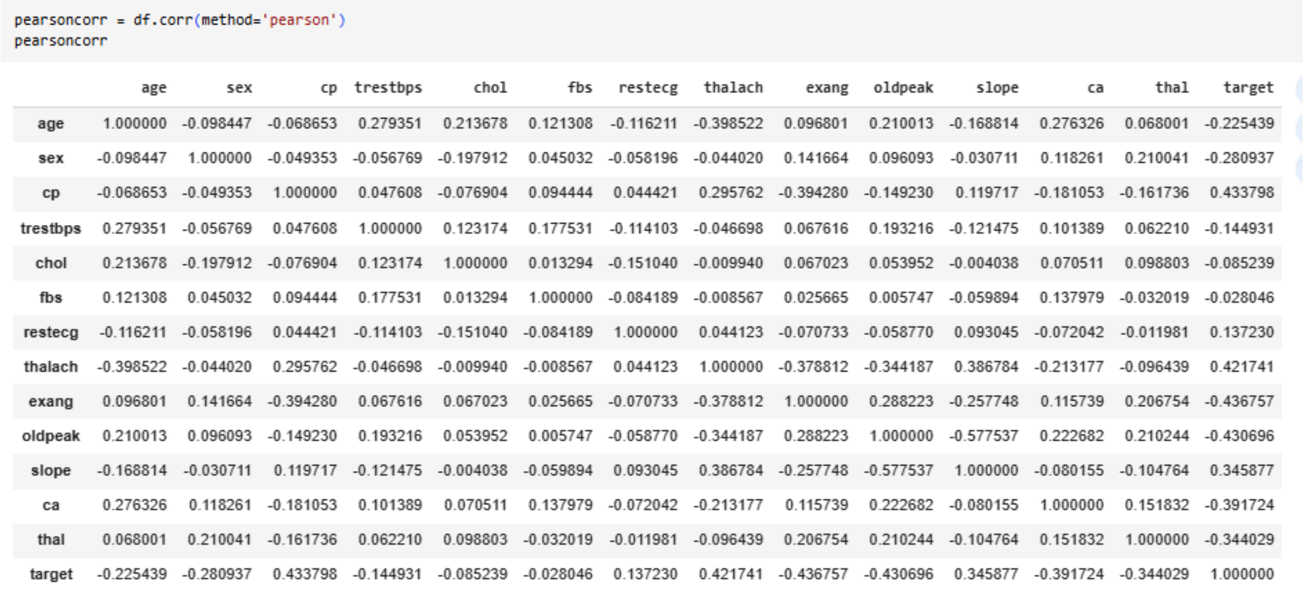
Experiment 4

Aim: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn.

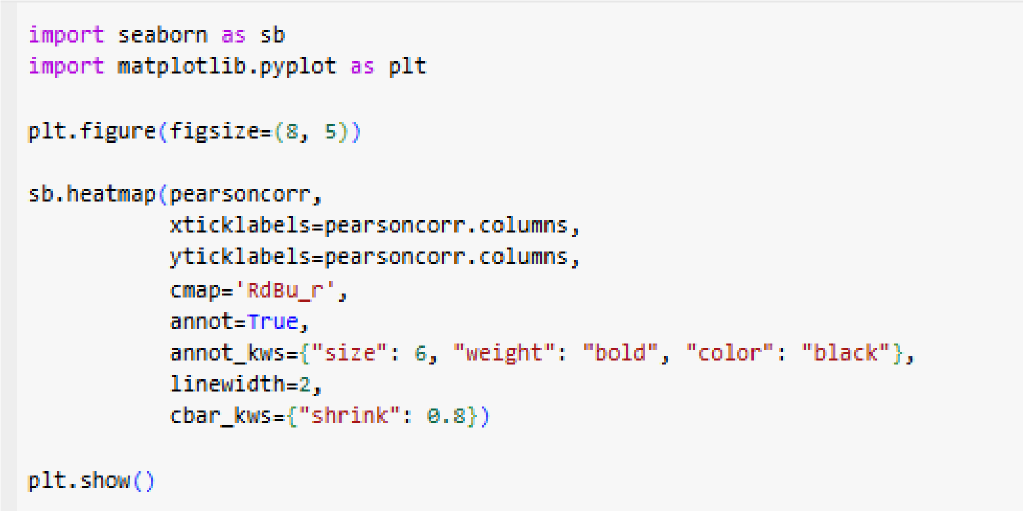
Theory:

1. **Pearson's Correlation**: Pearson's correlation measures the linear relationship between two continuous variables. A coefficient close to 1 or -1 indicates a strong linear relationship, while a value near 0 suggests no linear association.

