Data Collection and Cleaning
***Objective**: Obtain, clean, and understand the dataset.*

1. Our set contains data which can be used to predict the likelihood that a patient will suffer a heart attack. It can be found at the link: Heart Attack Analysis & Prediction Dataset
2. Here are the first 5 rows of the data as imported:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

3. There were no missing values in any of the columns.
4. I renamed the columns for clarity:

	age	sex	chest_pain_type	resting_bp	cholesterol	fasting_blood_sugar	resting_ecg	max_heart_rate
0	63	1	3	145	233	1	0	150
1	37	1	2	130	250	0	1	187
2	41	0	1	130	204	0	0	172
3	56	1	1	120	236	0	1	178
4	57	0	0	120	354	0	1	163

exercise_induced_angina	previous_peak	slope	num_major_vessels	thal_rate	risk_level
0	2.3	0	0	1	1
0	3.5	0	0	2	1
0	1.4	2	0	2	1
0	0.8	2	0	2	1
1	0.6	2	0	2	1

5. The target variable, Risk Level, has values 0 (less chance of heart attack) and 1 (more chance of heart attack)
6. The shape of the dataset is : (303, 14)
7. I separated the categorical, continuous, and target columns:
    a. The categorical columns are: 'sex', 'exercise_induced_angina', 'num_major_vessels', 'chest_pain_type', 'fasting_blood_sugar', 'resting_ecg', 'slope', 'thal_rate'
    b. The continuous columns are: 'age', 'resting_bp', 'cholesterol', 'max_heart_rate', 'previous_peak'
    c. The target column is: 'risk_level'

8. Key for the categorical features:

   - Sex: 0 = female, 1 = male
   - Exercise-induced angina: 0 = no, 1 = yes
   - Number major vessels: 0 to 4
   - Chest pain type: 0 = typical angina, 1 = atypical angina, 2 = non-anginal, 3 = asymptomatic
   - Fasting blood sugar: 0 = less than or equal to 120 mg/dl, 1 = greater than 120 mg/dl
   - Resting ECG: 0 = normal, 1 = ST-T wave normality, 2 = left ventricular hypertrophy
   - Slope = 0 to 2
   - Thal rate = Thalium stress test result 0 - 3

9. Summary statistics for the numerical features:

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
resting_bp	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
cholesterol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
max_heart_rate	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
previous_peak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2

