Aim:- To understand and apply data transformation and data discretization techniques on a given dataset using Python.

Software Used: IDLE

Theory :

Dataset (dataset.csv):

```
Age,Income,Education_Level
23,50000,Bachelor
45,65000,Master
25,48000,Bachelor
34,52000,PhD
65,70000,PhD
42,62000,Master
21,45000,Bachelor
35,51000,PhD
32,59000,Master
40,60000,Bachelor
23,52000,Master
52,68000,PhD
30,58000,Master
45,62000,Bachelor
54,71000,PhD
24,49000,Bachelor
28,53000,Master
36,54000,PhD
```

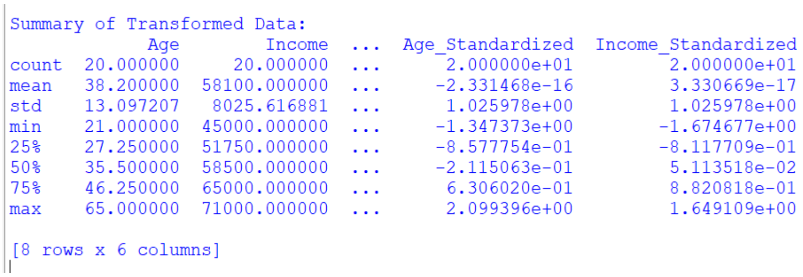
```
50,65000,Bachelor
60,68000,Master
```

Code:

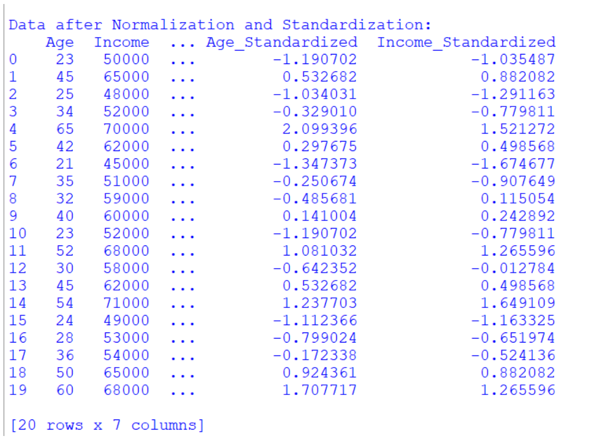
```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler, StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
df = pd.read_csv('dataset.csv')

# Display the original data
print("Original Data:")
print(df)

# Step 3: Data Transformation
# Normalization
scaler = MinMaxScaler()
df[['Age_Normalized', 'Income_Normalized']] = scaler.fit_transform(df[['Age', 'Income']])

# Standardization
scaler = StandardScaler()
df[['Age_Standardized', 'Income_Standardized']] = scaler.fit_transform(df[['Age', 'Income']])

print("\nData after Normalization and Standardization:")
print(df)

# Step 4: Data Discretization
df['Age_Bins'] = pd.cut(df['Age'], bins=[20, 30, 40, 50, 60, 70], labels=['20-30', '30-40', '40-50', '50-60', '60-70'])
df['Income_Bins'] = pd.cut(df['Income'], bins=4, labels=['Low', 'Medium', 'High', 'Very High'])

print("\nData after Discretization:")
print(df)

# Step 5: Summary and Visualization
print("\nSummary of Transformed Data:")
print(df.describe())

# Visualization
plt.figure(figsize=(12, 6))
```

```
# Age distribution
plt.subplot(1, 2, 1)
sns.histplot(df['Age'], bins=5, kde=True)
plt.title('Age Distribution')

# Income distribution
plt.subplot(1, 2, 2)
sns.histplot(df['Income'], bins=4, kde=True)
plt.title('Income Distribution')

plt.tight_layout()
plt.show()

plt.figure(figsize=(12, 6))

# Age Bins
plt.subplot(1, 2, 1)
sns.countplot(x='Age_Bins', data=df)
plt.title('Age Bins')

# Income Bins
plt.subplot(1, 2, 2)
sns.countplot(x='Income_Bins', data=df)
plt.title('Income Bins')

plt.tight_layout()
plt.show()
```

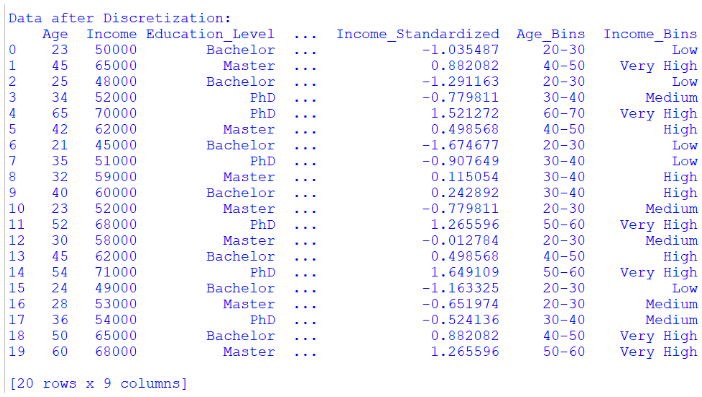
Output-



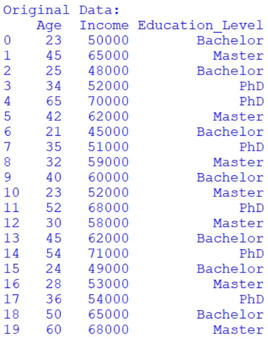
```
Summary of Transformed Data:
             Age        Income  ...  Age_Standardized  Income_Standardized
count  20.000000     20.000000  ...      2.000000e+01         2.000000e+01
mean   38.200000  58100.000000  ...     -2.331468e-16         3.330669e-17
std    13.097207   8025.616881  ...      1.025978e+00         1.025978e+00
min    21.000000  45000.000000  ...     -1.347373e+00        -1.674677e+00
25%    27.250000  51750.000000  ...     -8.577754e-01        -8.117709e-01
50%    35.500000  58500.000000  ...     -2.115063e-01         5.113518e-02
75%    46.250000  65000.000000  ...      6.306020e-01         8.820818e-01
max    65.000000  71000.000000  ...      2.099396e+00         1.649109e+00

[8 rows x 6 columns]
```



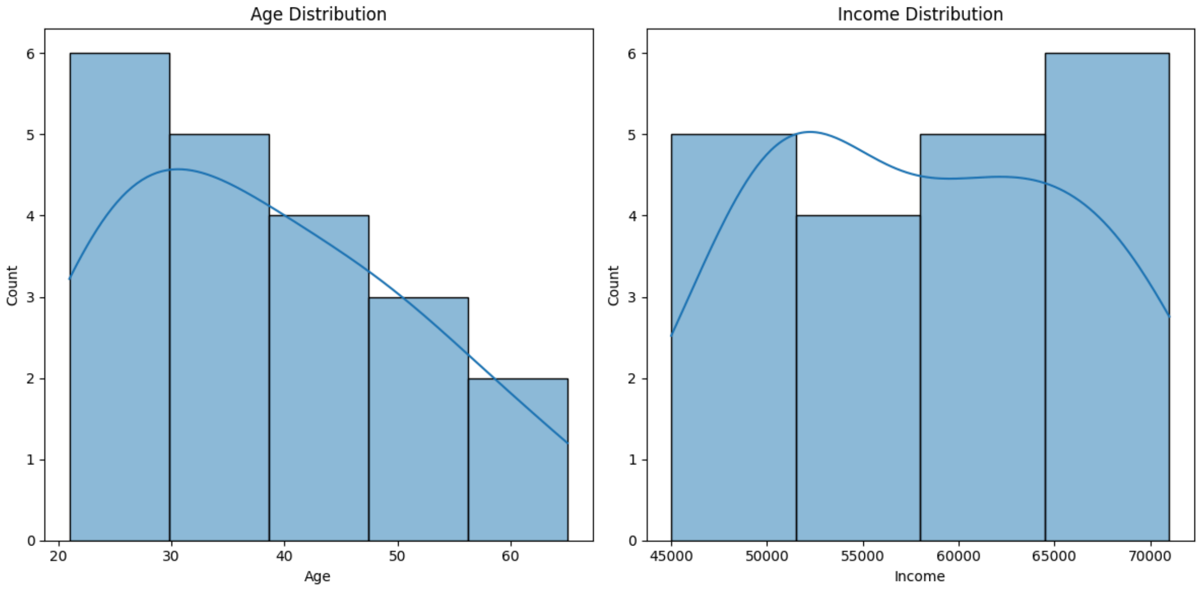
```
Data after Normalization and Standardization:
    Age  Income  ... Age_Standardized  Income_Standardized
0    23   50000  ...        -1.190702            -1.035487
1    45   65000  ...         0.532682             0.882082
2    25   48000  ...        -1.034031            -1.291163
3    34   52000  ...        -0.329010            -0.779811
4    65   70000  ...         2.099396             1.521272
5    42   62000  ...         0.297675             0.498568
6    21   45000  ...        -1.347373            -1.674677
7    35   51000  ...        -0.250674            -0.907649
8    32   59000  ...        -0.485681             0.115054
9    40   60000  ...         0.141004             0.242892
10   23   52000  ...        -1.190702            -0.779811
11   52   68000  ...         1.081032             1.265596
12   30   58000  ...        -0.642352            -0.012784
13   45   62000  ...         0.532682             0.498568
14   54   71000  ...         1.237703             1.649109
15   24   49000  ...        -1.112366            -1.163325
16   28   53000  ...        -0.799024            -0.651974
17   36   54000  ...        -0.172338            -0.524136
18   50   65000  ...         0.924361             0.882082
19   60   68000  ...         1.707717             1.265596

[20 rows x 7 columns]
```



```
Data after Discretization:
    Age  Income Education_Level  ...  Income_Standardized  Age_Bins  Income_Bins
0    23   50000        Bachelor  ...            -1.035487     20-30          Low
1    45   65000          Master  ...             0.882082     40-50    Very High
2    25   48000        Bachelor  ...            -1.291163     20-30          Low
3    34   52000             PhD  ...            -0.779811     30-40       Medium
4    65   70000             PhD  ...             1.521272     60-70    Very High
5    42   62000          Master  ...             0.498568     40-50         High
6    21   45000        Bachelor  ...            -1.674677     20-30          Low
7    35   51000             PhD  ...            -0.907649     30-40          Low
8    32   59000          Master  ...             0.115054     30-40         High
9    40   60000        Bachelor  ...             0.242892     30-40         High
10   23   52000          Master  ...            -0.779811     20-30       Medium
11   52   68000             PhD  ...             1.265596     50-60    Very High
12   30   58000          Master  ...            -0.012784     20-30       Medium
13   45   62000        Bachelor  ...             0.498568     40-50         High
14   54   71000             PhD  ...             1.649109     50-60    Very High
15   24   49000        Bachelor  ...            -1.163325     20-30          Low
16   28   53000          Master  ...            -0.651974     20-30       Medium
17   36   54000             PhD  ...            -0.524136     30-40       Medium
18   50   65000        Bachelor  ...             0.882082     40-50    Very High
19   60   68000          Master  ...             1.265596     50-60    Very High

[20 rows x 9 columns]
```



```
Original Data:
    Age  Income Education_Level
0    23   50000        Bachelor
1    45   65000          Master
2    25   48000        Bachelor
3    34   52000             PhD
4    65   70000             PhD
5    42   62000          Master
6    21   45000        Bachelor
7    35   51000             PhD
8    32   59000          Master
9    40   60000        Bachelor
10   23   52000          Master
11   52   68000             PhD
12   30   58000          Master
13   45   62000        Bachelor
14   54   71000             PhD
15   24   49000        Bachelor
16   28   53000          Master
17   36   54000             PhD
18   50   65000        Bachelor
19   60   68000          Master
```


Age Distribution
Count
Age
Income Distribution
Count
Income