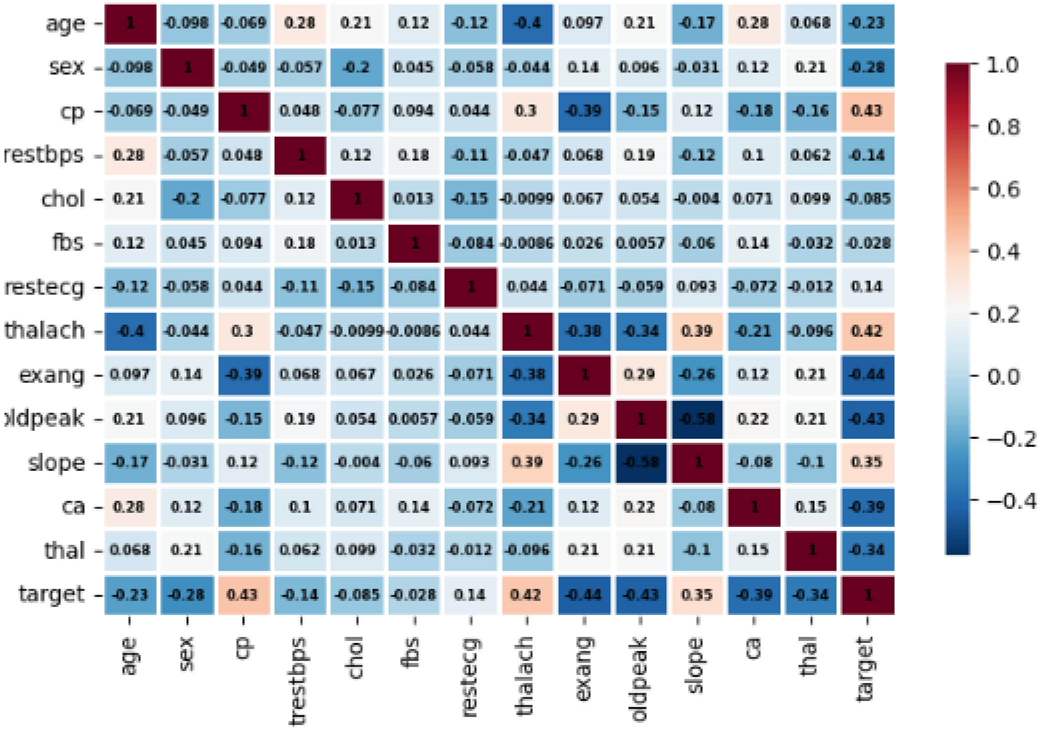


```
pearsoncorr = df.corr(method='pearson')
pearsoncorr
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

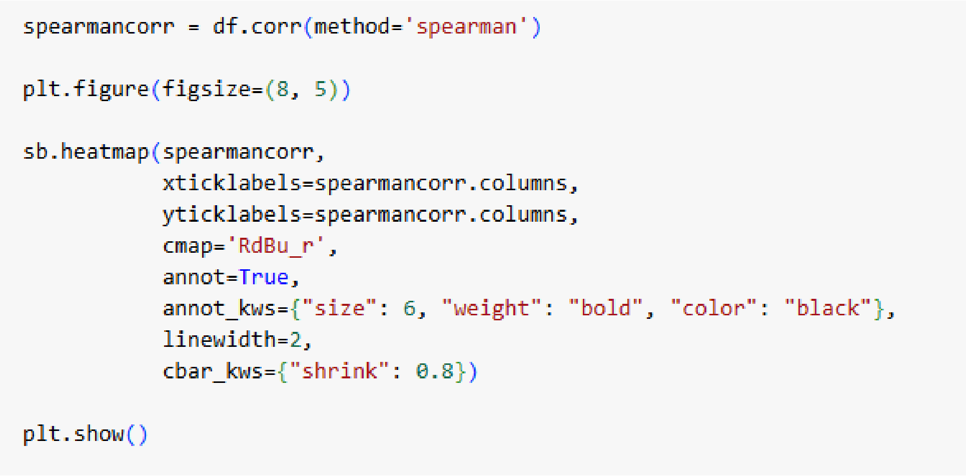


```
import seaborn as sb
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 5))

sb.heatmap(pearsoncorr,
           xticklabels=pearsoncorr.columns,
           yticklabels=pearsoncorr.columns,
           cmap='RdBu_r',
           annot=True,
           annot_kws={"size": 6, "weight": "bold", "color": "black"},
           linewidth=2,
           cbar_kws={"shrink": 0.8})

plt.show()
```

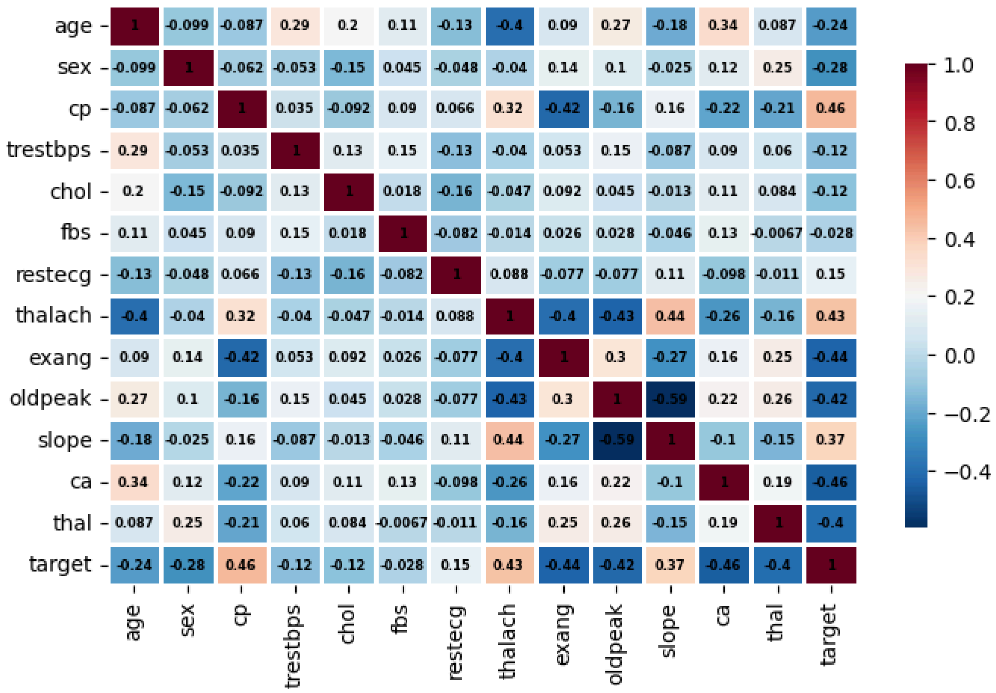


Result: There is a moderate positive relationship between cp and heart disease (target), with a correlation of 0.43.

2. **Spearman's Rank Correlation**: Spearman's rank correlation assesses the monotonic relationship between variables, relying on their ranks rather than raw data. It is suitable for ordinal or non-linear relationships.



```
spearmancorr = df.corr(method='spearman')

plt.figure(figsize=(8, 5))

sb.heatmap(spearmancorr,
           xticklabels=spearmancorr.columns,
           yticklabels=spearmancorr.columns,
           cmap='RdBu_r',
           annot=True,
           annot_kws={"size": 6, "weight": "bold", "color": "black"},
           linewidth=2,
           cbar_kws={"shrink": 0.8})

plt.show()
```

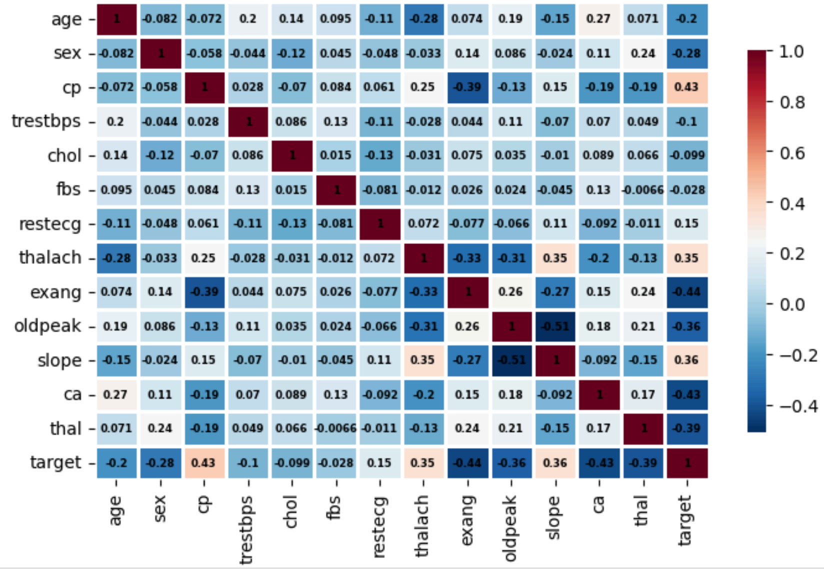


Results: The correlation between cp (chest pain type) and target is 0.46, indicating a moderate positive association.

3. **Kendall's Rank Correlation**: Kendall's Tau measures the strength of the ordinal relationship between two variables by comparing the ranks of pairs. It is more robust to ties in data.

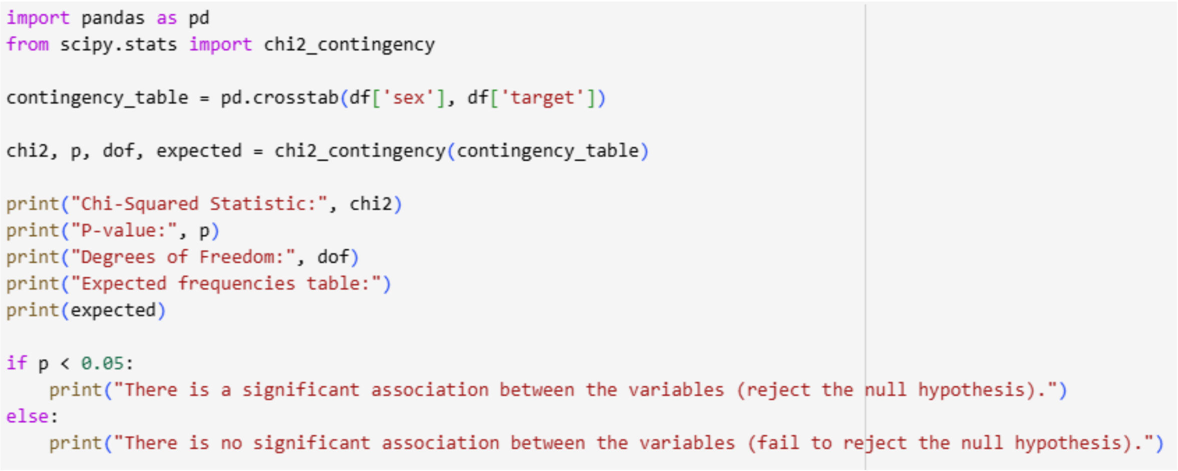
```
from scipy.stats import pearsonr, spearmanr, kendalltau, chi2_contingency
```