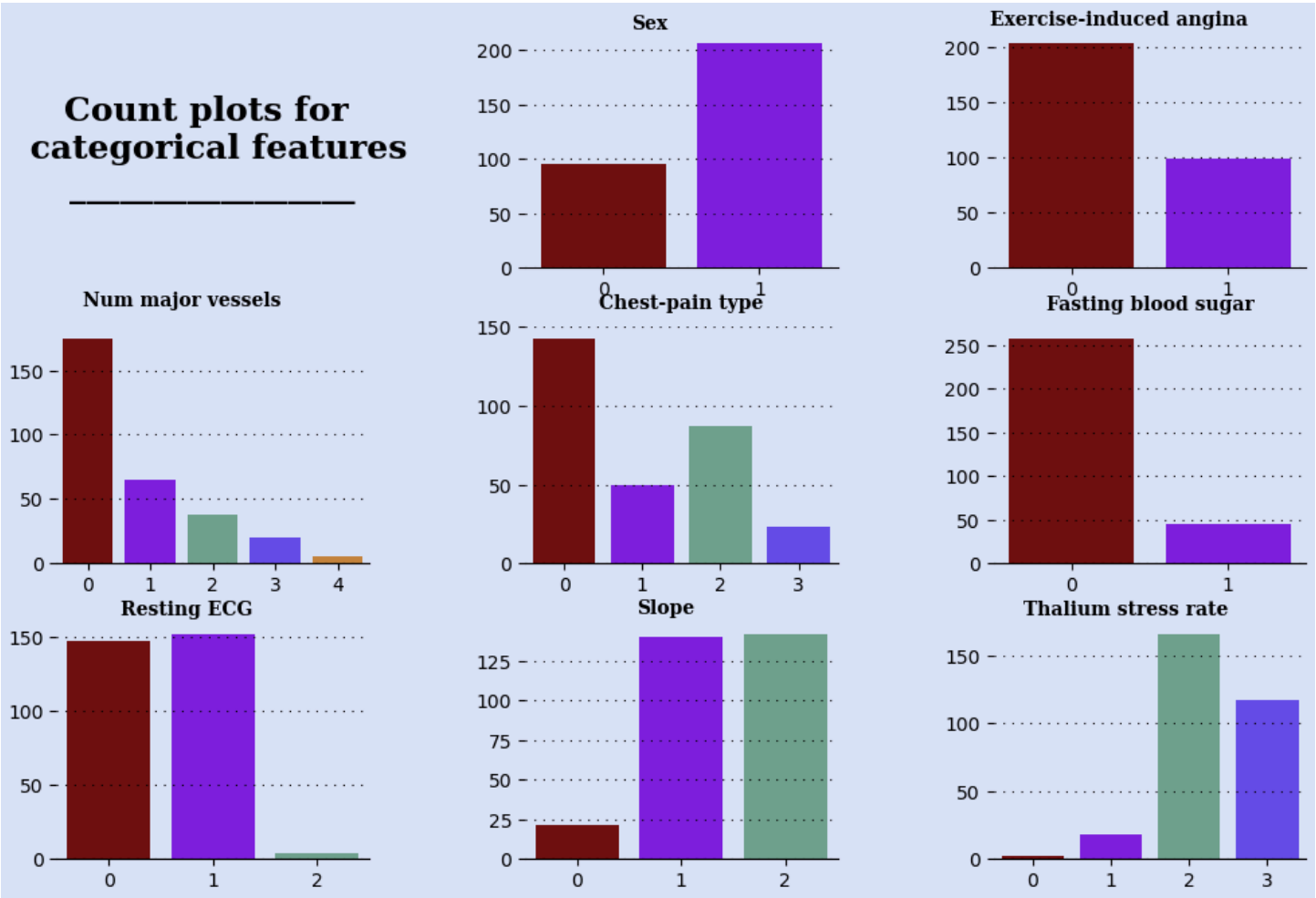
10. Data types for all features:

```
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   age                      303 non-null    int64
 1   sex                      303 non-null    int64
 2   chest_pain_type          303 non-null    int64
 3   resting_bp               303 non-null    int64
 4   cholesterol              303 non-null    int64
 5   fasting_blood_sugar      303 non-null    int64
 6   resting_ecg              303 non-null    int64
 7   max_heart_rate           303 non-null    int64
 8   exercise_induced_angina  303 non-null    int64
 9   previous_peak            303 non-null    float64
 10  slope                    303 non-null    int64
 11  num_major_vessels        303 non-null    int64
 12  thal_rate                303 non-null    int64
 13  risk_level               303 non-null    int64
dtypes: float64(1), int64(13)
```

11. Unique values for each feature:

	unique count
age	41
sex	2
chest_pain_type	4
resting_bp	49
cholesterol	152
fasting_blood_sugar	2
resting_ecg	3
max_heart_rate	91
exercise_induced_angina	2

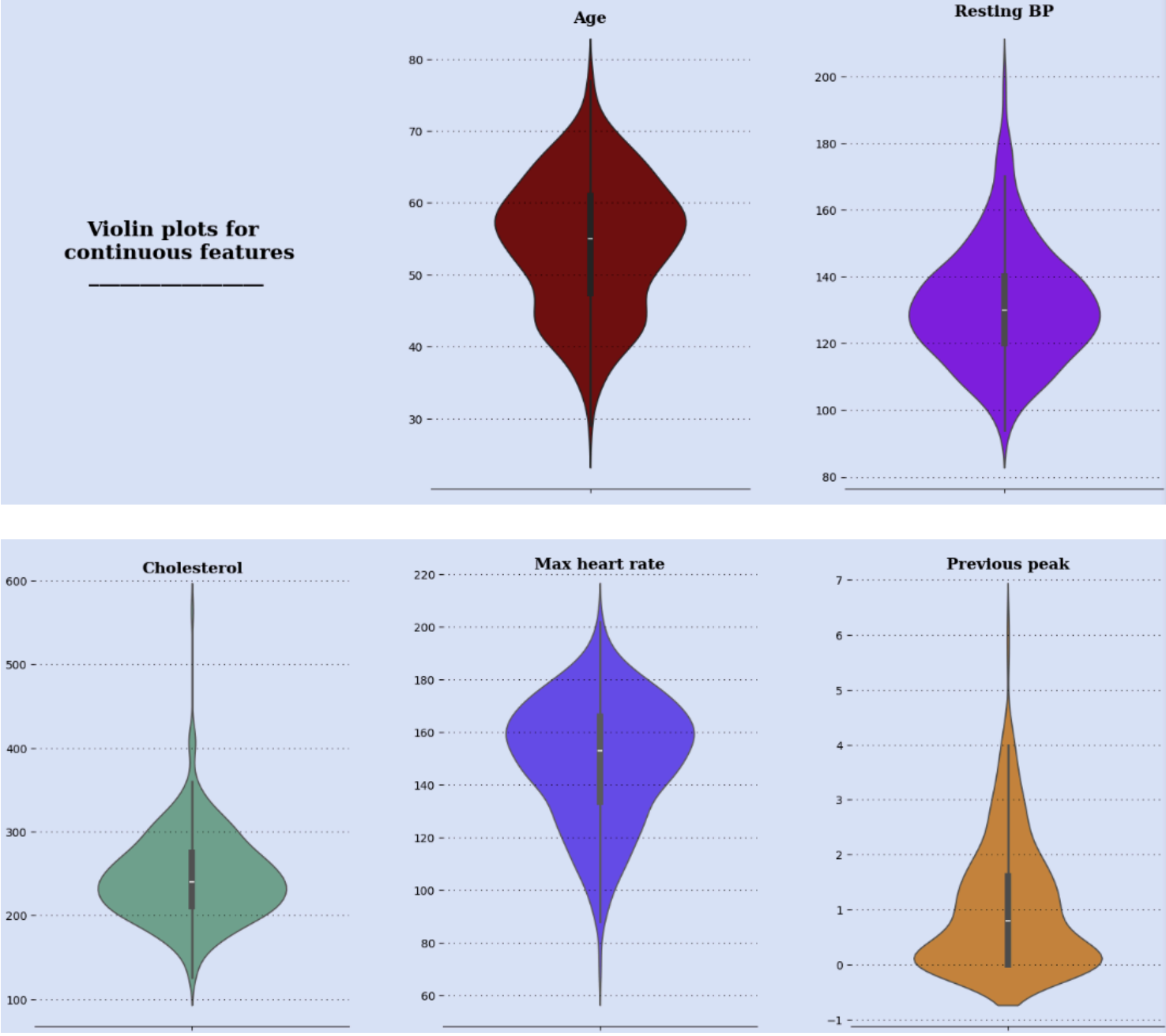
12. Count plots for categorical features:



13. Observations from the above count plots:
    a. Our data includes about twice as many males as females.
    b. Twice as many patients experienced no exercise-induced angina compared with those who did experience it.

c. Number of major vessels is very right-skewed.
d. Chest pain type is mostly typical angina, with non-anginal pain the next most frequent.
e. The vast majority of patients had fasting blood sugar below or equal to 120 mg/dl.
f. Nearly all patients had normal or ST-T wave normal resting ECGs.
g. Slopes were most often split between 1 and 2, with few 0s. This refers to the rate of increase of heart rate in stress tests.
h. Thalium stress rate is left-skewed, with most patients having higher rates, and almost none with a rate of 0.

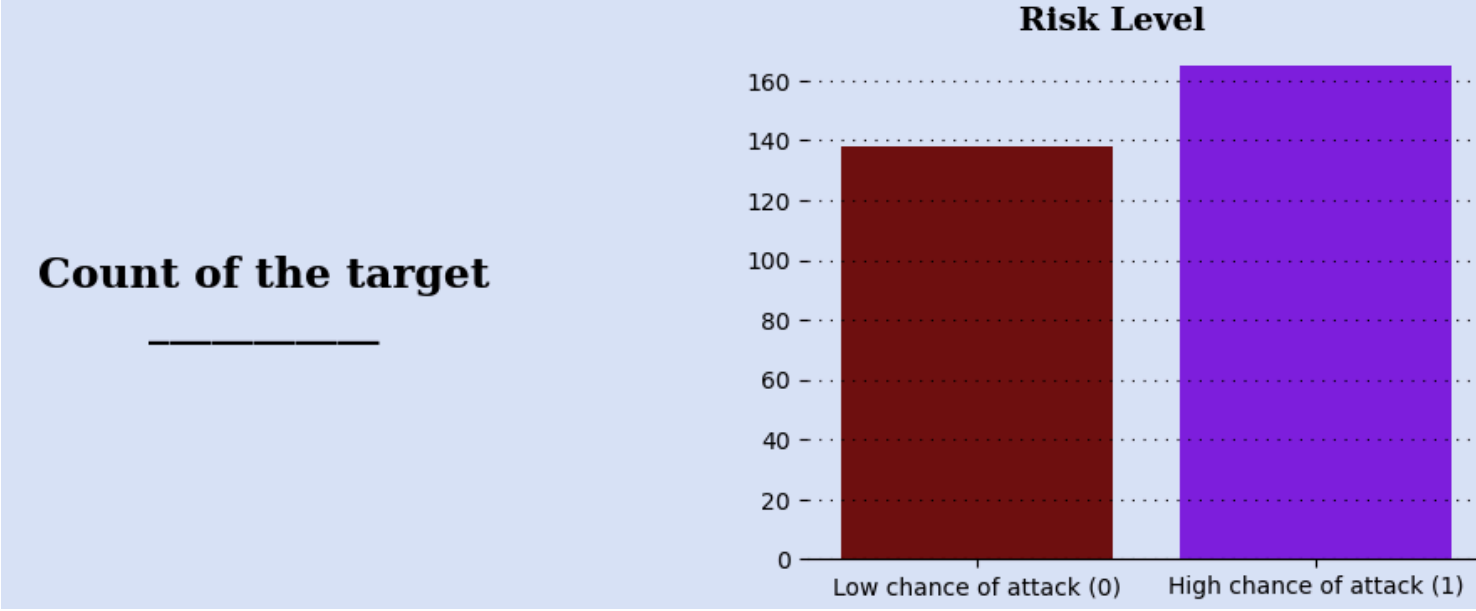
14. Violin plots of continuous features:



15. Observations from the above violin plots:
a. Most patients are between 40 and 65 years old, with outliers as low as the low 20s and the mid 80s.
b. Most patients' resting blood pressures (systolic) are between 115 and 145, with outliers as low as near 80 and as high as over 200.

c. Most patients' cholesterol levels are between 175 and 300, with outliers as low as the 90s and as high as 600.
d. The max heart rates are mostly between 120 and 180, with outliers as low as below 60 and as high as nearly 220.
e. The previous peak data is very right-skewed, with most data from near -1 to about 2.5. There outliers as high as 7.

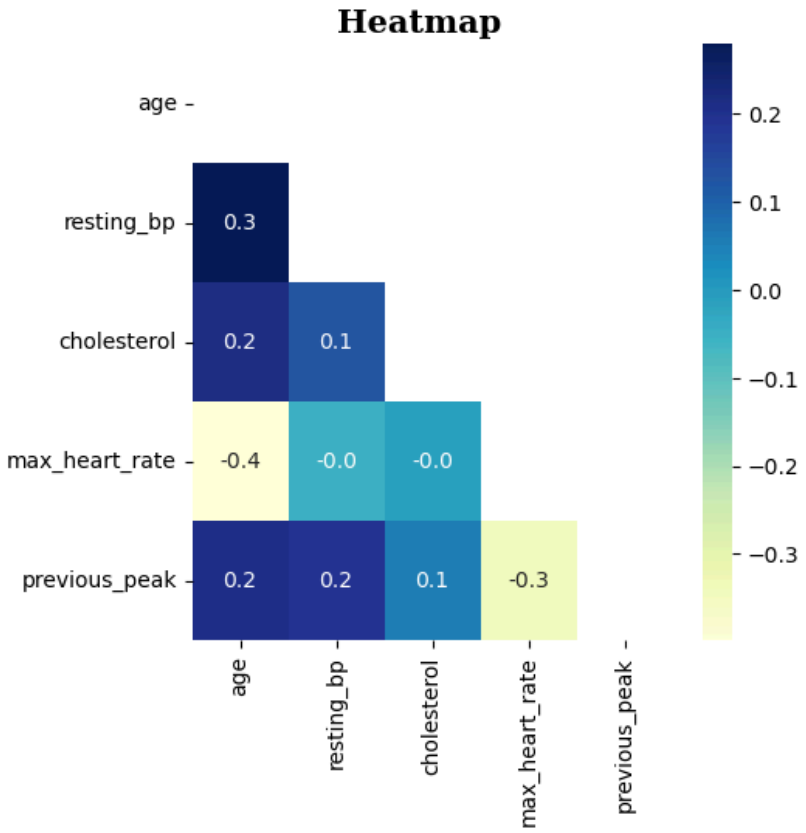
16. Count plot of the target, Risk Level:



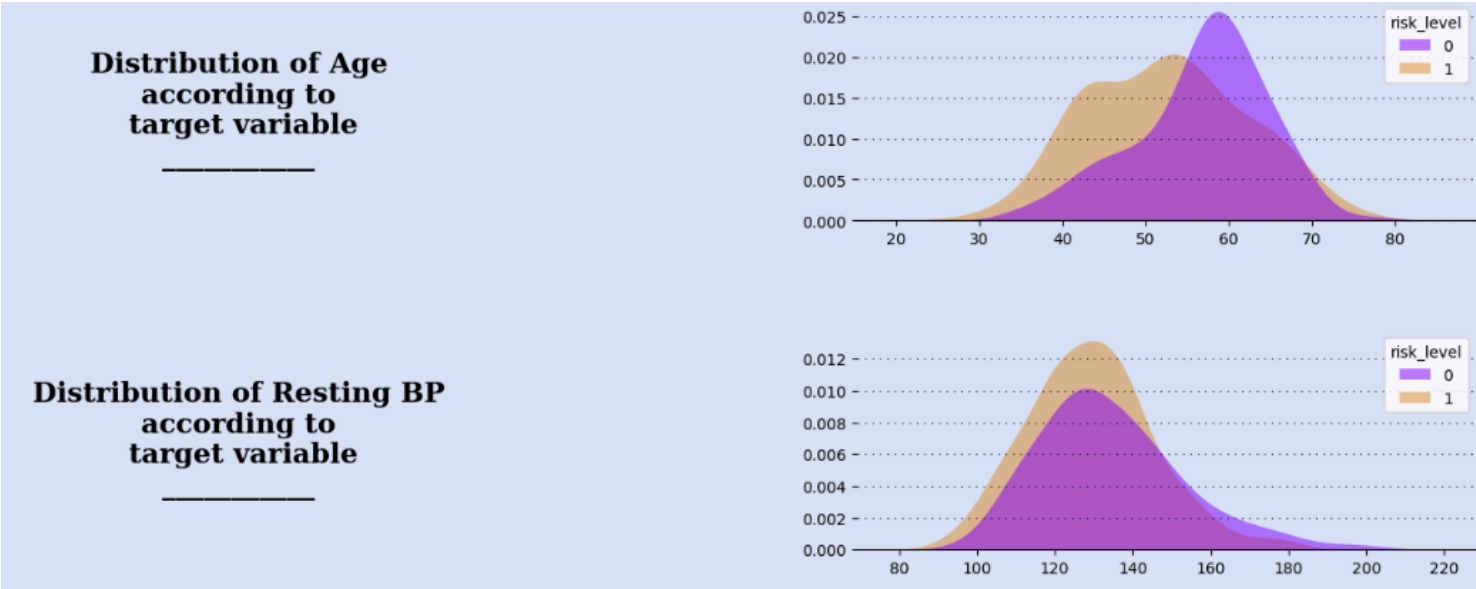
17. Observation: The data is nearly evenly split with respect to the target variable, heart attack risk level.
18. Correlation matrix of continuous features:

	age	resting_bp	cholesterol	max_heart_rate	previous_peak
age	1.000000	0.279351	0.213678	-0.398522	0.210013
resting_bp	0.279351	1.000000	0.123174	-0.046698	0.193216
cholesterol	0.213678	0.123174	1.000000	-0.009940	0.053952
max_heart_rate	-0.398522	-0.046698	-0.009940	1.000000	-0.344187
previous_peak	0.210013	0.193216	0.053952	-0.344187	1.000000

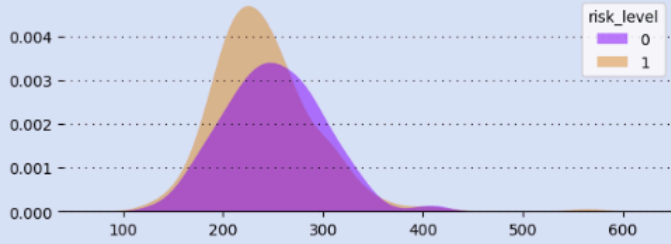
19. Heatmap of continuous features:



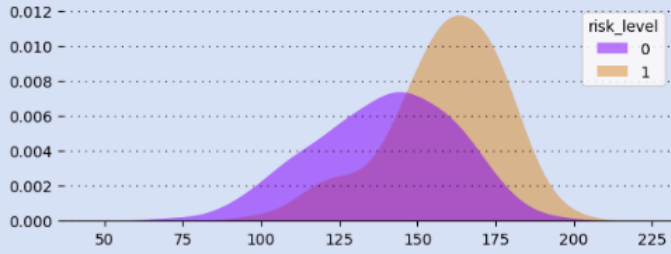
20. Observation from the correlation matrix and heatmap: There do not appear to be any significant correlations between pairs of continuous features.
21. Distribution of continuous features according to target variable:



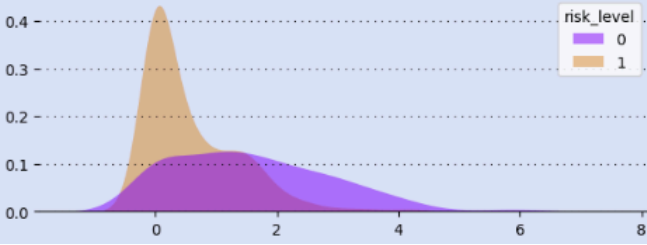
Distribution of Cholesterol according to target variable



Distribution of Max Heart Rate according to target variable



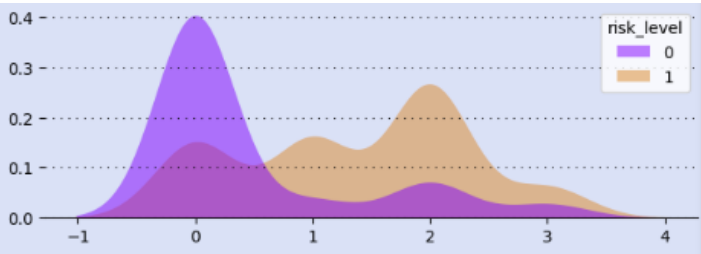
Distribution of Previous Peak according to target variable



22. Observations from the above graphs:
    a. For patients with low risk level, the graph's mode is 60 with a marked left-skew shape. For high-risk patients, the graph is more mound-shaped and with a slightly lower mode. The ranges of the graphs are similar.
    b. The resting BP graphs are very similar for both risk levels.
    c. The cholesterol graphs are also very similar, with high-risk patients actually having the peak at a lower level.
    d. High-risk patients have noticeably higher max heart rates than low-risk patients.
    e. There is a marked difference in the graphs of previous peak. Most high-risk patients score between 0 and 1. The graph for low-risk patients is mound-shaped, with a fairly uniform distribution between 0 and 3.
23. Plots of other features with respect to risk level:

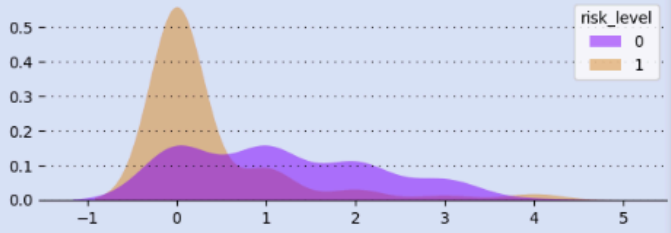
Chest pain distribution and risk level

0 - Typical Angina
1 - Atypical Angina
2 - Non-anginal Pain
3 - Asymptomatic



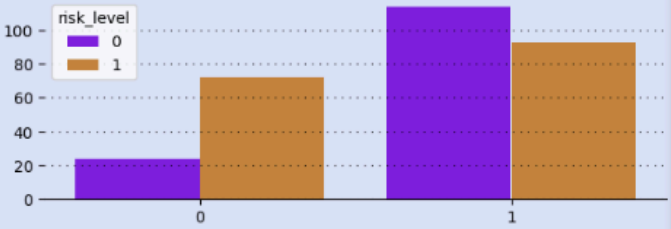
Number of major vessels and risk level

0 vessels
1 vessel
2 vessels
3 vessels
4vessels



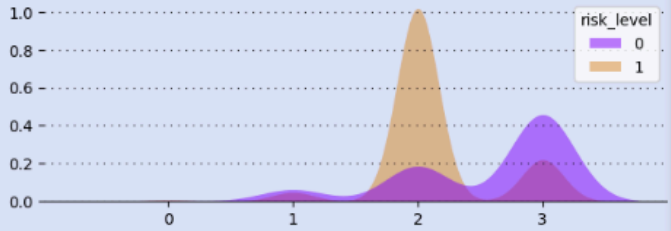
Risk level according to sex

0 - Female
1 - Male



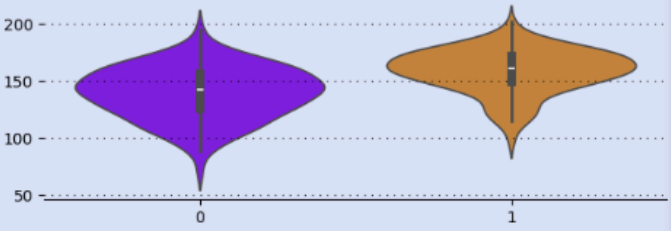
Distribution of stress test according to risk level

Thalium Stress Test Result
0, 1, 2, 3



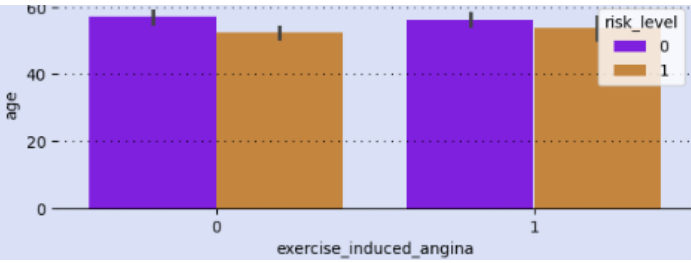
Violin plot of max heart rate and risk level

Maximum heart rate achieved



Bar plot of Exercise angina vs age

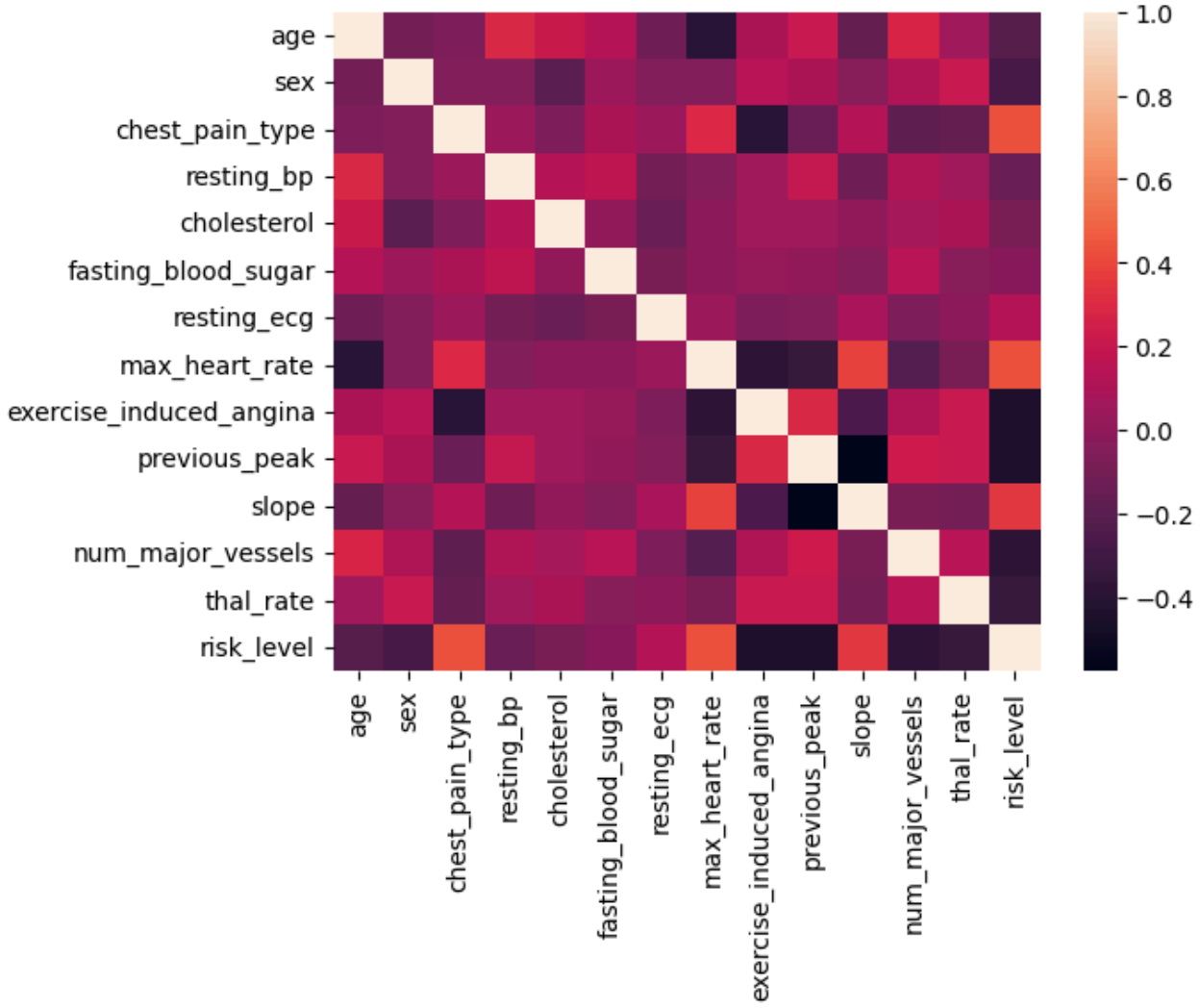
Exercise induced angina
0 - No
1 - Yes



24. Observations from the above graphs:
    a. Typical angina seems to be strongly associated with low risk, while chest pain of all types is associated with high risk, especially non-anginal pain.
    b. Similarly, 0 vessel involvement is highly associated with low risk, while high-risk seems to be only slightly right-skewed with respect to vessels.
    c. Females have three times high risk as low risk, while males are more nearly evenly distributed between high and low.

d. A thalium stress test result of 2 is highly associated with high risk, while 3 is less so. Low-risk levels increase as stress test scores increase, but only slightly until a score of 3 is reached.
e. The distributions for max heart rate achieved are similar, though the high-risk group has a higher median and a smaller range.
f. Exercise-induced angina appears independent of age and risk-level.

25. Heatmap of all features and target:



26. Observation from the heatmap: Sex, exercise-induced angina, previous peak, and number of major vessels have the most correlation with risk-level.