In [1]:
```
import pandas as pd
import numpy as np
import matplotlib as mlt
```

In [2]:
```
df = pd.read_csv(r"E:\unified mentor projects\Heart Disease Diagnostic Analysis\
df
```

Out[2]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	52	1	0	125	212	0	1	168	0	1.0	2	2
1	53	1	0	140	203	1	0	155	1	3.1	0	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1
4	62	0	0	138	294	1	1	106	0	1.9	1	3
...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0
1021	60	1	0	125	258	0	0	141	1	2.8	1	1
1022	47	1	0	110	275	0	0	118	1	1.0	1	1
1023	50	0	0	110	254	0	0	159	0	0.0	2	0
1024	54	1	0	120	188	0	1	113	0	1.4	1	1

1025 rows × 14 columns

In [3]:
```
# Print the column names to identify the exact names
print(df.columns)
```

```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

In [4]:
```
df.shape
```

Out[4]:
```
(1025, 14)
```

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

In [6]:
```
df.duplicated().sum()
```

Out[6]:
```
723
```

In [7]:
```
df = df.drop_duplicates()
df.head()
```

Out[7]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2

In [8]:
```
column = 'oldpeak'
q1 = df[column].quantile(0.25)
q3 = df[column].quantile(0.75)
iqr = q3 - q1

lower_bound = q1 -1.5 * iqr
upper_bound = q3 + 1.5 * iqr

df = df[(df[column]>= lower_bound) & (df[column] <= upper_bound)]
df.head()
```

Out[8]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2

In [9]:
```
df['Patient ID'] = df.index+1
```

In [10]:
```
df['Cholesterol Score'] = pd.cut(df['chol'],bins=[0,199,239,float('inf')], label

df['Exercise Capcity Score'] = df['thalach'] - df['age']

df['Heart Disease Prevalence'] = df['target'].map({0: 'No Heart Disease', 1: 'He

df_cleaned = df
```

In [11]:
```
df_cleaned
```

Out[11]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	tl
0	52	1	0	125	212	0	1	168	0	1.0	2	2	
1	53	1	0	140	203	1	0	155	1	3.1	0	0	
2	70	1	0	145	174	0	1	125	1	2.6	0	0	
3	61	1	0	148	203	0	1	161	0	0.0	2	1	
4	62	0	0	138	294	1	1	106	0	1.9	1	3	
...	...	...	...	...	...	...	...	...	...	...	...	...	
723	68	0	2	120	211	0	0	115	0	1.5	1	0	
733	44	0	2	108	141	0	1	175	0	0.6	1	0	
739	52	1	0	128	255	0	1	161	1	0.0	2	1	
843	59	1	3	160	273	0	0	125	0	0.0	2	0	
878	54	1	0	120	188	0	1	113	0	1.4	1	1	

297 rows × 18 columns

In [12]:
```
df['sex'].replace(0,'Female', inplace = True)
df['sex'].replace(1,'Male', inplace = True)
```

In [13]:
```
df_cleaned.describe()
```

Out[13]:

	age	cp	trestbps	chol	fbs	restecg	thala
count	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.0000
mean	54.377104	0.969697	131.353535	246.383838	0.151515	0.521886	149.8552
std	9.104826	1.027867	17.381051	51.668389	0.359155	0.520225	22.9195
min	29.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.0000
25%	47.000000	0.000000	120.000000	211.000000	0.000000	0.000000	134.0000
50%	55.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.0000
75%	61.000000	2.000000	140.000000	274.000000	0.000000	1.000000	166.0000
max	77.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.0000

In [14]:
```
df_cleaned.head()
```

Out[14]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	52	Male	0	125	212	0	1	168	0	1.0	2	2
1	53	Male	0	140	203	1	0	155	1	3.1	0	0
2	70	Male	0	145	174	0	1	125	1	2.6	0	0
3	61	Male	0	148	203	0	1	161	0	0.0	2	1
4	62	Female	0	138	294	1	1	106	0	1.9	1	3

In [15]:
```
df_cleaned.columns
```

Out[15]:
```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target', 'Patient ID',
       'Cholesterol Score', 'Exercise Capcity Score',
       'Heart Disease Prevalence'],
      dtype='object')
```

In [16]:
```
df_cleaned.to_csv('HeartDiseaseData_Cleaned.csv', index=False)
df_cleaned
```

Out[16]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	52	Male	0	125	212	0	1	168	0	1.0	2	2
1	53	Male	0	140	203	1	0	155	1	3.1	0	0
2	70	Male	0	145	174	0	1	125	1	2.6	0	0
3	61	Male	0	148	203	0	1	161	0	0.0	2	1
4	62	Female	0	138	294	1	1	106	0	1.9	1	3
...	...	...	...	...	...	...	...	...	...	...	...	...
723	68	Female	2	120	211	0	0	115	0	1.5	1	0
733	44	Female	2	108	141	0	1	175	0	0.6	1	0
739	52	Male	0	128	255	0	1	161	1	0.0	2	1
843	59	Male	3	160	273	0	0	125	0	0.0	2	0
878	54	Male	0	120	188	0	1	113	0	1.4	1	1

297 rows × 18 columns

In [17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 297 entries, 0 to 878
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       297 non-null    int64
 1   sex                       297 non-null    object
 2   cp                        297 non-null    int64
 3   trestbps                  297 non-null    int64
 4   chol                      297 non-null    int64
 5   fbs                       297 non-null    int64
 6   restecg                   297 non-null    int64
 7   thalach                   297 non-null    int64
 8   exang                     297 non-null    int64
 9   oldpeak                   297 non-null    float64
 10  slope                     297 non-null    int64
 11  ca                        297 non-null    int64
 12  thal                      297 non-null    int64
 13  target                    297 non-null    int64
 14  Patient ID                297 non-null    int64
 15  Cholesterol Score         297 non-null    category
 16  Exercise Capcity Score    297 non-null    int64
 17  Heart Disease Prevalence  297 non-null    object
dtypes: category(1), float64(1), int64(14), object(2)
memory usage: 42.2+ KB
```

In [ ]: