gender_inequality_new

May 14, 2024

[2]:
```
#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
```

```
Requirement already satisfied: ipython-sql in
c:\users\qlhmysrh\anaconda3\lib\site-packages (0.5.0)
Requirement already satisfied: prettytable in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython-sql) (3.10.0)
Requirement already satisfied: ipython in c:\users\qlhmysrh\anaconda3\lib\site-
packages (from ipython-sql) (8.20.0)
Requirement already satisfied: sqlalchemy>=2.0 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython-sql) (2.0.25)
Requirement already satisfied: sqlparse in c:\users\qlhmysrh\anaconda3\lib\site-
packages (from ipython-sql) (0.5.0)
Requirement already satisfied: six in c:\users\qlhmysrh\anaconda3\lib\site-
packages (from ipython-sql) (1.16.0)
Requirement already satisfied: ipython-genutils in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython-sql) (0.2.0)
Requirement already satisfied: typing-extensions>=4.6.0 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from sqlalchemy>=2.0->ipython-
sql) (4.9.0)
Requirement already satisfied: greenlet!=0.4.17 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from sqlalchemy>=2.0->ipython-
sql) (3.0.1)
Requirement already satisfied: decorator in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(5.1.1)
Requirement already satisfied: jedi>=0.16 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(0.18.1)
Requirement already satisfied: matplotlib-inline in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(0.1.6)
Requirement already satisfied: prompt-toolkit<3.1.0,>=3.0.41 in
```

```
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(3.0.43)
Requirement already satisfied: pygments>=2.4.0 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(2.15.1)
Requirement already satisfied: stack-data in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(0.2.0)
Requirement already satisfied: traitlets>=5 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from ipython->ipython-sql)
(5.7.1)
Requirement already satisfied: colorama in c:\users\qlhmysrh\anaconda3\lib\site-
packages (from ipython->ipython-sql) (0.4.6)
Requirement already satisfied: wcwidth in c:\users\qlhmysrh\anaconda3\lib\site-
packages (from prettytable->ipython-sql) (0.2.5)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from
jedi>=0.16->ipython->ipython-sql) (0.8.3)
Requirement already satisfied: executing in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from stack-
data->ipython->ipython-sql) (0.8.3)
Requirement already satisfied: asttokens in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from stack-
data->ipython->ipython-sql) (2.0.5)
Requirement already satisfied: pure-eval in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from stack-
data->ipython->ipython-sql) (0.2.2)
Requirement already satisfied: sqlalchemy in
c:\users\qlhmysrh\anaconda3\lib\site-packages (2.0.25)
Requirement already satisfied: typing-extensions>=4.6.0 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from sqlalchemy) (4.9.0)
Requirement already satisfied: greenlet!=0.4.17 in
c:\users\qlhmysrh\anaconda3\lib\site-packages (from sqlalchemy) (3.0.1)
Requirement already satisfied: psycopg2 in c:\users\qlhmysrh\anaconda3\lib\site-
packages (2.9.9)
```

[3]:
```
#since we are using SQL magic commands in the notebook
%reload_ext sql
```

[4]:
```
from sqlalchemy import create_engine
```

[5]:
```
import pandas as pd
```

[6]:
```
#connecting to PostgreSQL databases from Python.
import psycopg2 as ps
```

[7]:
```
#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
```

[8]:
```
conn=ps.connect(dbname="gender_inequality_old",
                 user="postgres", password="12345", host="localhost",
                 port="5432")
```

[9]:
```
#to retrieve information about table from database
sql="""SELECT * FROM pg_catalog.pg_tables"""
```

[10]:
```
sql="""SELECT * FROM allage"""
```

[11]:
```
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)
```

[11]:
```
     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]
```

[12]:

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

[12]:

```
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
```

[13]:

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

[13]: 0

[14]:
```
shape_allage=df_1.shape
shape_allage
```

[14]:
```
(173, 12)
```

[15]:
```
#drop unnecessary column
drop_column_df1 = df_1.drop(columns=['numbering'], inplace=True)
```

[16]:
```
# 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)
```

[17]:
```
df_1['unemployment_rate'] = df_1['unemployment_rate'].round(2)
```

[18]:
```
print(df_1)
```

```
     major_code                                        major_name  \
0        1100.0                               GENERAL AGRICULTURE
1        1101.0             AGRICULTURE PRODUCTION AND MANAGEMENT
2        1102.0                            AGRICULTURAL ECONOMICS
3        1103.0                                  ANIMAL SCIENCES
4        1104.0                                     FOOD SCIENCE
..          …                                                …
168      6211.0                           HOSPITALITY MANAGEMENT
169      6212.0         MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
170      6299.0  MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION
171      6402.0                                          HISTORY
172      6403.0                            UNITED STATES HISTORY

                        major_course  total_students  employed_grad  \
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_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]
```

[19]:

```
sql="""SELECT * FROM gradstudent"""
```

[20]:

```
df_2=sqlio.read_sql_query(sql,conn)
df_2
```

```
C:\Users\Qlhmysrh\AppData\Local\Temp\ipykernel_24472\398108606.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_2=sqlio.read_sql_query(sql,conn)
```

[20]:

```
     numbering  major_code                                 major  \
0          0.0      5601.0                 CONSTRUCTION SERVICES
1          1.0      6004.0        COMMERCIAL ART AND GRAPHIC DESIGN
2          2.0      6211.0                HOSPITALITY MANAGEMENT
3          3.0      2201.0   COSMETOLOGY SERVICES AND CULINARY ARTS
4          4.0      2001.0              COMMUNICATION TECHNOLOGIES
..         …           …                                      …
168      168.0      5203.0                 COUNSELING PSYCHOLOGY
```

```
169      169.0      5202.0                         CLINICAL PSYCHOLOGY
170      170.0      6106.0     HEALTH AND MEDICAL PREPARATORY PROGRAMS
171      171.0      2303.0                   SCHOOL STUDENT COUNSELING
172      172.0      2301.0  EDUCATIONAL ADMINISTRATION AND SUPERVISION

                          major_category  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  …  nongrad_total  nongrad_employed  \
0                  0.087543  …        86062.0           73607.0
1                  0.057756  …       461977.0          347166.0
2                  0.073867  …       179335.0          145597.0
3                  0.080901  …        37575.0           29738.0
4                  0.058411  …        53819.0           43163.0
..                      …  …            …                …
168                0.035600  …        16781.0           12377.0
169                0.044958  …         6519.0            4368.0
170                0.021687  …        26320.0           16221.0
171                0.051400  …         2232.0            1328.0
172                0.016761  …         4003.0            3079.0

     nongrad_full_time_year_round  nongrad_unemployed  \
0                         62435.0              3928.0
1                        250596.0             25484.0
```

```
2                    113579.0             7409.0
3                     23249.0             1661.0
4                     34231.0             3389.0
..                        …                  …
168                    8502.0              835.0
169                    3033.0              357.0
170                   12185.0             1012.0
171                     980.0              169.0
172                    2434.0                0.0

     nongrad_unemployment_rate  nongrad_median  nongrad_p25  nongrad_p75  \
0                     0.050661         65000.0      47000.0      98000.0
1                     0.068386         48000.0      34000.0      71000.0
2                     0.048423         50000.0      35000.0      75000.0
3                     0.052900         41600.0      29000.0      60000.0
4                     0.072800         52000.0      36000.0      78000.0
..                        …               …            …            …
168                   0.063200         40000.0      25000.0      50000.0
169                   0.075556         46000.0      30000.0      70000.0
170                   0.058725         51000.0      35000.0      87000.0
171                   0.112892         42000.0      27000.0      51000.0
172                   0.000000         58000.0      45000.0      79000.0

     grad_share  grad_premium
0      0.096320      0.153846
1      0.104420      0.250000
2      0.119837      0.300000
3      0.125878      0.129808
4      0.144753      0.096154
..          …             …
168    0.755354      0.250000
169    0.777014      0.521739
170    0.813718      1.647059
171    0.898881      0.333333
172    0.931175      0.120690

[173 rows x 23 columns]
```

```
[21]: check_null_2=df_2.isnull().sum()
      check_null_2
```

```
[21]: numbering                    0
      major_code                   0
      major                        0
      major_category               0
      grad_total                   0
      grad_sample_size             0
```

```
grad_employed                  0
grad_full_time_year_round      0
grad_unemployed                0
grad_unemployment_rate         0
grad_median                    0
grad_p25                       0
grad_p75                       0
nongrad_total                  0
nongrad_employed               0
nongrad_full_time_year_round   0
nongrad_unemployed             0
nongrad_unemployment_rate      0
nongrad_median                 0
nongrad_p25                    0
nongrad_p75                    0
grad_share                     0
grad_premium                   0
dtype: int64
```

[22]:

```
check_duplicate_2=df_2.duplicated().sum()
check_duplicate_2
```

[22]:

```
0
```

[23]:

```
shape_gradstudent_2=df_2.shape
shape_gradstudent_2
```

[23]:

```
(173, 23)
```

[24]:

```
# change the column names from the original names to new names
new_column_names_2 = {'major': 'major_name', 'grad_premium':␣
 ↪'diff_salary','major_category':'major_course','grad_median':
 ↪'grad_median_salary','grad_p25':'grad_p25th_salary','grad_p75':
 ↪'grad_p75th_salary','nongrad_median':'nongrad_median_salary','nongrad_p25':
 ↪'nongrad_p25th_salary','nongrad_p75':'nongrad_p75th_salary'}

# Use the rename() method to change the column names
df_2.rename(columns=new_column_names_2, inplace=True)
```

[25]:

```
#drop unnecessary column
drop_column_df2 = df_2.drop(columns=['numbering'], inplace=True)
```

[26]:

```
df_2['grad_unemployment_rate'] = df_2['grad_unemployment_rate'].round(2)
```

[27]:

```
df_2['nongrad_unemployment_rate'] = df_2['nongrad_unemployment_rate'].round(2)
```

[28]:

```
df_2['grad_share'] = df_2['grad_share'].round(2)
```

[29]:

```
df_2['diff_salary'] = df_2['diff_salary'].round(4)
```

[30]:

```
print(df_2)
```

```
     major_code                               major_name  \
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]
```

[31]:

```
sql="""SELECT * FROM recentlygrads"""
```

[32]:

```
df_3=sqlio.read_sql_query(sql,conn)
df_3
```

```
C:\Users\Qlhmysrh\AppData\Local\Temp\ipykernel_24472\515624067.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_3=sqlio.read_sql_query(sql,conn)
```

[32]:

```
     numbering  popularity_rank  major_code  \
0          0.0              1.0      2419.0
1          1.0              2.0      2416.0
2          2.0              3.0      2415.0
3          3.0              4.0      2417.0
4          4.0              5.0      2405.0
..           …                …           …
168      168.0            169.0      3609.0
169      169.0            170.0      5201.0
170      170.0            171.0      5202.0
171      171.0            172.0      5203.0
172      172.0            173.0      3501.0

                                        major            major_category  \
0                       PETROLEUM ENGINEERING               Engineering
1                 MINING AND MINERAL ENGINEERING               Engineering
2                    METALLURGICAL ENGINEERING               Engineering
3     NAVAL ARCHITECTURE AND MARINE ENGINEERING               Engineering
4                        CHEMICAL ENGINEERING               Engineering
..                                          …                         …
168                                   ZOOLOGY    Biology & Life Science
169                     EDUCATIONAL PSYCHOLOGY  Psychology & Social Work
170                        CLINICAL PSYCHOLOGY  Psychology & Social Work
171                      COUNSELING PSYCHOLOGY  Psychology & Social Work
172                            LIBRARY SCIENCE                 Education

       total  sample_size      men    women  sharewomen  …  part_time  \
0     2339.0         36.0   2057.0    282.0    0.120564  …      270.0
1      756.0          7.0    679.0     77.0    0.101852  …      170.0
2      856.0          3.0    725.0    131.0    0.153037  …      133.0
3     1258.0         16.0   1123.0    135.0    0.107313  …      150.0
4    32260.0        289.0  21239.0  11021.0    0.341631  …     5180.0
..         …            …        …        …           …  …          …
168   8409.0         47.0   3050.0   5359.0    0.637293  …     2190.0
169   2854.0          7.0    522.0   2332.0    0.817099  …      572.0
170   2838.0         13.0    568.0   2270.0    0.799859  …      648.0
171   4626.0         21.0    931.0   3695.0    0.798746  …      965.0
```

```
172   1098.0          2.0    134.0    964.0    0.877960  …      237.0

     full_time_year_round  unemployed  unemployment_rate    median    p25th  \
0                  1207.0        37.0           0.018381  110000.0  95000.0
1                   388.0        85.0           0.117241   75000.0  55000.0
2                   340.0        16.0           0.024096   73000.0  50000.0
3                   692.0        40.0           0.050125   70000.0  43000.0
4                 16697.0      1672.0           0.061098   65000.0  50000.0
..                    …           …                 …         …        …
168                3602.0       304.0           0.046320   26000.0  20000.0
169                1211.0       148.0           0.065112   25000.0  24000.0
170                1293.0       368.0           0.149048   25000.0  25000.0
171                2738.0       214.0           0.053621   23400.0  19200.0
172                 410.0        87.0           0.104946   22000.0  20000.0

        p75th  college_jobs  non_college_jobs  low_wage_jobs
0    125000.0        1534.0             364.0          193.0
1     90000.0         350.0             257.0           50.0
2    105000.0         456.0             176.0            0.0
3     80000.0         529.0             102.0            0.0
4     75000.0       18314.0            4440.0          972.0
..        …             …                 …             …
168   39000.0        2771.0            2947.0          743.0
169   34000.0        1488.0             615.0           82.0
170   40000.0         986.0             870.0          622.0
171   26000.0        2403.0            1245.0          308.0
172   22000.0         288.0             338.0          192.0

[173 rows x 22 columns]
```

[33]:
```
check_null_3=df_3.isnull().sum()
check_null_3
```

[33]:
```
numbering               0
popularity_rank         0
major_code              0
major                   0
major_category          0
total                   0
sample_size             0
men                     0
women                   0
sharewomen              0
employed                0
full_time               0
part_time               0
full_time_year_round    0
```

```
unemployed              0
unemployment_rate       1
median                  0
p25th                   0
p75th                   0
college_jobs            0
non_college_jobs        0
low_wage_jobs           0
dtype: int64
```

[35]:

```
# Fill null values in 'unemployment_rate' column with integer 0
df_3['unemployment_rate'].fillna(0, inplace=True)

# Round the values in 'unemployment_rate' column to 2 decimal places
df_3['unemployment_rate'] = df_3['unemployment_rate'].round(2)
```

[36]:

```
df_3['sharewomen'] = df_3['sharewomen'].round(2)
```

[37]:

```
check_null_3=df_3.isnull().sum()
check_null_3
```

[37]:

```
numbering               0
popularity_rank         0
major_code              0
major                   0
major_category          0
total                   0
sample_size             0
men                     0
women                   0
sharewomen              0
employed                0
full_time               0
part_time               0
full_time_year_round    0
unemployed              0
unemployment_rate       0
median                  0
p25th                   0
p75th                   0
college_jobs            0
non_college_jobs        0
low_wage_jobs           0
dtype: int64
```

[38]:

```
check_duplicate_3=df_3.duplicated().sum()
check_duplicate_3
```

[38]: 0

[39]:
```
shape_majorlist_3=df_3.shape
shape_majorlist_3
```

[39]: (173, 22)

[40]:
```
# change the column names from the original names to new names
new_column_names_3 = {'major': 'major_name','major_category':
 ↪'major_course','total':'total_students','employed':
 ↪'employed_grad','unemployed':'unemployed_grad','median':
 ↪'median_salary','p25th':'p25th_salary','p75th':'p75th_salary'}

# Use the rename() method to change the column names
df_3.rename(columns=new_column_names_3, inplace=True)
```

[41]:
```
#drop unnecessary column
drop_column_df3 = df_3.drop(columns=['numbering'], inplace=True)
```

[42]:
```
print(df_3)
```

```
     popularity_rank  major_code                               major_name  \
0                1.0      2419.0                     PETROLEUM ENGINEERING
1                2.0      2416.0            MINING AND MINERAL ENGINEERING
2                3.0      2415.0                 METALLURGICAL ENGINEERING
3                4.0      2417.0 NAVAL ARCHITECTURE AND MARINE ENGINEERING
4                5.0      2405.0                      CHEMICAL ENGINEERING
..               …         …                                          …
168            169.0      3609.0                                   ZOOLOGY
169            170.0      5201.0                    EDUCATIONAL PSYCHOLOGY
170            171.0      5202.0                       CLINICAL PSYCHOLOGY
171            172.0      5203.0                     COUNSELING PSYCHOLOGY
172            173.0      3501.0                           LIBRARY SCIENCE

                major_course  total_students  sample_size      men    women  \
0                Engineering          2339.0         36.0   2057.0    282.0
1                Engineering           756.0          7.0    679.0     77.0
2                Engineering           856.0          3.0    725.0    131.0
3                Engineering          1258.0         16.0   1123.0    135.0
4                Engineering         32260.0        289.0  21239.0  11021.0
..                       …             …           …      …      …
168    Biology & Life Science          8409.0         47.0   3050.0   5359.0
169  Psychology & Social Work          2854.0          7.0    522.0   2332.0
170  Psychology & Social Work          2838.0         13.0    568.0   2270.0
171  Psychology & Social Work          4626.0         21.0    931.0   3695.0
172                Education          1098.0          2.0    134.0    964.0

     sharewomen  employed_grad  …  part_time  full_time_year_round  \
0          0.12        1976.0  …     270.0                    1207.0
1          0.10         640.0  …     170.0                     388.0
2          0.15         648.0  …     133.0                     340.0
3          0.11         758.0  …     150.0                     692.0
4          0.34       25694.0  …    5180.0                   16697.0
..          …           …   …       …                   …
168        0.64        6259.0  …    2190.0                    3602.0
169        0.82        2125.0  …     572.0                    1211.0
170        0.80        2101.0  …     648.0                    1293.0
171        0.80        3777.0  …     965.0                    2738.0
172        0.88         742.0  …     237.0                     410.0

     unemployed_grad  unemployment_rate  median_salary  p25th_salary  \
0               37.0               0.02       110000.0       95000.0
1               85.0               0.12        75000.0       55000.0
2               16.0               0.02        73000.0       50000.0
3               40.0               0.05        70000.0       43000.0
4             1672.0               0.06        65000.0       50000.0
..               …                 …               …             …
168            304.0               0.05        26000.0       20000.0
169            148.0               0.07        25000.0       24000.0
170            368.0               0.15        25000.0       25000.0
171            214.0               0.05        23400.0       19200.0
172             87.0               0.10        22000.0       20000.0

     p75th_salary  college_jobs  non_college_jobs  low_wage_jobs
0        125000.0        1534.0             364.0          193.0
1         90000.0         350.0             257.0           50.0
2        105000.0         456.0             176.0            0.0
3         80000.0         529.0             102.0            0.0
4         75000.0       18314.0            4440.0          972.0
..            …             …                …              …
168       39000.0        2771.0            2947.0          743.0
169       34000.0        1488.0             615.0           82.0
170       40000.0         986.0             870.0          622.0
171       26000.0        2403.0            1245.0          308.0
172       22000.0         288.0             338.0          192.0

[173 rows x 21 columns]
```

[43]:
```
sql="""SELECT * FROM womensstem"""
```

[44]:
```
df_4=sqlio.read_sql_query(sql,conn)
df_4
```

```
C:\Users\Qlhmysrh\AppData\Local\Temp\ipykernel_24472\113174976.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_4=sqlio.read_sql_query(sql,conn)
```

[44]:

```
    numbering  popularity_rank  major_code  \
0         0.0              1.0      2419.0
1         1.0              2.0      2416.0
2         2.0              3.0      2415.0
3         3.0              4.0      2417.0
4         4.0              5.0      2418.0
..        …                …          …
71       71.0             72.0      3604.0
72       72.0             73.0      6109.0
73       73.0             74.0      6100.0
74       74.0             75.0      6102.0
75       75.0             76.0      3609.0

                                       major         major_category  \
0                      PETROLEUM ENGINEERING            Engineering
1             MINING AND MINERAL ENGINEERING            Engineering
2                  METALLURGICAL ENGINEERING            Engineering
3    NAVAL ARCHITECTURE AND MARINE ENGINEERING          Engineering
4                        NUCLEAR ENGINEERING            Engineering
..                                         …                  …
71                                   ECOLOGY  Biology & Life Science
72             TREATMENT THERAPY PROFESSIONS                 Health
73          GENERAL MEDICAL AND HEALTH SERVICES              Health
74  COMMUNICATION DISORDERS SCIENCES AND SERVICES             Health
75                                   ZOOLOGY  Biology & Life Science

      total      men    women  sharewomen    median
0    2339.0   2057.0    282.0    0.120564  110000.0
1     756.0    679.0     77.0    0.101852   75000.0
2     856.0    725.0    131.0    0.153037   73000.0
3    1258.0   1123.0    135.0    0.107313   70000.0
4    2573.0   2200.0    373.0    0.144967   65000.0
..      …        …        …          …        …
71   9154.0   3878.0   5276.0    0.576360   33000.0
72  48491.0  13487.0  35004.0    0.721866   33000.0
73  33599.0   7574.0  26025.0    0.774577   32400.0
74  38279.0   1225.0  37054.0    0.967998   28000.0
75   8409.0   3050.0   5359.0    0.637293   26000.0

[76 rows x 10 columns]
```

[45]:

```python
check_null_4=df_4.isnull().sum()
check_null_4
```

[45]:

```
numbering          0
popularity_rank    0
major_code         0
major              0
major_category     0
total              0
men                0
women              0
sharewomen         0
median             0
dtype: int64
```

[46]:

```
check_duplicate_4=df_4.duplicated().sum()
check_duplicate_4
```

[46]:

```
0
```

[47]:

```
shape_majorlist_4=df_4.shape
shape_majorlist_4
```

[47]:

```
(76, 10)
```

[48]:

```
# change the column names from the original names to new names
new_column_names_4 = {'major': 'major_name','major_category':
 ↪'major_course','total':'total_students','median':'median_salary'}

# Use the rename() method to change the column names
df_4.rename(columns=new_column_names_4, inplace=True)
```

[49]:

```
#drop unnecessary column
drop_column_df4 = df_4.drop(columns=['numbering'], inplace=True)
```

[51]:

```
df_4['sharewomen'] = df_4['sharewomen'].round(2)
```

[52]:

```
print(df_4)
```

```
    popularity_rank  major_code  \
0               1.0      2419.0
1               2.0      2416.0
2               3.0      2415.0
3               4.0      2417.0
4               5.0      2418.0
..              …          …
71             72.0      3604.0
72             73.0      6109.0
73             74.0      6100.0
74             75.0      6102.0
75             76.0      3609.0
```

```
                                       major_name            major_course  \
0                           PETROLEUM ENGINEERING             Engineering
1                  MINING AND MINERAL ENGINEERING             Engineering
2                       METALLURGICAL ENGINEERING             Engineering
3       NAVAL ARCHITECTURE AND MARINE ENGINEERING             Engineering
4                             NUCLEAR ENGINEERING             Engineering
..                                            …                     …
71                                        ECOLOGY  Biology & Life Science
72                  TREATMENT THERAPY PROFESSIONS                  Health
73            GENERAL MEDICAL AND HEALTH SERVICES                  Health
74  COMMUNICATION DISORDERS SCIENCES AND SERVICES                  Health
75                                        ZOOLOGY  Biology & Life Science

    total_students       men    women  sharewomen  median_salary
0           2339.0    2057.0    282.0        0.12       110000.0
1            756.0     679.0     77.0        0.10        75000.0
2            856.0     725.0    131.0        0.15        73000.0
3           1258.0    1123.0    135.0        0.11        70000.0
4           2573.0    2200.0    373.0        0.14        65000.0
..             …      …      …         …            …
71          9154.0    3878.0   5276.0        0.58        33000.0
72         48491.0   13487.0  35004.0        0.72        33000.0
73         33599.0    7574.0  26025.0        0.77        32400.0
74         38279.0    1225.0  37054.0        0.97        28000.0
75          8409.0    3050.0   5359.0        0.64        26000.0

[76 rows x 9 columns]
```

[53]:

```python
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
}
```

```python
# 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)
```

[ ]: