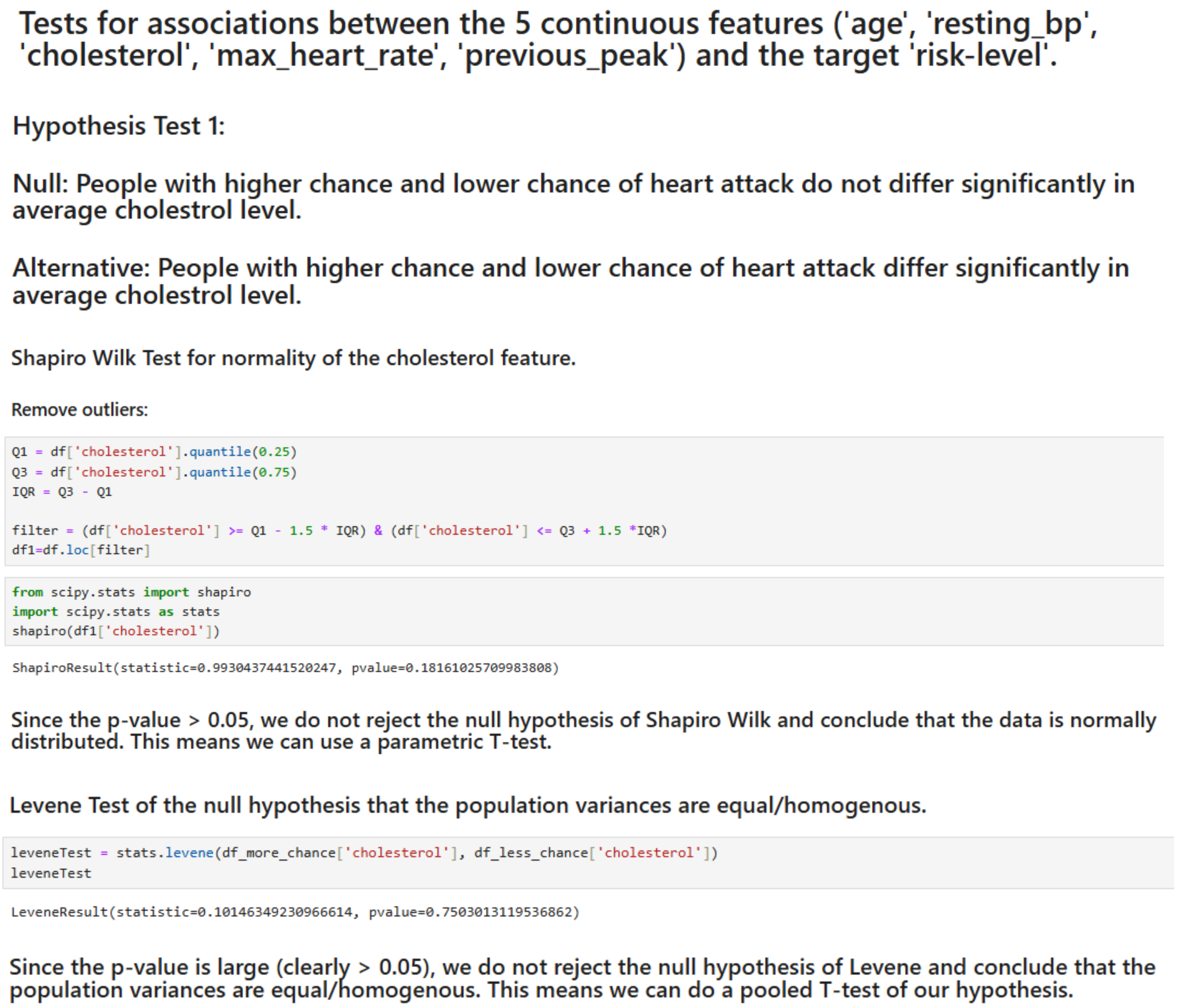
Formulating Hypotheses
Objective*: Identify key variables and formulate hypotheses to test.*

Statistical Testing
Objective*: Conduct statistical tests to validate hypotheses.*

Note: Because of the relatively small number of variables in this data set, I tested, for each one, whether the given variable was associated with the target variable, which is the risk-level of having a heart attack (high- or low-risk).

Part 1: Continuous Features



Tests for associations between the 5 continuous features ('age', 'resting_bp', 'cholesterol', 'max_heart_rate', 'previous_peak') and the target 'risk-level'.

Hypothesis Test 1:

Null: People with higher chance and lower chance of heart attack do not differ significantly in average cholestrol level.

Alternative: People with higher chance and lower chance of heart attack differ significantly in average cholestrol level.

Shapiro Wilk Test for normality of the cholesterol feature.

Remove outliers:

```
Q1 = df['cholesterol'].quantile(0.25)
Q3 = df['cholesterol'].quantile(0.75)
IQR = Q3 - Q1

filter = (df['cholesterol'] >= Q1 - 1.5 * IQR) & (df['cholesterol'] <= Q3 + 1.5 *IQR)
df1=df.loc[filter]
```

```
from scipy.stats import shapiro
import scipy.stats as stats
shapiro(df1['cholesterol'])
```

```
ShapiroResult(statistic=0.9930437441520247, pvalue=0.18161025709983808)
```

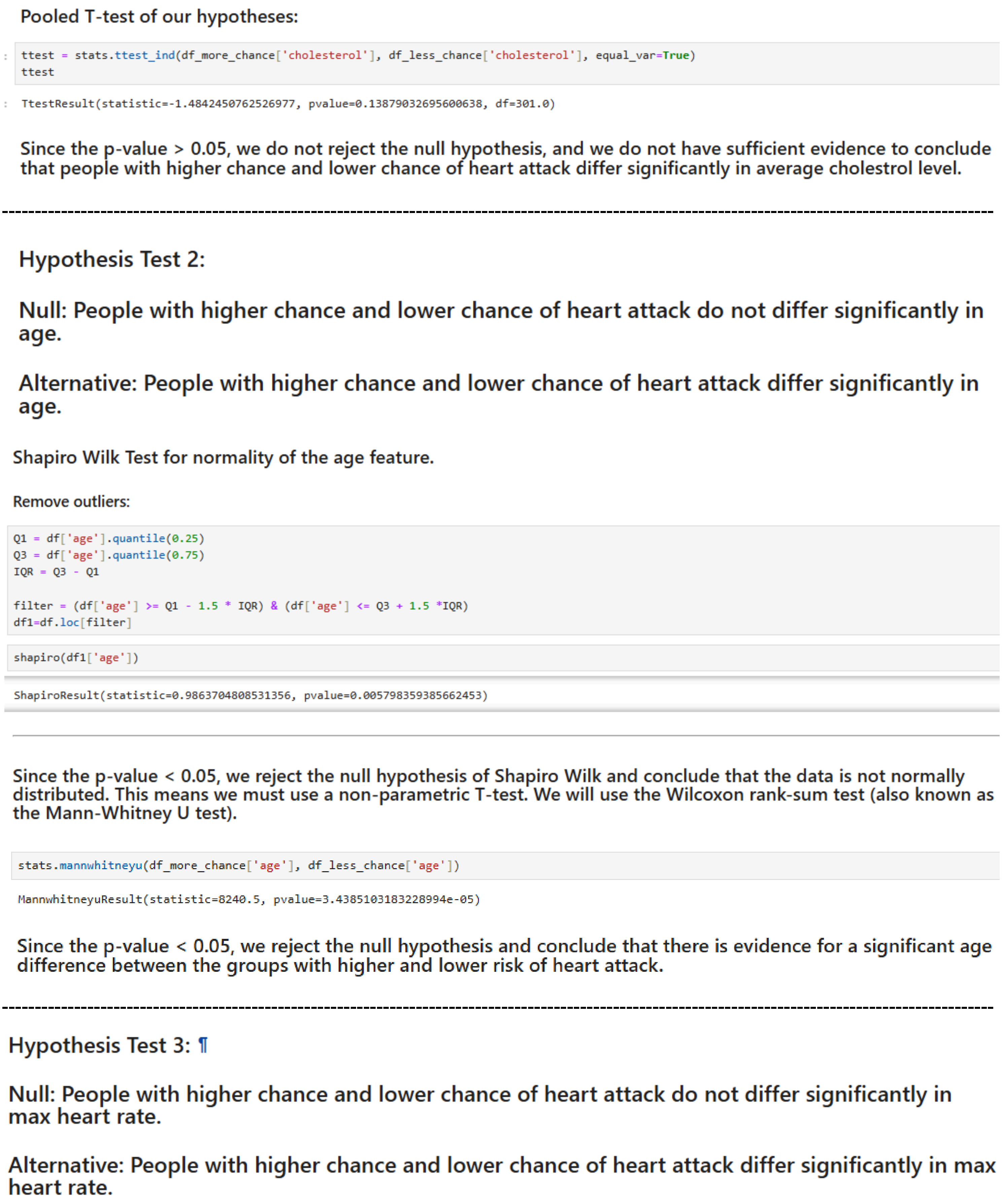
Since the p-value > 0.05, we do not reject the null hypothesis of Shapiro Wilk and conclude that the data is normally distributed. This means we can use a parametric T-test.

