.4.4. Table 3 in our dataset consists of 2 columns with string values and 20 columns with numerical data types. In total, this table has 22 columns. Gradstudents table contains detailed data about graduates and non-graduates from various academic majors. It focuses on their employment status and salary outcomes.



```
In [11]:  womensstem_df.dtypes

Out[11]:  popularity_rank     float64
          major_code          float64
          major_name           object
          major_course         object
          total_students      float64
          men                 float64
          women               float64
          sharewomen          float64
          median_salary       float64
          dtype: object
```



```
[12]: womensstem_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 76 entries, 0 to 75
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   popularity_rank  76 non-null     float64
 1   major_code       76 non-null     float64
 2   major_name       76 non-null     object 
 3   major_course     76 non-null     object 
 4   total_students   76 non-null     float64
 5   men              76 non-null     float64
 6   women            76 non-null     float64
 7   sharewomen       76 non-null     float64
 8   median_salary    76 non-null     float64
dtypes: float64(7), object(2)
memory usage: 5.5+ KB
```

Figure 1.4.5 womensstem Table

Figure 1.4.5 shows the data schemas of the last table in the database. Womensstem table provides information about the demographics and median salaries of graduates from various STEM majors, with a focus on the number of male and female students and the percentage of female students. There are 9 columns in total with 2 strings and 7 numerical data types.

2.0 Architecture

2.1 Pipeline Structure

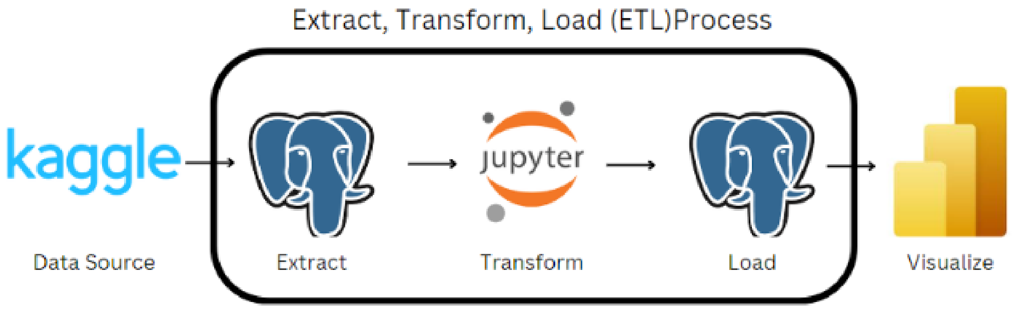


Figure 2.1.1

We are using Kimball's Approach to design our project with the title "Unveiling Gender Disparities: A Data-Driven Exploration of College Majors and STEM Fields". This approach offers us the ability to construct customized data marts, guaranteeing valuable insights can be achieved in mean time. By creating different data marts for variables such as enrolment figures, graduation rates, gender ratios, and career outcomes, we can effectively find specific components and identify any existing disparities. Kimball's approach is advantage for small teams, it only requires minimal adjustments, and significantly improving query performance for better analysis and reporting. The used of Kimball's technique useful to us to design targeted data marts, achieve faster results, and conduct a thorough investigation into gender disparities within educational field.

As shown in Figure 2.1.1, our dataset was obtained from Kaggle. It consists of four tables, namely, allage, womensstem, gradstudent and recentlygrads. To start our work, we build a database and tables in PostgreSQL based on our tables. Then, we imported the data and use Jupyter Notebook for further data operations.

In Jupyter Notebook, our data transformation process started with the installation of the required libraries, which simplified those data cleaning, loading, and saving process. We loaded multiple datasets, confirmed the data types and checked for any missing values by deleting any incomplete and unnecessary data. The tables were then merged to create an overview analysis, and the cleaned datasets were saved as new CSV files. These files were next re-imported into PostgreSQL in a new database.

After completing the cleaning and transformation processes, we executed the OLAP operations and PowerBI for multidimensional analysis, visualization and answering some of our objectives. We used OLAP operations in PostgreSQL. The techniques we used are such as slicing, and pivot as we want to a gained deeper insight into the data.

Finally, we used Power BI. Power BI provided a strong platform for creating visualizations that help us to observe carefully the expected results. Its interactive and customizable functionalities allowed the creation of informative visual representations of the data, easy to understand and faster decision-making based on the analyzed findings.

2.2 ETL Pipeline

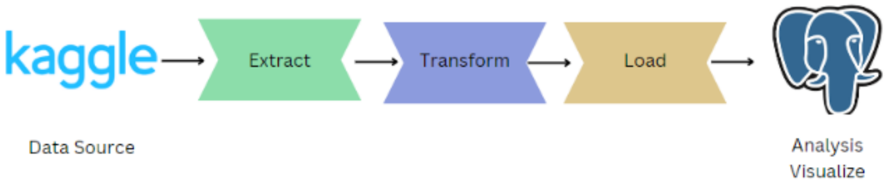


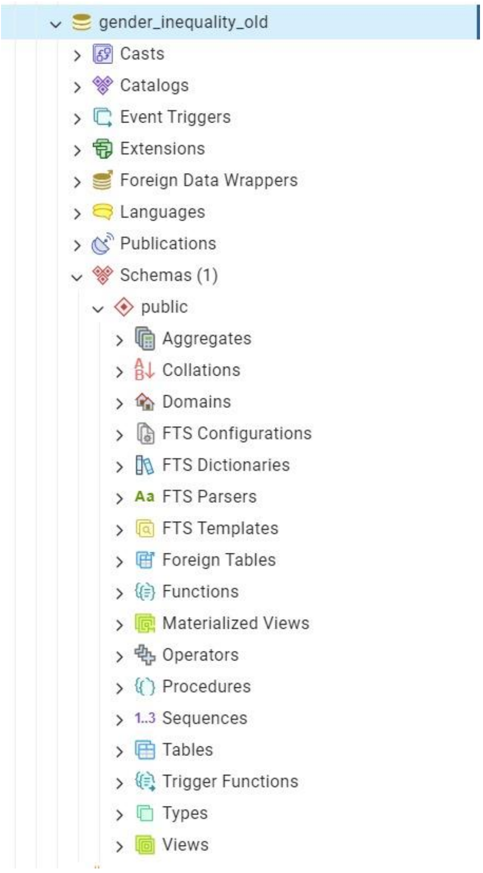
Figure 2.2.1

Figure 2.2.1 displays the ETL pipeline for the dataset. The process involves extracting data from a source, transforming it, and loading as Data Warehouse System. For this project, in details, we use PostgreSQL to extract data from a CSV file, transformed it using Python in Jupyter Notebook connected to PostgreSQL, loaded the clean data back into PostgreSQL, and finally visualized the data using OLAP and Power BI. In total we have 4 tables in a database, so the ETL process is repeated to those 4 tables and the data is ready to be visualized and analysed.

2.3 ETL Process

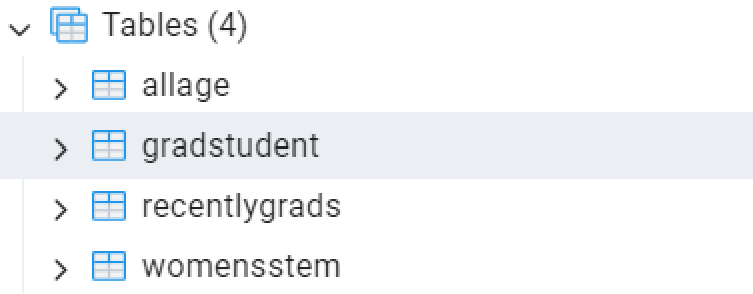
Extract:

To begin the ETL process, it is necessary to store the datasets in a PostgreSQL database. Firstly, we create a new database and do the query to create those 4 tables and import them from 4 different csv.



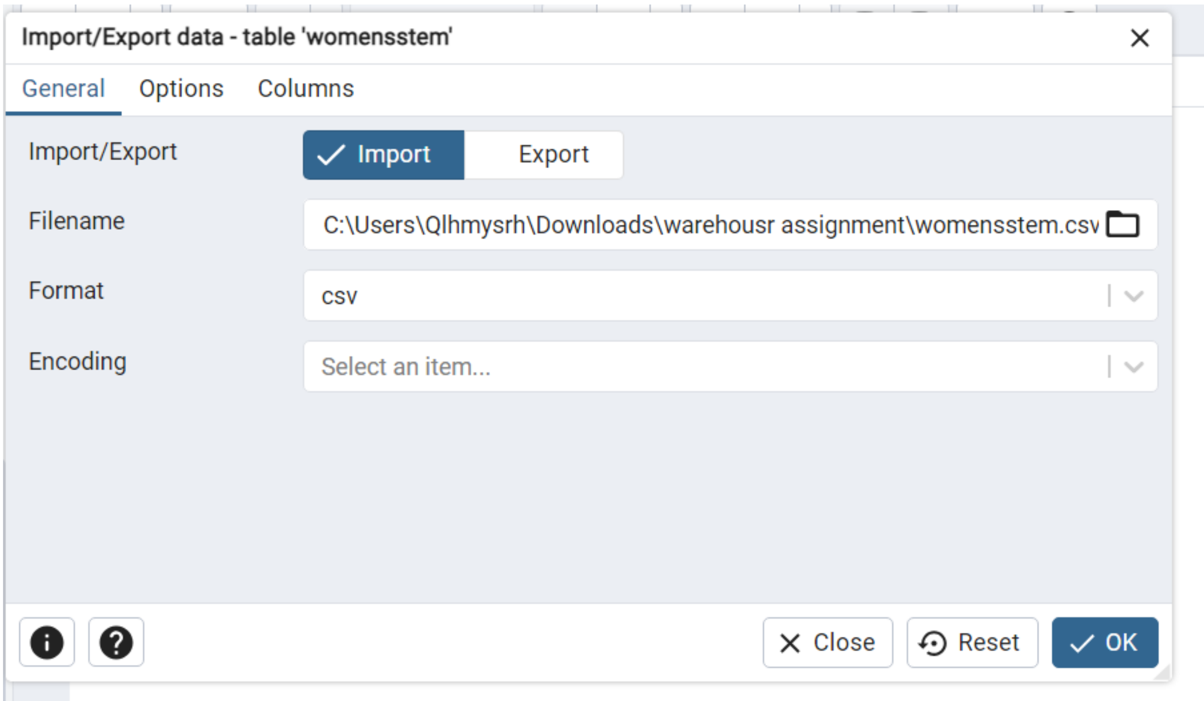
gender_inequality_old
Casts
Catalogs
Event Triggers
Extensions
Foreign Data Wrappers
Languages
Publications
Schemas (1)
public
Aggregates
Collations
Domains
FTS Configurations
FTS Dictionaries
FTS Parsers
FTS Templates
Foreign Tables
Functions
Materialized Views
Operators
Procedures
Sequences
Tables
Trigger Functions
Types
Views

In PostgreSQL, we have successfully created a database called 'gender_inequality_old'



Tables (4)
allage
gradstudent
recentlygrads
womensstem

Tables created.



Import/Export data - table 'womensstem'

Import/Export: ✓ Import | Export

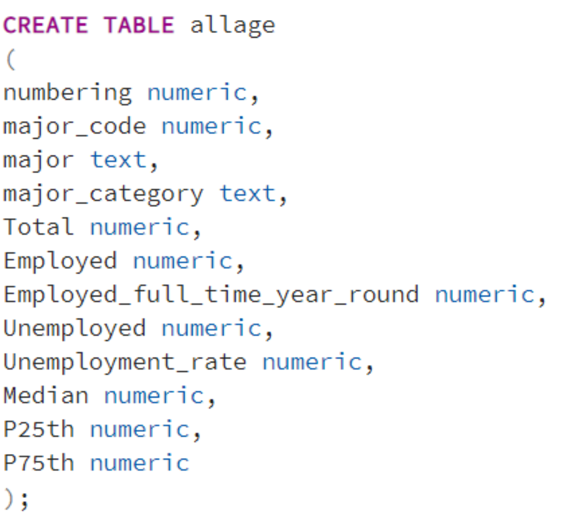
Filename: C:\Users\Qlhmysrh\Downloads\warehousr assignment\womensstem.csv

Format: csv

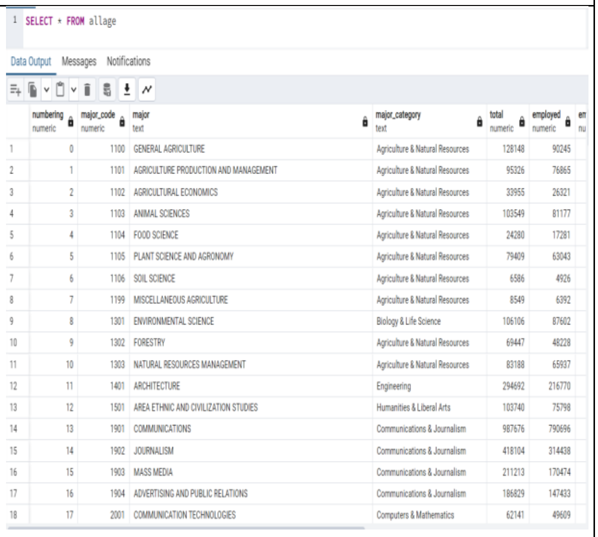
Encoding: Select an item...

Import csv file into table, repeat to the other 3 csv file

Create Table	Output

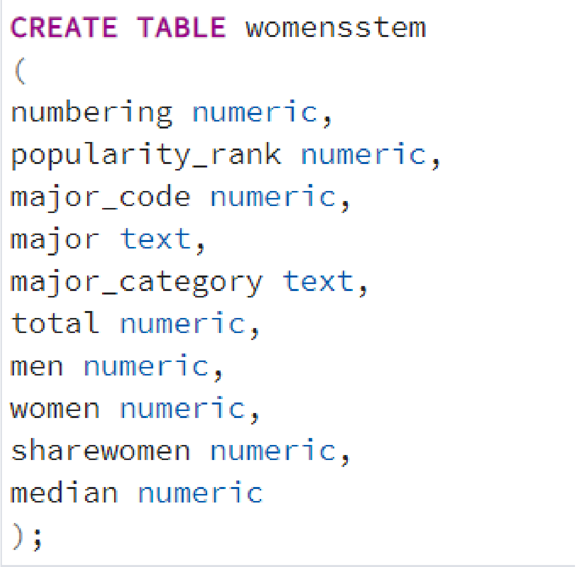


```
CREATE TABLE allage
(
numbering numeric,
major_code numeric,
major text,
major_category text,
Total numeric,
Employed numeric,
Employed_full_time_year_round numeric,
Unemployed numeric,
Unemployment_rate numeric,
Median numeric,
P25th numeric,
P75th numeric
);
```

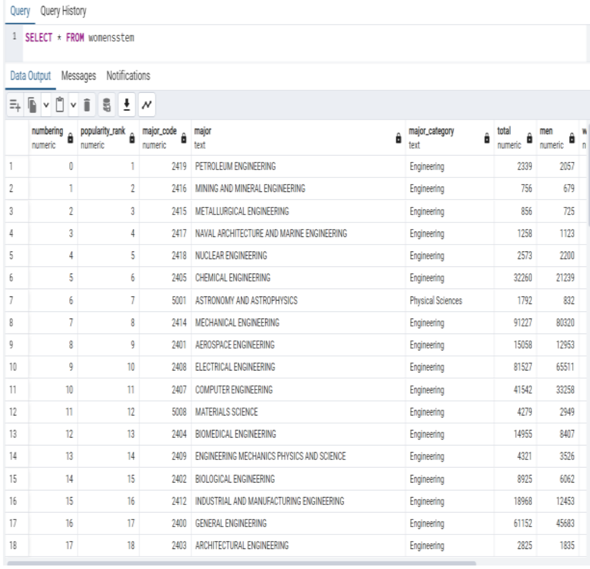


```
SELECT * FROM allage
```

	numbering numeric	major_code numeric	major text	major_category text	total numeric	employed numeric	em nu
1	0	1100	GENERAL AGRICULTURE	Agriculture & Natural Resources	128148	90245	
2	1	1101	AGRICULTURE PRODUCTION AND MANAGEMENT	Agriculture & Natural Resources	95326	76865	
3	2	1102	AGRICULTURAL ECONOMICS	Agriculture & Natural Resources	33955	26321	
4	3	1103	ANIMAL SCIENCES	Agriculture & Natural Resources	103549	81177	
5	4	1104	FOOD SCIENCE	Agriculture & Natural Resources	24280	17281	
6	5	1105	PLANT SCIENCE AND AGRONOMY	Agriculture & Natural Resources	79409	63043	
7	6	1106	SOIL SCIENCE	Agriculture & Natural Resources	6586	4926	
8	7	1199	MISCELLANEOUS AGRICULTURE	Agriculture & Natural Resources	8549	6392	
9	8	1301	ENVIRONMENTAL SCIENCE	Biology & Life Science	106106	87602	
10	9	1302	FORESTRY	Agriculture & Natural Resources	69447	48228	
11	10	1303	NATURAL RESOURCES MANAGEMENT	Agriculture & Natural Resources	83188	65937	
12	11	1401	ARCHITECTURE	Engineering	294692	216770	
13	12	1501	AREA ETHNIC AND CIVILIZATION STUDIES	Humanities & Liberal Arts	103740	75798	
14	13	1901	COMMUNICATIONS	Communications & Journalism	987676	790696	
15	14	1902	JOURNALISM	Communications & Journalism	418104	314438	
16	15	1903	MASS MEDIA	Communications & Journalism	211213	170474	
17	16	1904	ADVERTISING AND PUBLIC RELATIONS	Communications & Journalism	186829	147433	
18	17	2001	COMMUNICATION TECHNOLOGIES	Computers & Mathematics	62141	49609	



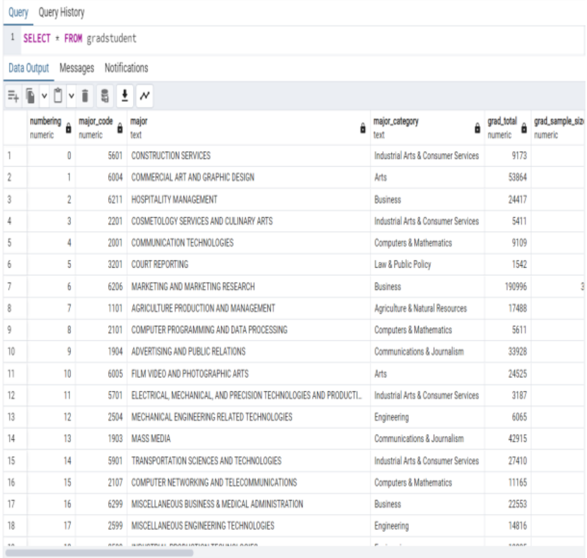
```
CREATE TABLE womensstem
(
numbering numeric,
popularity_rank numeric,
major_code numeric,
major text,
major_category text,
total numeric,
men numeric,
women numeric,
sharewomen numeric,
median numeric
);
```



```
SELECT * FROM womensstem
```

	numbering numeric	popularity_rank numeric	major_code numeric	major text	major_category text	total numeric	men numeric	w n
1	0	1	2419	PETROLEUM ENGINEERING	Engineering	2339	2057	
2	1	2	2416	MINING AND MINERAL ENGINEERING	Engineering	756	679	
3	2	3	2415	METALLURGICAL ENGINEERING	Engineering	856	725	
4	3	4	2417	NAVAL ARCHITECTURE AND MARINE ENGINEERING	Engineering	1258	1123	
5	4	5	2418	NUCLEAR ENGINEERING	Engineering	2573	2200	
6	5	6	2405	CHEMICAL ENGINEERING	Engineering	32260	21239	
7	6	7	5001	ASTRONOMY AND ASTROPHYSICS	Physical Sciences	1792	832	
8	7	8	2414	MECHANICAL ENGINEERING	Engineering	91227	80320	
9	8	9	2401	AEROSPACE ENGINEERING	Engineering	15058	12953	
10	9	10	2408	ELECTRICAL ENGINEERING	Engineering	81527	65511	
11	10	11	2407	COMPUTER ENGINEERING	Engineering	41542	33258	
12	11	12	5008	MATERIALS SCIENCE	Engineering	4279	2949	
13	12	13	2404	BIOMEDICAL ENGINEERING	Engineering	14955	8407	
14	13	14	2409	ENGINEERING MECHANICS PHYSICS AND SCIENCE	Engineering	4321	3526	
15	14	15	2402	BIOLOGICAL ENGINEERING	Engineering	8925	6062	
16	15	16	2412	INDUSTRIAL AND MANUFACTURING ENGINEERING	Engineering	18968	12453	
17	16	17	2400	GENERAL ENGINEERING	Engineering	61152	45683	
18	17	18	2403	ARCHITECTURAL ENGINEERING	Engineering	2825	1835	

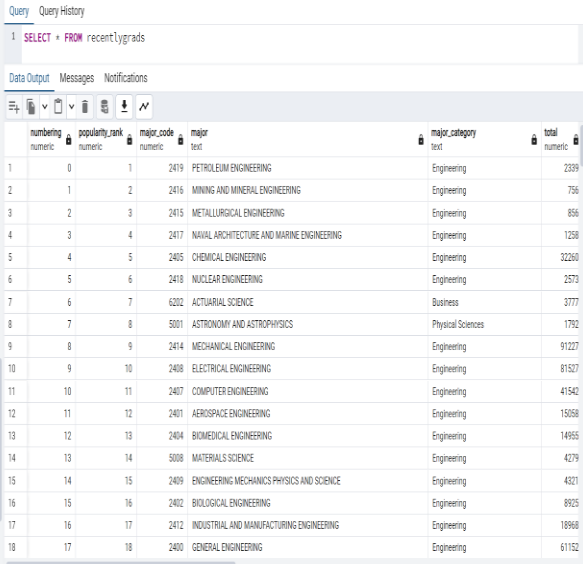
```
CREATE TABLE gradstudent
(
numbering numeric,
major_code numeric,
major text,
major_category text,
Grad_total numeric,
Grad_sample_size numeric,
Grad_employed numeric,
Grad_full_time_year_round numeric,
Grad_unemployed numeric,
Grad_unemployment_rate numeric,
Grad_median numeric,
Grad_P25 numeric,
Grad_P75 numeric,
Nongrad_total numeric,
Nongrad_employed numeric,
Nongrad_full_time_year_round numeric,
Nongrad_unemployed numeric,
Nongrad_unemployment_rate numeric,
Nongrad_median numeric,
Nongrad_P25 numeric,
Nongrad_P75 numeric,
Grad_share numeric,
Grad_premium numeric
);
```



```
1 SELECT * FROM gradstudent
```

	numbering numeric	major_code numeric	major text	major_category text	grad_total numeric	grad_sample_siz numeric
1	0	5601	CONSTRUCTION SERVICES	Industrial Arts & Consumer Services	9173	
2	1	6004	COMMERCIAL ART AND GRAPHIC DESIGN	Arts	53864	
3	2	6211	HOSPITALITY MANAGEMENT	Business	24417	
4	3	2201	COSMETOLOGY SERVICES AND CULINARY ARTS	Industrial Arts & Consumer Services	5411	
5	4	2001	COMMUNICATION TECHNOLOGIES	Computers & Mathematics	9109	
6	5	3201	COURT REPORTING	Law & Public Policy	1542	
7	6	6206	MARKETING AND MARKETING RESEARCH	Business	190996	3
8	7	1101	AGRICULTURE PRODUCTION AND MANAGEMENT	Agriculture & Natural Resources	17488	
9	8	2101	COMPUTER PROGRAMMING AND DATA PROCESSING	Computers & Mathematics	5611	
10	9	1904	ADVERTISING AND PUBLIC RELATIONS	Communications & Journalism	33928	
11	10	6005	FILM VIDEO AND PHOTOGRAPHIC ARTS	Arts	24525	
12	11	5701	ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTI...	Industrial Arts & Consumer Services	3187	
13	12	2504	MECHANICAL ENGINEERING RELATED TECHNOLOGIES	Engineering	6065	
14	13	1903	MASS MEDIA	Communications & Journalism	42915	
15	14	5901	TRANSPORTATION SCIENCES AND TECHNOLOGIES	Industrial Arts & Consumer Services	27410	
16	15	2107	COMPUTER NETWORKING AND TELECOMMUNICATIONS	Computers & Mathematics	11165	
17	16	6299	MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	Business	22553	
18	17	2599	MISCELLANEOUS ENGINEERING TECHNOLOGIES	Engineering	14816	

```
CREATE TABLE recentlygrads
(
numbering numeric,
popularity_rank numeric,
major_code numeric,
major text,
major_category text,
Total numeric,
Sample_size numeric,
Men numeric,
Women numeric,
ShareWomen numeric,
employed numeric,
full_time numeric,
part_time numeric,
Full_time_year_round numeric,
unemployed numeric,
unemployment_rate numeric,
median numeric,
P25th numeric,
P75th numeric,
College_jobs numeric,
Non_college_jobs numeric,
Low_wage_jobs numeric
);
```



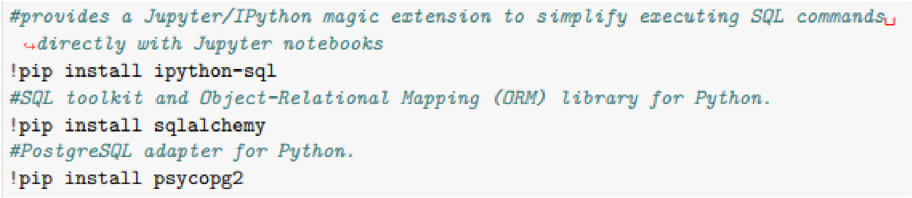
```
1 SELECT * FROM recentlygrads
```

	numbering numeric	popularity_rank numeric	major_code numeric	major text	major_category text	total numeric
1	0	1	2419	PETROLEUM ENGINEERING	Engineering	2339
2	1	2	2416	MINING AND MINERAL ENGINEERING	Engineering	756
3	2	3	2415	METALLURGICAL ENGINEERING	Engineering	856
4	3	4	2417	NAVAL ARCHITECTURE AND MARINE ENGINEERING	Engineering	1258
5	4	5	2405	CHEMICAL ENGINEERING	Engineering	32260
6	5	6	2418	NUCLEAR ENGINEERING	Engineering	2573
7	6	7	6202	ACTUARIAL SCIENCE	Business	3777
8	7	8	5001	ASTRONOMY AND ASTROPHYSICS	Physical Sciences	1792
9	8	9	2414	MECHANICAL ENGINEERING	Engineering	91227
10	9	10	2408	ELECTRICAL ENGINEERING	Engineering	81527
11	10	11	2407	COMPUTER ENGINEERING	Engineering	41542
12	11	12	2401	AEROSPACE ENGINEERING	Engineering	15058
13	12	13	2404	BIOMEDICAL ENGINEERING	Engineering	14955
14	13	14	5008	MATERIALS SCIENCE	Engineering	4279
15	14	15	2409	ENGINEERING MECHANICS PHYSICS AND SCIENCE	Engineering	4321
16	15	16	2402	BIOLOGICAL ENGINEERING	Engineering	8925
17	16	17	2412	INDUSTRIAL AND MANUFACTURING ENGINEERING	Engineering	18968
18	17	18	2400	GENERAL ENGINEERING	Engineering	61152

This table shows the query to create the tables and the outputs for each table.

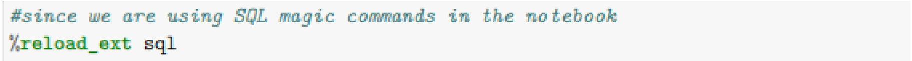
Transform:

Once the raw data has been transferred to postgreSQL, our next task is to create a connection between postgreSQL and Jupyter Notebook. This connection is important for us to move forward with the data transformation process.



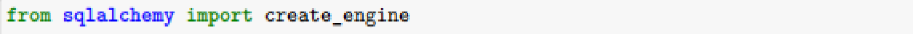
```
#provides a Jupyter/IPython magic extension to simplify executing SQL commands␣
 ↪directly with Jupyter notebooks
!pip install ipython-sql
#SQL toolkit and Object-Relational Mapping (ORM) library for Python.
!pip install sqlalchemy
#PostgreSQL adapter for Python.
!pip install psycopg2
```

This figure shows the packages that we installed.



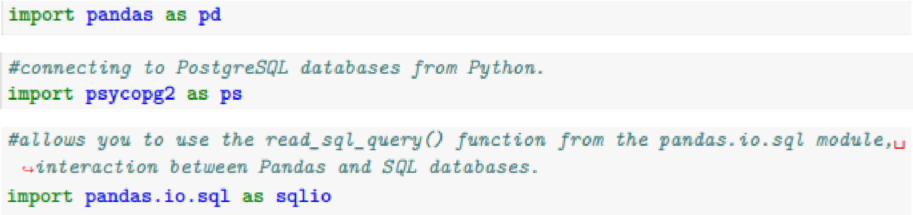
```
#since we are using SQL magic commands in the notebook
%reload_ext sql
```

This figure shows the load of ipython-sql.



```
from sqlalchemy import create_engine
```

This figure shows a call to create engine.

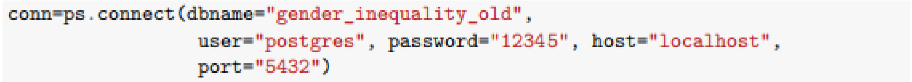


```
import pandas as pd
```

```
#connecting to PostgreSQL databases from Python.
import psycopg2 as ps
```

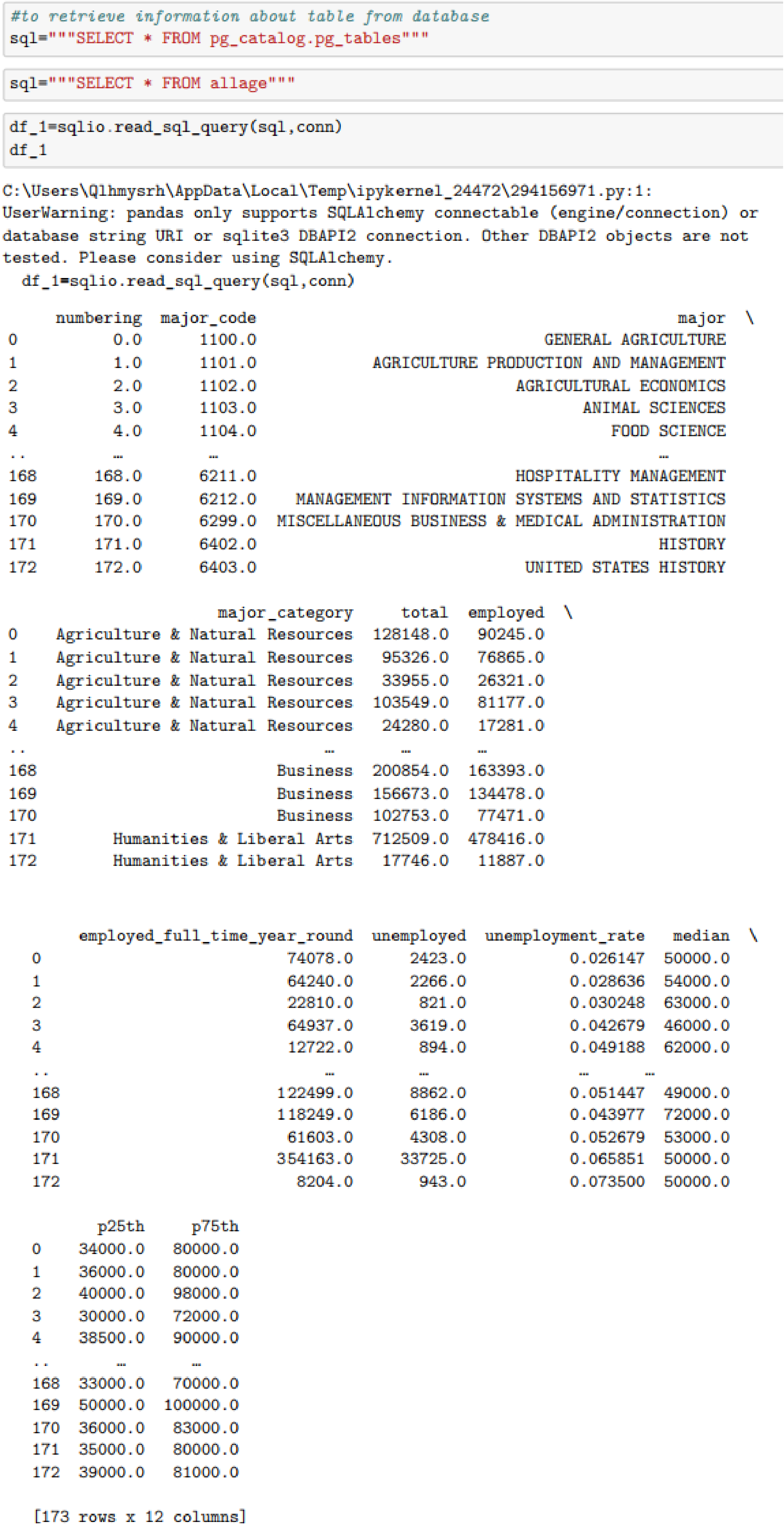
```
#allows you to use the read_sql_query() function from the pandas.io.sql module,␣
 ↪interaction between Pandas and SQL databases.
import pandas.io.sql as sqlio
```

Import necessary libraries for the ETL process



```
conn=ps.connect(dbname="gender_inequality_old",
                user="postgres", password="12345", host="localhost",
                port="5432")
```

Connect the PgAdmin with Jupyter Notebook



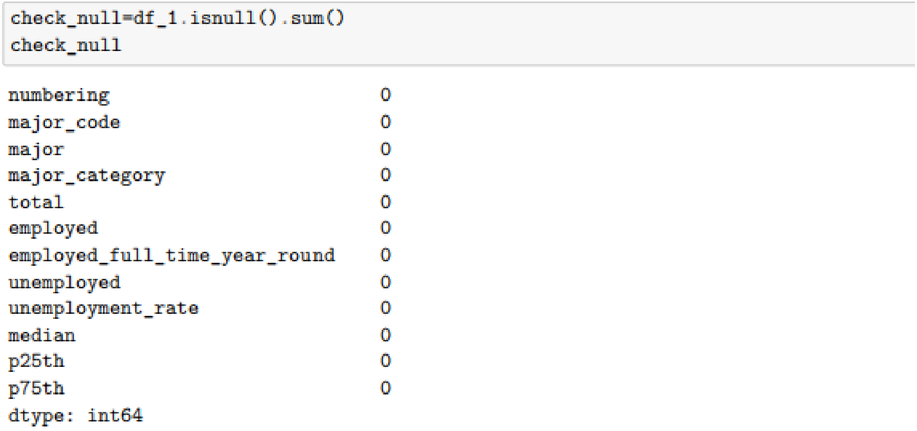
```
#to retrieve information about table from database
sql="""SELECT * FROM pg_catalog.pg_tables"""
```

```
sql="""SELECT * FROM allage"""
```

```
df_1=sqlio.read_sql_query(sql,conn)
df_1
```

```
C:\Users\Qlhmysrh\AppData\Local\Temp\ipykernel_24472\294156971.py:1:
UserWarning: pandas only supports SQLAlchemy connectable (engine/connection) or
database string URI or sqlite3 DBAPI2 connection. Other DBAPI2 objects are not
tested. Please consider using SQLAlchemy.
  df_1=sqlio.read_sql_query(sql,conn)

     numbering  major_code                                      major  \
0          0.0      1100.0                         GENERAL AGRICULTURE
1          1.0      1101.0         AGRICULTURE PRODUCTION AND MANAGEMENT
2          2.0      1102.0                      AGRICULTURAL ECONOMICS
3          3.0      1103.0                             ANIMAL SCIENCES
4          4.0      1104.0                                FOOD SCIENCE
..         ...         ...                                         ...
168      168.0      6211.0                      HOSPITALITY MANAGEMENT
169      169.0      6212.0     MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
170      170.0      6299.0  MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION
171      171.0      6402.0                                     HISTORY
172      172.0      6403.0                       UNITED STATES HISTORY

                 major_category     total  employed  \
0    Agriculture & Natural Resources  128148.0   90245.0
1    Agriculture & Natural Resources   95326.0   76865.0
2    Agriculture & Natural Resources   33955.0   26321.0
3    Agriculture & Natural Resources  103549.0   81177.0
4    Agriculture & Natural Resources   24280.0   17281.0
..                              ...       ...       ...
168                         Business  200854.0  163393.0
169                         Business  156673.0  134478.0
170                         Business  102753.0   77471.0
171          Humanities & Liberal Arts  712509.0  478416.0
172          Humanities & Liberal Arts   17746.0   11887.0


     employed_full_time_year_round  unemployed  unemployment_rate   median  \
0                          74078.0      2423.0           0.026147  50000.0
1                          64240.0      2266.0           0.028636  54000.0
2                          22810.0       821.0           0.030248  63000.0
3                          64937.0      3619.0           0.042679  46000.0
4                          12722.0       894.0           0.049188  62000.0
..                             ...         ...                ...      ...
168                       122499.0      8862.0           0.051447  49000.0
169                       118249.0      6186.0           0.043977  72000.0
170                        61603.0      4308.0           0.052679  53000.0
171                       354163.0     33725.0           0.065851  50000.0
172                         8204.0       943.0           0.073500  50000.0

        p25th     p75th
0     34000.0   80000.0
1     36000.0   80000.0
2     40000.0   98000.0
3     30000.0   72000.0
4     38500.0   90000.0
..        ...       ...
168   33000.0   70000.0
169   50000.0  100000.0
170   36000.0   83000.0
171   35000.0   80000.0
172   39000.0   81000.0

[173 rows x 12 columns]
```

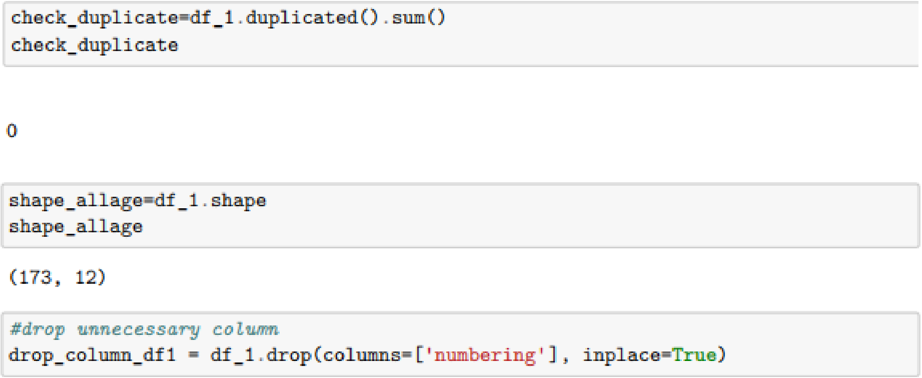
Extract data from PgAdmin to Jupyter Notebook



```
check_null=df_1.isnull().sum()
check_null
```

```
numbering                        0
major_code                       0
major                            0
major_category                   0
total                            0
employed                         0
employed_full_time_year_round    0
unemployed                       0
unemployment_rate                0
median                           0
p25th                            0
p75th                            0
dtype: int64
```

Checking null



```
check_duplicate=df_1.duplicated().sum()
check_duplicate
```

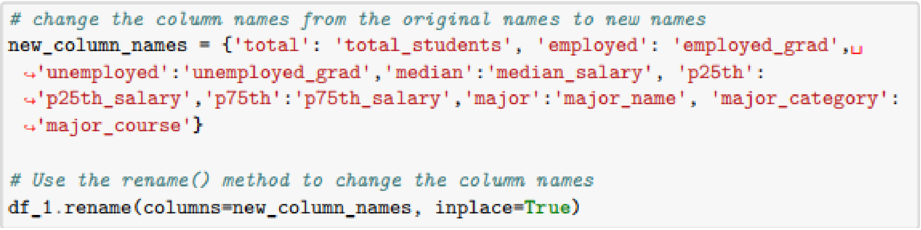
```
0
```

```
shape_allage=df_1.shape
shape_allage
```

```
(173, 12)
```

```
#drop unnecessary column
drop_column_df1 = df_1.drop(columns=['numbering'], inplace=True)
```

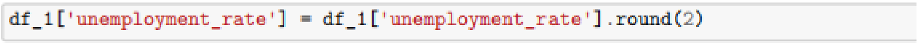
Drop and check null values.



```
# change the column names from the original names to new names
new_column_names = {'total': 'total_students', 'employed': 'employed_grad',␣
 ↪'unemployed':'unemployed_grad','median':'median_salary', 'p25th':
 ↪'p25th_salary','p75th':'p75th_salary','major':'major_name', 'major_category':
 ↪'major_course'}

# Use the rename() method to change the column names
df_1.rename(columns=new_column_names, inplace=True)
```

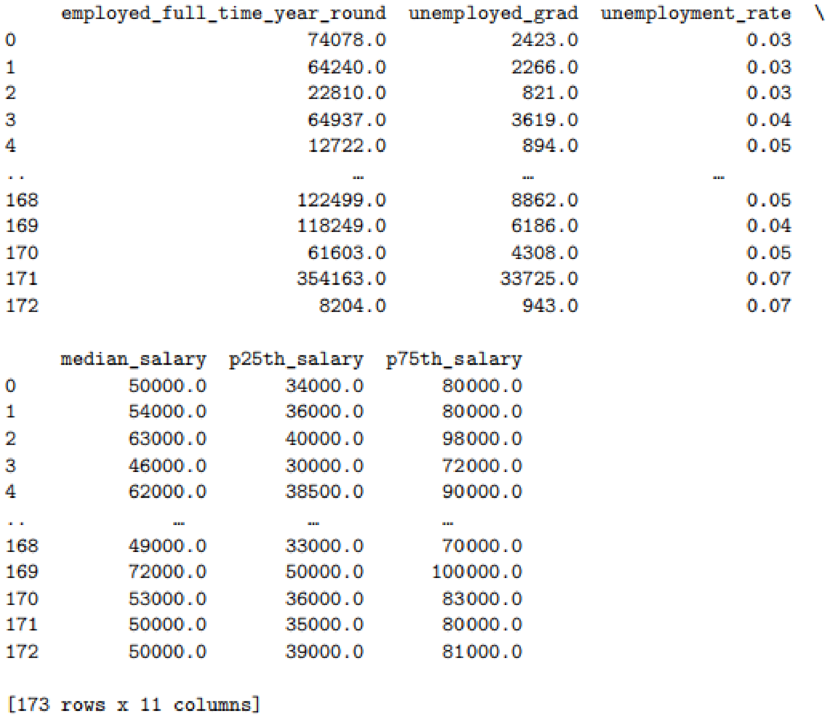
Rename the column.

Once we have checked the null values, it is important to proceed examine the primary key for any duplicates. It is important to make sure that the primary key in the dataset remains unique after the cleaning procedure, as this allows us to use the software to create connections between datasets and create a relational model.



```
df_1['unemployment_rate'] = df_1['unemployment_rate'].round(2)
```

Round of numerical data to easy analysize



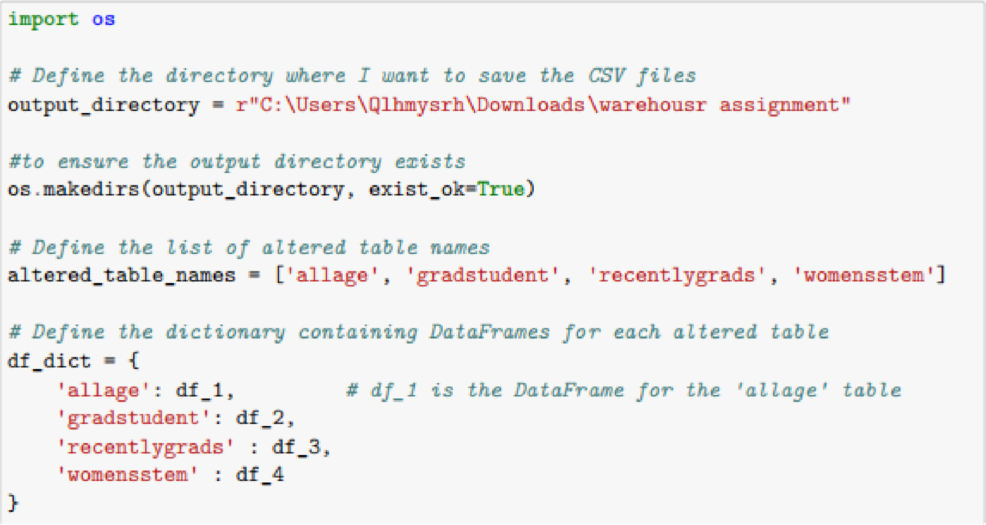
```
     employed_full_time_year_round  unemployed_grad  unemployment_rate  \
0                          74078.0           2423.0               0.03
1                          64240.0           2266.0               0.03
2                          22810.0            821.0               0.03
3                          64937.0           3619.0               0.04
4                          12722.0            894.0               0.05
..                             ...              ...                ...
168                       122499.0           8862.0               0.05
169                       118249.0           6186.0               0.04
170                        61603.0           4308.0               0.05
171                       354163.0          33725.0               0.07
172                         8204.0            943.0               0.07

     median_salary  p25th_salary  p75th_salary
0          50000.0       34000.0       80000.0
1          54000.0       36000.0       80000.0
2          63000.0       40000.0       98000.0
3          46000.0       30000.0       72000.0
4          62000.0       38500.0       90000.0
..             ...           ...           ...
168        49000.0       33000.0       70000.0
169        72000.0       50000.0      100000.0
170        53000.0       36000.0       83000.0
171        50000.0       35000.0       80000.0
172        50000.0       39000.0       81000.0

[173 rows x 11 columns]
```

The cleaned version data frame.

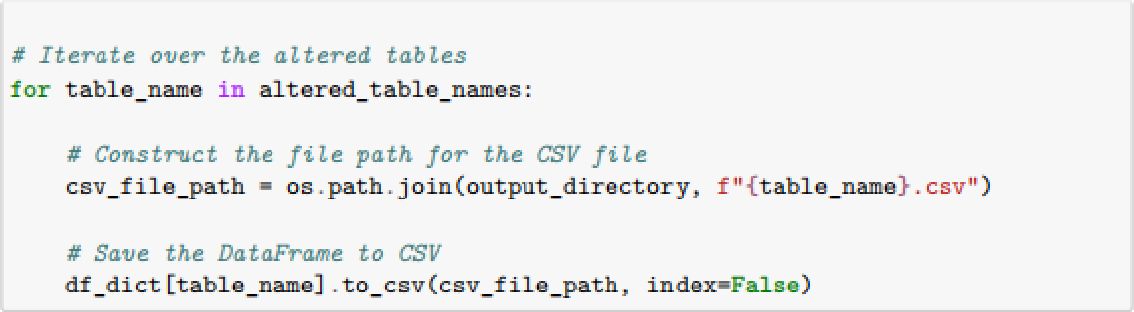
This process is repeated for all the other 3 data frames.

Load:

Once we have finished cleaning our data, the next step is to transfer it into PostgreSQL. We can achieve this by creating a database and table in postgreSQL. By using the following code, we can get cleaned dataset imported effortlessly to our desktop, and then it is our job to import the cleaned csv file in the database for each tables:



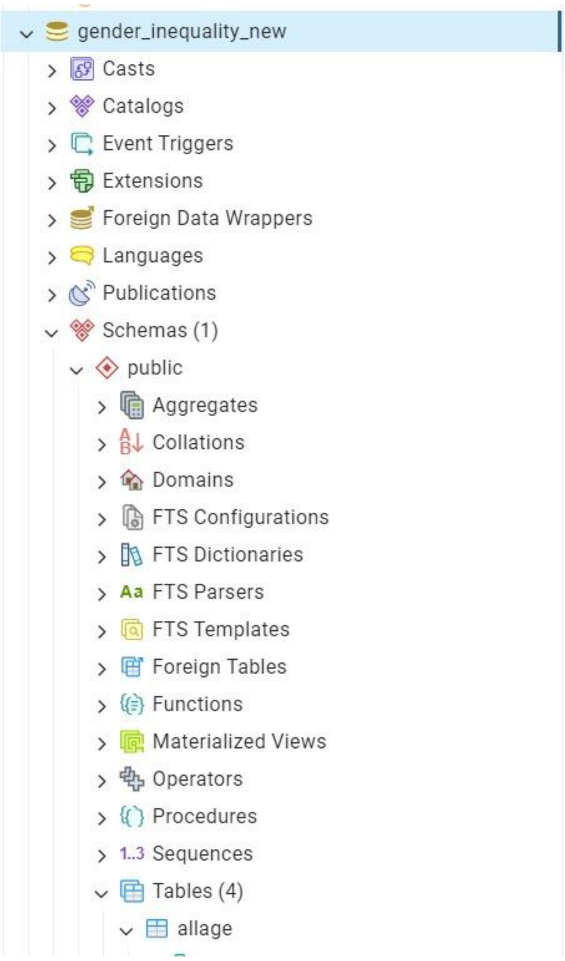
```
import os

# Define the directory where I want to save the CSV files
output_directory = r"C:\Users\Qlhmysrh\Downloads\warehousr assignment"

#to ensure the output directory exists
os.makedirs(output_directory, exist_ok=True)

# Define the list of altered table names
altered_table_names = ['allage', 'gradstudent', 'recentlygrads', 'womensstem']

# Define the dictionary containing DataFrames for each altered table
df_dict = {
    'allage': df_1,         # df_1 is the DataFrame for the 'allage' table
    'gradstudent': df_2,
    'recentlygrads' : df_3,
    'womensstem' : df_4
}
```



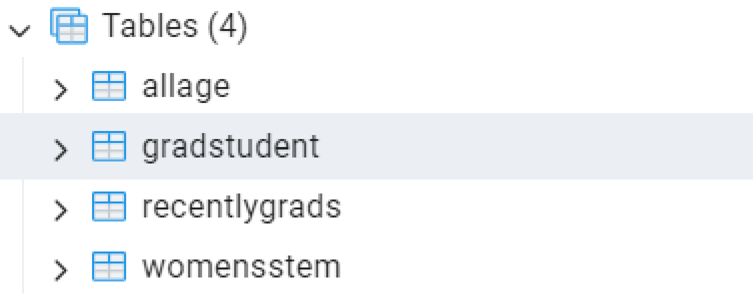
```
# Iterate over the altered tables
for table_name in altered_table_names:

    # Construct the file path for the CSV file
    csv_file_path = os.path.join(output_directory, f"{table_name}.csv")

    # Save the DataFrame to CSV
    df_dict[table_name].to_csv(csv_file_path, index=False)
```

Data loaded into desktop.

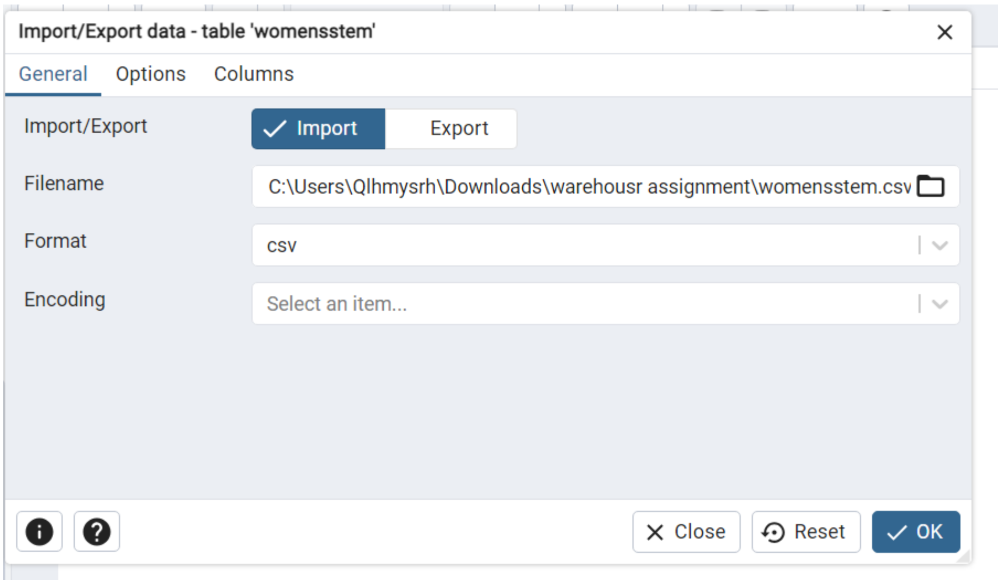
This is what we are doing for our new and cleaned csv files:



Create a new database



Create a new table



Import/Export data - table 'womensstem'

General | Options | Columns

Import/Export: [✓ Import] Export

Filename: C:\Users\Qlhmysrh\Downloads\warehousr assignment\womensstem.csv

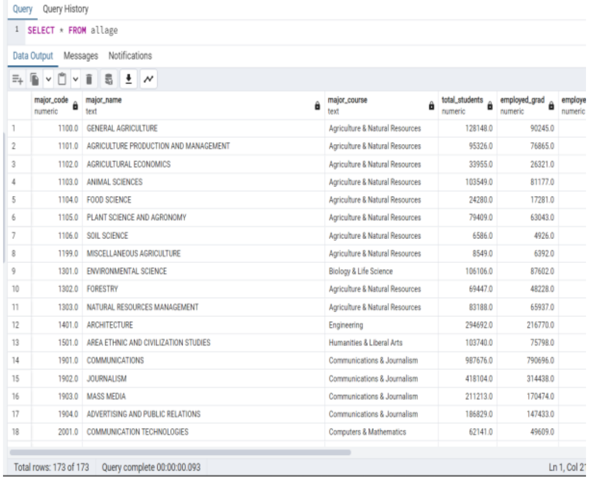
Format: csv

Encoding: Select an item...

Import csv file into table, repeat to the other 3 csv file

Create Table	Output

```sql
CREATE TABLE allage
(
major_code numeric,
major_name text,
major_course text,
Total_students numeric,
Employed_grad numeric,
Employed_full_time_year_round numeric,
Unemployed_grad numeric,
Unemployment_rate numeric,
Median_salary numeric,
P25th_salary numeric,
P75th_salary numeric
);
```

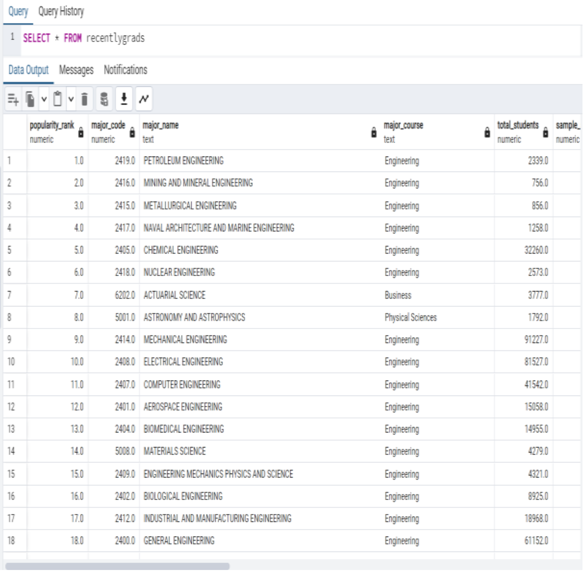


```sql
SELECT * FROM allage
```

	major_code numeric	major_name text	major_course text	total_students numeric	employed_grad numeric	employe numeric
1	1100.0	GENERAL AGRICULTURE	Agriculture & Natural Resources	128148.0	90245.0	
2	1101.0	AGRICULTURE PRODUCTION AND MANAGEMENT	Agriculture & Natural Resources	95326.0	76865.0	
3	1102.0	AGRICULTURAL ECONOMICS	Agriculture & Natural Resources	33955.0	26321.0	
4	1103.0	ANIMAL SCIENCES	Agriculture & Natural Resources	103549.0	81177.0	
5	1104.0	FOOD SCIENCE	Agriculture & Natural Resources	24280.0	17281.0	
6	1105.0	PLANT SCIENCE AND AGRONOMY	Agriculture & Natural Resources	79409.0	63043.0	
7	1106.0	SOIL SCIENCE	Agriculture & Natural Resources	6586.0	4926.0	
8	1199.0	MISCELLANEOUS AGRICULTURE	Agriculture & Natural Resources	8549.0	6392.0	
9	1301.0	ENVIRONMENTAL SCIENCE	Biology & Life Science	106106.0	87602.0	
10	1302.0	FORESTRY	Agriculture & Natural Resources	69447.0	48228.0	
11	1303.0	NATURAL RESOURCES MANAGEMENT	Agriculture & Natural Resources	83188.0	65937.0	
12	1401.0	ARCHITECTURE	Engineering	294692.0	216770.0	
13	1501.0	AREA ETHNIC AND CIVILIZATION STUDIES	Humanities & Liberal Arts	103740.0	75798.0	
14	1901.0	COMMUNICATIONS	Communications & Journalism	987676.0	790696.0	
15	1902.0	JOURNALISM	Communications & Journalism	418104.0	314438.0	
16	1903.0	MASS MEDIA	Communications & Journalism	211213.0	170474.0	
17	1904.0	ADVERTISING AND PUBLIC RELATIONS	Communications & Journalism	186829.0	147433.0	
18	2001.0	COMMUNICATION TECHNOLOGIES	Computers & Mathematics	62141.0	49609.0	

Total rows: 173 of 173 Query complete 00:00:00.093 Ln 1, Col 2

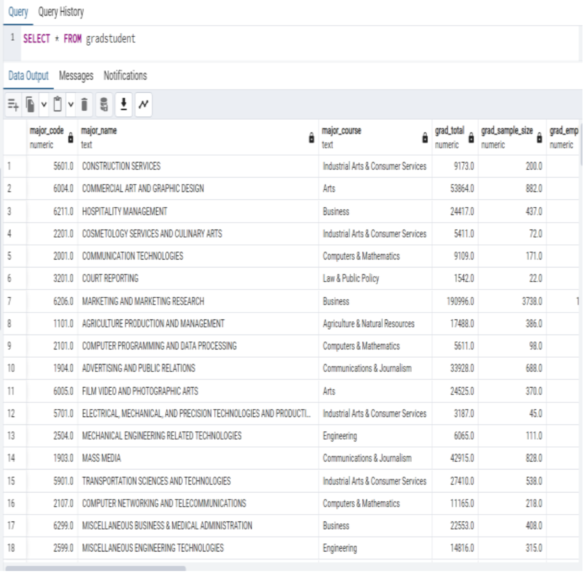
```sql
CREATE TABLE recentlygrads
(
popularity_rank numeric,
major_code numeric,
major_name text,
major_course text,
total_students numeric,
Sample_size numeric,
Men numeric,
Women numeric,
ShareWomen numeric,
employed_grad numeric,
full_time numeric,
part_time numeric,
Full_time_year_round numeric,
unemployed_grad numeric,
unemployment_rate numeric,
median_salary numeric,
P25th_salary numeric,
P75th_salary numeric,
College_jobs numeric,
Non_college_jobs numeric,
Low_wage_jobs numeric
);
```



```sql
1 SELECT * FROM recentlygrads
```

	popularity_rank	major_code	major_name	major_course	total_students	sample_
	numeric	numeric	text	text	numeric	numeric
1	1.0	2419.0	PETROLEUM ENGINEERING	Engineering	2339.0	
2	2.0	2416.0	MINING AND MINERAL ENGINEERING	Engineering	756.0	
3	3.0	2415.0	METALLURGICAL ENGINEERING	Engineering	856.0	
4	4.0	2417.0	NAVAL ARCHITECTURE AND MARINE ENGINEERING	Engineering	1258.0	
5	5.0	2405.0	CHEMICAL ENGINEERING	Engineering	32260.0	
6	6.0	2418.0	NUCLEAR ENGINEERING	Engineering	2573.0	
7	7.0	6202.0	ACTUARIAL SCIENCE	Business	3777.0	
8	8.0	5001.0	ASTRONOMY AND ASTROPHYSICS	Physical Sciences	1792.0	
9	9.0	2414.0	MECHANICAL ENGINEERING	Engineering	91227.0	
10	10.0	2408.0	ELECTRICAL ENGINEERING	Engineering	81527.0	
11	11.0	2407.0	COMPUTER ENGINEERING	Engineering	41542.0	
12	12.0	2401.0	AEROSPACE ENGINEERING	Engineering	15058.0	
13	13.0	2404.0	BIOMEDICAL ENGINEERING	Engineering	14955.0	
14	14.0	5008.0	MATERIALS SCIENCE	Engineering	4279.0	
15	15.0	2409.0	ENGINEERING MECHANICS PHYSICS AND SCIENCE	Engineering	4321.0	
16	16.0	2402.0	BIOLOGICAL ENGINEERING	Engineering	8925.0	
17	17.0	2412.0	INDUSTRIAL AND MANUFACTURING ENGINEERING	Engineering	18968.0	
18	18.0	2400.0	GENERAL ENGINEERING	Engineering	61152.0	

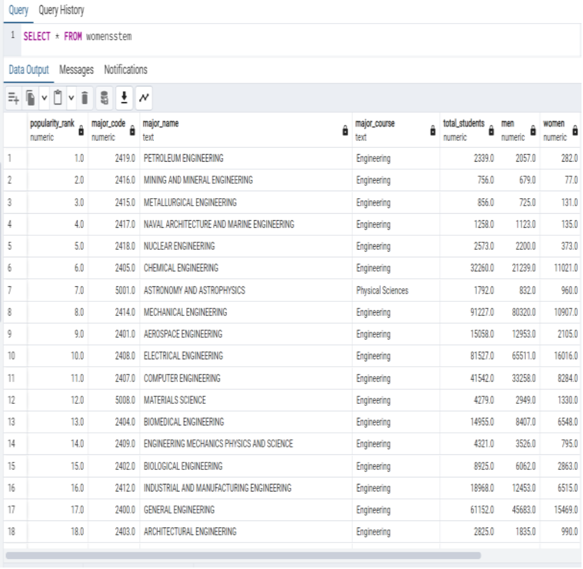
```sql
CREATE TABLE gradstudent
(
major_code numeric,
major_name text,
major_course text,
Grad_total numeric,
Grad_sample_size numeric,
Grad_employed numeric,
Grad_full_time_year_round numeric,
Grad_unemployed numeric,
Grad_unemployment_rate numeric,
Grad_median_salary numeric,
Grad_P25_salary numeric,
Grad_P75_salary numeric,
Nongrad_total numeric,
Nongrad_employed numeric,
Nongrad_full_time_year_round numeric,
Nongrad_unemployed numeric,
Nongrad_unemployment_rate numeric,
Nongrad_median_salary numeric,
Nongrad_P25_salary numeric,
Nongrad_P75_salary numeric,
Grad_share numeric,
diff_salary numeric
);
```



```sql
1 SELECT * FROM gradstudent
```

	major_code	major_name	major_course	grad_total	grad_sample_size	grad_emp
	numeric	text	text	numeric	numeric	numeric
1	5601.0	CONSTRUCTION SERVICES	Industrial Arts & Consumer Services	9173.0	200.0	
2	6004.0	COMMERCIAL ART AND GRAPHIC DESIGN	Arts	53864.0	882.0	
3	6211.0	HOSPITALITY MANAGEMENT	Business	24417.0	437.0	
4	2201.0	COSMETOLOGY SERVICES AND CULINARY ARTS	Industrial Arts & Consumer Services	5411.0	72.0	
5	2001.0	COMMUNICATION TECHNOLOGIES	Computers & Mathematics	9109.0	171.0	
6	3201.0	COURT REPORTING	Law & Public Policy	1542.0	22.0	
7	6206.0	MARKETING AND MARKETING RESEARCH	Business	190996.0	3738.0	1
8	1101.0	AGRICULTURE PRODUCTION AND MANAGEMENT	Agriculture & Natural Resources	17488.0	386.0	
9	2101.0	COMPUTER PROGRAMMING AND DATA PROCESSING	Computers & Mathematics	5611.0	98.0	
10	1904.0	ADVERTISING AND PUBLIC RELATIONS	Communications & Journalism	33928.0	688.0	
11	6005.0	FILM VIDEO AND PHOTOGRAPHIC ARTS	Arts	24525.0	370.0	
12	5701.0	ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTI...	Industrial Arts & Consumer Services	3187.0	45.0	
13	2504.0	MECHANICAL ENGINEERING RELATED TECHNOLOGIES	Engineering	6065.0	111.0	
14	1903.0	MASS MEDIA	Communications & Journalism	42915.0	828.0	
15	5901.0	TRANSPORTATION SCIENCES AND TECHNOLOGIES	Industrial Arts & Consumer Services	27410.0	538.0	
16	2107.0	COMPUTER NETWORKING AND TELECOMMUNICATIONS	Computers & Mathematics	11165.0	218.0	
17	6299.0	MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION	Business	22553.0	408.0	
18	2599.0	MISCELLANEOUS ENGINEERING TECHNOLOGIES	Engineering	14816.0	315.0	

```sql
CREATE TABLE womensstem
(
popularity_rank numeric,
major_code numeric,
major_name text,
major_course text,
total_students numeric,
men numeric,
women numeric,
sharewomen numeric,
median_salary numeric
);
```



```sql
1 SELECT * FROM womensstem
```

	popularity_rank	major_code	major_name	major_course	total_students	men	women
	numeric	numeric	text	text	numeric	numeric	numeric
1	1.0	2419.0	PETROLEUM ENGINEERING	Engineering	2339.0	2057.0	282.0
2	2.0	2416.0	MINING AND MINERAL ENGINEERING	Engineering	756.0	679.0	77.0
3	3.0	2415.0	METALLURGICAL ENGINEERING	Engineering	856.0	725.0	131.0
4	4.0	2417.0	NAVAL ARCHITECTURE AND MARINE ENGINEERING	Engineering	1258.0	1123.0	135.0
5	5.0	2418.0	NUCLEAR ENGINEERING	Engineering	2573.0	2200.0	373.0
6	6.0	2405.0	CHEMICAL ENGINEERING	Engineering	32260.0	21239.0	11021.0
7	7.0	5001.0	ASTRONOMY AND ASTROPHYSICS	Physical Sciences	1792.0	832.0	960.0
8	8.0	2414.0	MECHANICAL ENGINEERING	Engineering	91227.0	80320.0	10907.0
9	9.0	2401.0	AEROSPACE ENGINEERING	Engineering	15058.0	12953.0	2105.0
10	10.0	2408.0	ELECTRICAL ENGINEERING	Engineering	81527.0	65511.0	16016.0
11	11.0	2407.0	COMPUTER ENGINEERING	Engineering	41542.0	33258.0	8284.0
12	12.0	5008.0	MATERIALS SCIENCE	Engineering	4279.0	2949.0	1330.0
13	13.0	2404.0	BIOMEDICAL ENGINEERING	Engineering	14955.0	8407.0	6548.0
14	14.0	2409.0	ENGINEERING MECHANICS PHYSICS AND SCIENCE	Engineering	4321.0	3526.0	795.0
15	15.0	2402.0	BIOLOGICAL ENGINEERING	Engineering	8925.0	6062.0	2863.0
16	16.0	2412.0	INDUSTRIAL AND MANUFACTURING ENGINEERING	Engineering	18968.0	12453.0	6515.0
17	17.0	2400.0	GENERAL ENGINEERING	Engineering	61152.0	45683.0	15469.0
18	18.0	2403.0	ARCHITECTURAL ENGINEERING	Engineering	2825.0	1835.0	990.0

3.0 Database

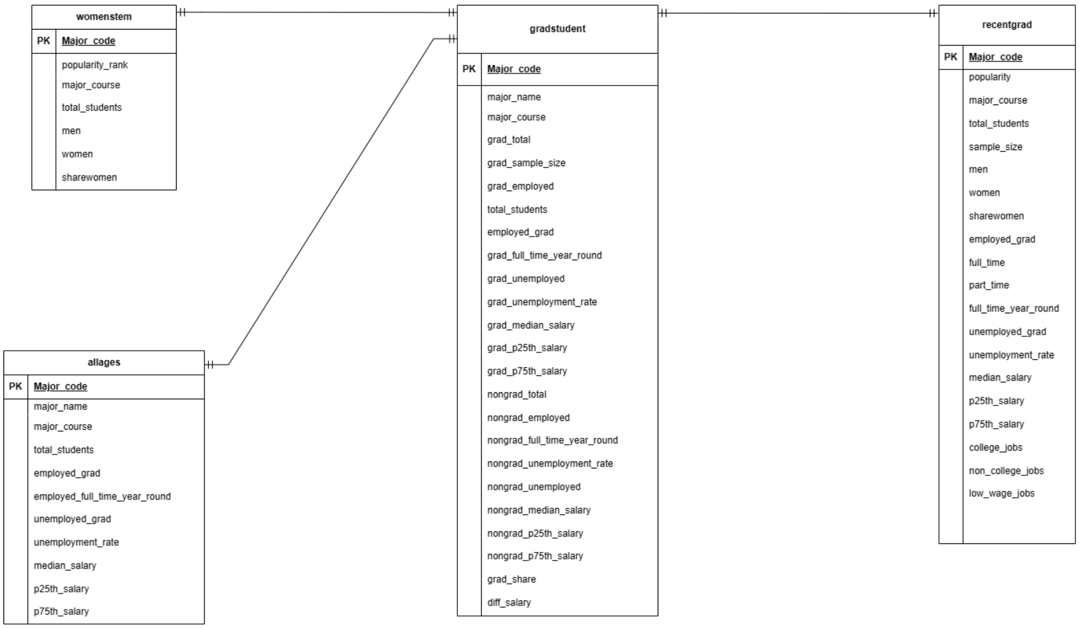


Figure 3.1 9

The entity-relationship diagram (ERD) for this project consists of four tables which is womenstem, allages, gradstudent, and recentgrad. The ERD represents a star schema, with the gradstudent table is the central fact table, and the other three tables are dimension tables.

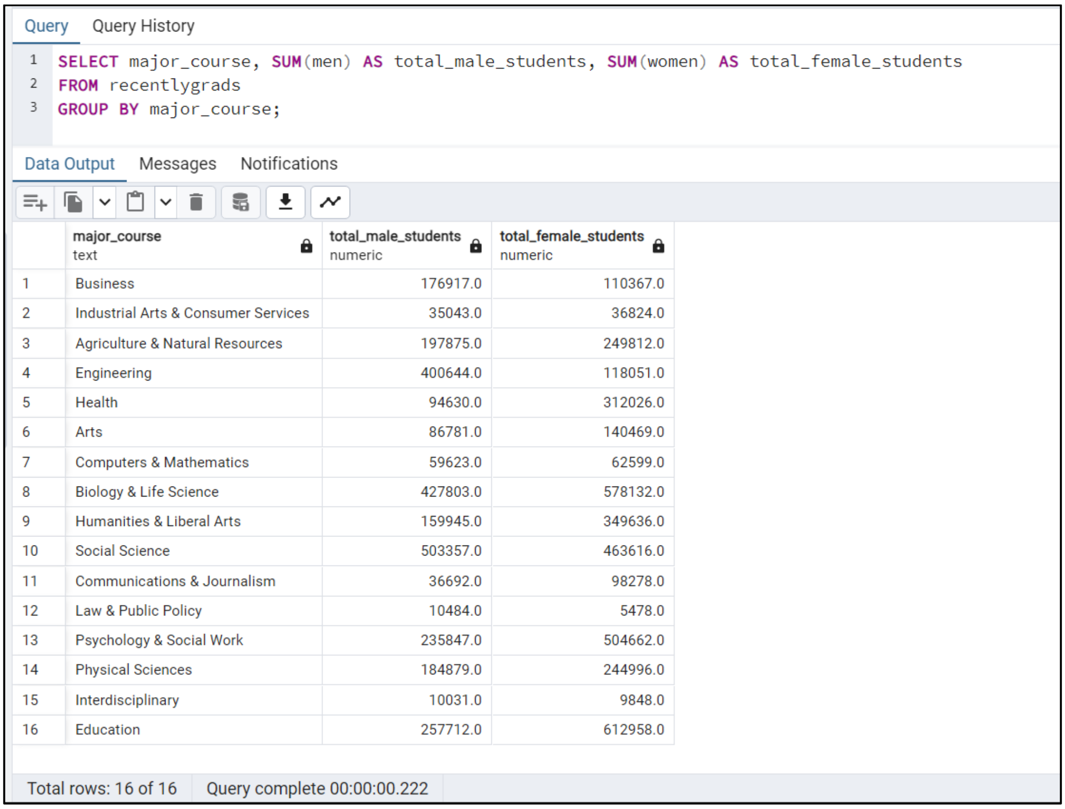
This ERD in this project is denormalized data structure to the third normal form. All relationships between the tables in this ERD are one-to-one type. Specifically, the gradstudent table has one-to-one relationships with the womenstem, allages, and recentgrad tables. Since this ERD is a star schema, it only requires simple joins, and it gives faster query result because of it. The dimension tables, which is tables womenstem, allages and recentgrad are not split into pieces. Data is redundant due its denormalized structure. Relationships between the tables are all in one-to-one relationship. Allages table has one-to-one relationship with the gradstudent table. Same goes for womenstem table has one-to-one relationship with gradstudent and lastly recentgrad has one-to-one relationship with gradstudent. This schema helps users look at graduate data in various ways, giving a clear understanding of factors related to job outcomes and gender differences in STEM fields.

4.0 Results and Data Analysis

After integrating the data, we analyzed it using Power BI for visualization. Then, we used PostgreSQL to perform OLAP tasks such as pivoting and slicing.

4.1 OLAP Coding

Slicing



```sql
SELECT major_course, SUM(men) AS total_male_students, SUM(women) AS total_female_students
FROM recentlygrads
GROUP BY major_course;
```

	major_course	total_male_students	total_female_students
	text	numeric	numeric
1	Business	176917.0	110367.0
2	Industrial Arts & Consumer Services	35043.0	36824.0
3	Agriculture & Natural Resources	197875.0	249812.0
4	Engineering	400644.0	118051.0
5	Health	94630.0	312026.0
6	Arts	86781.0	140469.0
7	Computers & Mathematics	59623.0	62599.0
8	Biology & Life Science	427803.0	578132.0
9	Humanities & Liberal Arts	159945.0	349636.0
10	Social Science	503357.0	463616.0
11	Communications & Journalism	36692.0	98278.0
12	Law & Public Policy	10484.0	5478.0
13	Psychology & Social Work	235847.0	504662.0
14	Physical Sciences	184879.0	244996.0
15	Interdisciplinary	10031.0	9848.0
16	Education	257712.0	612958.0

Total rows: 16 of 16 Query complete 00:00:00.222

Figure 4.1.1 shows the slicing operation conducted to examine the ratio of female to male graduates across majors

Based on figure 4.1.1, the objective here is to examine the ratio of female to male graduates across different college majors. By examining the ratio of female to male graduates, we can identify fields where one gender is leading or where there is a more balanced representation.

Slicing means picking one specific layer or section from a data cube to look at and summarizing the data within that layer. In this case, the column we are interested in is the "major_course" , which represents the different major fields. By slicing the data along this column, we can see the total number of male and female students within each major, giving us understanding into the gender distribution across majors.

Detailed Explaination:

Business: For every male student, there are about 0.62 female students. Business is mostly chosen by men, but there are also many women involved.

Industrial Arts & Consumer Services: There are almost an equal number of male and female students, but there are a bit more females than males.

Agriculture & Natural Resources: There are about 1.26 female students showing that this field is mostly filled with women

Engineering: This course mostly has men, with about 0.29 female for every men.

Health: In this fields, there are a lot more female than men, with about 3.30 female for every men.

Arts: Arts have a higher number of female students, with a ratio about 1.62 female students for every male student.

Computers & Mathematics: There are nearly the same number for both gender, with slightly more female students than male students.

Biology & Life Science: There are more female studying Biology & Life Science, with about 1.35 female for every men.

Humanities & Liberal Arts: Humanities & Liberal Arts have a strong female presence, with about 2.19 female students for every male student.

Social Science: Social Science is nearly balanced, with slightly more male students than female students.

Communications & Journalism: More female study Communications & Journalism, with about 2.68 female for every men.

Law & Public Policy: There are more men studying in this field, with about 0.52 female for every men

Psychology & Social Work: Lots of female study in this field, with about 2.14 female for every men.

Physical Sciences: Physical Sciences have more female students, with a ratio of about 1.33 female students for every male student.

Interdisciplinary: Both gender in Interdisciplinary fields is almost equal.

Education: This field has a lot more female student, with about 2.38 female for every men.

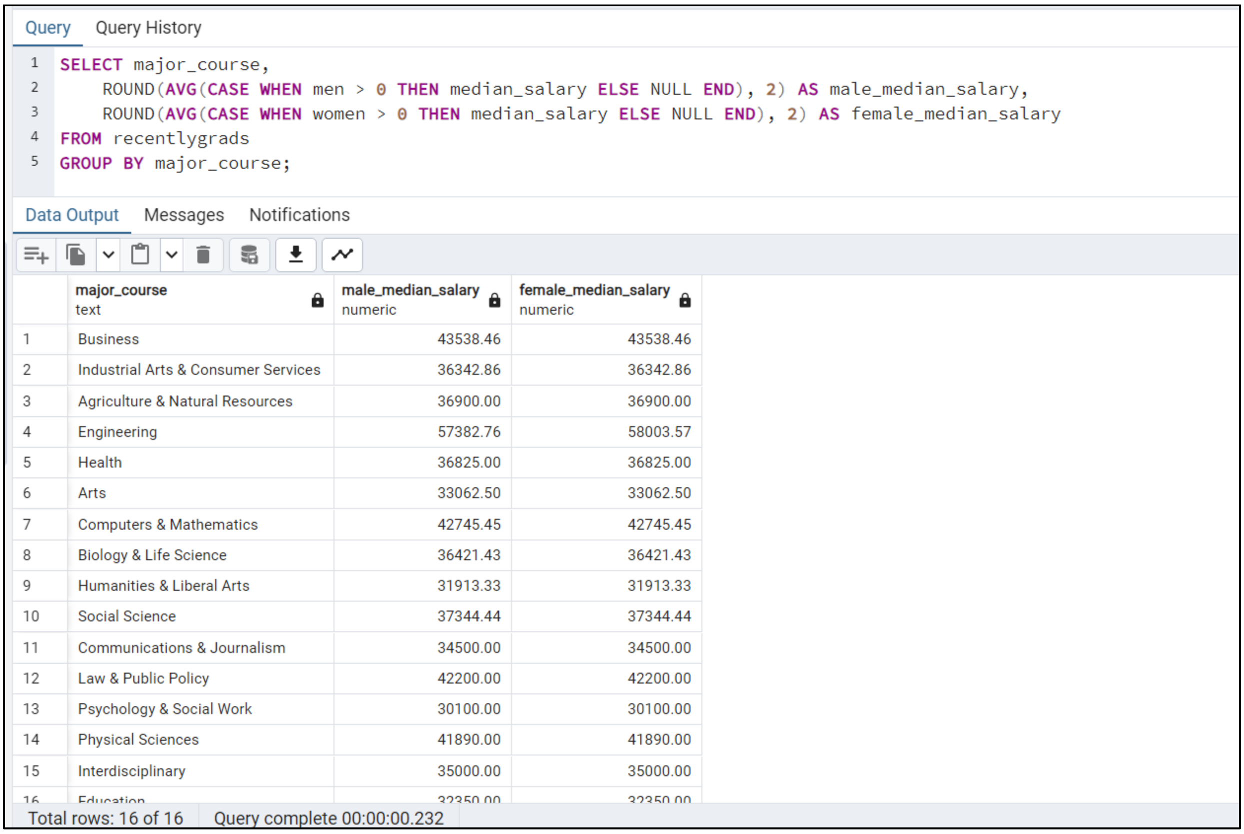
Summary:

Female-Dominated Fields: Health, Humanities & Liberal Arts, Communications & Journalism, Psychology & Social Work, and Education are mostly filled with female students.

Balanced Fields: Industrial Arts & Consumer Services, Computers & Mathematics, Social Science, and Interdisciplinary fields have almost equal numbers of men and female students.

Male-Dominated Fields: Business, Engineering, and Law & Public Policy have a lot more men than female students.

Pivot



```sql
SELECT major_course,
    ROUND(AVG(CASE WHEN men > 0 THEN median_salary ELSE NULL END), 2) AS male_median_salary,
    ROUND(AVG(CASE WHEN women > 0 THEN median_salary ELSE NULL END), 2) AS female_median_salary
FROM recentlygrads
GROUP BY major_course;
```

	major_course text	male_median_salary numeric	female_median_salary numeric
1	Business	43538.46	43538.46
2	Industrial Arts & Consumer Services	36342.86	36342.86
3	Agriculture & Natural Resources	36900.00	36900.00
4	Engineering	57382.76	58003.57
5	Health	36825.00	36825.00
6	Arts	33062.50	33062.50
7	Computers & Mathematics	42745.45	42745.45
8	Biology & Life Science	36421.43	36421.43
9	Humanities & Liberal Arts	31913.33	31913.33
10	Social Science	37344.44	37344.44
11	Communications & Journalism	34500.00	34500.00
12	Law & Public Policy	42200.00	42200.00
13	Psychology & Social Work	30100.00	30100.00
14	Physical Sciences	41890.00	41890.00
15	Interdisciplinary	35000.00	35000.00
16	Education	32350.00	32350.00

Total rows: 16 of 16 Query complete 00:00:00.232

Figure 4.1.2 displays the pivot operation used to determine whether the median salary is one of the factors leading to fewer women choosing STEM fields.

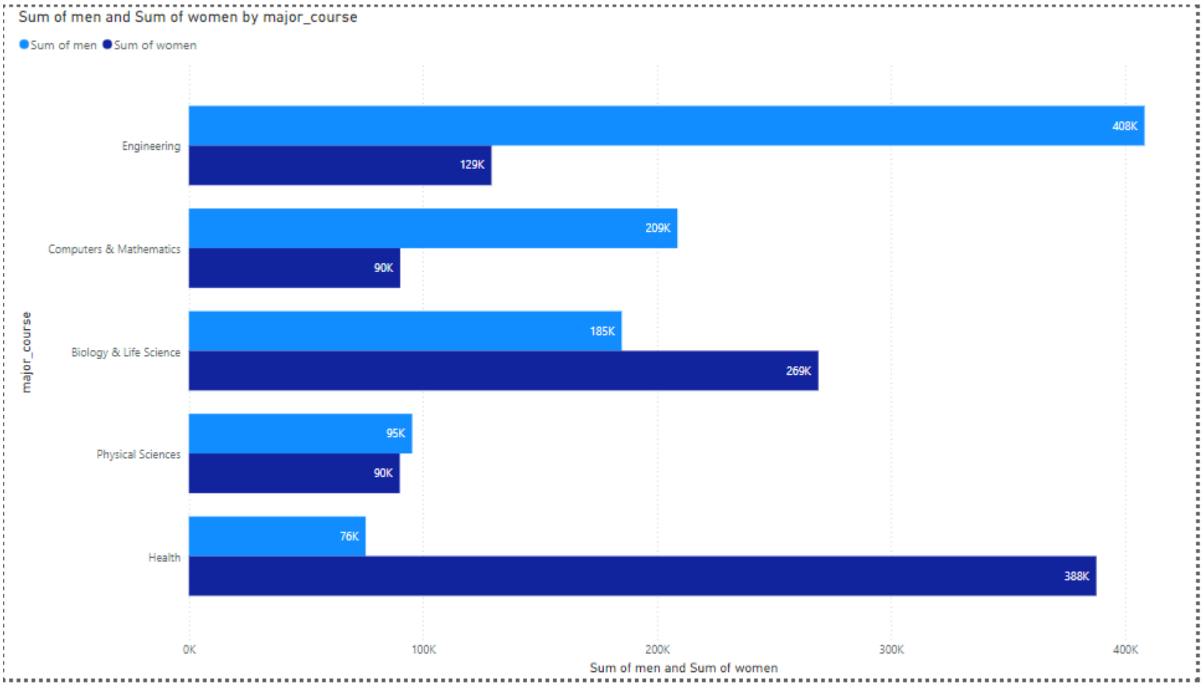
Based on figure 4.1.2, the objective here is to check whether the median salaries between male and female graduates in different STEM fields. This allows us easily to see if there are any differences in average salaries that might affect the jobs people choose.

Pivot operations mean turning around the data axes to show the information in a different way. In this situation, we are pivoting the data to compare the median salaries for male and female graduates separately within each STEM field. This means we will have separate columns for male median salaries and female median salaries.

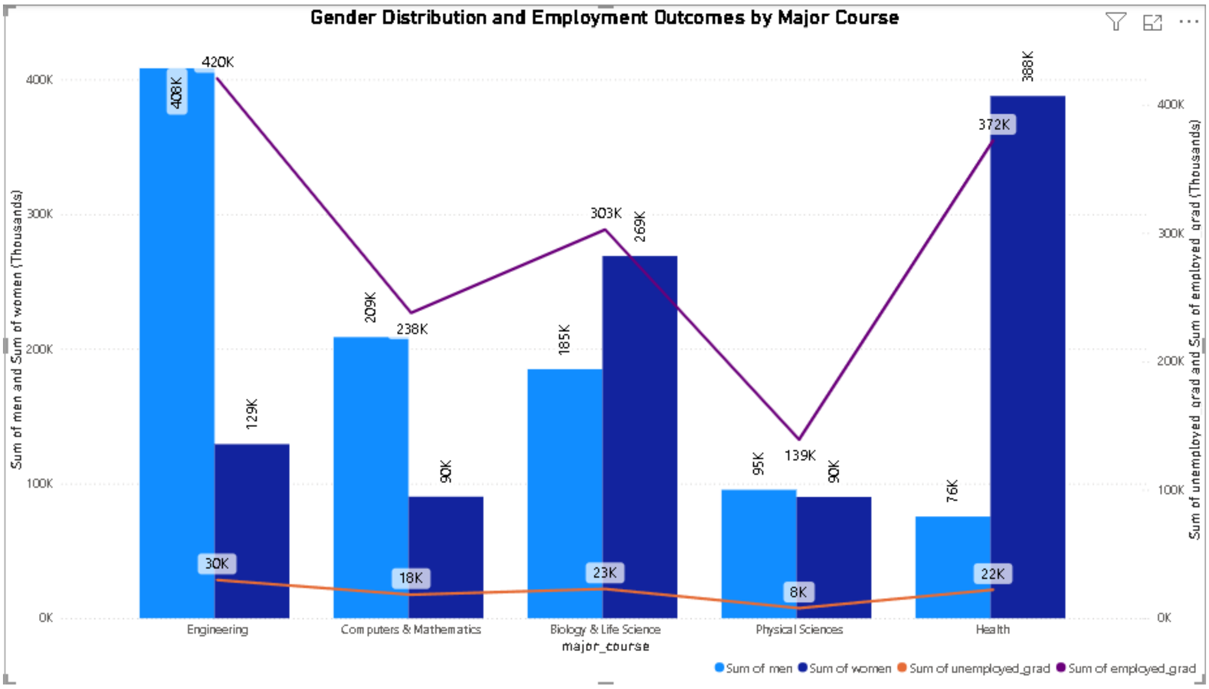
Pivoting the data allows us to compare the median salaries between genders within each STEM field directly. By presenting the data in this format, we can easily spot any different in median salaries based on gender within specific fields of study. This helps address the objective of determining salaries whether the median salary is one of the factors leading to fewer women choosing STEM fields.

The output displays that the median salaries for men and women are equal across all STEM fields. This consistency shows that there is no salary difference based on gender in all the STEM fields. Since the median salary is the same for both men and women, differences in pay are not causing fewer women to pick STEM fields. By focusing on median salary through the pivot operation, we can say that salary is not a factor in why there are fewer women in STEM fields.

4.2 Data Visualisation



The bar chart shows the sum of men and women in different STEM major. In engineering, there are many more men than women, showing that it is harder for women to get into or succeed in this field. Computers and mathematics also have more men than women, but the difference is smaller. In biology and life sciences, and health, there are more women than men, meaning these fields are easier for women to enter. Physical sciences have almost the same number of men and women, but still a few more men. Overall, the chart shows that some STEM fields have more men, and some have more women, pointing out the need for more gender equality in all STEM areas.



Above visualization shows that the relationship between gender and employment outcomes in STEM fields. As we can see, in engineering major, men are higher than women, it shows the ongoing barrier for women in this field, and we know that engineering is a most men choices in study. Additionally, the increasing number of unemployed graduates in engineering shows a potential issue with job market demand and unbalanced skills in them, women maybe face even more difficulties in finding job in this field. To promote gender equality, more thing needs to be done to encourage and support women in engineering field. Steps like mentorship, scholarships and work-life balance could help create a more supportive environment for women in this field.

Next, the Health and Biology & Life Science field, with a higher number of female graduates, shows that some STEM areas are more successful in achieving gender balance. The visualization also shows that women in those major not only graduate in large numbers but also have strong employment graduate in industries. Both major are better aligned with job market needs from those interpretation above. To achieve gender equality across all STEM fields, it's important to identify what makes fields like that successful for women and apply these strategies to other STEM areas. This approach can help ensure that women not only enter but also succeed in various STEM careers.

5.0 Conclusion

In the end, our objectives for this study in topic of gender inequality issue in STEM fields have been achieved and answered. We start from seeing in a wider sight. For overall graduates of all college majors, we can say that we couldn't highlight the gender inequality issues since not most of the course are being led by a gender. There are some courses that being led by men, and some led by women. We can say that both genders are trying their best and involve in all courses. They grab their opportunity in all courses.

When we start to narrow our sight to the STEM fields. We can say that, even though our issue is to solve the gender inequality in STEM field, but based on the data we have visualised, we can see that, not all the courses in STEM field are led by men. There are still some courses that led by women. In addition, we figured out that salary is not one of the factors that affect the number of women less taking STEM fields. It is because the salary of women and men across all STEM fields are equal. Both findings tell us that in every STEM field, men are women are treated equally.

Based on the employment student graduates in STEM field, it shows high statistics, for overall course is fine since not all the courses in STEM field are led by men but when we are focusing on every course, there is a problem. Since the employment are high, the chances for all men are women are there, to continue in STEM fields after graduated. Women need to be more outstanding to get into those courses like engineering and computers and mathematics. It is because both of those course shows a very high different between men and women ratio. They need to get the benefit and treated equally in this field.

Men are not left behind in this topic, there are still some courses in STEM field that women ratio are higher than men. Men need to get the same benefit in Health field, since they are so left behind. After all, most of the other courses they are doing well, it just a little different in number of men and women ratio. Men or woman can improve themselves and try to achieve this gender equality in STEM fields by getting mentorship, search for scholarships and applying work-life balance could help create a more supportive environment for women and men in this field.

In conclusion for the process throughout the project, we figure out that our project focused on converting and visualizing the datasets. We start with the process importing raw dataset into a data warehouse (postgres) and performing data cleaning and combining operations in Jupyter, resulting in a new, tidy CSV file. and transfer it back into the PostgreSQL. Last step is visualizing in PowerBI, and reporting.

Throughout the project, we faced some challenges, which shows our dedication to learn and improve our problem-solving skills. One of the big problems was finding relevant datasets from Kaggle, it was difficult to find one that perfectly matched our project's objectives. Additionally, we find the difficulties in connecting PostgreSQL with Jupyter for the ETL process, as we want to show the power of both tools. Making sure our coding is accurate also one of one problem, since a single mistake can delay our progress. During the analysis phase, we heavily discuss whether to analyze each table separately or after combined them. However, in the end, we do find out the answer and solve the problem together.

To summarised, our project shows the use of data warehouses and databases to convert raw datasets into organized formats. Even though we were facing challenges throughout the process, our team's efforts allowed us to overcome these issues. This project provided us with valuable experience in data transformation, teamwork, and problem-solving, greatly improving our skills in data management, time management and analysis.

6.0 References

Bismi, I. (2023, February 21). OLAP Operations in SQL - IQRA BisMi - Medium. Medium. https://medium.com/@iqra.bismi/olap-operations-in-sql-1293793d811e

Charlesworth, T. E., & Banaji, M. R. (2019). Gender in science, Technology, engineering, and Mathematics: Issues, causes, solutions. the Journal of Neuroscience/the Journal of Neuroscience, 39(37), 7228–7243. https://doi.org/10.1523/jneurosci.0475-18.2019

ERA Portal Austria – Eurostat: 41% of people employed as scientists and engineers are women. (n.d.). Era.gv.at. Retrieved May 27, 2024, from https://era.gv.at/news-items/eurostat-41-of-people-employed-as-scientists-and-engineers-are-women/

Jeferson.Zambrano. (2023, October 2). Women in Leadership Positions | MIT Professional Education. MIT Professional Education. https://professionalprograms.mit.edu/blog/leadership/the-gender-gap-in-stem/

Team, D., & Bothma, J. (2022, April 15). Power BI Tutorial for Beginners. https://www.datacamp.com/tutorial/tutorial-power-bi-for-beginners

Tiwari, N. (2023, December 6). Data cleaning using Pandas in Python – complete guide for beginners. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/data-cleaning-using-pandas/

What is an ETL Pipeline? (n.d.). Snowflake. https://www.snowflake.com/guides/etl-pipeline/

7.0 Appendix

Google drive:

-Raw dataset before cleaned process

-Cleaned dataset after cleaned process

-Coding for insert raw and cleaned tables in postgreSQL

-Coding for transformation process

https://drive.google.com/drive/folders/13Z3oSXxvPdRDucEjSgoyHjOYqn0ecs4o?usp=sharing

Source of dataset used:

-Kaggle

https://www.kaggle.com/datasets/thedevastator/uncovering-insights-to-college-majors-and-their?select=women-stem.csv