NAME : SUBHAM BEURA

Id – B521060

Branch : CE

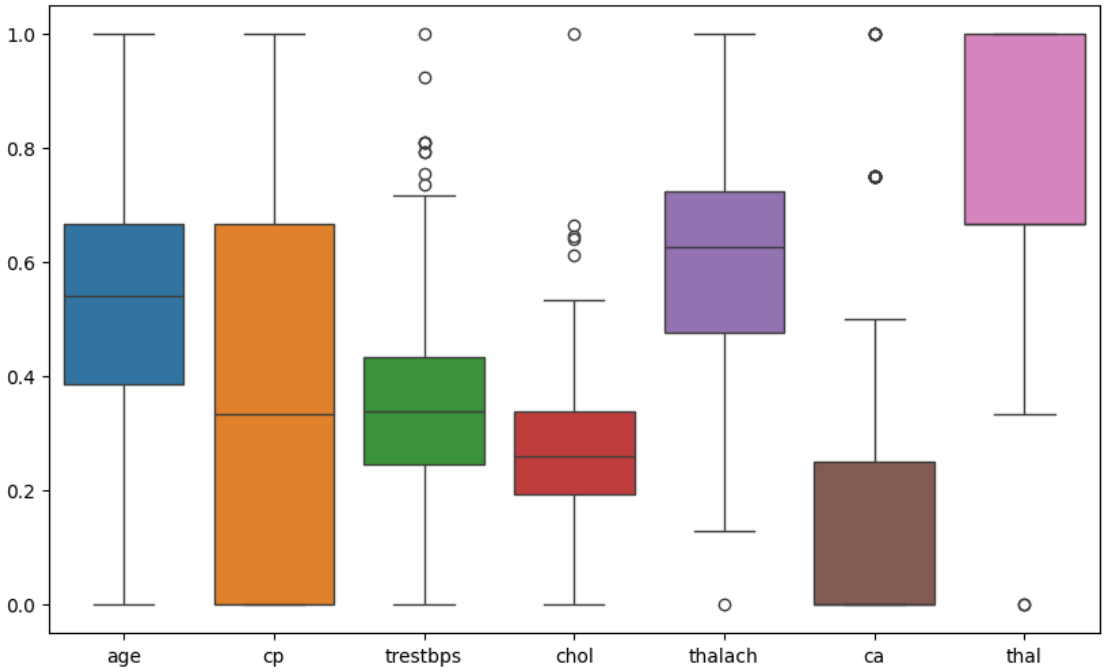
ML Lab - 3

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

import scipy.stats as stats
import seaborn as sns

df = pd.read_csv('data.csv')
x = df.iloc[:, :-1].values
y = df.iloc[:, -1].values

print("Checking for Missing values in dataset:\n",df.isnull().sum())
```

```
Checking for Missing values in dataset:
 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
```

```
df1 = df.drop(['restecg', 'fbs', 'exang', 'slope', 'oldpeak', 'sex',
'target'], axis=1, inplace=False)
df_scaled = (df1-np.min(df1, axis=0))/(np.max(df1, axis=0)-np.min(df1,
axis=0)).values
plt.figure(figsize=(10,6))
sns.boxplot(data=df_scaled)
```

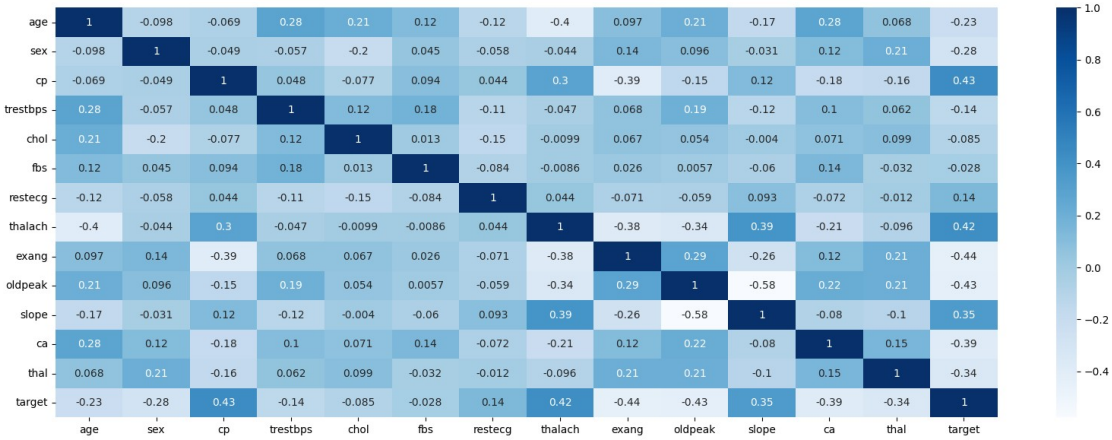
```
<Axes: >
```



“ There is a strong correlation between thal , cp, ca and target value being either positive or negative I.e if positively correlated then there is chance of having heart disease and vice versa. This is important in relation to calculate the t-test.“

