LABS

Data Mining

Eng: Malek Almosanif

Lab 5 => Data Encoding And Correlation by Eng:Malek Almosanif

Sklearn Lib

Data Sets on Sklearns:

1.1.1 Iris data 1.1.2 Digits data 1.1.3 Boston data 1.1.4 Wine data 1.1.5 Breast cancer data 1.1.6 Diabetes data 1.1.7 Sample Regression 1.1.8 Sample Classification 1.1.9 Sample images

```
from sklearn.datasets import load_breast_cancer
from sklearn.datasets import load_diabetes

dataSets=load_breast_cancer()
X=dataSets.data
y=dataSets.target

print(dataSets.feature_names)
print('****************************')
print(dataSets.target_names)
print('************************************')
x=X[:20,:]
```

```
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
****************************
['malignant' 'benign']
************************************
```

```
dataset=load_diabetes()
X=dataset.data
y=dataset.target
print(dataset.feature_names)
print('****************************')
# print(dataset.target_names)
print('************************************')
x=X[:20,:]
x.reshape(20,-1)
x
```

```
['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']
**************************
**********************************

array([[ 3.80759064e-02,  5.06801187e-02,  6.16962065e-02,
         2.18723855e-02, -4.42234984e-02, -3.48207628e-02,
        -4.34008457e-02, -2.59226200e-03,  1.99074862e-02,
        -1.76461252e-02],
       [-1.88201653e-03, -4.46416365e-02, -5.14740612e-02,
        -2.63275281e-02, -8.44872411e-03, -1.91633397e-02,
         7.44115641e-02, -3.94933829e-02, -6.83315471e-02,
        -9.22040496e-02],
       [ 8.52989063e-02,  5.06801187e-02,  4.44512133e-02,
        -5.67042229e-03, -4.55994513e-02, -3.41944659e-02,
        -3.23559322e-02, -2.59226200e-03,  2.86130929e-03,
        -2.59303390e-02],
       [-8.90629394e-02, -4.46416365e-02, -1.15950145e-02,
        -3.66560811e-02,  1.21905688e-02,  2.49905934e-02,
        -3.60375700e-02,  3.43088589e-02,  2.26877450e-02,
        -9.36191133e-03],
       [ 5.38306037e-03, -4.46416365e-02, -3.63846922e-02,
         2.18723855e-02,  3.93485161e-03,  1.55961395e-02,
         8.14208361e-03, -2.59226200e-03, -3.19876395e-02,
        -4.66408736e-02],
       [-9.26954778e-02, -4.46416365e-02, -4.06959405e-02,
        -1.94418262e-02, -6.89906499e-02, -7.92878444e-02,
         4.12768238e-02, -7.63945038e-02, -4.11761669e-02,
        -9.63461565e-02],
       [-4.54724779e-02,  5.06801187e-02, -4.71628129e-02,
        -1.59989752e-02, -4.00956398e-02, -2.48000121e-02,
         7.78807997e-04, -3.94933829e-02, -6.29168791e-02,
        -3.83566597e-02],
       [ 6.35036756e-02,  5.06801187e-02, -1.89470584e-03,
         6.66294482e-02,  9.06198817e-02,  1.08914381e-01,
         2.28686348e-02,  1.77033545e-02, -3.58161926e-02,
         3.06440941e-03],
       [ 4.17084449e-02,  5.06801187e-02,  6.16962065e-02,
        -4.00989321e-02, -1.39525355e-02,  6.20168566e-03,
        -2.86742944e-02, -2.59226200e-03, -1.49596938e-02,
         1.13486232e-02],
       [-7.09002471e-02, -4.46416365e-02,  3.90621530e-02,
        -3.32132301e-02, -1.25765827e-02, -3.45076144e-02,
        -2.49926566e-02, -2.59226200e-03,  6.77370531e-02,
        -1.35040182e-02],
       [-9.63280163e-02, -4.46416365e-02, -8.38084235e-02,
         8.10098161e-03, -1.03389471e-01, -9.05611890e-02,
        -1.39477432e-02, -7.63945038e-02, -6.29168791e-02,
        -3.42145528e-02],
       [ 2.71782911e-02,  5.06801187e-02,  1.75059115e-02,
        -3.32132301e-02, -7.07277125e-03,  4.59715403e-02,
```

```
        -6.54906725e-02,  7.12099798e-02, -9.64349499e-02,
        -5.90671943e-02],
       [ 1.62806757e-02, -4.46416365e-02, -2.88400077e-02,
        -9.11327327e-03, -4.32086554e-03, -9.76888589e-03,
         4.49584616e-02, -3.94933829e-02, -3.07479175e-02,
        -4.24987666e-02],
       [ 5.38306037e-03,  5.06801187e-02, -1.89470584e-03,
         8.10098161e-03, -4.32086554e-03, -1.57187067e-02,
        -2.90282981e-03, -2.59226200e-03,  3.83939283e-02,
        -1.35040182e-02],
       [ 4.53409833e-02, -4.46416365e-02, -2.56065715e-02,
        -1.25561242e-02,  1.76943802e-02, -6.12835791e-05,
         8.17748397e-02, -3.94933829e-02, -3.19876395e-02,
        -7.56356220e-02],
       [-5.27375548e-02,  5.06801187e-02, -1.80618869e-02,
         8.04008521e-02,  8.92439288e-02,  1.07661787e-01,
        -3.97192078e-02,  1.08111101e-01,  3.60603340e-02,
        -4.24987666e-02],
       [-5.51455498e-03, -4.46416365e-02,  4.22955892e-02,
         4.94151933e-02,  2.45741445e-02, -2.38605667e-02,
         7.44115641e-02, -3.94933829e-02,  5.22769910e-02,
         2.79170509e-02],
       [ 7.07687525e-02,  5.06801187e-02,  1.21168511e-02,
         5.63008953e-02,  3.42058145e-02,  4.94161734e-02,
        -3.97192078e-02,  3.43088589e-02,  2.73640491e-02,
        -1.07769750e-03],
       [-3.82074010e-02, -4.46416365e-02, -1.05172024e-02,
        -3.66560811e-02, -3.73437341e-02, -1.94764882e-02,
        -2.86742944e-02, -2.59226200e-03, -1.81136923e-02,
        -1.76461252e-02],
       [-2.73097857e-02, -4.46416365e-02, -1.80618869e-02,
        -4.00989321e-02, -2.94491268e-03, -1.13346282e-02,
         3.75951860e-02, -3.94933829e-02, -8.94339609e-03,
        -5.49250874e-02]])
```

Clean Data With Sklearn

```
from sklearn.impute import SimpleImputer
import pandas as pd
import numpy as np

mydataSet=pd.read_csv('marketing_campaign_Dataset.csv')

mydataSet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 27 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
```

```
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Response             2240 non-null   int64
dtypes: float64(1), int64(23), object(3)
memory usage: 472.6+ KB
```

```
mydataSet.isna().sum()
```

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
```

```
NumStorePurchases       0
NumWebVisitsMonth       0
AcceptedCmp3            0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Response                0
dtype: int64
```

```
cleaner=SimpleImputer(missing_values=np.nan,strategy='most_frequent')
# most_frequent,mean,median
my_numerical_data = mydataSet.select_dtypes(include=['int64',
'float64'])

my_numerical_data.head()
```

```
     ID  Year_Birth   Income  Kidhome  Teenhome  Recency  MntWines
MntFruits  \
0  5524        1957  58138.0        0         0       58       635
88
1  2174        1954  46344.0        1         1       38        11
1
2  4141        1965  71613.0        0         0       26       426
49
3  6182        1984  26646.0        1         0       26        11
4
4  5324        1981  58293.0        1         0       94       173
43

   MntMeatProducts  MntFishProducts  ...  NumCatalogPurchases  \
0              546              172  ...                   10
1                6                2  ...                    1
2              127              111  ...                    2
3               20               10  ...                    0
4              118               46  ...                    3

   NumStorePurchases  NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4  \
0                  4                  7             0             0
1                  2                  5             0             0
2                 10                  4             0             0
3                  4                  6             0             0
4                  6                  5             0             0

   AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response
0             0             0             0         0         1
1             0             0             0         0         0
2             0             0             0         0         0
3             0             0             0         0         0
```

```
4              0              0              0         0         0

[5 rows x 24 columns]
```

```
my_numerical_data[my_numerical_data["Income"].isna()]
```

```
          ID  Year_Birth  Income  Kidhome  Teenhome  Recency  MntWines
\
10      1994        1983     NaN        1         0       11         5
27      5255        1986     NaN        1         0       19         5
43      7281        1959     NaN        0         0       80        81
48      7244        1951     NaN        2         1       96        48
58      8557        1982     NaN        1         0       57        11
71     10629        1973     NaN        1         0       25        25
90      8996        1957     NaN        2         1        4       230
91      9235        1957     NaN        1         1       45         7
92      5798        1973     NaN        0         0       87       445
128     8268        1961     NaN        0         1       23       352
133     1295        1963     NaN        0         1       96       231
312     2437        1989     NaN        0         0       69       861
319     2863        1970     NaN        1         2       67       738
1379   10475        1970     NaN        0         1       39       187
1382    2902        1958     NaN        1         1       87        19
1383    4345        1964     NaN        1         1       49         5
1386    3769        1972     NaN        1         0       17        25
2059    7187        1969     NaN        1         1       52       375
2061    1612        1981     NaN        1         0       82        23
2078    5079        1971     NaN        1         1       82        71
2079   10339        1954     NaN        0         1       83       161
2081    3117        1955     NaN        0         1       95       264
```

```
2084   5250         1943      NaN         0          0        75         532

2228   8720         1978      NaN         0          0        53          32


      MntFruits  MntMeatProducts  MntFishProducts  ...
NumCatalogPurchases  \
10            5                6                0  ...
0
27            1                3                3  ...
0
43           11               50                3  ...
3
48            5               48                6  ...
1
58            3               22                2  ...
0
71            3               43               17  ...
0
90           42              192               49  ...
2
91            0                8                2  ...
0
92           37              359               98  ...
4
128           0               27               10  ...
1
133          65              196               38  ...
5
312         138              461               60  ...
5
319          20              172               52  ...
3
1379          5               65               26  ...
2
1382          4               12                2  ...
0
1383          1                9                2  ...
0
1386          1               13                0  ...
0
2059         42               48               94  ...
10
2061          0               15                0  ...
0
2078          1               16                0  ...
1
2079          0               22                0  ...
1
2081          0               21               12  ...
```

```
1
2084        126              490              164  ...
5
2228          2             1607               12  ...
0

      NumStorePurchases  NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4
\
10                    2                  7             0             0
27                    0                  1             0             0
43                    4                  2             0             0
48                    4                  6             0             0
58                    3                  6             0             0
71                    3                  8             0             0
90                    8                  9             0             0
91                    2                  7             0             0
92                    8                  1             0             0
128                   7                  6             0             0
133                   7                  4             0             0
312                  12                  3             0             1
319                  10                  7             0             1
1379                  6                  5             0             0
1382                  3                  5             0             0
1383                  2                  7             0             0
1386                  3                  7             0             0
2059                  4                  3             0             0
2061                  3                  6             0             0
2078                  3                  8             0             0
2079                  4                  6             0             0
2081                  5                  7             0             0
```

```
2084                  11                   1             0             0

2228                   1                   0             0             1


      AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response
10               0             0             0         0         0
27               0             0             0         0         0
43               0             0             0         0         0
48               0             0             0         0         0
58               0             0             0         0         0
71               0             0             0         0         0
90               0             0             0         0         0
91               0             0             0         0         0
92               0             0             0         0         0
128              0             0             0         0         0
133              0             0             0         0         0
312              0             1             0         0         0
319              0             1             0         0         0
1379             0             0             0         0         0
1382             0             0             0         0         0
1383             0             0             0         0         0
1386             0             0             0         0         0
2059             0             0             0         0         0
2061             0             0             0         0         0
2078             0             0             0         0         0
2079             0             0             0         0         0
2081             0             0             0         0         0
2084             1             0             0         0         1
2228             0             0             0         0         0

[24 rows x 24 columns]

my_numerical_data.describe()

                 ID   Year_Birth         Income      Kidhome
Teenhome  \
count   2240.000000  2240.000000    2216.000000  2240.000000
2240.000000
mean    5592.159821  1968.805804   52247.251354     0.444196
0.506250
std     3246.662198    11.984069   25173.076661     0.538398
0.544538
min        0.000000  1893.000000    1730.000000     0.000000
0.000000
25%     2828.250000  1959.000000   35303.000000     0.000000
0.000000
50%     5458.500000  1970.000000   51381.500000     0.000000
0.000000
```

```
75%     8427.750000  1977.000000   68522.000000     1.000000
1.000000
max    11191.000000  1996.000000  666666.000000     2.000000
2.000000

           Recency     MntWines    MntFruits  MntMeatProducts  \
count  2240.000000  2240.000000  2240.000000      2240.000000
mean     49.109375   303.935714    26.302232       166.950000
std      28.962453   336.597393    39.773434       225.715373
min       0.000000     0.000000     0.000000         0.000000
25%      24.000000    23.750000     1.000000        16.000000
50%      49.000000   173.500000     8.000000        67.000000
75%      74.000000   504.250000    33.000000       232.000000
max      99.000000  1493.000000   199.000000      1725.000000

       MntFishProducts  ...  NumCatalogPurchases  NumStorePurchases  \
count      2240.000000  ...          2240.000000        2240.000000
mean         37.525446  ...             2.662054           5.790179
std          54.628979  ...             2.923101           3.250958
min           0.000000  ...             0.000000           0.000000
25%           3.000000  ...             0.000000           3.000000
50%          12.000000  ...             2.000000           5.000000
75%          50.000000  ...             4.000000           8.000000
max         259.000000  ...            28.000000          13.000000

       NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4  AcceptedCmp5  \
count        2240.000000   2240.000000   2240.000000   2240.000000
mean            5.316518      0.072768      0.074554      0.072768
std             2.426645      0.259813      0.262728      0.259813
min             0.000000      0.000000      0.000000      0.000000
25%             3.000000      0.000000      0.000000      0.000000
50%             6.000000      0.000000      0.000000      0.000000
75%             7.000000      0.000000      0.000000      0.000000
max            20.000000      1.000000      1.000000      1.000000

       AcceptedCmp1  AcceptedCmp2     Complain     Response
count   2240.000000   2240.000000  2240.000000  2240.000000
mean       0.064286      0.013393     0.009375     0.149107
std        0.245316      0.114976     0.096391     0.356274
min        0.000000      0.000000     0.000000     0.000000
25%        0.000000      0.000000     0.000000     0.000000
50%        0.000000      0.000000     0.000000     0.000000
75%        0.000000      0.000000     0.000000     0.000000
max        1.000000      1.000000     1.000000     1.000000

[8 rows x 24 columns]
```

```
cleaner_fit=cleaner.fit(my_numerical_data)
```

```
myCleanedData=cleaner_fit.transform(my_numerical_data)
```

```
myCleanedData = pd.DataFrame(myCleanedData,
columns=my_numerical_data.columns)

myCleanedData.head(11)
```

```
         ID  Year_Birth   Income  Kidhome  Teenhome  Recency  MntWines
\
0   5524.0      1957.0  58138.0      0.0       0.0     58.0     635.0

1   2174.0      1954.0  46344.0      1.0       1.0     38.0      11.0

2   4141.0      1965.0  71613.0      0.0       0.0     26.0     426.0

3   6182.0      1984.0  26646.0      1.0       0.0     26.0      11.0

4   5324.0      1981.0  58293.0      1.0       0.0     94.0     173.0

5   7446.0      1967.0  62513.0      0.0       1.0     16.0     520.0

6    965.0      1971.0  55635.0      0.0       1.0     34.0     235.0

7   6177.0      1985.0  33454.0      1.0       0.0     32.0      76.0

8   4855.0      1974.0  30351.0      1.0       0.0     19.0      14.0

9   5899.0      1950.0   5648.0      1.0       1.0     68.0      28.0

10  1994.0      1983.0   7500.0      1.0       0.0     11.0       5.0


    MntFruits  MntMeatProducts  MntFishProducts  ...
NumCatalogPurchases  \
0        88.0            546.0            172.0  ...
10.0
1         1.0              6.0              2.0  ...
1.0
2        49.0            127.0            111.0  ...
2.0
3         4.0             20.0             10.0  ...
0.0
4        43.0            118.0             46.0  ...
3.0
5        42.0             98.0              0.0  ...
4.0
6        65.0            164.0             50.0  ...
3.0
7        10.0             56.0              3.0  ...
0.0
8         0.0             24.0              3.0  ...
0.0
9         0.0              6.0              1.0  ...
```

```
0.0
10        5.0                6.0                0.0  ...
0.0

    NumStorePurchases  NumWebVisitsMonth  AcceptedCmp3
AcceptedCmp4  \
0                 4.0                7.0           0.0           0.0

1                 2.0                5.0           0.0           0.0

2                10.0                4.0           0.0           0.0

3                 4.0                6.0           0.0           0.0

4                 6.0                5.0           0.0           0.0

5                10.0                6.0           0.0           0.0

6                 7.0                6.0           0.0           0.0

7                 4.0                8.0           0.0           0.0

8                 2.0                9.0           0.0           0.0

9                 0.0               20.0           1.0           0.0

10                2.0                7.0           0.0           0.0


    AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response
0            0.0           0.0           0.0       0.0       1.0
1            0.0           0.0           0.0       0.0       0.0
2            0.0           0.0           0.0       0.0       0.0
3            0.0           0.0           0.0       0.0       0.0
4            0.0           0.0           0.0       0.0       0.0
5            0.0           0.0           0.0       0.0       0.0
6            0.0           0.0           0.0       0.0       0.0
7            0.0           0.0           0.0       0.0       0.0
8            0.0           0.0           0.0       0.0       1.0
9            0.0           0.0           0.0       0.0       0.0
10           0.0           0.0           0.0       0.0       0.0

[11 rows x 24 columns]
```

```
myCleanedData.isna().sum()
```

```
ID                     0
Year_Birth             0
Income                 0
Kidhome                0
Teenhome               0
```

```
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Response               0
dtype: int64
```

```
# myCleanedData[np.isclose(myCleanedData["Income"], 52247.251354)]
#mean
# myCleanedData[np.isclose(myCleanedData["Income"], 51381.500000)]
#meadian
myCleanedData[np.isclose(myCleanedData["Income"], 7500.0)]
#most_frequent
```

```
           ID  Year_Birth  Income  Kidhome  Teenhome  Recency
MntWines  \
10     1994.0      1983.0  7500.0      1.0       0.0     11.0
5.0
11      387.0      1976.0  7500.0      0.0       0.0     59.0
6.0
27     5255.0      1986.0  7500.0      1.0       0.0     19.0
5.0
43     7281.0      1959.0  7500.0      0.0       0.0     80.0
81.0
44     2139.0      1975.0  7500.0      1.0       0.0     19.0
3.0
46     9909.0      1996.0  7500.0      0.0       0.0     24.0
3.0
48     7244.0      1951.0  7500.0      2.0       1.0     96.0
48.0
58     8557.0      1982.0  7500.0      1.0       0.0     57.0
11.0
71    10629.0      1973.0  7500.0      1.0       0.0     25.0
25.0
90     8996.0      1957.0  7500.0      2.0       1.0      4.0
230.0
```

```
91     9235.0      1957.0  7500.0      1.0       1.0     45.0
7.0
92     5798.0      1973.0  7500.0      0.0       0.0     87.0
445.0
128    8268.0      1961.0  7500.0      0.0       1.0     23.0
352.0
133    1295.0      1963.0  7500.0      0.0       1.0     96.0
231.0
238    7297.0      1973.0  7500.0      1.0       0.0     54.0
5.0
312    2437.0      1989.0  7500.0      0.0       0.0     69.0
861.0
319    2863.0      1970.0  7500.0      1.0       2.0     67.0
738.0
439     456.0      1986.0  7500.0      1.0       0.0     96.0
1.0
724    4692.0      1976.0  7500.0      1.0       0.0     19.0
7.0
862    9553.0      1987.0  7500.0      0.0       0.0     94.0
0.0
1152  10710.0      1979.0  7500.0      0.0       1.0     61.0
5.0
1246   4136.0      1992.0  7500.0      1.0       0.0     63.0
10.0
1299  10641.0      1978.0  7500.0      1.0       1.0      5.0
6.0
1379  10475.0      1970.0  7500.0      0.0       1.0     39.0
187.0
1382   2902.0      1958.0  7500.0      1.0       1.0     87.0
19.0
1383   4345.0      1964.0  7500.0      1.0       1.0     49.0
5.0
1386   3769.0      1972.0  7500.0      1.0       0.0     17.0
25.0
2029  10001.0      1985.0  7500.0      1.0       0.0     98.0
5.0
2059   7187.0      1969.0  7500.0      1.0       1.0     52.0
375.0
2061   1612.0      1981.0  7500.0      1.0       0.0     82.0
23.0
2078   5079.0      1971.0  7500.0      1.0       1.0     82.0
71.0
2079  10339.0      1954.0  7500.0      0.0       1.0     83.0
161.0
2081   3117.0      1955.0  7500.0      0.0       1.0     95.0
264.0
2084   5250.0      1943.0  7500.0      0.0       0.0     75.0
532.0
2222  10659.0      1979.0  7500.0      1.0       0.0      7.0
```

```
2.0
2228   8720.0      1978.0  7500.0      0.0       0.0     53.0
32.0

      MntFruits  MntMeatProducts  MntFishProducts  ...
NumCatalogPurchases  \
10          5.0              6.0              0.0  ...
0.0
11         16.0             11.0             11.0  ...
0.0
27          1.0              3.0              3.0  ...
0.0
43         11.0             50.0              3.0  ...
3.0
44          1.0             10.0              3.0  ...
0.0
46         18.0             14.0             15.0  ...
1.0
48          5.0             48.0              6.0  ...
1.0
58          3.0             22.0              2.0  ...
0.0
71          3.0             43.0             17.0  ...
0.0
90         42.0            192.0             49.0  ...
2.0
91          0.0              8.0              2.0  ...
0.0
92         37.0            359.0             98.0  ...
4.0
128         0.0             27.0             10.0  ...
1.0
133        65.0            196.0             38.0  ...
5.0
238         3.0             10.0             12.0  ...
1.0
312       138.0            461.0             60.0  ...
5.0
319        20.0            172.0             52.0  ...
3.0
439        11.0              5.0              4.0  ...
0.0
724         0.0             12.0             13.0  ...
1.0
862         2.0              3.0              4.0  ...
0.0
1152        2.0              3.0              3.0  ...
0.0
1246       17.0             18.0              8.0  ...
```

```
2.0
1299         5.0              4.0             13.0  ...
1.0
1379         5.0             65.0             26.0  ...
2.0
1382         4.0             12.0              2.0  ...
0.0
1383         1.0              9.0              2.0  ...
0.0
1386         1.0             13.0              0.0  ...
0.0
2029        17.0             17.0             13.0  ...
1.0
2059        42.0             48.0             94.0  ...
10.0
2061         0.0             15.0              0.0  ...
0.0
2078         1.0             16.0              0.0  ...
1.0
2079         0.0             22.0              0.0  ...
1.0
2081         0.0             21.0             12.0  ...
1.0
2084       126.0            490.0            164.0  ...
5.0
2222         8.0             11.0              3.0  ...
2.0
2228         2.0           1607.0             12.0  ...
0.0

      NumStorePurchases  NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4
\
10                  2.0                7.0           0.0           0.0

11                  3.0                8.0           0.0           0.0

27                  0.0                1.0           0.0           0.0

43                  4.0                2.0           0.0           0.0

44                  3.0                5.0           0.0           0.0

46                  3.0                9.0           0.0           0.0

48                  4.0                6.0           0.0           0.0

58                  3.0                6.0           0.0           0.0

71                  3.0                8.0           0.0           0.0
```

90	8.0	9.0	0.0	0.0
91	2.0	7.0	0.0	0.0
92	8.0	1.0	0.0	0.0
128	7.0	6.0	0.0	0.0
133	7.0	4.0	0.0	0.0
238	3.0	7.0	0.0	0.0
312	12.0	3.0	0.0	1.0
319	10.0	7.0	0.0	1.0
439	3.0	8.0	0.0	0.0
724	2.0	9.0	1.0	0.0
862	3.0	6.0	0.0	0.0
1152	2.0	8.0	0.0	0.0
1246	2.0	9.0	0.0	0.0
1299	3.0	6.0	0.0	0.0
1379	6.0	5.0	0.0	0.0
1382	3.0	5.0	0.0	0.0
1383	2.0	7.0	0.0	0.0
1386	3.0	7.0	0.0	0.0
2029	3.0	9.0	0.0	0.0
2059	4.0	3.0	0.0	0.0
2061	3.0	6.0	0.0	0.0
2078	3.0	8.0	0.0	0.0
2079	4.0	6.0	0.0	0.0
2081	5.0	7.0	0.0	0.0
2084	11.0	1.0	0.0	0.0
2222	2.0	7.0	0.0	0.0

```
2228                1.0                 0.0           0.0            1.0


      AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response
10             0.0           0.0           0.0       0.0       0.0
11             0.0           0.0           0.0       0.0       0.0
27             0.0           0.0           0.0       0.0       0.0
43             0.0           0.0           0.0       0.0       0.0
44             0.0           0.0           0.0       0.0       0.0
46             0.0           0.0           0.0       0.0       1.0
48             0.0           0.0           0.0       0.0       0.0
58             0.0           0.0           0.0       0.0       0.0
71             0.0           0.0           0.0       0.0       0.0
90             0.0           0.0           0.0       0.0       0.0
91             0.0           0.0           0.0       0.0       0.0
92             0.0           0.0           0.0       0.0       0.0
128            0.0           0.0           0.0       0.0       0.0
133            0.0           0.0           0.0       0.0       0.0
238            0.0           0.0           0.0       0.0       0.0
312            0.0           1.0           0.0       0.0       0.0
319            0.0           1.0           0.0       0.0       0.0
439            0.0           0.0           0.0       0.0       0.0
724            0.0           0.0           0.0       0.0       1.0
862            0.0           0.0           0.0       0.0       0.0
1152           0.0           0.0           0.0       0.0       0.0
1246           0.0           0.0           0.0       0.0       0.0
1299           0.0           0.0           0.0       0.0       0.0
1379           0.0           0.0           0.0       0.0       0.0
1382           0.0           0.0           0.0       0.0       0.0
1383           0.0           0.0           0.0       0.0       0.0
1386           0.0           0.0           0.0       0.0       0.0
2029           0.0           0.0           0.0       0.0       0.0
2059           0.0           0.0           0.0       0.0       0.0
2061           0.0           0.0           0.0       0.0       0.0
2078           0.0           0.0           0.0       0.0       0.0
2079           0.0           0.0           0.0       0.0       0.0
2081           0.0           0.0           0.0       0.0       0.0
2084           1.0           0.0           0.0       0.0       1.0
2222           0.0           0.0           0.0       0.0       0.0
2228           0.0           0.0           0.0       0.0       0.0

[36 rows x 24 columns]
```

Data Encoding

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder

mydataset=pd.read_csv('Intgreted_data_set.csv')
```

```
mydataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 762 entries, 0 to 761
Data columns (total 13 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Unnamed: 0.1                 762 non-null    int64  
 1   Unnamed: 0                   762 non-null    int64  
 2   std_ID                       762 non-null    int64  
 3   Sex                          762 non-null    int64  
 4   race_ethnicity               762 non-null    object 
 5   parental_level_of_education  762 non-null    object 
 6   lunch                        762 non-null    object 
 7   test_preparation_course      762 non-null    object 
 8   math_digree                  762 non-null    int64  
 9   reading_digree               762 non-null    float64
 10  writing_score                762 non-null    int64  
 11  Sumation                     762 non-null    int64  
 12  Avarge                       762 non-null    float64
dtypes: float64(2), int64(7), object(4)
memory usage: 77.5+ KB
```

```
mydataset.describe(include='all')
```

```
        Unnamed: 0.1  Unnamed: 0       std_ID          Sex  \
race_ethnicity  \
count     762.000000  762.000000    762.000000  762.000000   
762
unique           NaN         NaN           NaN         NaN   
5
top              NaN         NaN           NaN         NaN   
Class C
freq             NaN         NaN           NaN         NaN   
224
mean      380.500000  380.500000   3533.076115    1.517060   
NaN
std       220.114743  220.114743   7194.122792    0.500037   
NaN
min         0.000000    0.000000      1.000000    1.000000   
NaN
25%       190.250000  190.250000    297.000000    1.000000   
NaN
50%       380.500000  380.500000    626.000000    2.000000   
NaN
75%       570.750000  570.750000    894.000000    2.000000   
NaN
max       761.000000  761.000000  21003.000000    2.000000   
NaN
```

```
       parental_level_of_education     lunch
test_preparation_course  \
count                            762       762                       762

unique                             6         2                         2

top                     some college  standard                      none

freq                             188       512                       500

mean                             NaN       NaN                       NaN

std                              NaN       NaN                       NaN

min                              NaN       NaN                       NaN

25%                              NaN       NaN                       NaN

50%                              NaN       NaN                       NaN

75%                              NaN       NaN                       NaN

max                              NaN       NaN                       NaN


        math_digree  reading_digree  writing_score    Sumation
Avarge
count    762.000000      762.000000     762.000000  762.000000
762.000000
unique          NaN             NaN            NaN         NaN
NaN
top             NaN             NaN            NaN         NaN
NaN
freq            NaN             NaN            NaN         NaN
NaN
mean      63.094488       69.877037      69.270341  202.241470
67.413955
std       24.140336       13.981493      14.218416   46.047447
15.349258
min        1.000000       31.000000      28.000000   79.000000
26.333333
25%       50.000000       60.000000      59.000000  171.000000
57.000000
50%       68.000000       70.000000      70.000000  209.000000
69.666667
75%       80.000000       80.000000      79.000000  235.000000
78.333333
max      100.000000      100.000000     100.000000  297.000000
99.000000

mydataset.head()
```

```
   Unnamed: 0.1  Unnamed: 0  std_ID   Sex race_ethnicity  \
0             0           0      663    2        Class C
1             1           1      287    1        Class B
2             2           2      626    1        Class B
3             3           3      686    1        Class E
4             4           4      773    1        Class C

  parental_level_of_education         lunch test_preparation_course  \
0                 high school      standard                    none
1           some high school      standard                    none
2         associate's degree  free/reduced               completed
3               some college      standard               completed
4          bachelor's degree  free/reduced                    none

   math_digree  reading_digree  writing_score  Sumation     Avarge
0           63            69.0             67       199  66.333333
1           62            89.0             82       233  77.666667
2           83            70.0             63       216  72.000000
3           85            75.0             68       228  76.000000
4           75            78.0             79       232  77.333333
```

LabelEncoding

Before Encoding

```
mydataset.race_ethnicity.unique()

array(['Class C', 'Class B', 'Class E', 'Class D', 'Class A'],
      dtype=object)

mydataset.parental_level_of_education.unique()

array(['high school', 'some high school', "associate's degree",
       'some college', "bachelor's degree", "master's degree"],
      dtype=object)

mydataset.lunch.unique()

array(['standard', 'free/reduced'], dtype=object)

mydataset.test_preparation_course.unique()

array(['none', 'completed'], dtype=object)
```

Applay Encoding

```
Encoder=LabelEncoder()

mydataset.race_ethnicity=Encoder.fit_transform(mydataset.race_ethnicit
y)
mydataset.parental_level_of_education=Encoder.fit_transform(mydataset.
parental_level_of_education)
```

```
mydataset.lunch=Encoder.fit_transform(mydataset.lunch)
mydataset.test_preparation_course=Encoder.fit_transform(mydataset.test
_preparation_course)
```

After Encoding

```
mydataset.race_ethnicity.unique()

array([2, 1, 4, 3, 0])

mydataset.parental_level_of_education.unique()

array([2, 5, 0, 4, 1, 3])

mydataset.lunch.unique()

array([1, 0])

mydataset.test_preparation_course.unique()

array([1, 0])

mydataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 762 entries, 0 to 761
Data columns (total 13 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Unnamed: 0.1                 762 non-null    int64
 1   Unnamed: 0                   762 non-null    int64
 2   std_ID                       762 non-null    int64
 3   Sex                          762 non-null    int64
 4   race_ethnicity               762 non-null    int32
 5   parental_level_of_education  762 non-null    int32
 6   lunch                        762 non-null    int32
 7   test_preparation_course      762 non-null    int32
 8   math_digree                  762 non-null    int64
 9   reading_digree               762 non-null    float64
 10  writing_score                762 non-null    int64
 11  Sumation                     762 non-null    int64
 12  Avarge                       762 non-null    float64
dtypes: float64(2), int32(4), int64(7)
memory usage: 65.6 KB

mydataset.describe()

       Unnamed: 0.1  Unnamed: 0       std_ID          Sex
race_ethnicity  \
count    762.000000  762.000000    762.000000  762.000000
762.000000
mean     380.500000  380.500000   3533.076115    1.517060
```

```
2.275591
std      220.114743  220.114743   7194.122792    0.500037
1.197965
min        0.000000    0.000000      1.000000    1.000000
0.000000
25%      190.250000  190.250000    297.000000    1.000000
1.000000
50%      380.500000  380.500000    626.000000    2.000000
2.000000
75%      570.750000  570.750000    894.000000    2.000000
3.000000
max      761.000000  761.000000  21003.000000    2.000000
4.000000

       parental_level_of_education       lunch
test_preparation_course  \
count                   762.000000  762.000000
762.000000
mean                      2.438320    0.671916
0.656168
std                       1.783116    0.469824
0.475298
min                       0.000000    0.000000
0.000000
25%                       1.000000    0.000000
0.000000
50%                       2.000000    1.000000
1.000000
75%                       4.000000    1.000000
1.000000
max                       5.000000    1.000000
1.000000

       math_digree  reading_digree  writing_score    Sumation
Avarge
count   762.000000      762.000000     762.000000  762.000000
762.000000
mean     63.094488       69.877037      69.270341  202.241470
67.413955
std      24.140336       13.981493      14.218416   46.047447
15.349258
min       1.000000       31.000000      28.000000   79.000000
26.333333
25%      50.000000       60.000000      59.000000  171.000000
57.000000
50%      68.000000       70.000000      70.000000  209.000000
69.666667
75%      80.000000       80.000000      79.000000  235.000000
78.333333
```

```
max     100.000000      100.000000     100.000000  297.000000
99.000000
```

one Hot Encoder

```
mydataset = pd.get_dummies(mydataset, columns=["race_ethnicity",
"parental_level_of_education", "lunch", "test_preparation_course"])

mydataset.columns

Index(['Unnamed: 0.1', 'Unnamed: 0', 'std_ID ', 'Sex', 'math_digree',
       'reading_digree', 'writing_score', 'Sumation', 'Avarge',
       'race_ethnicity_Class A', 'race_ethnicity_Class B',
       'race_ethnicity_Class C', 'race_ethnicity_Class D',
       'race_ethnicity_Class E',
       'parental_level_of_education_associate's degree',
       'parental_level_of_education_bachelor's degree',
       'parental_level_of_education_high school',
       'parental_level_of_education_master's degree',
       'parental_level_of_education_some college',
       'parental_level_of_education_some high school',
'lunch_free/reduced',
       'lunch_standard', 'test_preparation_course_completed',
       'test_preparation_course_none'],
      dtype='object')

mydataset.head()

   Unnamed: 0.1  Unnamed: 0  std_ID   Sex  math_digree  reading_digree
\
0             0           0      663    2           63            69.0

1             1           1      287    1           62            89.0

2             2           2      626    1           83            70.0

3             3           3      686    1           85            75.0

4             4           4      773    1           75            78.0


   writing_score  Sumation     Avarge  race_ethnicity_Class A  ...  \
0             67       199  66.333333                   False  ...
1             82       233  77.666667                   False  ...
2             63       216  72.000000                   False  ...
3             68       228  76.000000                   False  ...
4             79       232  77.333333                   False  ...

   parental_level_of_education_associate's degree  \
0                                           False
1                                           False
```

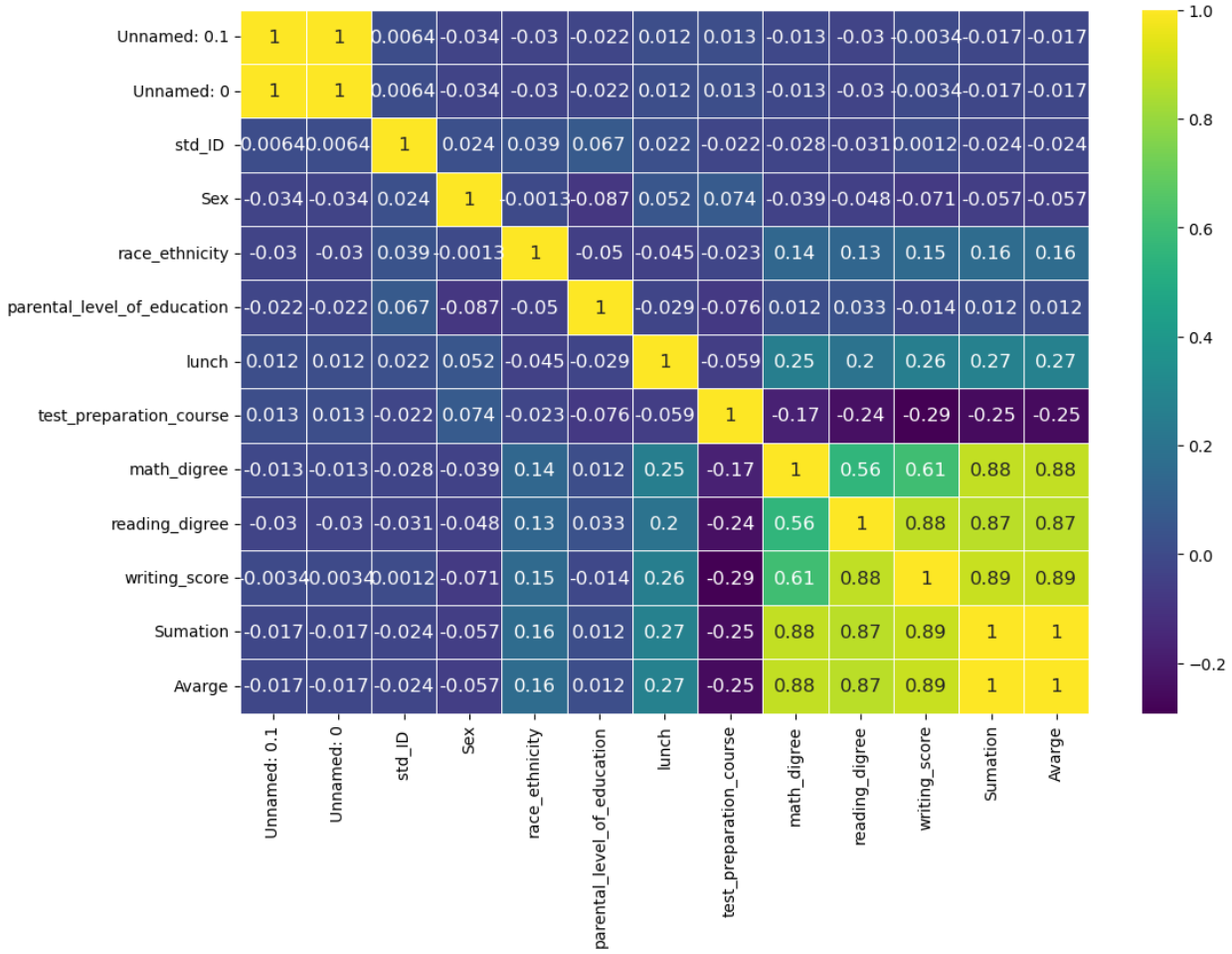
```
2                                                 True
3                                                False
4                                                False

   parental_level_of_education_bachelor's degree  \
0                                           False
1                                           False
2                                           False
3                                           False
4                                            True

   parental_level_of_education_high school  \
0                                     True
1                                    False
2                                    False
3                                    False
4                                    False

   parental_level_of_education_master's degree  \
0                                        False
1                                        False
2                                        False
3                                        False
4                                        False

   parental_level_of_education_some college  \
0                                     False
1                                     False
2                                     False
3                                      True
4                                     False

   parental_level_of_education_some high school  lunch_free/reduced  \
0                                         False               False
1                                          True               False
2                                         False                True
3                                         False               False
4                                         False                True

   lunch_standard  test_preparation_course_completed  \
0            True                              False
1            True                              False
2           False                               True
3            True                               True
4           False                              False

   test_preparation_course_none
0                          True
1                          True
2                         False
```

```
3                          False
4                           True

[5 rows x 24 columns]
```

Data correlation

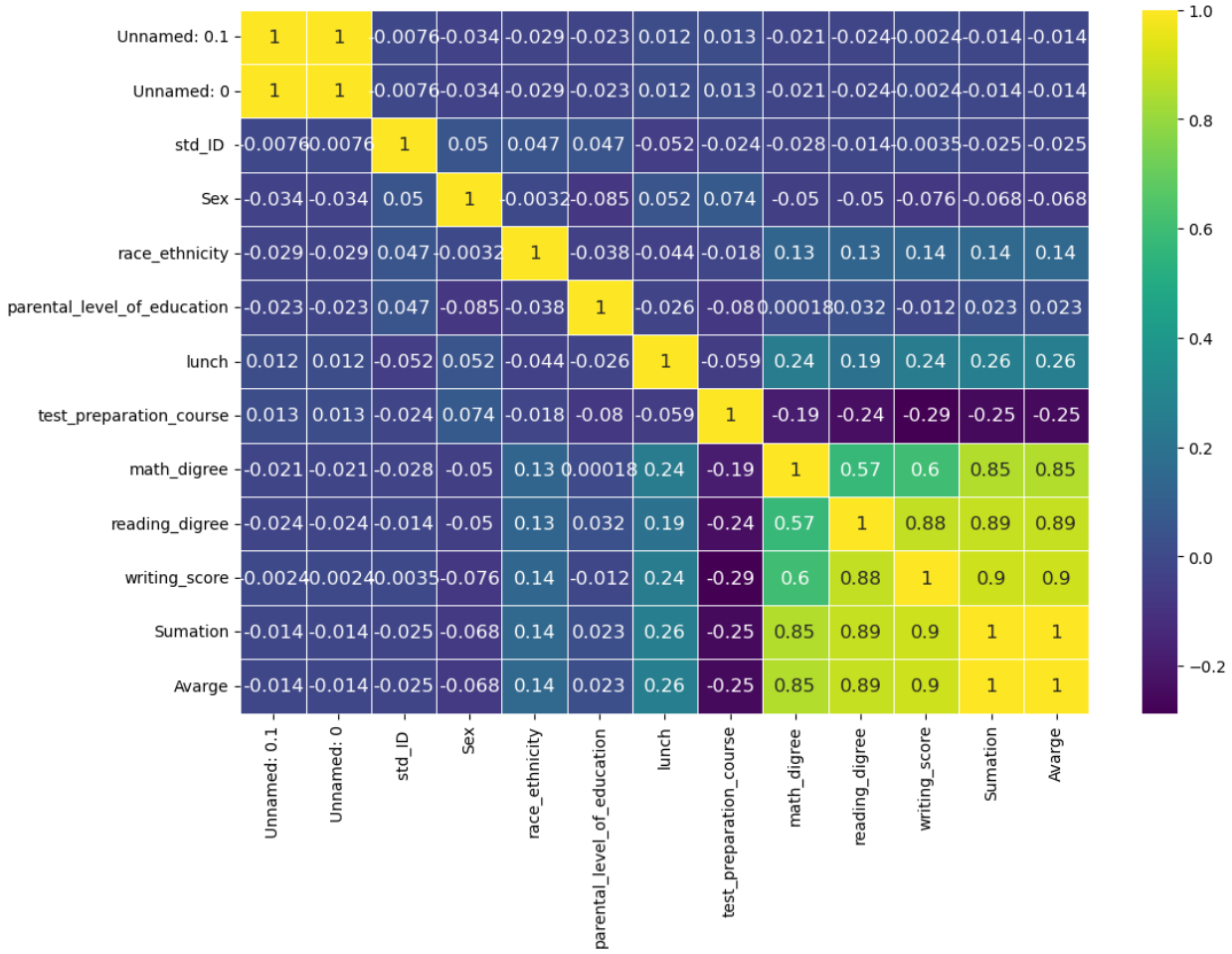
Pearson

```
import seaborn as sns
import matplotlib.pyplot as plt

correlation_matrix =mydataset.corr()

plt.figure(figsize=(12, 8))  # ضبط حجم الشكل (العرض x الارتفاع)
sns.heatmap(
    correlation_matrix,
    annot=True,
    cmap='viridis',
    annot_kws={"size": 12},  # تكبير حجم الأرقام
    linewidths=0.5,          # رسم خطوط فاصلة بين المربعات
    linecolor='white'        # لون الخطوط
)
plt.show()
```



spearman

```
correlation_matrix =mydataset.corr(method='spearman')

plt.figure(figsize=(12, 8))  # ضبط حجم الشكل (العرض x الارتفاع)
sns.heatmap(
    correlation_matrix,
    annot=True,
    cmap='viridis',
    annot_kws={"size": 12},  # تكبير حجم الأرقام
    linewidths=0.5,          # رسم خطوط فاصلة بين المربعات
    linecolor='white'        # لون الخطوط
)
plt.show()
```


Unnamed: 0.1 1 1 -0.0076 -0.034 -0.029 -0.023 0.012 0.013 -0.021 -0.024 0.0024 -0.014 -0.014
Unnamed: 0 1 1 -0.0076 -0.034 -0.029 -0.023 0.012 0.013 -0.021 -0.024 0.0024 -0.014 -0.014
std_ID -0.0076 -0.0076 1 0.05 0.047 0.047 -0.052 -0.024 -0.028 -0.014 -0.0035 -0.025 -0.025
Sex -0.034 -0.034 0.05 1 -0.0032 -0.085 0.052 0.074 -0.05 -0.05 -0.076 -0.068 -0.068
race_ethnicity -0.029 -0.029 0.047 -0.0032 1 -0.038 -0.044 -0.018 0.13 0.13 0.14 0.14 0.14
parental_level_of_education -0.023 -0.023 0.047 -0.085 -0.038 1 -0.026 -0.08 0.00018 0.032 -0.012 0.023 0.023
lunch 0.012 0.012 -0.052 0.052 -0.044 -0.026 1 -0.059 0.24 0.19 0.24 0.26 0.26
test_preparation_course 0.013 0.013 -0.024 0.074 -0.018 -0.08 -0.059 1 -0.19 -0.24 -0.29 -0.25 -0.25
math_digree -0.021 -0.021 -0.028 -0.05 0.13 0.00018 0.24 -0.19 1 0.57 0.6 0.85 0.85
reading_digree -0.024 -0.024 -0.014 -0.05 0.13 0.032 0.19 -0.24 0.57 1 0.88 0.89 0.89
writing_score -0.0024 -0.0024 -0.0035 -0.076 0.14 -0.012 0.24 -0.29 0.6 0.88 1 0.9 0.9
Sumation -0.014 -0.014 -0.025 -0.068 0.14 0.023 0.26 -0.25 0.85 0.89 0.9 1 1
Avarge -0.014 -0.014 -0.025 -0.068 0.14 0.023 0.26 -0.25 0.85 0.89 0.9 1 1
Unnamed: 0.1
Unnamed: 0
std_ID
Sex
race_ethnicity
parental_level_of_education
lunch
test_preparation_course
math_digree
reading_digree
writing_score
Sumation
Avarge
1.0
0.8
0.6
0.4
0.2
0.0
−0.2