Serenay_Goler_week1

August 9, 2025

[138]:

```
# Importing Libraries

import numpy as np  # Linear algebra operations
import pandas as pd  # Data processing and analysis
```

[140]:

```
# Upload dataset

df = pd.read_csv('/Users/serenaygoler/heart disease.csv')

df.head() # Displays the first 5 rows.
```

[140]:

```
   Age Sex ChestPainType  RestingBP  Cholesterol  FastingBS RestingECG  MaxHR  \
0   40   M           ATA        140          289          0     Normal    172
1   49   F           NAP        160          180          0     Normal    156
2   37   M           ATA        130          283          0         ST     98
3   48   F           ASY        138          214          0     Normal    108
4   54   M           NAP        150          195          0     Normal    122

  ExerciseAngina  Oldpeak ST_Slope  HeartDisease
0              N      0.0       Up             0
1              N      1.0     Flat             1
2              N      0.0       Up             0
3              Y      1.5     Flat             1
4              N      0.0       Up             0
```

[142]:

```
df.info() # Shows data types and counts of non-missing values.
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

[144]:
```
df.shape # Displays the number of rows and columns in the dataset.
```

[144]:
```
(918, 12)
```

[146]:
```
df.isna().sum() # Counts missing values in each column.
```

[146]:
```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

[148]:
```
df.nunique() # Shows the number of unique values per column.
```

[148]:
```
Age                50
Sex                 2
ChestPainType       4
RestingBP          67
Cholesterol       222
FastingBS           2
RestingECG          3
MaxHR             119
ExerciseAngina      2
Oldpeak            53
ST_Slope            3
HeartDisease        2
dtype: int64
```

[150]:
```
df.duplicated().sum() # Counts the number of duplicate rows.
```

[150]:
```
0
```

[199]:

```
# Provides summary statistics for numeric columns, rounded to 2 decimals and␣
  ↪transposedfor readability.
df.describe().round(2).T
```

[199]:

```
              count    mean     std   min     25%    50%    75%    max
Age           918.0   53.51    9.43  28.0   47.00   54.0   60.0   77.0
RestingBP     918.0  132.40   18.51   0.0  120.00  130.0  140.0  200.0
Cholesterol   918.0  198.80  109.38   0.0  173.25  223.0  267.0  603.0
FastingBS     918.0    0.23    0.42   0.0    0.00    0.0    0.0    1.0
MaxHR         918.0  136.81   25.46  60.0  120.00  138.0  156.0  202.0
Oldpeak       918.0    0.89    1.07  -2.6    0.00    0.6    1.5    6.2
HeartDisease  918.0    0.55    0.50   0.0    0.00    1.0    1.0    1.0
```

[154]:

```
df['HeartDisease'].value_counts() # Shows the count of each class in the target␣
  ↪variable
```

[154]:

```
HeartDisease
1    508
0    410
Name: count, dtype: int64
```

[157]:

```
df['FastingBS'].value_counts()
```

[157]:

```
FastingBS
0    704
1    214
Name: count, dtype: int64
```

[159]:

```
df['Sex'].value_counts()
```

[159]:

```
Sex
M    725
F    193
Name: count, dtype: int64
```

[57]:

```
df['ChestPainType'].value_counts()
```

[57]:

```
ChestPainType
ASY    496
NAP    203
ATA    173
TA      46
Name: count, dtype: int64
```

[61]:

```
df['RestingECG'].value_counts()
```

[61]:

```
RestingECG
Normal    552
```

```
LVH       188
ST        178
Name: count, dtype: int64
```

[63]:

```
df['ExerciseAngina'].value_counts()
```

[63]:

```
ExerciseAngina
N    547
Y    371
Name: count, dtype: int64
```

[71]:

```
df['ST_Slope'].value_counts()
```

[71]:

```
ST_Slope
Flat    460
Up      395
Down     63
Name: count, dtype: int64
```

[201]:

```
perc_dis =df['HeartDisease'].sum()/ len(df)
print('Percentage of patients with heart disease in the dataset:',␣
 ↪round(perc_dis, 4))
```

```
Percentage of patients with heart disease in the dataset: 0.5534
```

[209]:

```
num_df = df.copy()
num_df['Oldpeak'] = num_df['Oldpeak'].abs()

num_df.describe().round(2).T
```

[209]:

```
              count    mean     std   min     25%    50%    75%    max
Age           918.0   53.51    9.43  28.0   47.00   54.0   60.0   77.0
RestingBP     918.0  132.40   18.51   0.0  120.00  130.0  140.0  200.0
Cholesterol   918.0  198.80  109.38   0.0  173.25  223.0  267.0  603.0
FastingBS     918.0    0.23    0.42   0.0    0.00    0.0    0.0    1.0
MaxHR         918.0  136.81   25.46  60.0  120.00  138.0  156.0  202.0
Oldpeak       918.0    0.92    1.04   0.0    0.00    0.6    1.5    6.2
HeartDisease  918.0    0.55    0.50   0.0    0.00    1.0    1.0    1.0
```

[213]:

```
num_df = df.assign(
    Sex = df['Sex'].map({'F': 0, 'M': 1}),
    ExerciseAngina = df['ExerciseAngina'].map({'N': 0, 'Y': 1}),
    ChestPainType = df['ChestPainType'].map({'ASY':0, 'NAP':1, 'ATA':2, 'TA':
 ↪3}),
    RestingECG = df['RestingECG'].map({'Normal':0, 'LVH':1, 'ST':2,}),
    ST_Slope = df['ST_Slope'].map({'Flat':0, 'Up':1, 'Down':2}))
```

[175]:

```
num_df.head()
```

```
[175]:    Age  Sex  ChestPainType  RestingBP  Cholesterol  FastingBS  RestingECG  \
       0   40    1              2        140          289          0           0
       1   49    0              1        160          180          0           0
       2   37    1              2        130          283          0           2
       3   48    0              0        138          214          0           0
       4   54    1              1        150          195          0           0

          MaxHR  ExerciseAngina  Oldpeak  ST_Slope  HeartDisease
       0    172               0      0.0         1             0
       1    156               0      1.0         0             1
       2     98               0      0.0         1             0
       3    108               1      1.5         0             1
       4    122               0      0.0         1             0
```

```
[ ]: num_df = df.copy()
     num_df['Oldpeak'] = num_df['Oldpeak'].abs()
```

```
[215]: num_df.corr()
```

```
[215]:                      Age       Sex  ChestPainType  RestingBP  Cholesterol  \
       Age             1.000000  0.055750      -0.165896   0.254399    -0.095282
       Sex             0.055750  1.000000      -0.168254   0.005133    -0.200092
       ChestPainType  -0.165896 -0.168254       1.000000  -0.022168     0.136139
       RestingBP       0.254399  0.005133      -0.022168   1.000000     0.100893
       Cholesterol    -0.095282 -0.200092       0.136139   0.100893     1.000000
       FastingBS       0.198039  0.120076      -0.116703   0.070193    -0.260974
       RestingECG      0.210498  0.038320      -0.065099   0.117206    -0.042595
       MaxHR          -0.382045 -0.189186       0.343654  -0.112135     0.235792
       ExerciseAngina  0.215793  0.190664      -0.416625   0.155101    -0.034166
       Oldpeak         0.258612  0.105734      -0.245027   0.164803     0.050148
       ST_Slope       -0.093424 -0.066831       0.202675  -0.083418     0.007110
       HeartDisease    0.282039  0.305445      -0.471354   0.107589    -0.232741

                       FastingBS  RestingECG     MaxHR  ExerciseAngina   Oldpeak  \
       Age              0.198039    0.210498 -0.382045        0.215793  0.258612
       Sex              0.120076    0.038320 -0.189186        0.190664  0.105734
       ChestPainType   -0.116703   -0.065099  0.343654       -0.416625 -0.245027
       RestingBP        0.070193    0.117206 -0.112135        0.155101  0.164803
       Cholesterol     -0.260974   -0.042595  0.235792       -0.034166  0.050148
       FastingBS        1.000000    0.120774 -0.131438        0.060451  0.052698
       RestingECG       0.120774    1.000000 -0.093379        0.098360  0.099935
       MaxHR           -0.131438   -0.093379  1.000000       -0.370425 -0.160691
       ExerciseAngina   0.060451    0.098360 -0.370425        1.000000  0.408752
       Oldpeak          0.052698    0.099935 -0.160691        0.408752  1.000000
       ST_Slope        -0.043534   -0.019403  0.246927       -0.253181 -0.097323
       HeartDisease     0.267291    0.107628 -0.400421        0.494282  0.403951
```

```
                ST_Slope  HeartDisease
Age            -0.093424      0.282039
Sex            -0.066831      0.305445
ChestPainType   0.202675     -0.471354
RestingBP      -0.083418      0.107589
Cholesterol     0.007110     -0.232741
FastingBS      -0.043534      0.267291
RestingECG     -0.019403      0.107628
MaxHR           0.246927     -0.400421
ExerciseAngina -0.253181      0.494282
Oldpeak        -0.097323      0.403951
ST_Slope        1.000000     -0.397802
HeartDisease   -0.397802      1.000000
```