Levene Test of the null hypothesis that the population variances are equal/homogenous.

```
leveneTest = stats.levene(df_more_chance['cholesterol'], df_less_chance['cholesterol'])
leveneTest
```

```
LeveneResult(statistic=0.10146349230966614, pvalue=0.7503013119536862)
```

Since the p-value is large (clearly > 0.05), we do not reject the null hypothesis of Levene and conclude that the population variances are equal/homogenous. This means we can do a pooled T-test of our hypothesis.



Pooled T-test of our hypotheses:

```
ttest = stats.ttest_ind(df_more_chance['cholesterol'], df_less_chance['cholesterol'], equal_var=True)
ttest
```

```
TtestResult(statistic=-1.4842450762526977, pvalue=0.13879032695600638, df=301.0)
```

Since the p-value > 0.05, we do not reject the null hypothesis, and we do not have sufficient evidence to conclude that people with higher chance and lower chance of heart attack differ significantly in average cholestrol level.

Hypothesis Test 2:

Null: People with higher chance and lower chance of heart attack do not differ significantly in age.

Alternative: People with higher chance and lower chance of heart attack differ significantly in age.

Shapiro Wilk Test for normality of the age feature.

Remove outliers:

```
Q1 = df['age'].quantile(0.25)
Q3 = df['age'].quantile(0.75)
IQR = Q3 - Q1

filter = (df['age'] >= Q1 - 1.5 * IQR) & (df['age'] <= Q3 + 1.5 *IQR)
df1=df.loc[filter]
```

```
shapiro(df1['age'])
```

```
ShapiroResult(statistic=0.9863704808531356, pvalue=0.005798359385662453)
```

Since the p-value < 0.05, we reject the null hypothesis of Shapiro Wilk and conclude that the data is not normally distributed. This means we must use a non-parametric T-test. We will use the Wilcoxon rank-sum test (also known as the Mann-Whitney U test).

```
stats.mannwhitneyu(df_more_chance['age'], df_less_chance['age'])
```

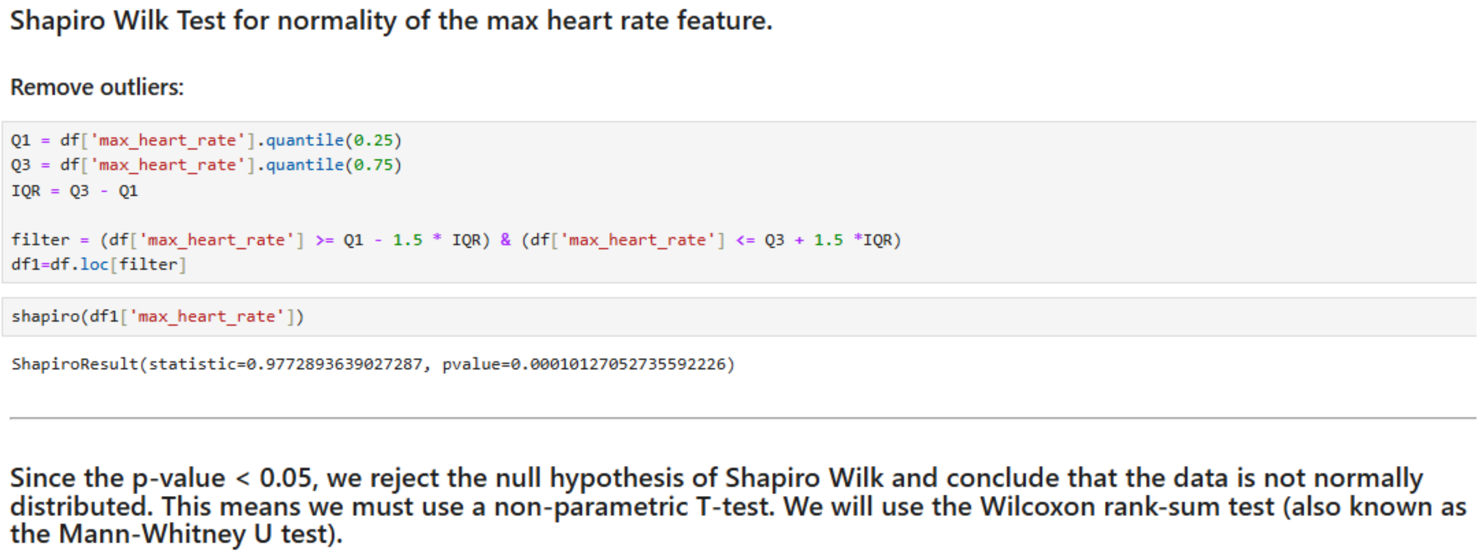
```
MannwhitneyuResult(statistic=8240.5, pvalue=3.4385103183228994e-05)
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is evidence for a significant age difference between the groups with higher and lower risk of heart attack.

Hypothesis Test 3: ¶

Null: People with higher chance and lower chance of heart attack do not differ significantly in max heart rate.

Alternative: People with higher chance and lower chance of heart attack differ significantly in max heart rate.



Shapiro Wilk Test for normality of the max heart rate feature.

Remove outliers:

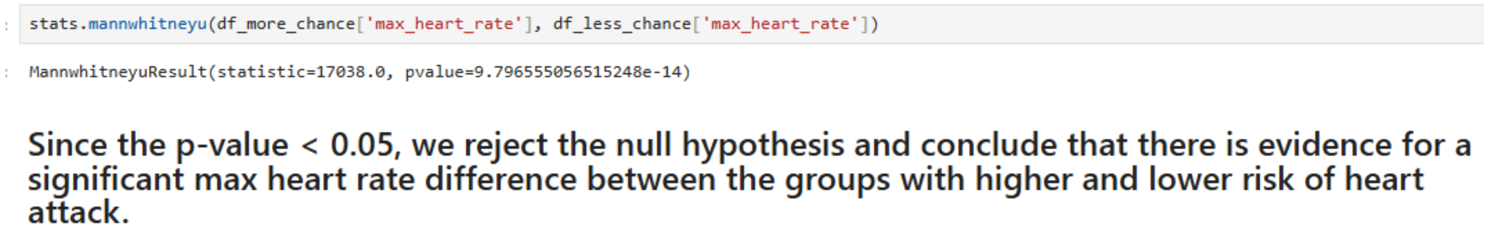
```
Q1 = df['max_heart_rate'].quantile(0.25)
Q3 = df['max_heart_rate'].quantile(0.75)
IQR = Q3 - Q1

filter = (df['max_heart_rate'] >= Q1 - 1.5 * IQR) & (df['max_heart_rate'] <= Q3 + 1.5 *IQR)
df1=df.loc[filter]
```

```
shapiro(df1['max_heart_rate'])
```

```
ShapiroResult(statistic=0.9772893639027287, pvalue=0.00010127052735592226)
```

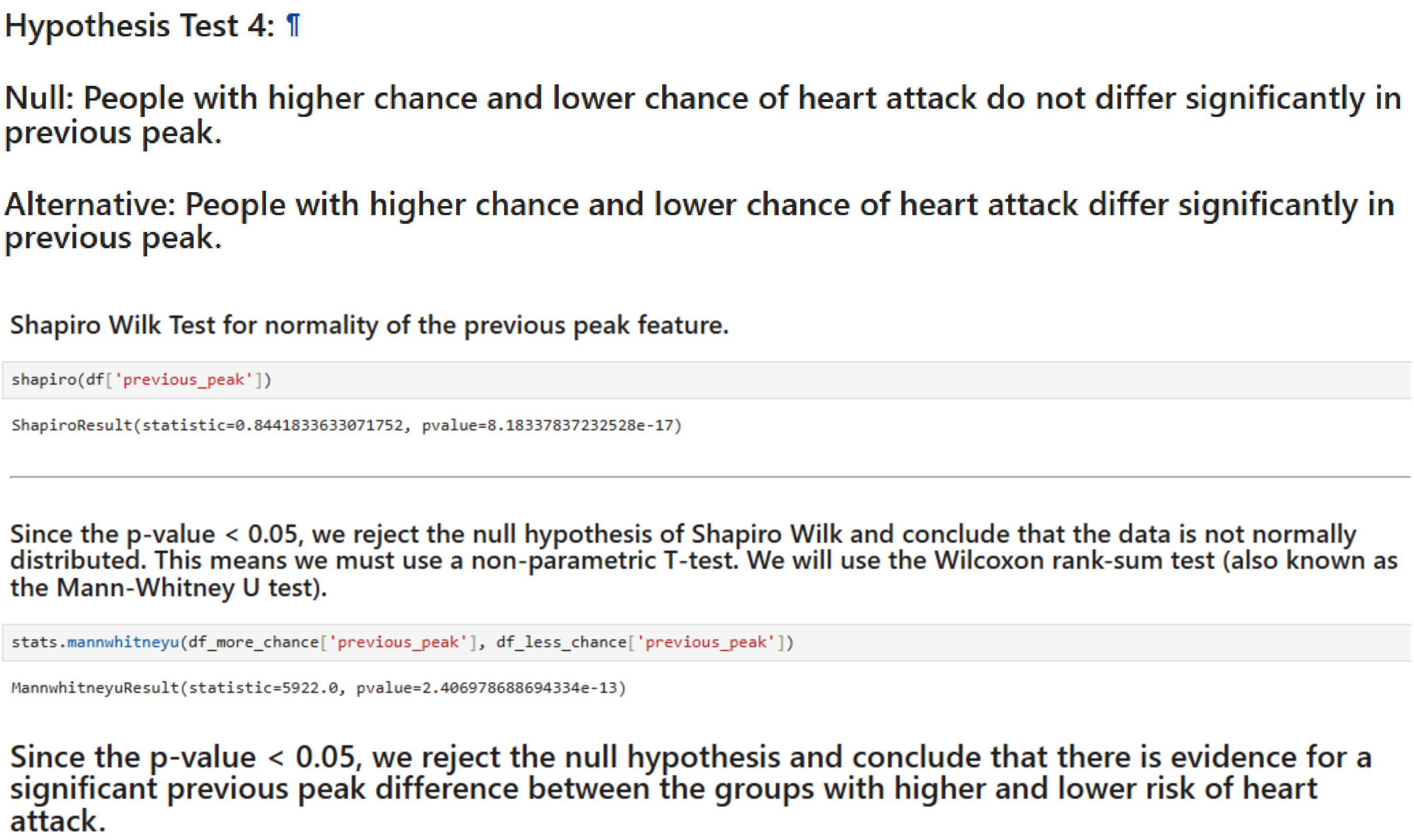
Since the p-value < 0.05, we reject the null hypothesis of Shapiro Wilk and conclude that the data is not normally distributed. This means we must use a non-parametric T-test. We will use the Wilcoxon rank-sum test (also known as the Mann-Whitney U test).



```
stats.mannwhitneyu(df_more_chance['max_heart_rate'], df_less_chance['max_heart_rate'])
```

```
MannwhitneyuResult(statistic=17038.0, pvalue=9.796555056515248e-14)
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is evidence for a significant max heart rate difference between the groups with higher and lower risk of heart attack.



Hypothesis Test 4: ¶

Null: People with higher chance and lower chance of heart attack do not differ significantly in previous peak.

Alternative: People with higher chance and lower chance of heart attack differ significantly in previous peak.

Shapiro Wilk Test for normality of the previous peak feature.

```
shapiro(df['previous_peak'])
```

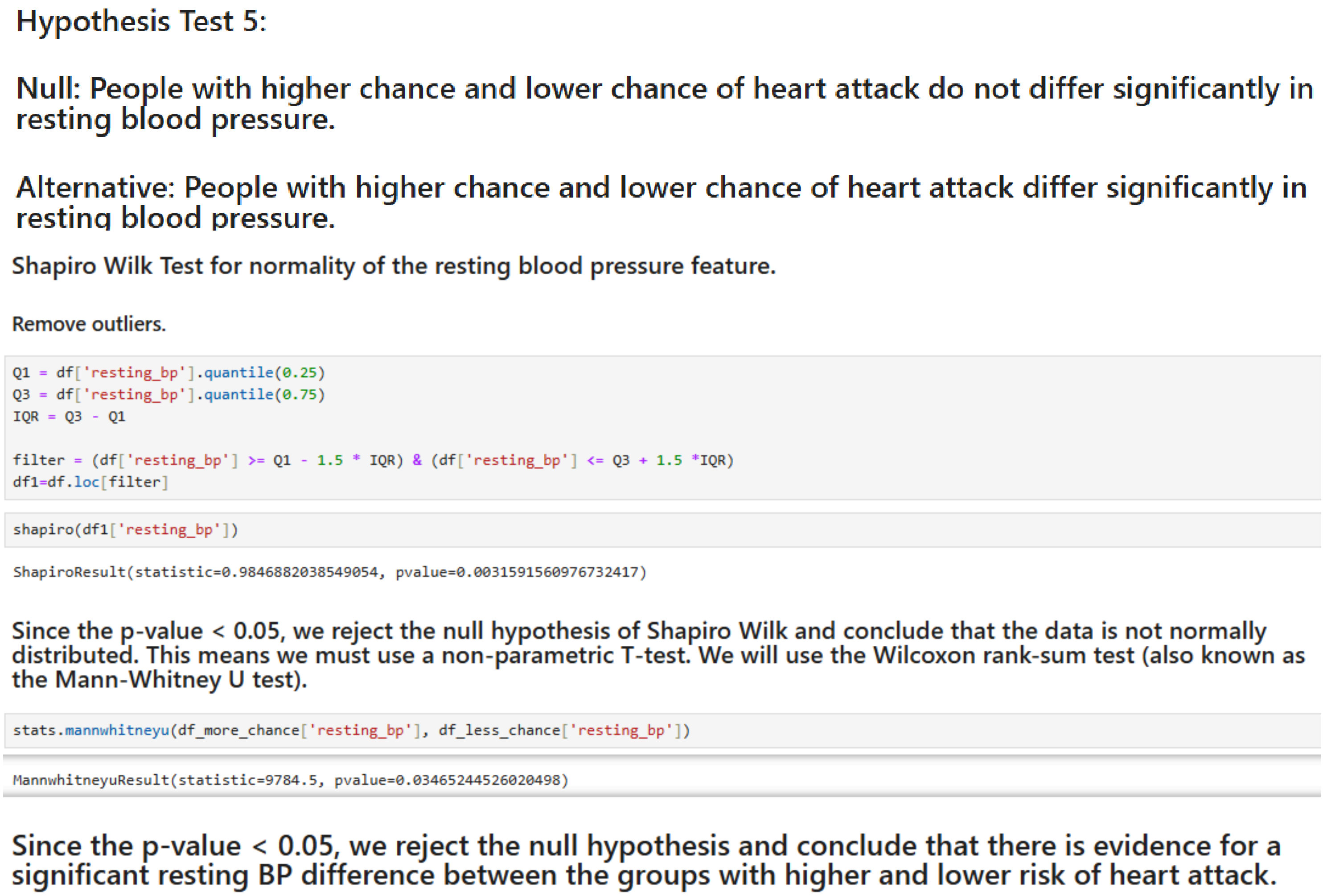
```
ShapiroResult(statistic=0.8441833633071752, pvalue=8.183378837232528e-17)
```

Since the p-value < 0.05, we reject the null hypothesis of Shapiro Wilk and conclude that the data is not normally distributed. This means we must use a non-parametric T-test. We will use the Wilcoxon rank-sum test (also known as the Mann-Whitney U test).

```
stats.mannwhitneyu(df_more_chance['previous_peak'], df_less_chance['previous_peak'])
```

```
MannwhitneyuResult(statistic=5922.0, pvalue=2.406978688694334e-13)
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is evidence for a significant previous peak difference between the groups with higher and lower risk of heart attack.



Hypothesis Test 5:

Null: People with higher chance and lower chance of heart attack do not differ significantly in resting blood pressure.

Alternative: People with higher chance and lower chance of heart attack differ significantly in resting blood pressure.

Shapiro Wilk Test for normality of the resting blood pressure feature.

Remove outliers.

```
Q1 = df['resting_bp'].quantile(0.25)
Q3 = df['resting_bp'].quantile(0.75)
IQR = Q3 - Q1

filter = (df['resting_bp'] >= Q1 - 1.5 * IQR) & (df['resting_bp'] <= Q3 + 1.5 *IQR)
df1=df.loc[filter]
```

```
shapiro(df1['resting_bp'])
```

```
ShapiroResult(statistic=0.9846882038549054, pvalue=0.0031591560976732417)
```

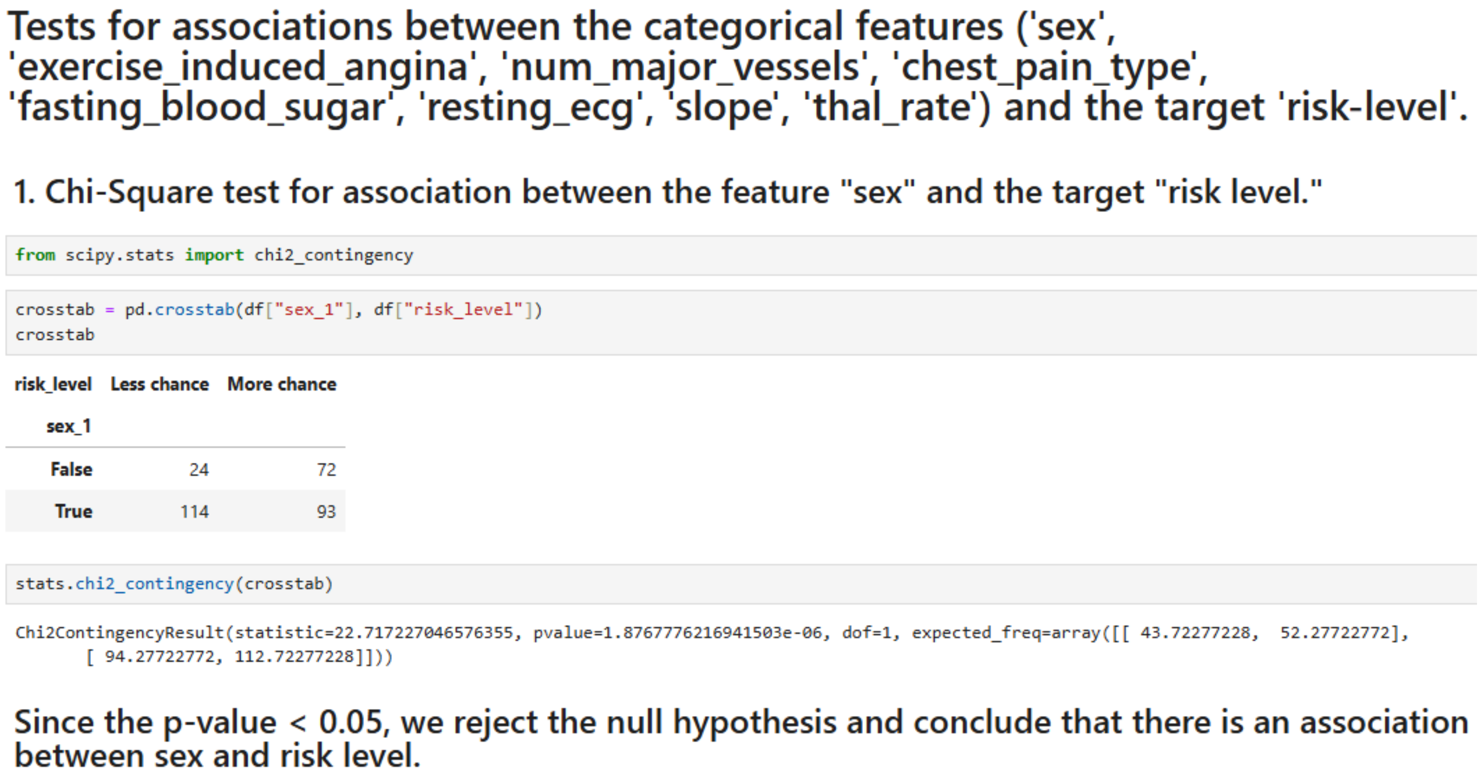
Since the p-value < 0.05, we reject the null hypothesis of Shapiro Wilk and conclude that the data is not normally distributed. This means we must use a non-parametric T-test. We will use the Wilcoxon rank-sum test (also known as the Mann-Whitney U test).

```
stats.mannwhitneyu(df_more_chance['resting_bp'], df_less_chance['resting_bp'])
```

```
MannwhitneyuResult(statistic=9784.5, pvalue=0.03465244526020498)
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is evidence for a significant resting BP difference between the groups with higher and lower risk of heart attack.

Part 2: Categorical Features



Tests for associations between the categorical features ('sex', 'exercise_induced_angina', 'num_major_vessels', 'chest_pain_type', 'fasting_blood_sugar', 'resting_ecg', 'slope', 'thal_rate') and the target 'risk-level'.

1. Chi-Square test for association between the feature "sex" and the target "risk level."

```
from scipy.stats import chi2_contingency
```

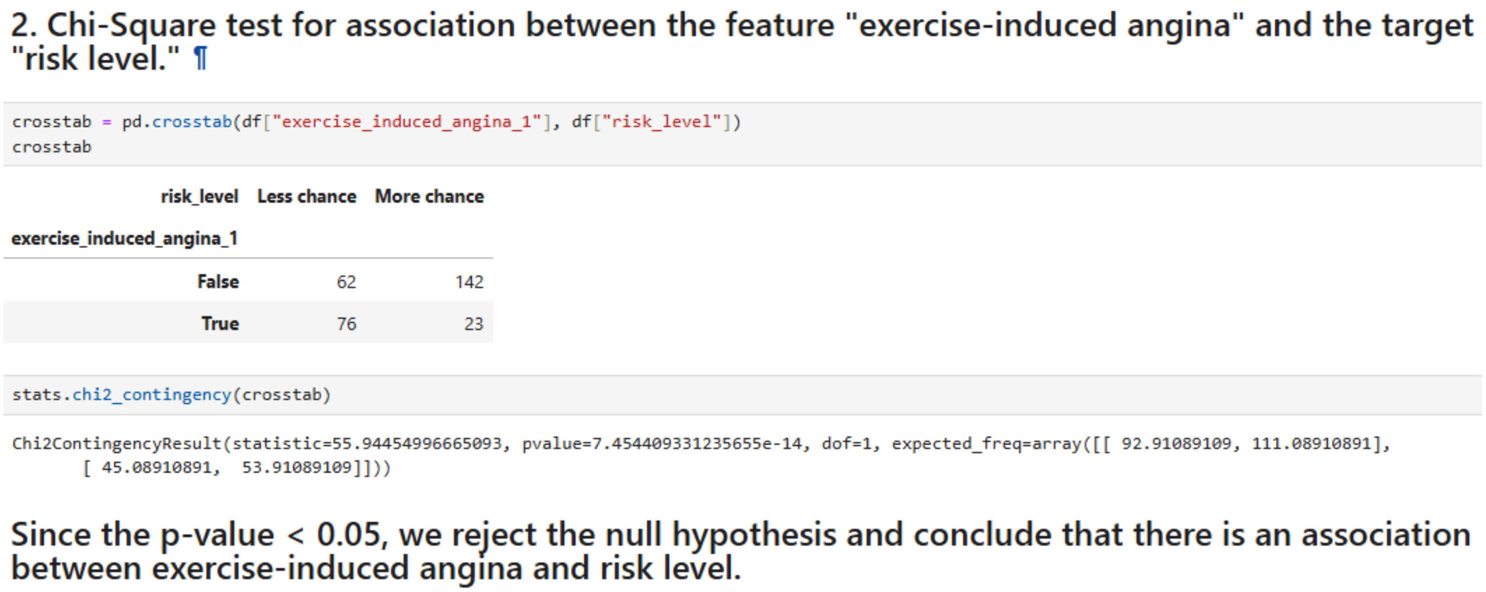
```
crosstab = pd.crosstab(df["sex_1"], df["risk_level"])
crosstab
```

risk_level	Less chance	More chance
sex_1		
False	24	72
True	114	93

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=22.717227046576355, pvalue=1.8767776216941503e-06, dof=1, expected_freq=array([[ 43.72277228,  52.27722772],
       [ 94.27722772, 112.72277228]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between sex and risk level.



2. Chi-Square test for association between the feature "exercise-induced angina" and the target "risk level." ¶

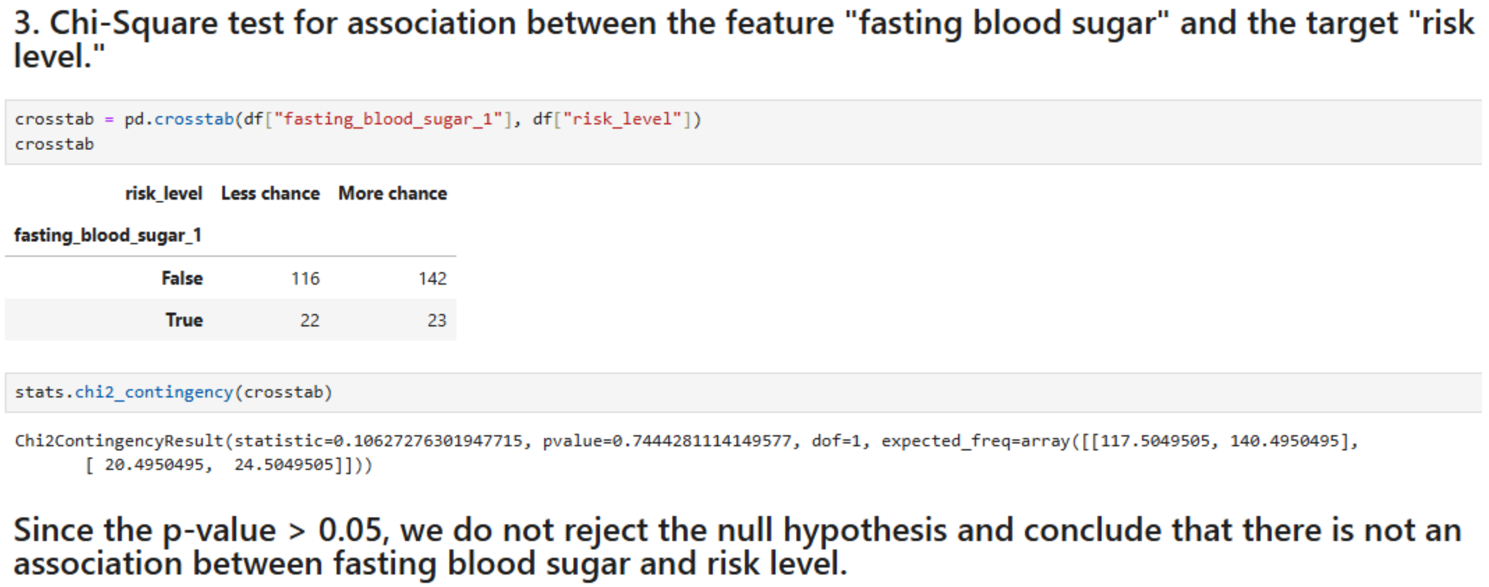
```
crosstab = pd.crosstab(df["exercise_induced_angina_1"], df["risk_level"])
crosstab
```

risk_level	Less chance	More chance
exercise_induced_angina_1		
False	62	142
True	76	23

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=55.94454996665093, pvalue=7.454409331235655e-14, dof=1, expected_freq=array([[ 92.91089109, 111.08910891],
       [ 45.08910891,  53.91089109]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between exercise-induced angina and risk level.



3. Chi-Square test for association between the feature "fasting blood sugar" and the target "risk level."

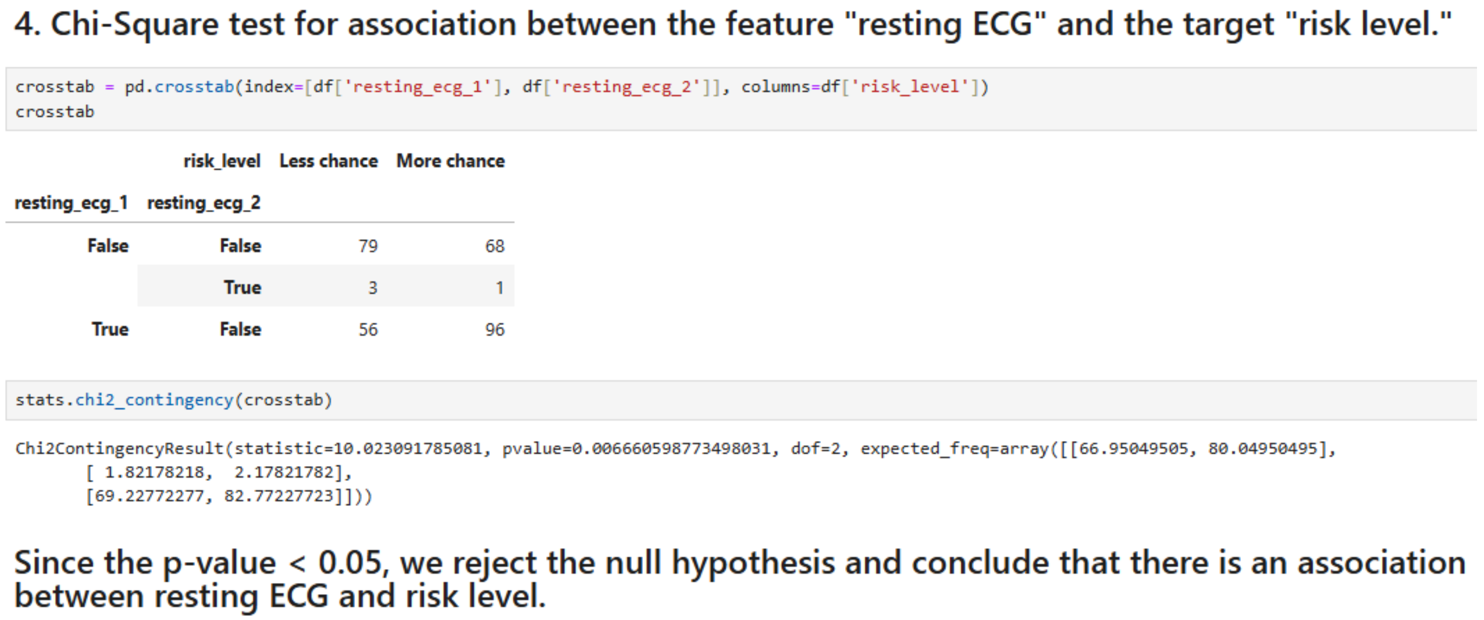
```
crosstab = pd.crosstab(df["fasting_blood_sugar_1"], df["risk_level"])
crosstab
```

risk_level	Less chance	More chance
fasting_blood_sugar_1		
False	116	142
True	22	23

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=0.10627276301947715, pvalue=0.7444281114149577, dof=1, expected_freq=array([[117.5049505, 140.4950495],
       [ 20.4950495,  24.5049505]]))
```

Since the p-value > 0.05, we do not reject the null hypothesis and conclude that there is not an association between fasting blood sugar and risk level.



4. Chi-Square test for association between the feature "resting ECG" and the target "risk level."

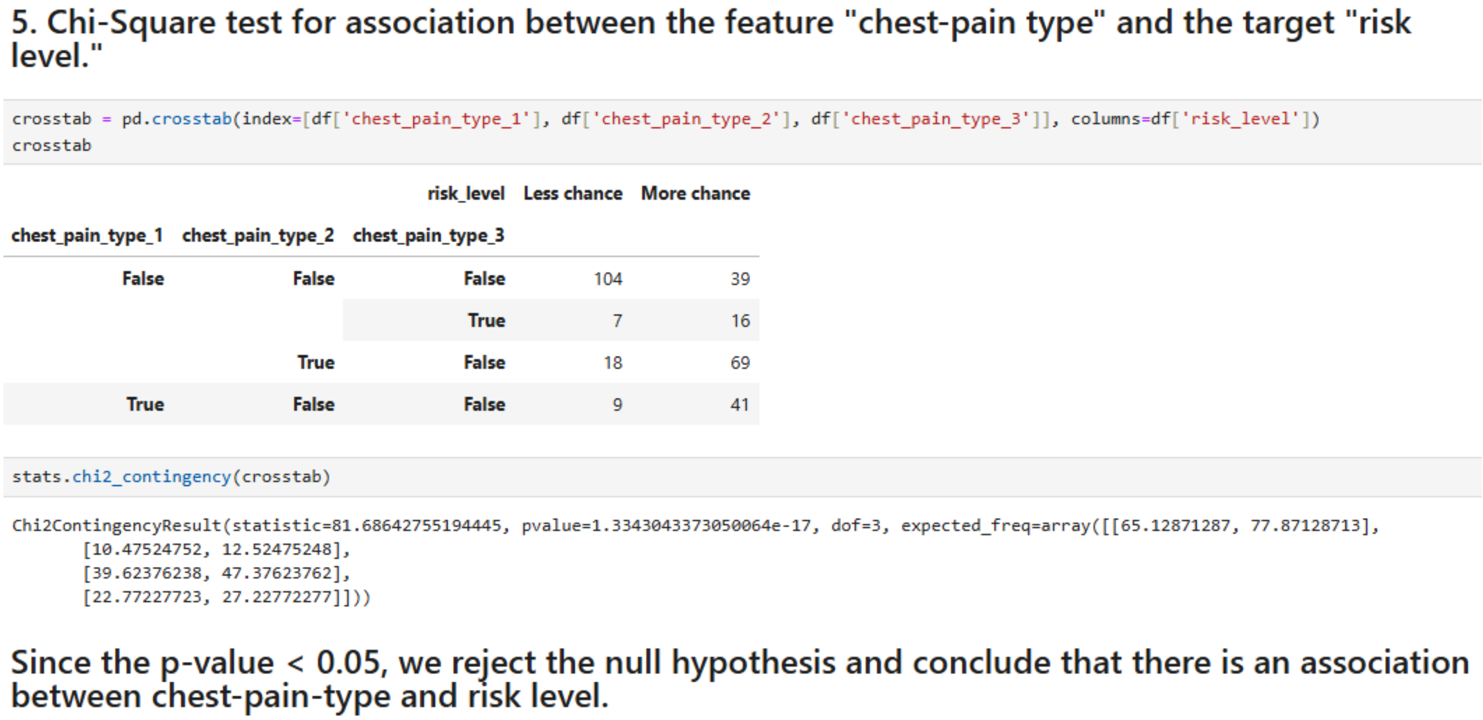
```
crosstab = pd.crosstab(index=[df['resting_ecg_1'], df['resting_ecg_2']], columns=df['risk_level'])
crosstab
```

	risk_level	Less chance	More chance
resting_ecg_1	**resting_ecg_2**		
False	**False**	79	68
	True	3	1
True	**False**	56	96

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=10.023091785081, pvalue=0.006660598773498031, dof=2, expected_freq=array([[66.95049505, 80.04950495],
       [ 1.82178218,  2.17821782],
       [69.22772277, 82.77227723]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between resting ECG and risk level.



5. Chi-Square test for association between the feature "chest-pain type" and the target "risk level."

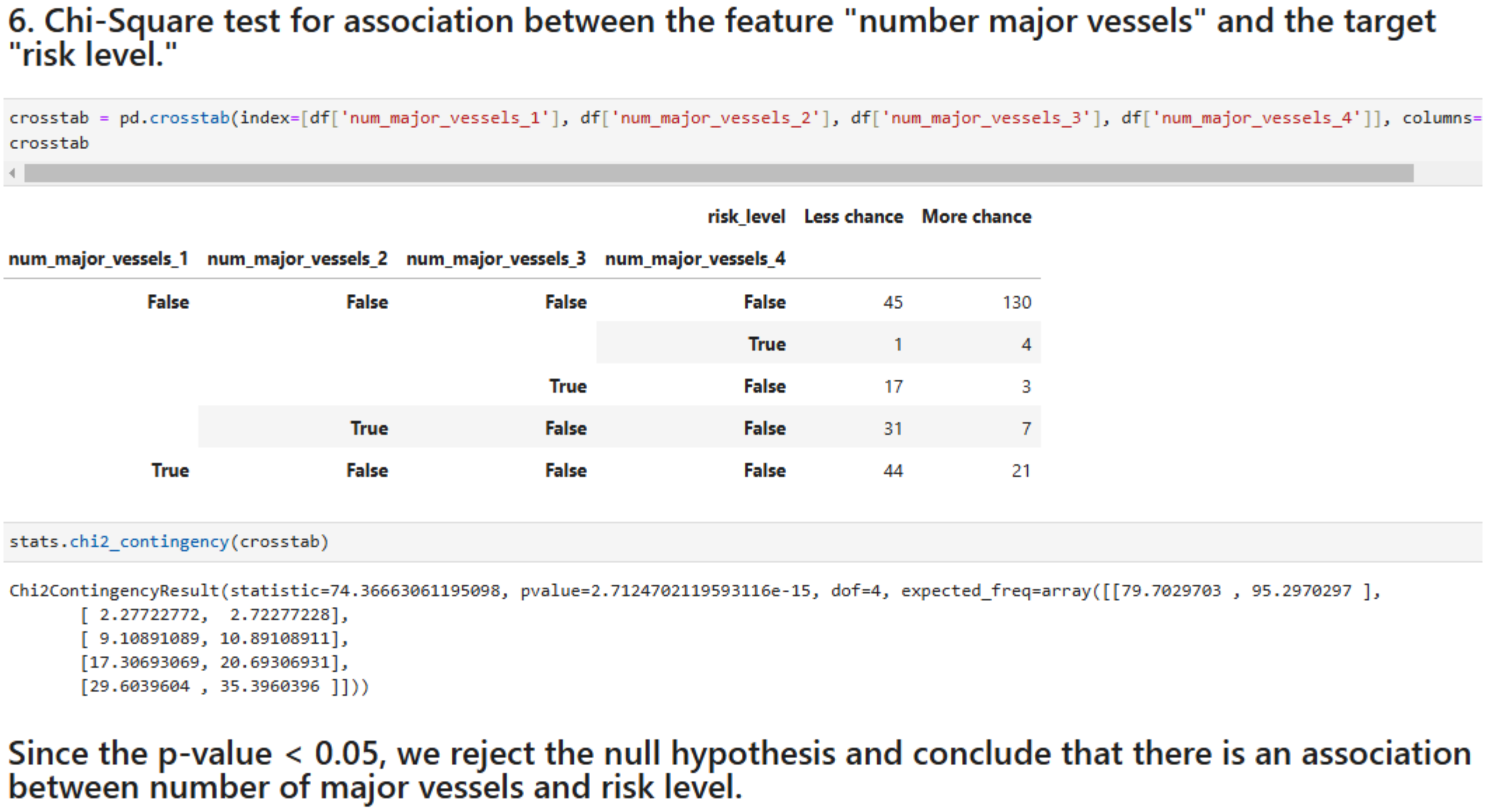
```
crosstab = pd.crosstab(index=[df['chest_pain_type_1'], df['chest_pain_type_2'], df['chest_pain_type_3']], columns=df['risk_level'])
crosstab
```

		risk_level	Less chance	More chance
chest_pain_type_1	chest_pain_type_2	chest_pain_type_3		
False	False	False	104	39
		True	7	16
	True	False	18	69
True	False	False	9	41

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=81.68642755194445, pvalue=1.3343043373050064e-17, dof=3, expected_freq=array([[65.12871287, 77.87128713],
       [10.47524752, 12.52475248],
       [39.62376238, 47.37623762],
       [22.77227723, 27.22772277]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between chest-pain-type and risk level.



6. Chi-Square test for association between the feature "number major vessels" and the target "risk level."

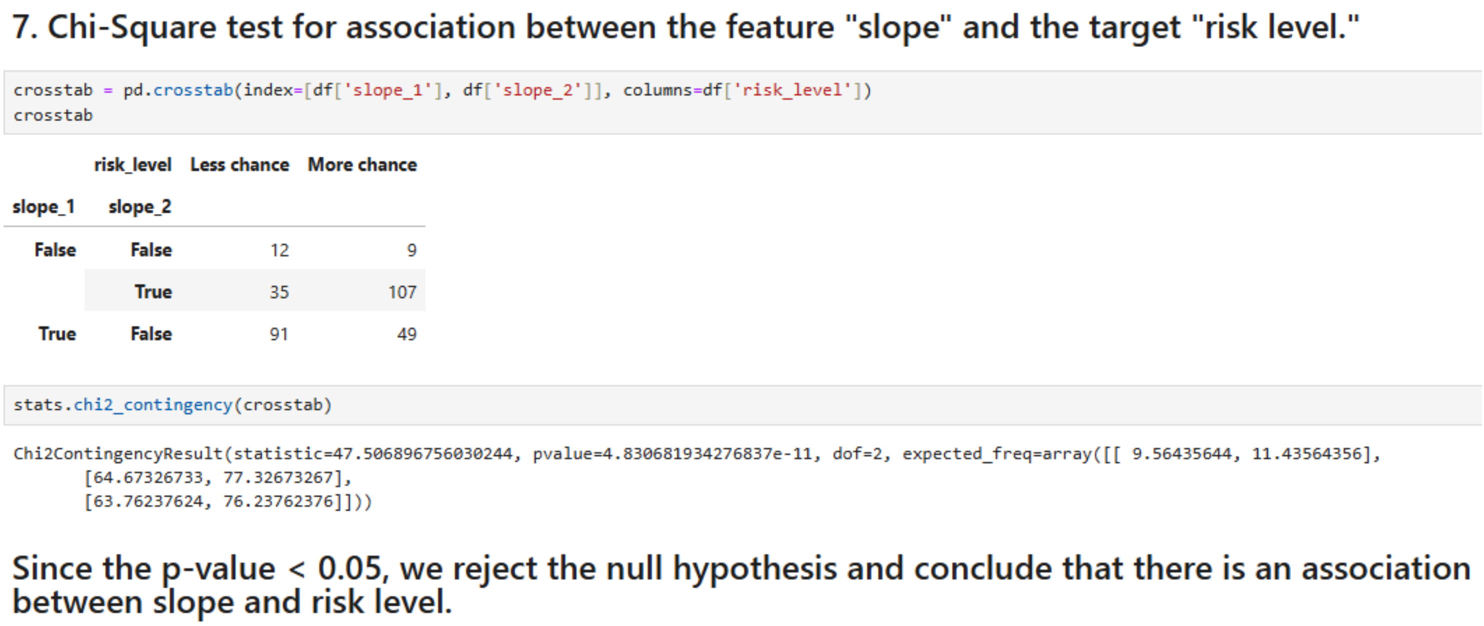
```
crosstab = pd.crosstab(index=[df['num_major_vessels_1'], df['num_major_vessels_2'], df['num_major_vessels_3'], df['num_major_vessels_4']], columns=
crosstab
```

			risk_level	Less chance	More chance
num_major_vessels_1	num_major_vessels_2	num_major_vessels_3	num_major_vessels_4		
False	False	False	False	45	130
			True	1	4
		True	False	17	3
	True	False	False	31	7
True	False	False	False	44	21

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=74.36663061195098, pvalue=2.7124702119593116e-15, dof=4, expected_freq=array([[79.7029703 , 95.2970297 ],
       [ 2.27722772,  2.72277228],
       [ 9.10891089, 10.89108911],
       [17.30693069, 20.69306931],
       [29.6039604 , 35.3960396 ]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between number of major vessels and risk level.



7. Chi-Square test for association between the feature "slope" and the target "risk level."

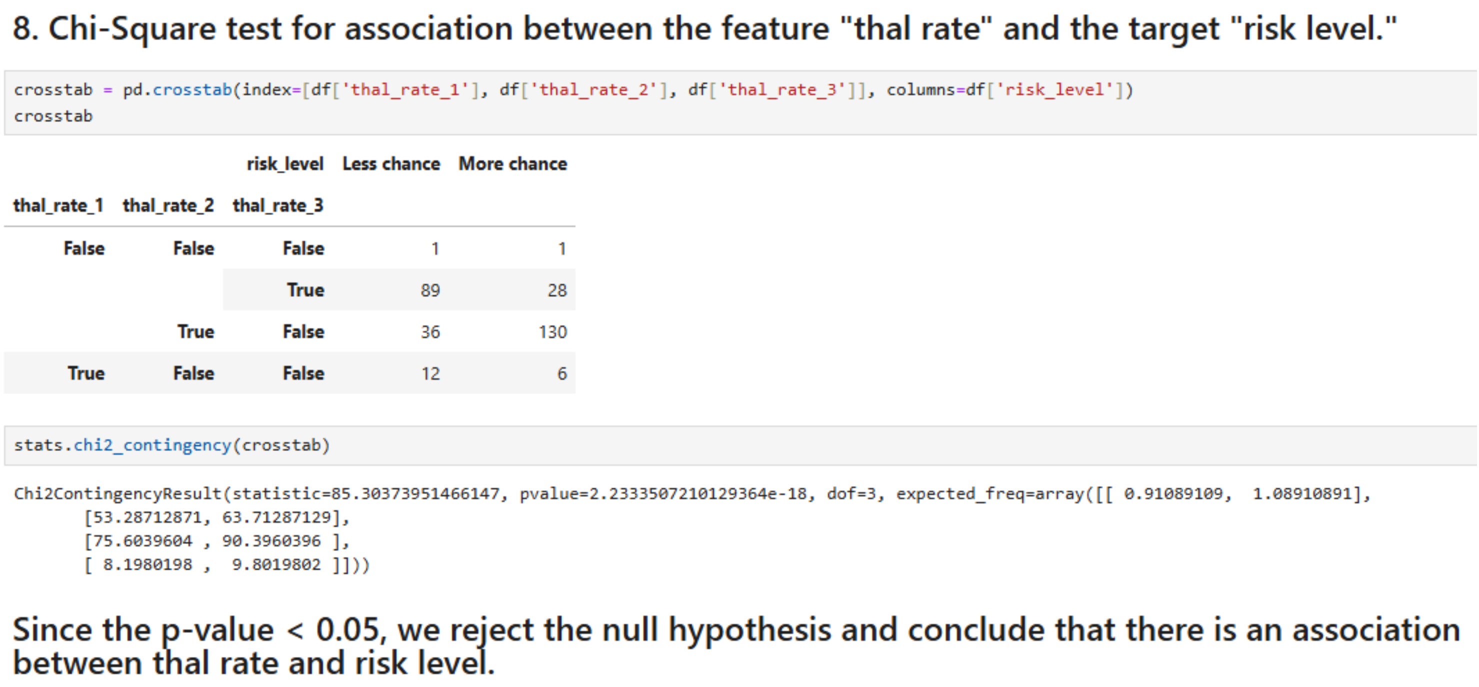
```
crosstab = pd.crosstab(index=[df['slope_1'], df['slope_2']], columns=df['risk_level'])
crosstab
```

	risk_level	Less chance	More chance
slope_1	slope_2		
False	False	12	9
	True	35	107
True	False	91	49

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=47.506896756030244, pvalue=4.830681934276837e-11, dof=2, expected_freq=array([[ 9.56435644, 11.43564356],
       [64.67326733, 77.32673267],
       [63.76237624, 76.23762376]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between slope and risk level.



8. Chi-Square test for association between the feature "thal rate" and the target "risk level."

```
crosstab = pd.crosstab(index=[df['thal_rate_1'], df['thal_rate_2'], df['thal_rate_3']], columns=df['risk_level'])
crosstab
```

		risk_level	Less chance	More chance
thal_rate_1	thal_rate_2	thal_rate_3		
False	False	False	1	1
		True	89	28
	True	False	36	130
True	False	False	12	6

```
stats.chi2_contingency(crosstab)
```

```
Chi2ContingencyResult(statistic=85.30373951466147, pvalue=2.2333507210129364e-18, dof=3, expected_freq=array([[ 0.91089109,  1.08910891],
       [53.28712871, 63.71287129],
       [75.6039604 , 90.3960396 ],
       [ 8.1980198 ,  9.8019802 ]]))
```

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is an association between thal rate and risk level.

Observation about the above tests: Only one continuous feature (cholesterol) and one categorical feature (blood sugar) showed no association with risk level, the target variable. All other features appear to have a significant association with risk level.