ANSH AWASTHI IT 3rd Year 2100290139002

In [1]: *#importing all required libraries*
```
import pandas as pd
import numpy as np
from scipy import stats as st
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]: *#reading the heart disease dataset*
```
df=pd.read_csv('D:/datasets_ML/heart.csv')
```

In [3]: *#displaying top 5 data*
```
df.head()
```

Out[3]:

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

In [4]: *#displaying the informtion of dataset*
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

In [5]: *#displaying bottom 5 data*
```
df.tail()
```

Out[5]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

In [6]: *#display the shape of dataset*
```
df.shape
```

Out[6]:(1025, 14)

In [7]: *#to display the value counts of age column*
```
#df['age'].value_counts()
```

In [8]: *#to display the value counts of sex column*
```
df['sex'].value_counts()
```

Out[8]:1   713
       0   312
       Name: sex, dtype: int64

In [9]: *#to display the value counts of chol column*
```
df['chol'].value_counts()
```

Out[9]:204   21
       234   21
       197   19
       212   18
       254   17
            ..
       164   3
       394   3
       215   3
       160   3
       141   3
       Name: chol, Length: 152, dtype: int64

In [10]: *#to display the statistcs of each numerical column*
         df.describe()

Out[10]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.38536
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.61775
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.00000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.00000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.00000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.00000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.00000

In [11]: *#finding mean of age column*
         print(np.mean(df['age']))
         #finding median of age column
         print(np.median(df['age']))
         #finding mode of age column
         print(st.mode(df['age']))

54.43414634146342
56.0
ModeResult(mode=array([58], dtype=int64), count=array([68]))

In [12]: *#finding mean of cp column*
         print(np.mean(df['cp']))
         #finding median of cp column
         print(np.median(df['cp']))
         #finding mode of cp column
         print(st.mode(df['cp']))

0.9424390243902439
1.0
ModeResult(mode=array([0], dtype=int64), count=array([497]))

In [13]: *#finding mean of trestbps column*
         print(np.mean(df['trestbps']))
         #finding median of trestbps column
         print(np.median(df['trestbps']))
         #finding mode of trestbps column
         print(st.mode(df['trestbps']))

131.61170731707318
130.0
ModeResult(mode=array([120], dtype=int64), count=array([128]))

In [14]: *#finding mean of chol column*
         print(np.mean(df['chol']))
         #finding median of chol column
         print(np.median(df['chol']))
         #finding mode of chol column
         print(st.mode(df['chol']))

246.0
240.0
ModeResult(mode=array([204], dtype=int64), count=array([21]))

In [15]: *#finding mean of fbs column*
         print(np.mean(df['fbs']))
         #finding median of fbs column
         print(np.median(df['fbs']))
         #finding mode of fbs column
         print(st.mode(df['fbs']))

0.14926829268292682
0.0
ModeResult(mode=array([0], dtype=int64), count=array([872]))

In [16]: *#finding mean of restecg column*
         print(np.mean(df['restecg']))

```
#finding median of restecg column
print(np.median(df['restecg']))
#finding mode of restecg column
print(st.mode(df['restecg']))
```

0.5297560975609756
1.0
ModeResult(mode=array([1], dtype=int64), count=array([513]))

In [17]:
```
#finding mean of thalach column
print(np.mean(df['thalach']))
#finding median of thalach column
print(np.median(df['thalach']))
#finding mode of thalach column
print(st.mode(df['thalach']))
```

149.11414634146342
152.0
ModeResult(mode=array([162], dtype=int64), count=array([35]))

In [18]:
```
#finding mean of exang column
print(np.mean(df['exang']))
#finding median of exang column
print(np.median(df['exang']))
#finding mode of exang column
print(st.mode(df['exang']))
```

0.33658536585365856
0.0
ModeResult(mode=array([0], dtype=int64), count=array([680]))

In [19]:
```
#finding mean of oldpeak column
print(np.mean(df['oldpeak']))
#finding median of oldpeak column
print(np.median(df['oldpeak']))
#finding mode of oldpeak column
print(st.mode(df['oldpeak']))
```

1.0715121951219524
0.8
ModeResult(mode=array([0.]), count=array([329]))

In [20]:
```
#finding mean of ca column
print(np.mean(df['ca']))
#finding median of ca column
print(np.median(df['ca']))
#finding mode of ca column
print(st.mode(df['ca']))
```

0.7541463414634146
0.0
ModeResult(mode=array([0], dtype=int64), count=array([578]))

In [21]:
```
#finding mean of thal column
print(np.mean(df['thal']))
#finding median of thal column
print(np.median(df['thal']))
#finding mode of thal column
print(st.mode(df['thal']))
```

2.32390243902439
2.0
ModeResult(mode=array([2], dtype=int64), count=array([544]))

In [22]:
```
#finding mean of target column
print(np.mean(df['target']))
#finding median of target column
print(np.median(df['target']))
#finding mode of target column
print(st.mode(df['target']))
```

0.5131707317073171
1.0
ModeResult(mode=array([1], dtype=int64), count=array([526]))

In [23]:
```
#checking for null values
df.isna()
```

Out[23]:

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 2 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 3 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 4 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1020 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1021 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1022 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1023 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1024 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |

1025 rows × 14 columns

In [24]:
```python
# Correlation matrix
corr_matrix = df.corr()
sns.heatmap(corr_matrix, annot=True, cmap='YlGnBu')
plt.show()
```
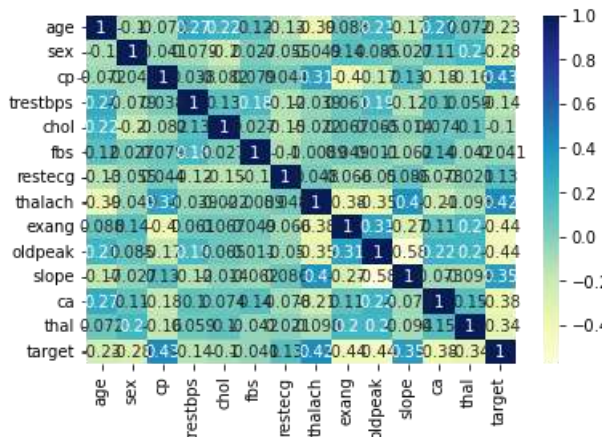

In [25]:
```python
# Histograms of all features
df.hist(bins=20, figsize=(15, 10))
plt.show()
```
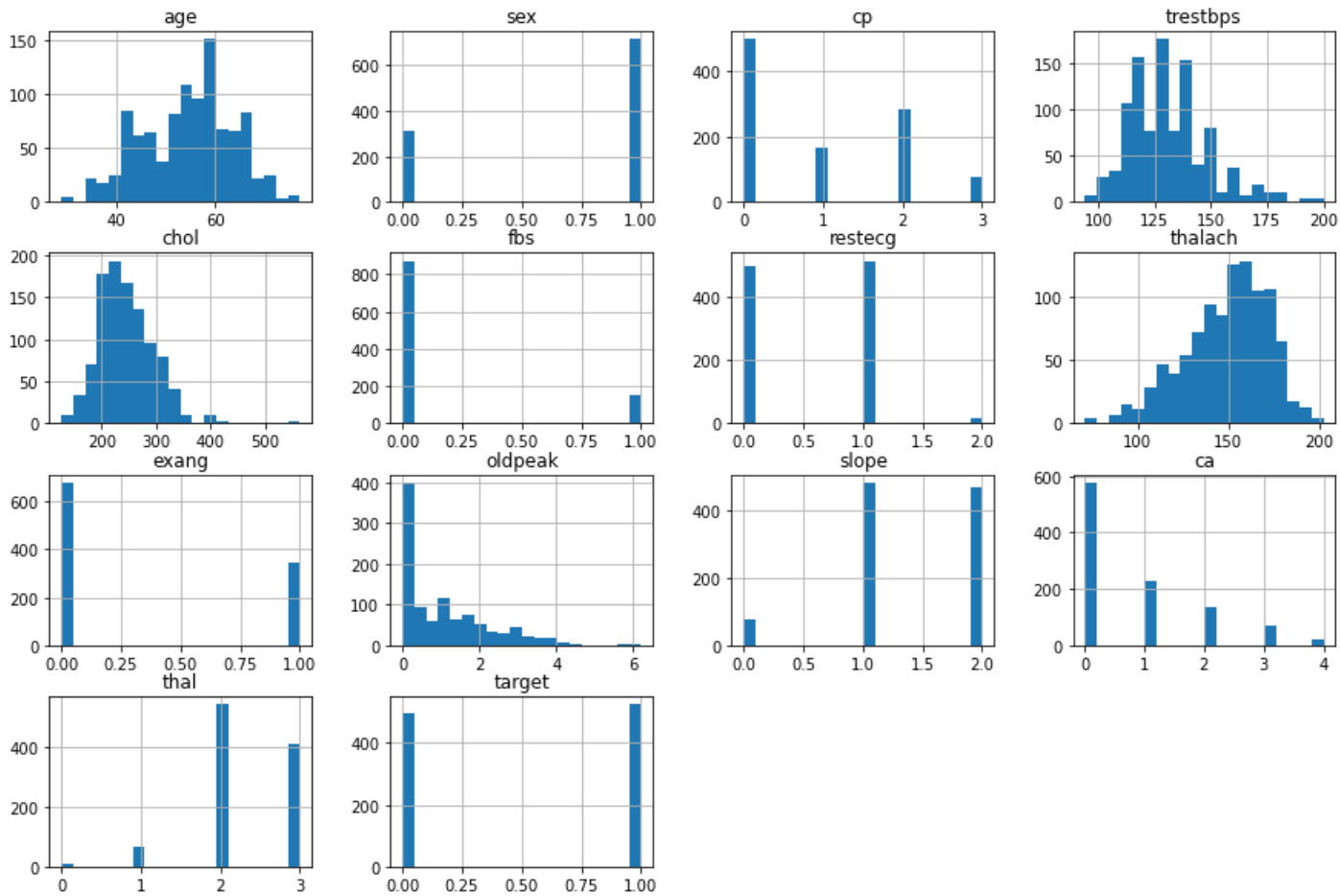
```
In [26]: # Boxplots of all features grouped by target
         fig, axes = plt.subplots(nrows=5, ncols=3, figsize=(15,15))
         for i, ax in enumerate(axes.flatten()):
             if i < len(df.columns)-1:
                 sns.boxplot(x='target', y=df.columns[i], data=df, ax=ax)
         plt.tight_layout()
         plt.show()
```