```
kendallcorr = df.corr(method='kendall')

plt.figure(figsize=(8, 5))

sb.heatmap(kendallcorr,
           xticklabels=kendallcorr.columns,
           yticklabels=kendallcorr.columns,
           cmap='RdBu_r',
           annot=True,
           annot_kws={"size": 6, "weight": "bold", "color": "black"},
           linewidth=2,
           cbar_kws={"shrink": 0.8})

plt.show()
```

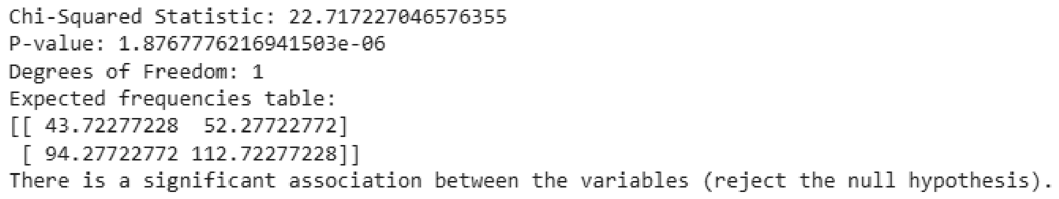


Results: cp and target have a Kendall's Tau of 0.43, showing a positive relationship. target and thalach show a weaker but still positive correlation (0.35), indicating heart rate's relevance in predicting heart disease.

4. **Chi-Squared Test**: The Chi-Squared test assesses the independence of two categorical variables. A significant p-value indicates that the variables are dependent (associated).



```python
import pandas as pd
from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['sex'], df['target'])

chi2, p, dof, expected = chi2_contingency(contingency_table)

print("Chi-Squared Statistic:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)
print("Expected frequencies table:")
print(expected)

if p < 0.05:
    print("There is a significant association between the variables (reject the null hypothesis).")
else:
    print("There is no significant association between the variables (fail to reject the null hypothesis).")
```



```
Chi-Squared Statistic: 22.717227046576355
P-value: 1.8767776216941503e-06
Degrees of Freedom: 1
Expected frequencies table:
[[ 43.72277228  52.27722772]
 [ 94.27722772 112.72277228]]
There is a significant association between the variables (reject the null hypothesis).
```

Result: The Chi-Squared statistic of 22.72 (p-value = 1.88e-06) indicates a significant association between categorical variables (e.g., sex and target), suggesting their role in heart disease prediction.

Conclusion:

In this analysis, four statistical tests were applied to assess the relationships between various features and heart disease:

1. Pearson's correlation showed a moderate positive relationship between cp and heart disease (target), with a correlation of 0.43.
2. Spearman's rank correlation confirmed a moderate positive relationship between cp (chest pain type) and target (0.46).
3. Kendall's Tau also revealed a moderate positive association between cp and target (0.43)
4. The Chi-Squared test showed a significant association between categorical variables, with a Chi-Squared statistic of 22.72 and a p-value of 1.88e-06.Since the p-value of 1.88e-06 is much smaller than the commonly used significance level of 0.05, we reject the null hypothesis.