```
import statsmodels.api as sm

plt.figure(figsize=(20,7))
sns.heatmap(df.corr(), annot = True, cmap="Blues")
```

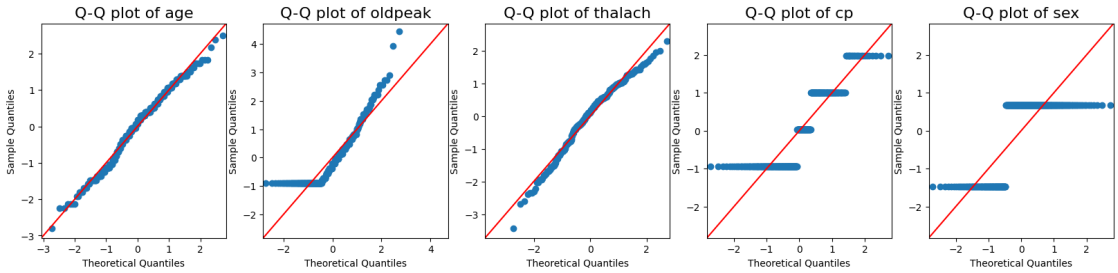
```
<Axes: >
```




```
from statsmodels.graphics.gofplots import qqplot
```

```
df2 = df[['age', 'oldpeak', 'thalach', 'cp', 'sex']]
```

```
fig, (ax1, ax2, ax3, ax4, ax5) = plt.subplots(ncols=5, figsize=(20,4))
axis = [ax1, ax2, ax3, ax4, ax5]
for idx, c in enumerate(df2.columns[:]):
  qqplot(df2[c], line='45', fit='True', ax=axis[idx])
  axis[idx].set_title("Q-Q plot of {}".format(c),fontsize=16)

plt.show()
```



“ The q-q plot of age shows it is normally distributed. “

```
yes_hd = df['age'][(df['target']==1)]
yes_hd_mean = df['age'][(df['target']==1)].mean()
print(yes_hd)
print("mean of patients with heart disease:",yes_hd_mean)
```

```
0      63
1      37
2      41
3      56
4      57
       ..
160    56
161    55
162    41
163    38
164    38
Name: age, Length: 165, dtype: int64
mean of patients with heart disease: 52.4969696969697
```

```
no_hd = df['age'][(df['target']==0)]
no_hd_mean = df['age'][(df['target']==0)].mean()
print(no_hd)
print("mean of patients with no heart disease:",no_hd_mean)
```

```
165    67
166    67
167    62
168    63
169    53
       ..
298    57
299    45
```

```
300    68
301    57
302    57
Name: age, Length: 138, dtype: int64
mean of patients with no heart disease: 56.60144927536232
```

```
t_statistic, p_value = stats.ttest_ind(yes_hd, no_hd)

alpha = 0.05
# Compute the degrees of freedom
dof = len(no_hd)+len(yes_hd)-2

# Calculate the critical t-value
# ppf is used to find the critical t-value for a two-tailed test
critical_t = stats.t.ppf(1 - alpha/2, dof)
print("T-statistic:", t_statistic)
print("P-value:", p_value)
print("Critical t-value:", critical_t)

print('With T-value')
if np.abs(t_statistic) >critical_t:
 print('There is significant difference between two groups')
else:
 print('No significant difference found between two groups')

print('With P-value')
if p_value >alpha:
 print('No evidence to reject the null hypothesis that a significant
difference between the two groups')
else:
 print('Evidence found to reject the null hypothesis, Hence there is a
significant difference between the two age groups')
```

Explanation:

```
T-statistic: -4.014560975148874
P-value: 7.524801303442373e-05
Critical t-value: 1.9678765312853974
With T-value
There is significant difference between two groups
With P-value
Evidence found to reject the null hypothesis, Hence
there is a significant difference between the two age
groups
```

```
yes_hd = df['age'][(df['sex']==1) & (df['target']==1)]
print(yes_hd)
```

```
0      63
1      37
3      56
5      57
7      44
       ..
159    56
160    56
162    41
163    38
164    38
Name: age, Length: 93, dtype: int64
```

```
yes_hd = df['age'][(df['sex']==1) & (df['target']==0)]
print(yes_hd)
```

```
165    67
166    67
167    62
168    63
169    53
       ..
298    57
299    45
300    68
301    57
302    57
Name: age, Length: 138, dtype: int64
```

```
# chi-square test
contingency_table = pd.crosstab(df['sex'], df['target'])
chi2_stat, p_val_chi2, dof, expected =
stats.chi2_contingency(contingency_table)

print("Chi-square statistic:", chi2_stat)
print("P-value:", p_val_chi2)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)

alpha = 0.05
print('With P-value')
if p_value >alpha:
    print('No evidence to reject the null hypothesis that a
significant difference between the two groups')
else:
    print('Evidence found to reject the null hypothesis, Hence there
is a significant difference between the two age groups')
```

```
Chi-square statistic: 22.717227046576355
P-value: 1.8767776216941503e-06
Degrees of freedom: 1
```

```
Expected frequencies:
 [[ 43.72277228  52.27722772]
 [ 94.27722772 112.72277228]]
With P-value
Evidence found to reject the null hypothesis, Hence there is a
significant difference between the two sex groups
```

Explanation :

The observed and expected frequencies in the contingency table of sex vs. target. A higher chi-square statistic indicates a larger difference between what was observed and what would be expected if there were no relationship between the variables.

Evidence found to reject the null hypothesis, Hence there is a significant difference between the two sex groups since p-value is much lower than significance level.

END