In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv('Heart2.csv')
df
```

Out[2]:

	Unnamed: 0	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slo
0	1	63	1	typical	145	233	1	2	150	0	2.3	
1	2	67	1	asymptomatic	160	286	0	2	108	1	1.5	
2	3	67	1	asymptomatic	120	229	0	2	129	1	2.6	
3	4	37	1	nonanginal	130	250	0	0	187	0	3.5	
4	5	41	0	nontypical	130	204	0	2	172	0	1.4	
...	...	...	...	...	...	...	...	...	...	...	...	
298	299	45	1	typical	110	264	0	0	132	0	1.2	
299	300	68	1	asymptomatic	144	193	1	0	141	0	3.4	
300	301	57	1	asymptomatic	130	131	0	0	115	1	1.2	
301	302	57	0	nontypical	130	236	0	2	174	0	0.0	
302	303	38	1	nonanginal	138	175	0	0	173	0	0.0	

303 rows × 15 columns

Data Cleaning

In [3]:
```
df.head()
```

Out[3]:

	Unnamed: 0	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope
0	1	63	1	typical	145	233	1	2	150	0	2.3	3
1	2	67	1	asymptomatic	160	286	0	2	108	1	1.5	2
2	3	67	1	asymptomatic	120	229	0	2	129	1	2.6	2
3	4	37	1	nonanginal	130	250	0	0	187	0	3.5	3
4	5	41	0	nontypical	130	204	0	2	172	0	1.4	1

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  303 non-null    int64  
 1   Age         303 non-null    int64  
 2   Sex         303 non-null    int64  
 3   ChestPain   303 non-null    object 
 4   RestBP      303 non-null    int64  
 5   Chol        303 non-null    int64  
 6   Fbs         303 non-null    int64  
 7   RestECG     303 non-null    int64  
 8   MaxHR       303 non-null    int64  
 9   ExAng       303 non-null    int64  
 10  Oldpeak     303 non-null    float64
 11  Slope       303 non-null    int64  
 12  Ca          299 non-null    float64
 13  Thal        301 non-null    object 
 14  AHD         303 non-null    object 
dtypes: float64(2), int64(10), object(3)
memory usage: 35.6+ KB
```

In [5]:
```
df.drop(['Ca','Thal'],axis = 1,inplace = True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  303 non-null    int64  
 1   Age         303 non-null    int64  
 2   Sex         303 non-null    int64  
 3   ChestPain   303 non-null    object 
 4   RestBP      303 non-null    int64  
 5   Chol        303 non-null    int64  
 6   Fbs         303 non-null    int64  
 7   RestECG     303 non-null    int64  
 8   MaxHR       303 non-null    int64  
 9   ExAng       303 non-null    int64  
 10  Oldpeak     303 non-null    float64
 11  Slope       303 non-null    int64  
 12  AHD         303 non-null    object 
dtypes: float64(1), int64(10), object(2)
memory usage: 30.9+ KB
```

Data Integration

In [6]:
```
sns.set_theme(style="whitegrid")
df.shape
```

Out[6]:
```
(303, 13)
```

In [7]:
```
Q1 = df.quantile(0.25) #first 25% of the data
Q3 = df.quantile(0.75) #first 75% of the data
IQR = Q3 - Q1 #IQR = InterQuartile Range
scale = 2 #For Normal Distributions, scale = 1.5
lower_lim = Q1 - scale*IQR
```

```
upper_lim = Q3 + scale*IQR
lower_outliers = (df[df.columns[2:13]] < lower_lim)
upper_outliers = (df[df.columns[2:13]] > upper_lim)
```

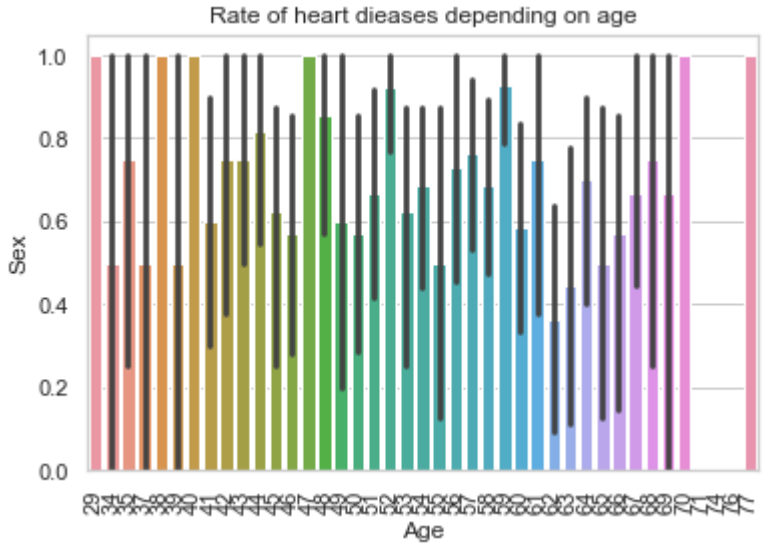
```
C:\Users\Sayuja\AppData\Local\Temp\ipykernel_11816\3886136910.py:7: FutureWarning: Au
tomatic reindexing on DataFrame vs Series comparisons is deprecated and will raise Va
lueError in a future version. Do `left, right = left.align(right, axis=1, copy=False)
` before e.g. `left == right`
  lower_outliers = (df[df.columns[2:13]] < lower_lim)
C:\Users\Sayuja\AppData\Local\Temp\ipykernel_11816\3886136910.py:8: FutureWarning: Au
tomatic reindexing on DataFrame vs Series comparisons is deprecated and will raise Va
lueError in a future version. Do `left, right = left.align(right, axis=1, copy=False)
` before e.g. `left == right`
  upper_outliers = (df[df.columns[2:13]] > upper_lim)
```

In [8]:
```
df[df.columns[2:13]][(lower_outliers | upper_outliers)].info()
```

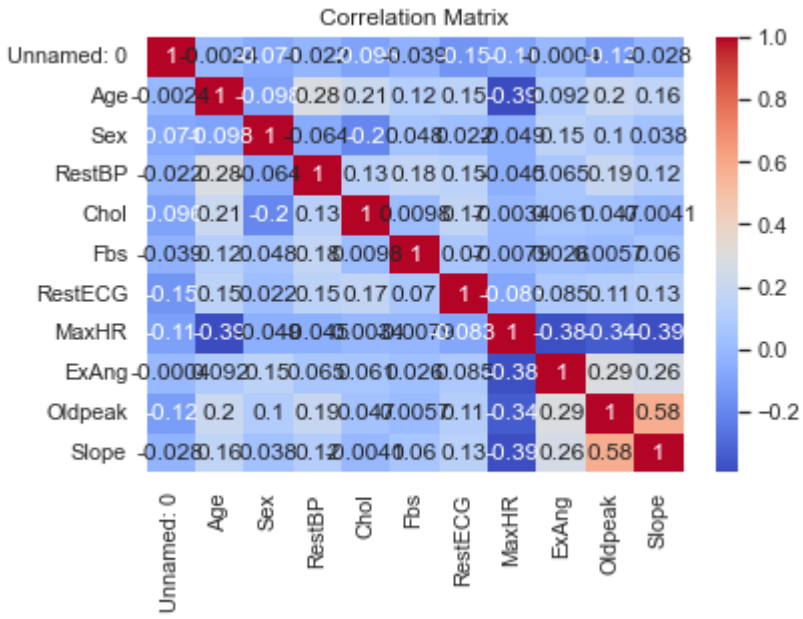
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Sex        0 non-null      float64
 1   ChestPain  0 non-null      object 
 2   RestBP     2 non-null      float64
 3   Chol       4 non-null      float64
 4   Fbs        45 non-null     float64
 5   RestECG    0 non-null      float64
 6   MaxHR      0 non-null      float64
 7   ExAng      0 non-null      float64
 8   Oldpeak    2 non-null      float64
 9   Slope      0 non-null      float64
 10  AHD        0 non-null      object 
dtypes: float64(9), object(2)
memory usage: 26.2+ KB
```

----------DATA TRANSFORMATION----------

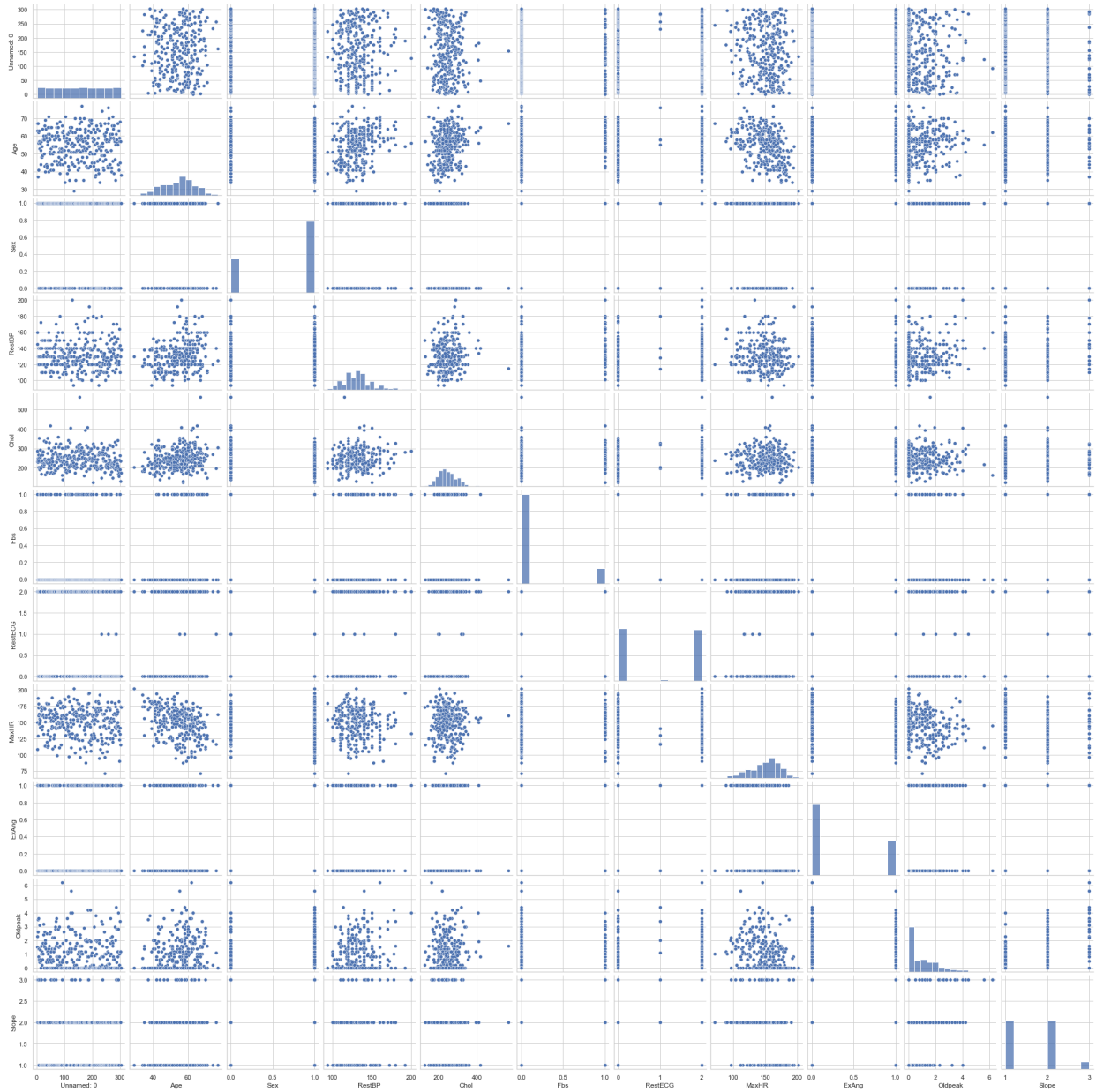
In [12]:
```
sns.barplot(x='Age',y='Sex', data=df.sort_values('Age'))
plt.title('Rate of heart dieases depending on age')
plt.xticks(rotation=90)
plt.show()
```



In [13]:
```
sns.heatmap(df.corr(),annot=True,cmap = 'coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



In [14]:
```
sns.pairplot(df)
plt.show()
```



----------Model BULIDING----------

In [15]:
```
from sklearn.model_selection import train_test_split
Y = df['Age'] #variável de predição
X = df.drop(['ChestPain','Slope'], axis=1)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
print(X_train.shape, X_test.shape)
```

```
(242, 11) (61, 11)
```

In [ ]: