💕💔 Heart Disease EDA & Prediction 🔮

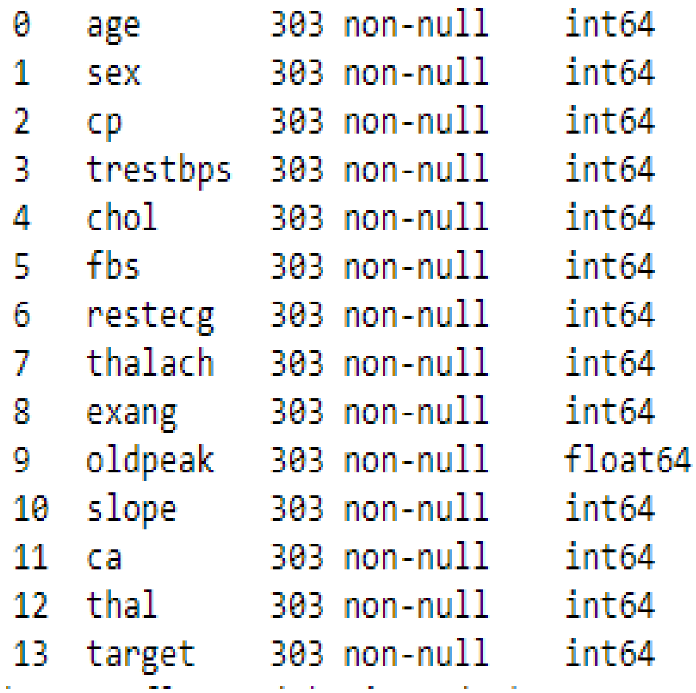
1. | INTODUCTION

DATASET PROBLEM

- This dataset contains information about diagnosis of heart disease patients. Machine learning model is needed to determine whether a person has heart disease or not

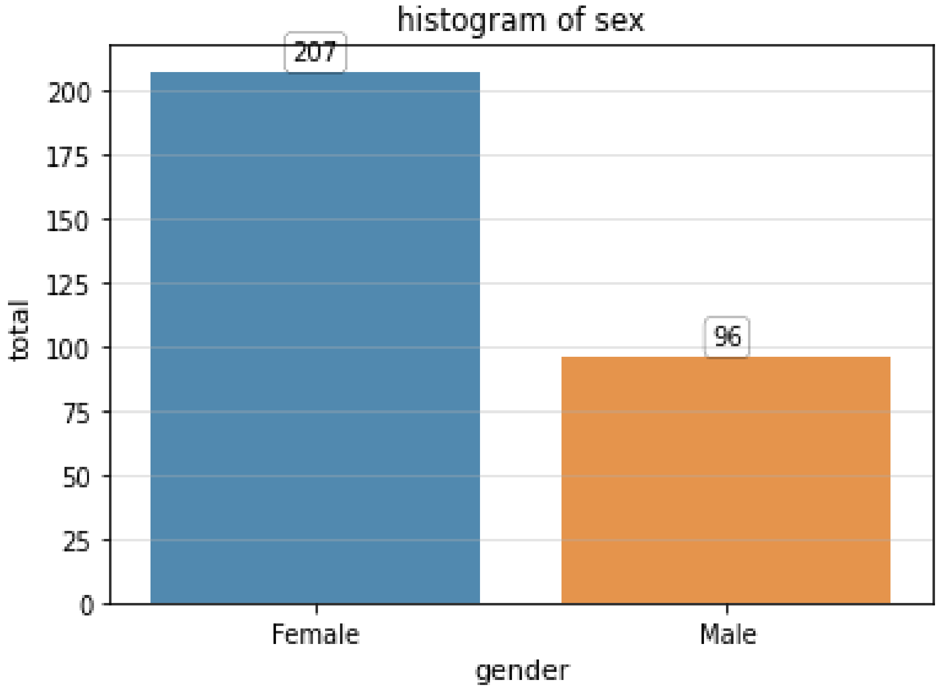
DATASET DISCRIPTION

- There are total 14 variables in this dataset
- 9 categorical variables, and
- 5 continuous variables.
- There are total 303 Rows.
- There are total 14 Columns.



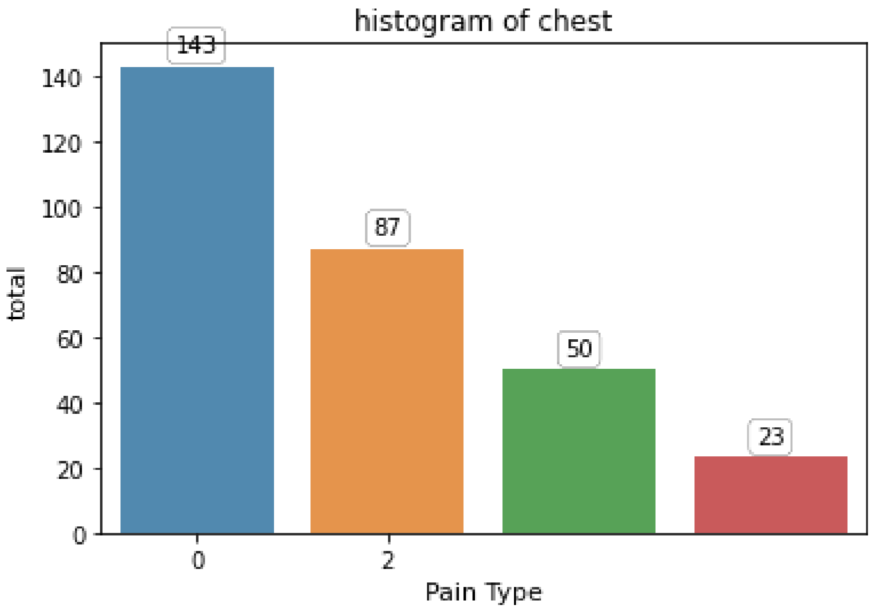
```
0   age       303 non-null    int64
1   sex       303 non-null    int64
2   cp        303 non-null    int64
3   trestbps  303 non-null    int64
4   chol      303 non-null    int64
5   fbs       303 non-null    int64
6   restecg   303 non-null    int64
7   thalach   303 non-null    int64
8   exang     303 non-null    int64
9   oldpeak   303 non-null    float64
10  slope     303 non-null    int64
11  ca        303 non-null    int64
12  thal      303 non-null    int64
13  target    303 non-null    int64
```

SEX



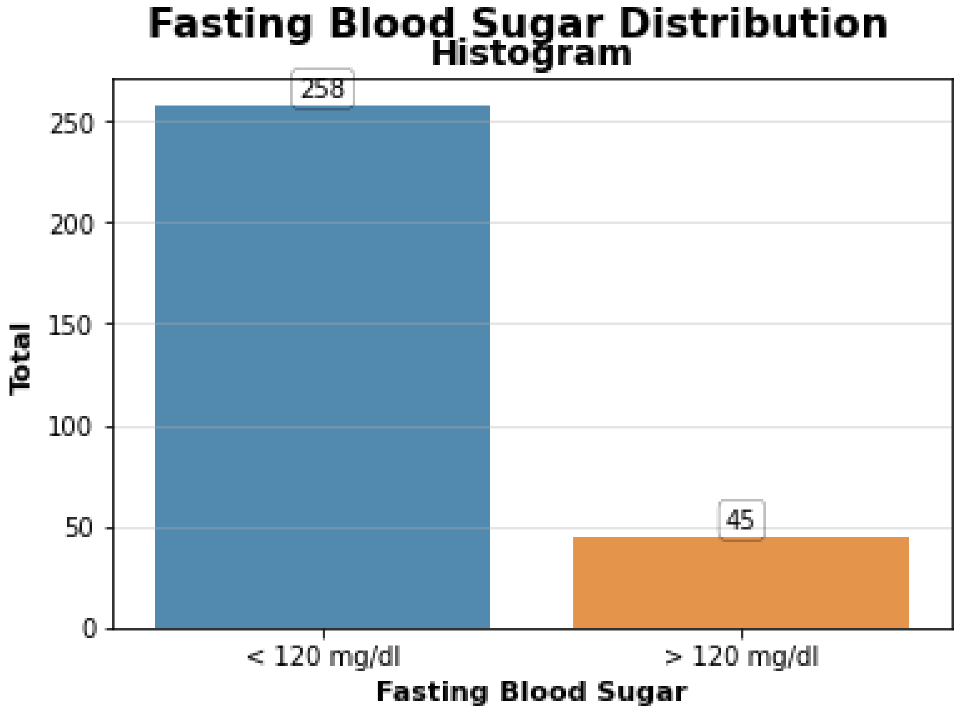
- Distribution of female patients are higher compared to male patients

CHEST



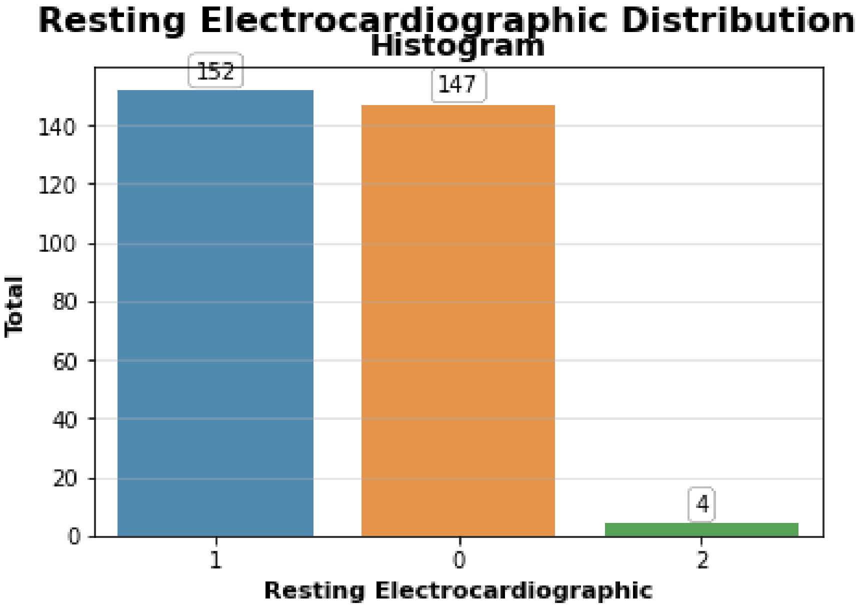
- Chest pain type 0 have the highest number compared to other types of chest pain.

fbs(FASTING BLOOD SUGAR)



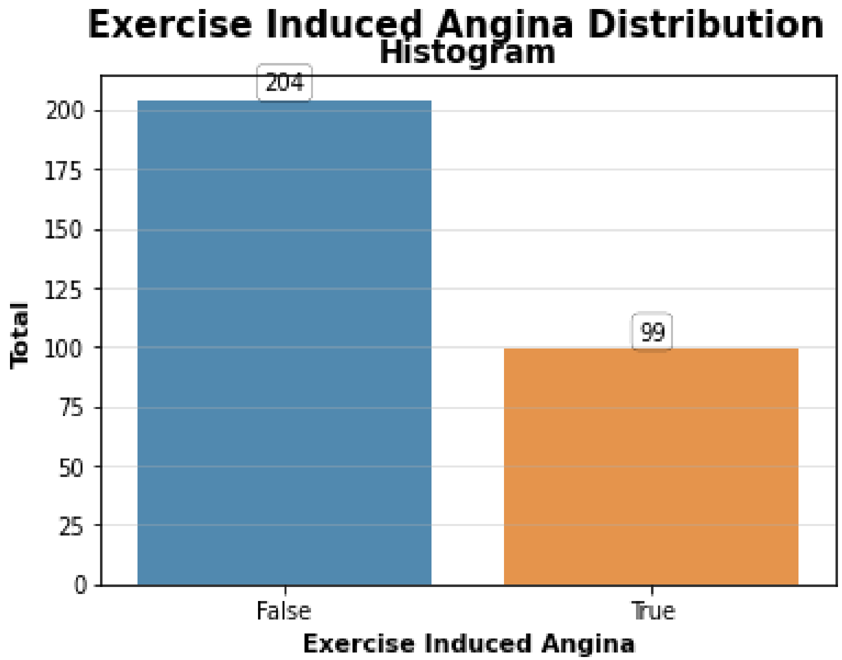
- It can be seen that the number of patients with fasting blood sugar less than 120 mg/dl have the highest numbers.

restecg(Resting Electrocardiographic Results)



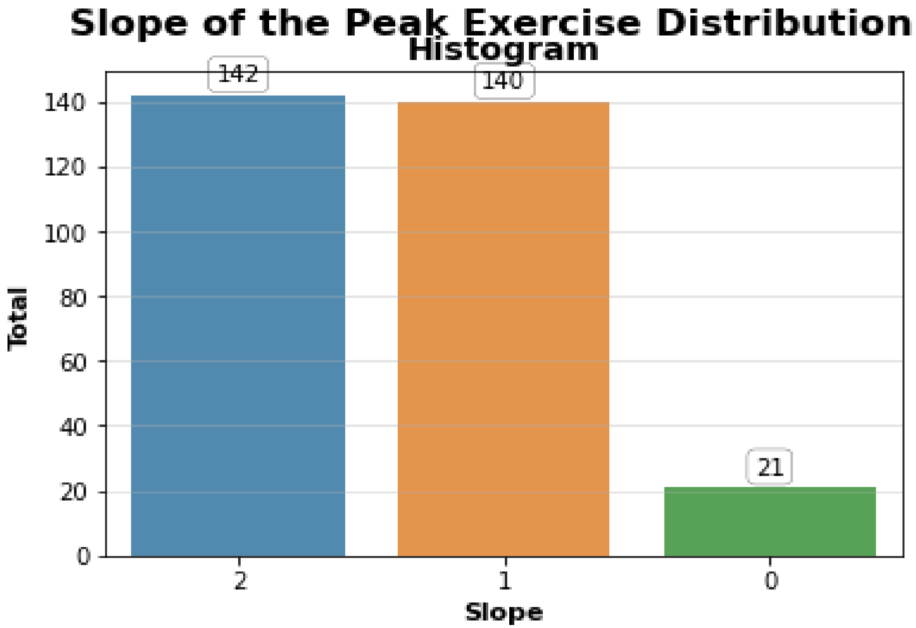
- Resting electrocardiographic with results 1 and 0 has a higher distribution than result 2
- In addition, result 1 has the highest distribution compared to other results.

Exang(exercise induced angina)



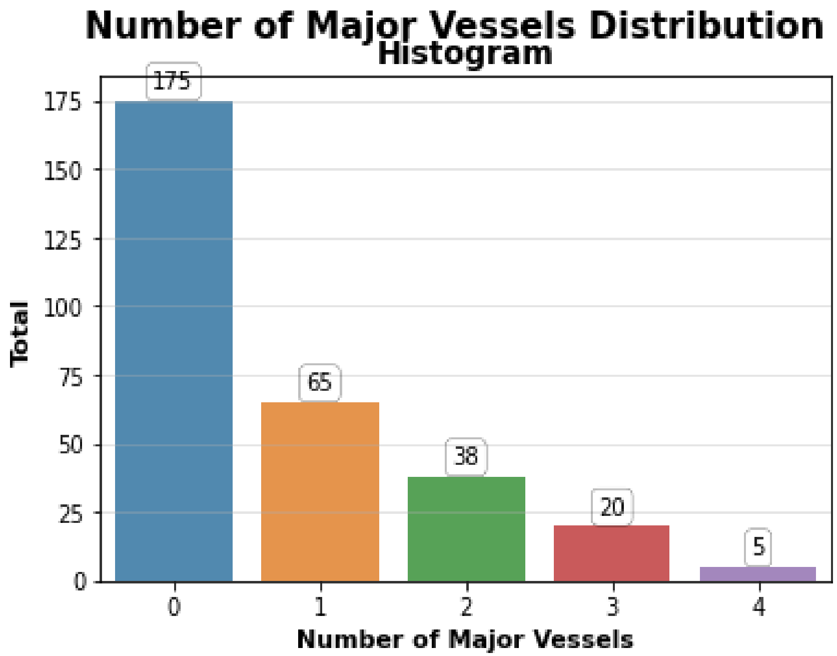
- Patients with no exercise induced angina are the highest compared to patients with exercise induced angina

Slope(slope of the peak exercise)



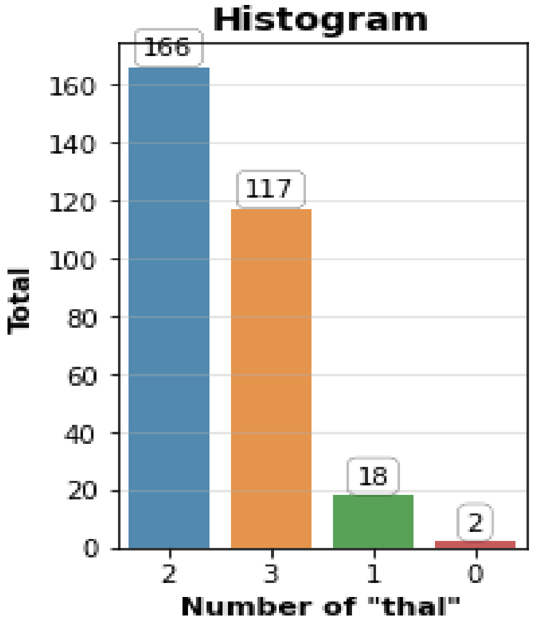
- The distribution of slope 1 and 2 are almost the same.
- Moreover, slope 2 has the highest distribution compared to others.

Ca(number of major vassels)



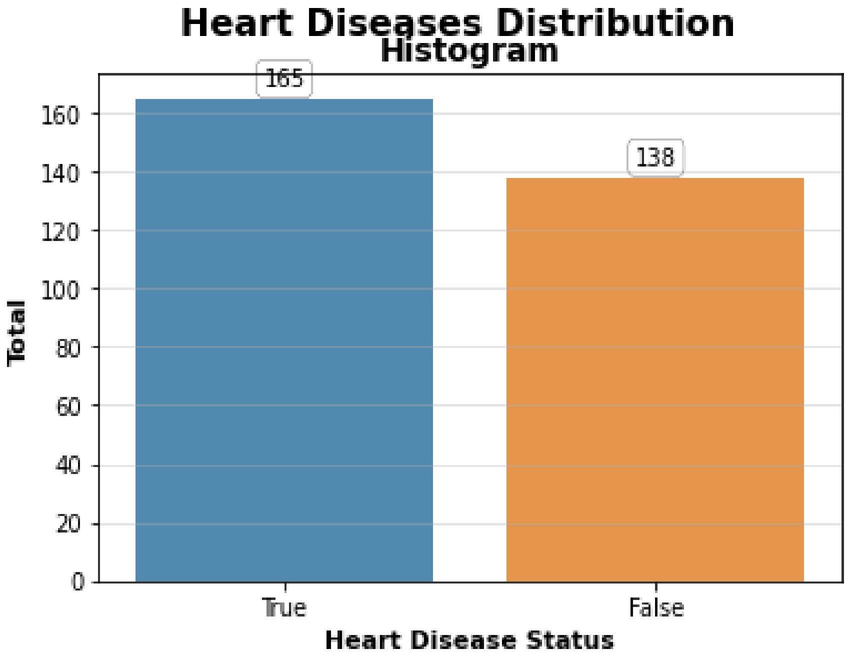
- People with 0 major vessel has the highest distribution compared to others

thal



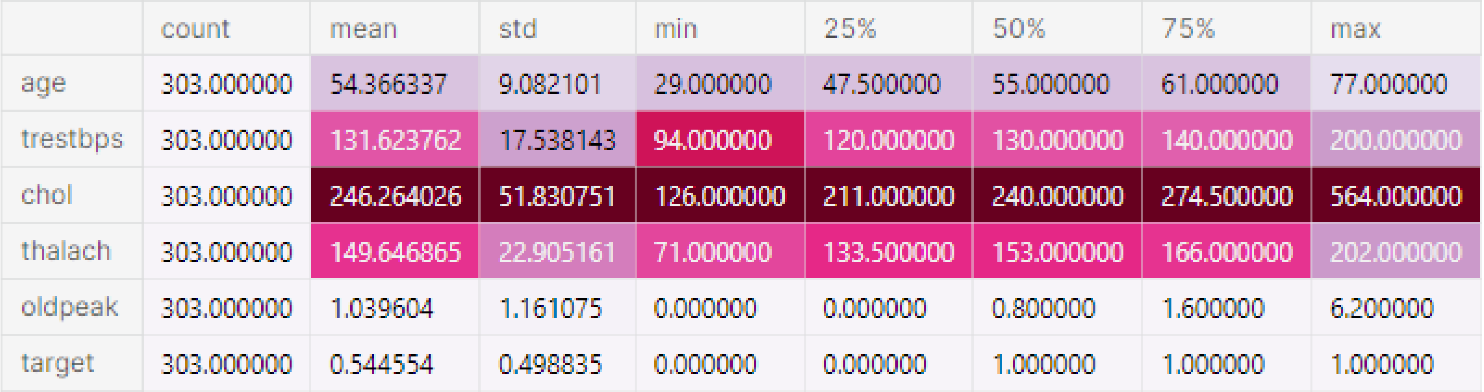
- Patients with 2 thal has the highest distribution compared to others.

Target(Heart diseases Status)



- The total number of patients that have heart diseases are higher than patients that have no heart diseases.

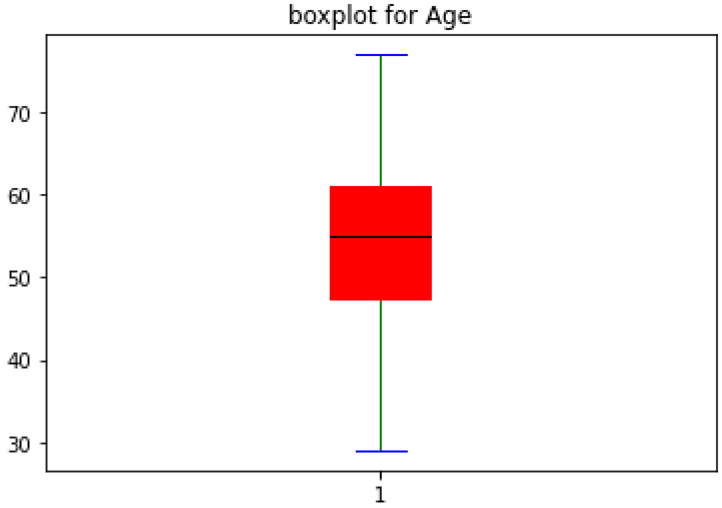
Descriptive statistics

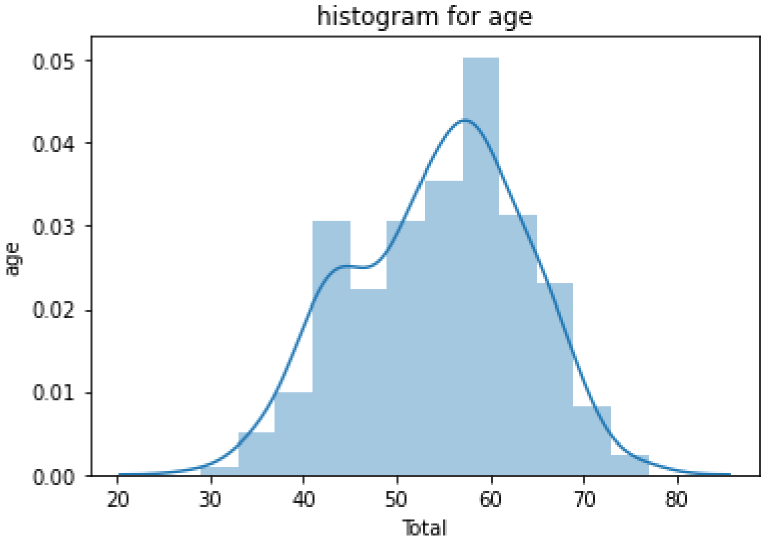


	count	mean	std	min	25%	50%	75%	max
age	303.000000	54.366337	9.082101	29.000000	47.500000	55.000000	61.000000	77.000000
trestbps	303.000000	131.623762	17.538143	94.000000	120.000000	130.000000	140.000000	200.000000
chol	303.000000	246.264026	51.830751	126.000000	211.000000	240.000000	274.500000	564.000000
thalach	303.000000	149.646865	22.905161	71.000000	133.500000	153.000000	166.000000	202.000000
oldpeak	303.000000	1.039604	1.161075	0.000000	0.000000	0.800000	1.600000	6.200000
target	303.000000	0.544554	0.498835	0.000000	0.000000	1.000000	1.000000	1.000000

- From the descriptive statistics it can be seen that age, resting blood pressure, cholestoral, and thalach are lack variation.

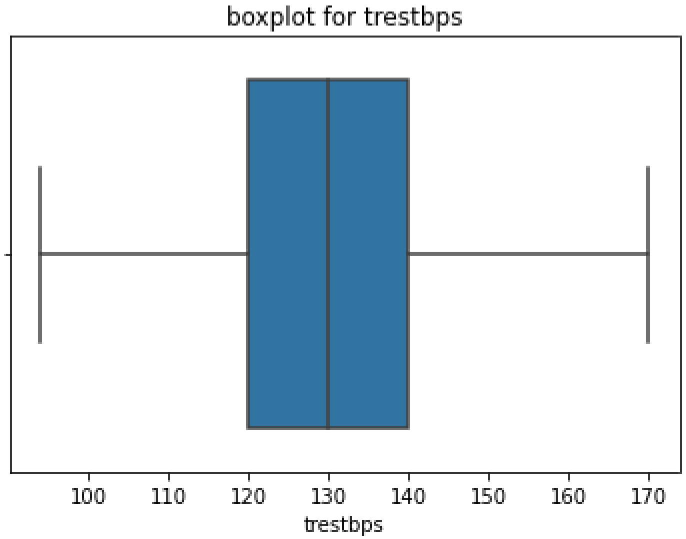
Age(patient Age)

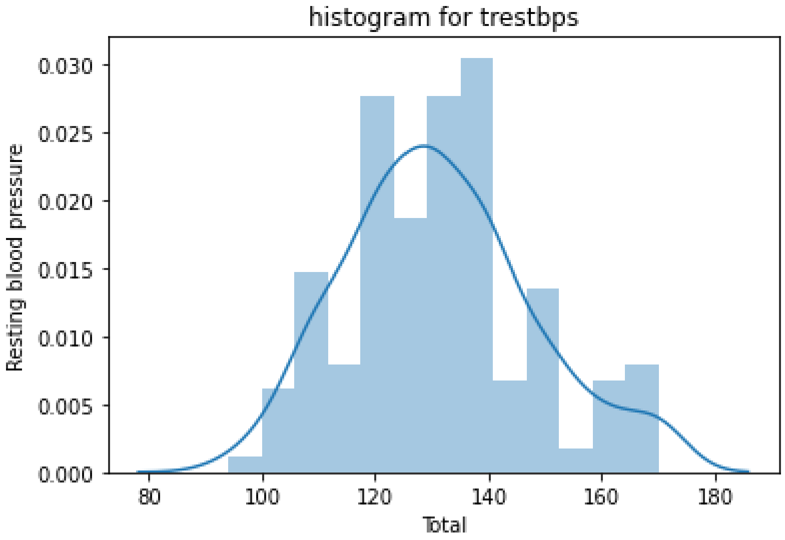




- From the histogram and boxplot, it can be seen that this column is normally distributed.

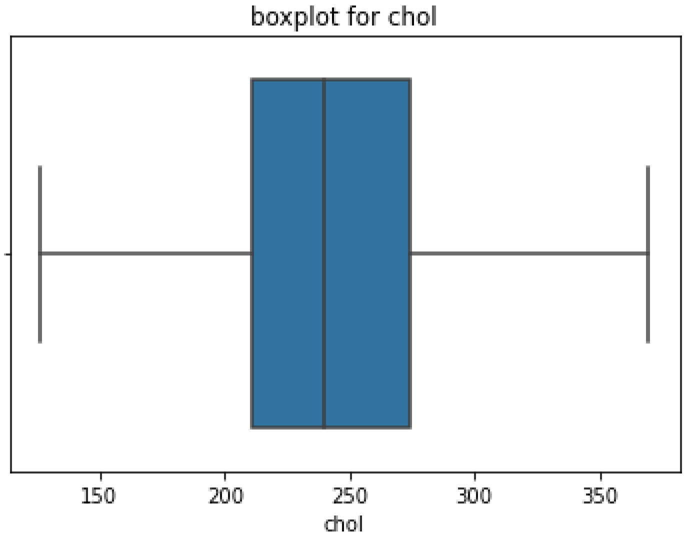
trestbps(Resting blood Pressure in mm Hg)

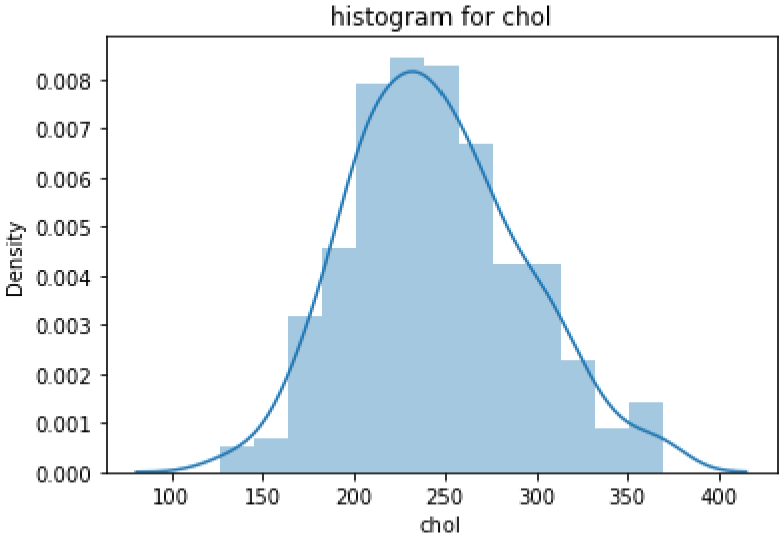




- From histogram it can be shown that this column is moderatly right skewed

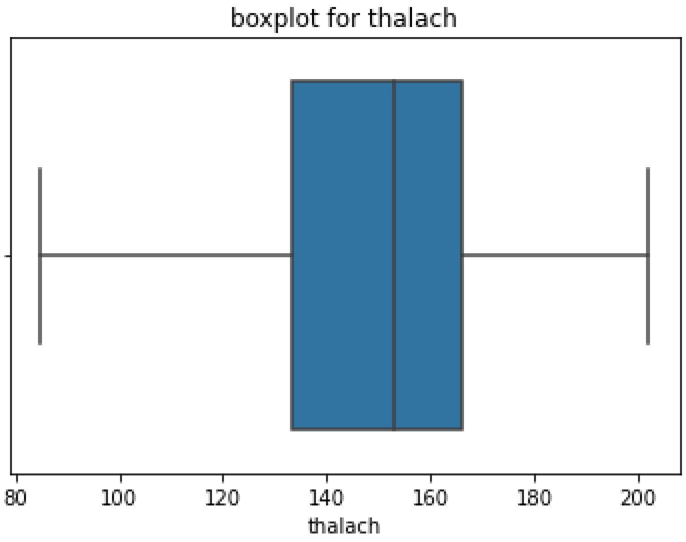
Chol(Serum Cholestral in mg/dl)

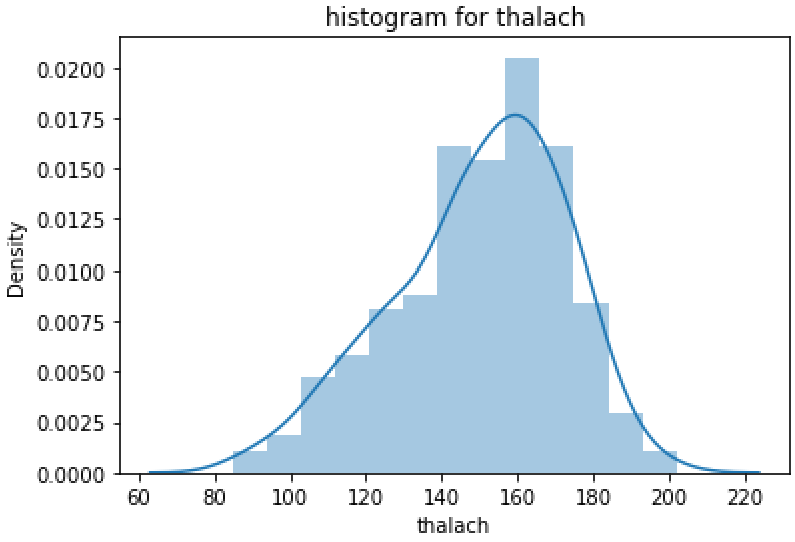




- From histogram it can be shown that these column is highly right skewed.

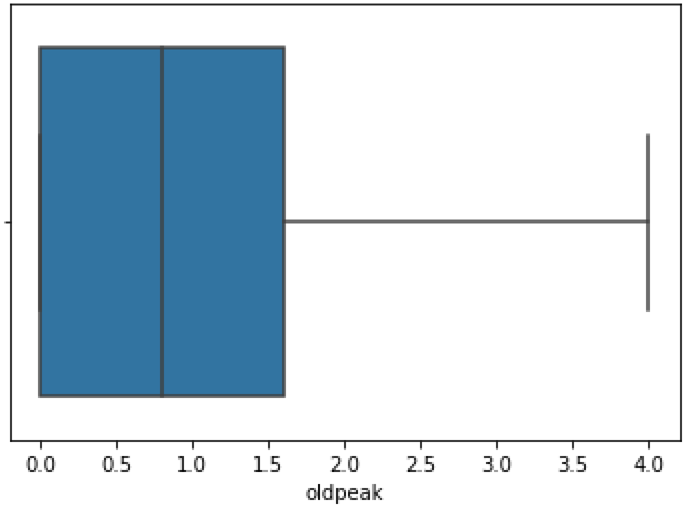
Thalach(Maximum heart rate)

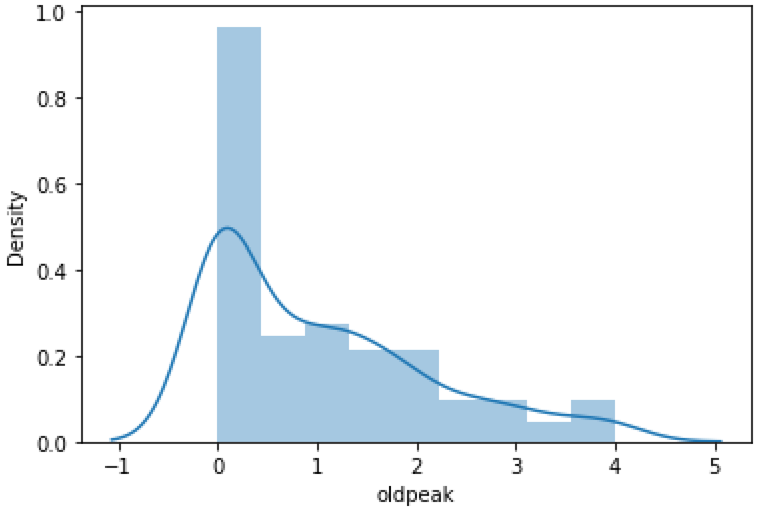




- From the histogram it can be seen that this column is moderatly left skewed.

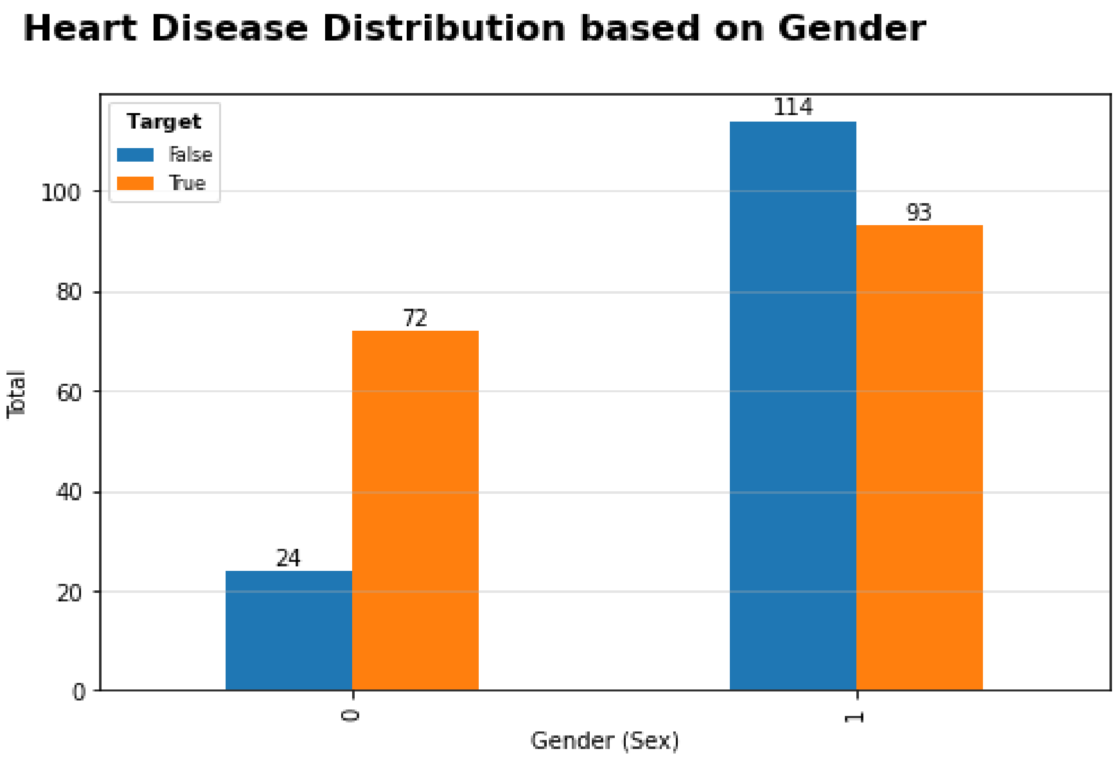
oldpeak



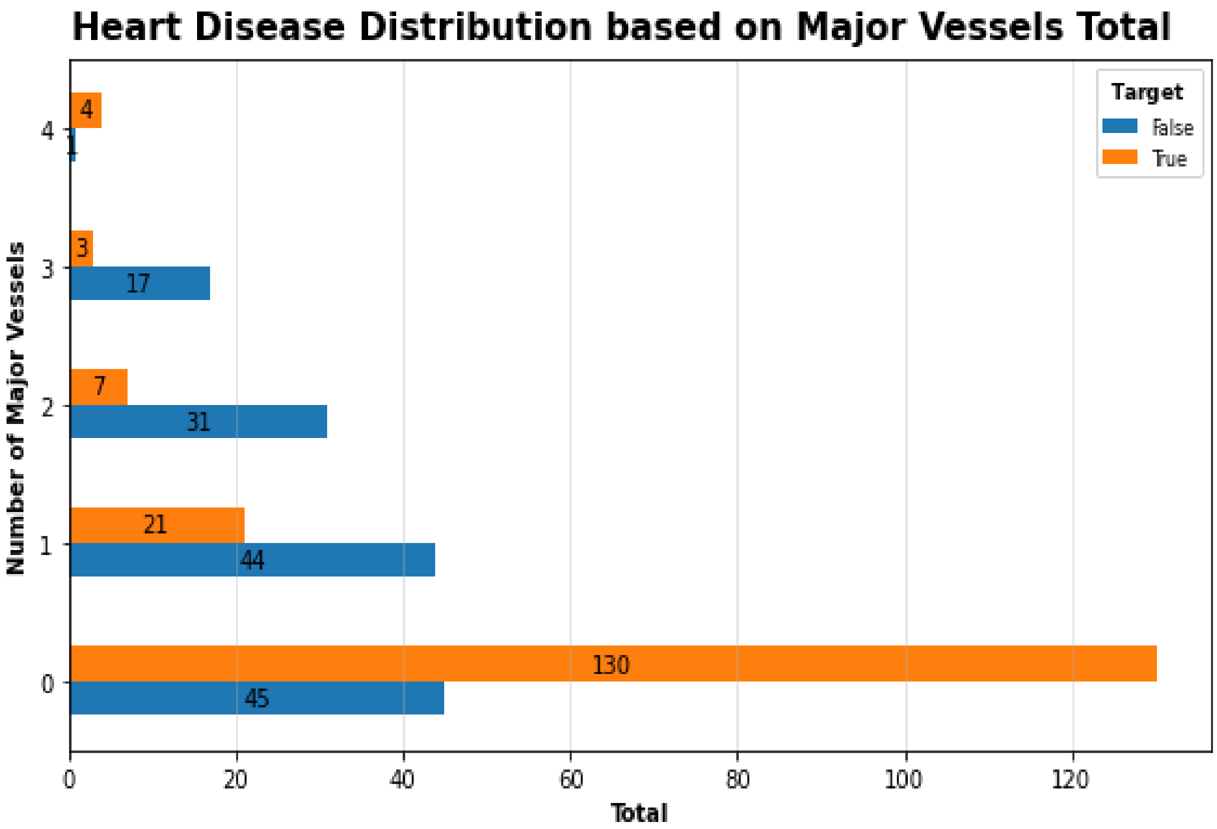


- From the histogram it can be seen that this column is highly right skewed.

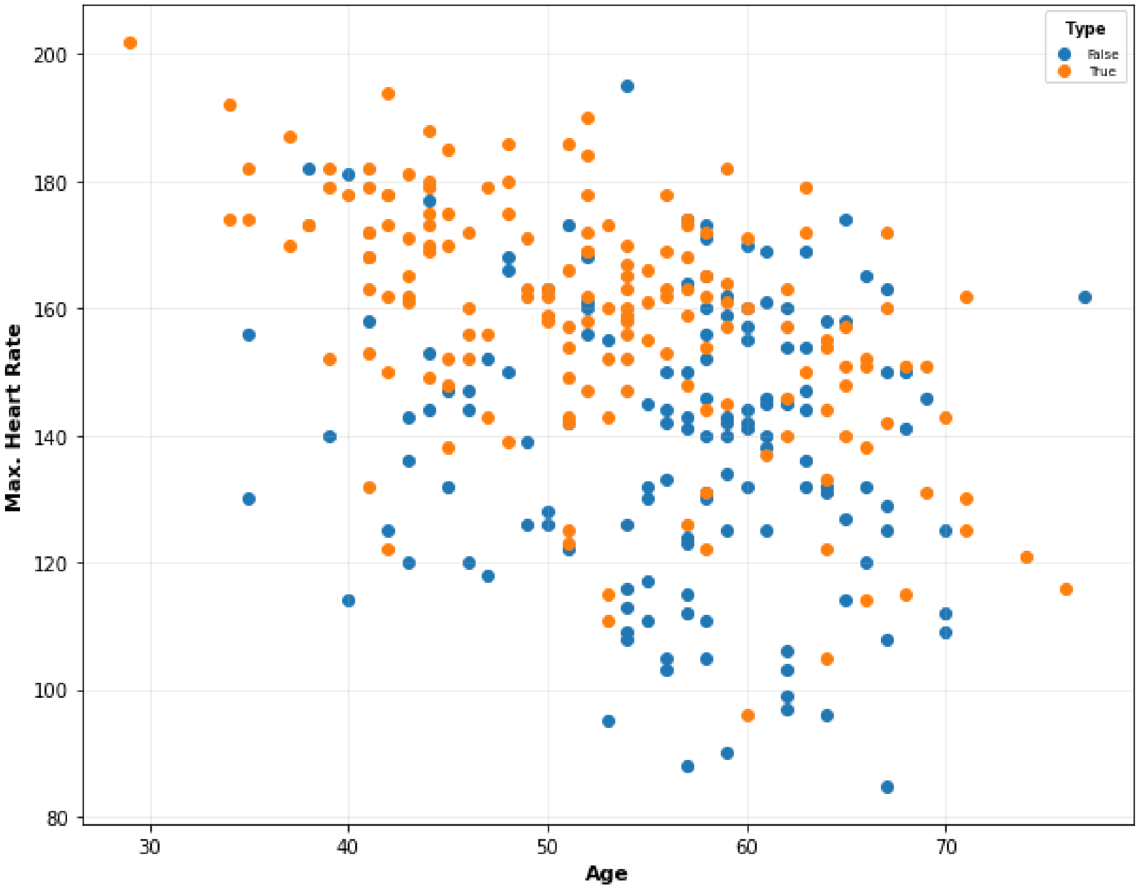
Heart disease distribution based on gender



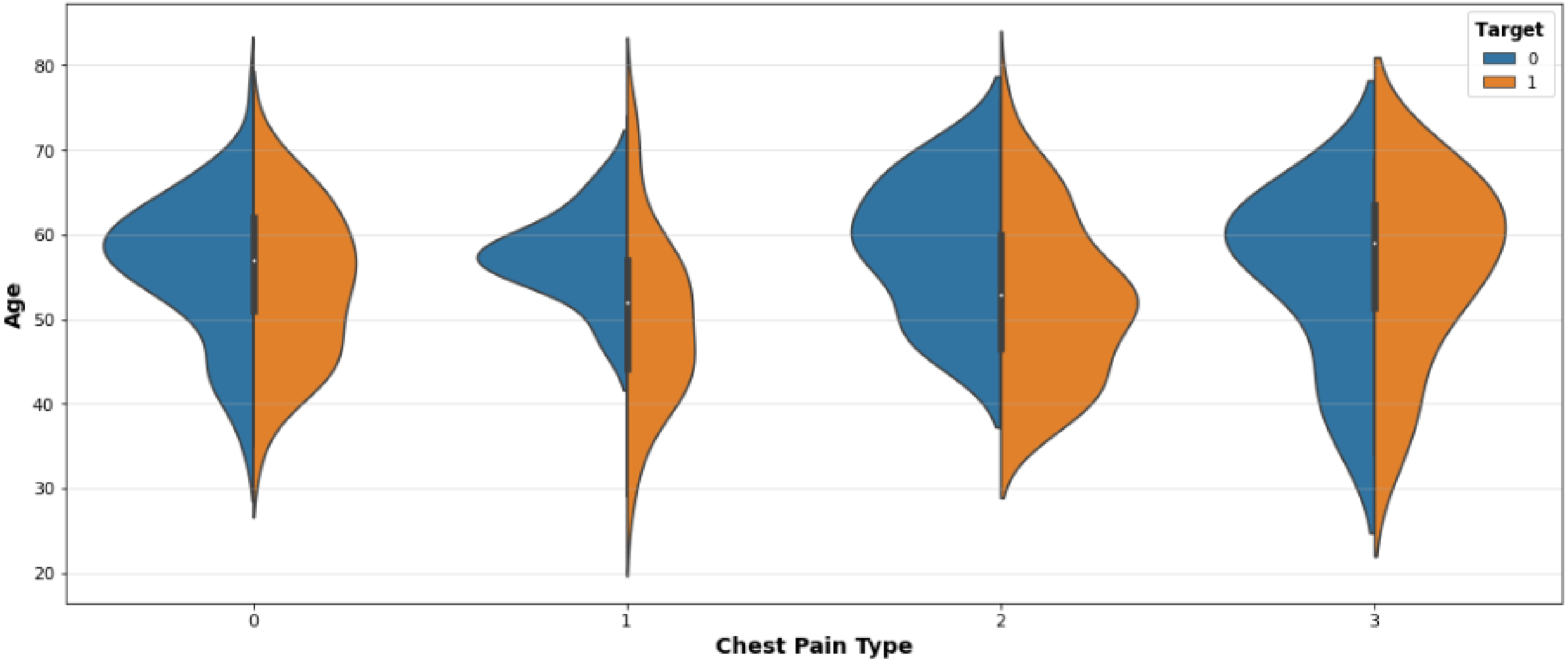
Heart disease distribution based on major vessel s total



Heart disease scatter plot based on age

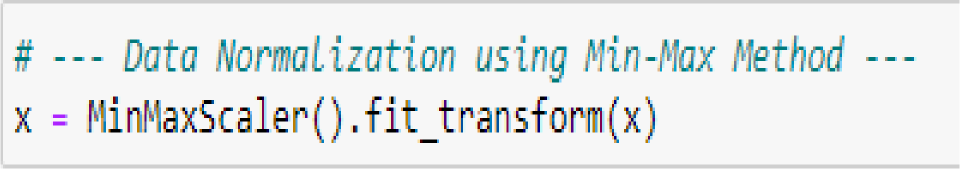


Chest pain type based on age



Data normalization

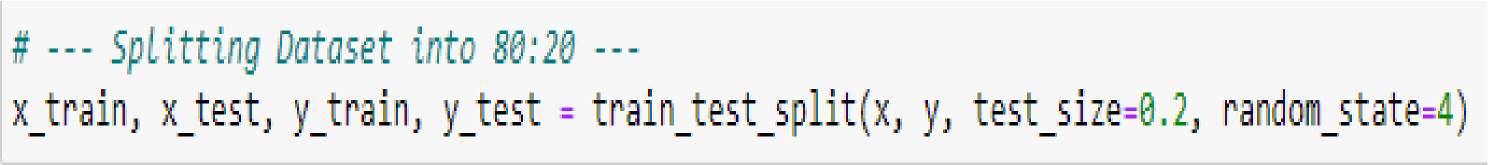
- In these section, data normalization will be performed to normalize the range of independent variables or featured data.
- Data normalization will use min-max normalization.



```
# --- Data Normalization using Min-Max Method ---
x = MinMaxScaler().fit_transform(x)
```

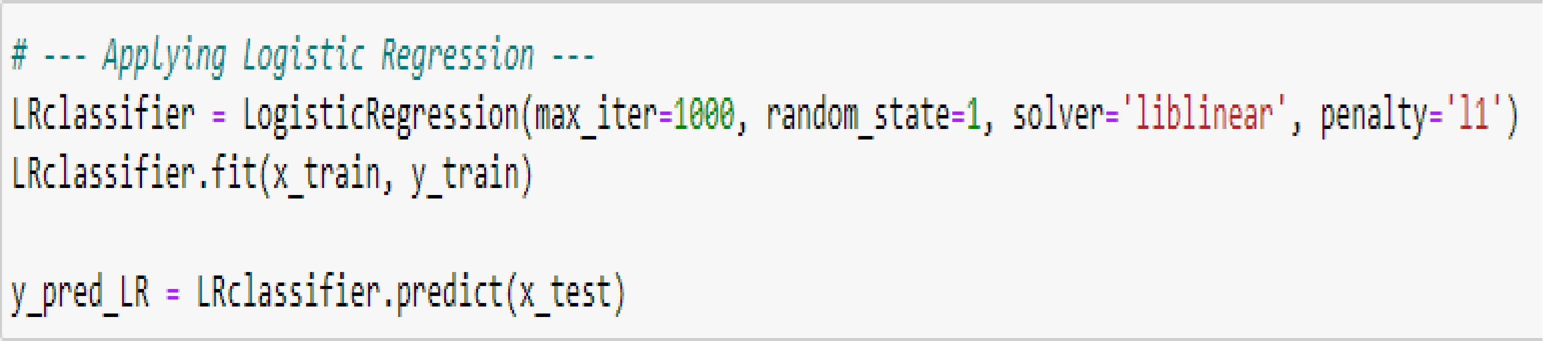
Splitting the dataset

- The dataset will be splitted into 80:20 (80% training and 20% testing)



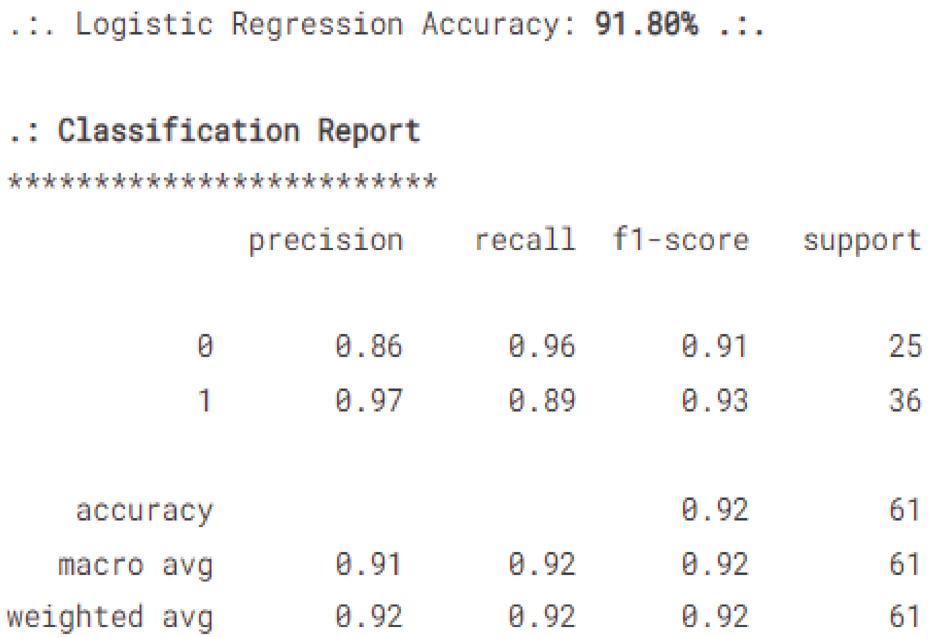
```
# --- Splitting Dataset into 80:20 ---
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=4)
```

Logistic regression



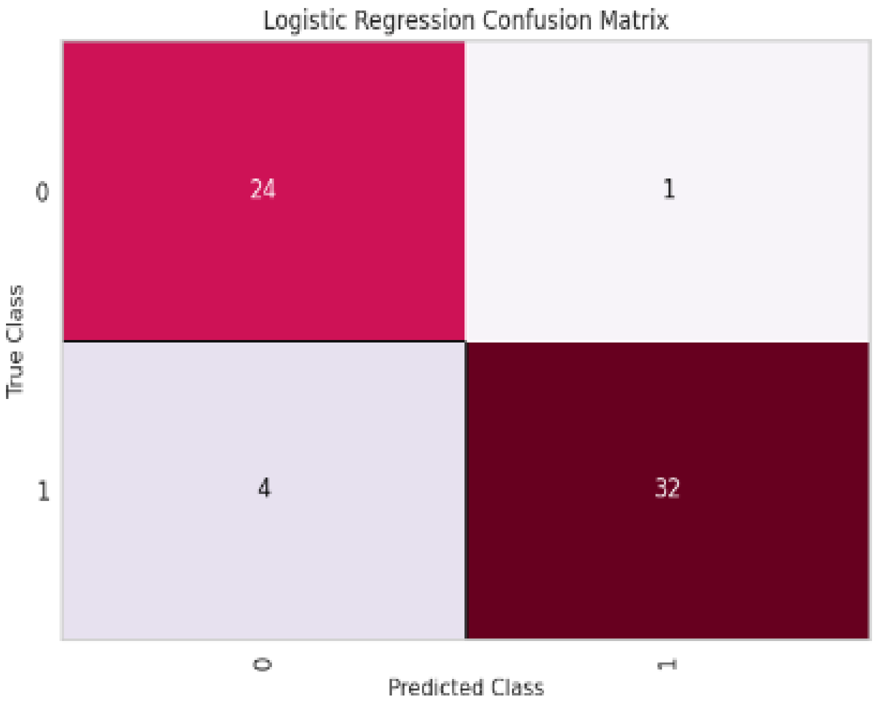
```
# --- Applying Logistic Regression ---
LRclassifier = LogisticRegression(max_iter=1000, random_state=1, solver='liblinear', penalty='l1')
LRclassifier.fit(x_train, y_train)

y_pred_LR = LRclassifier.predict(x_test)
```

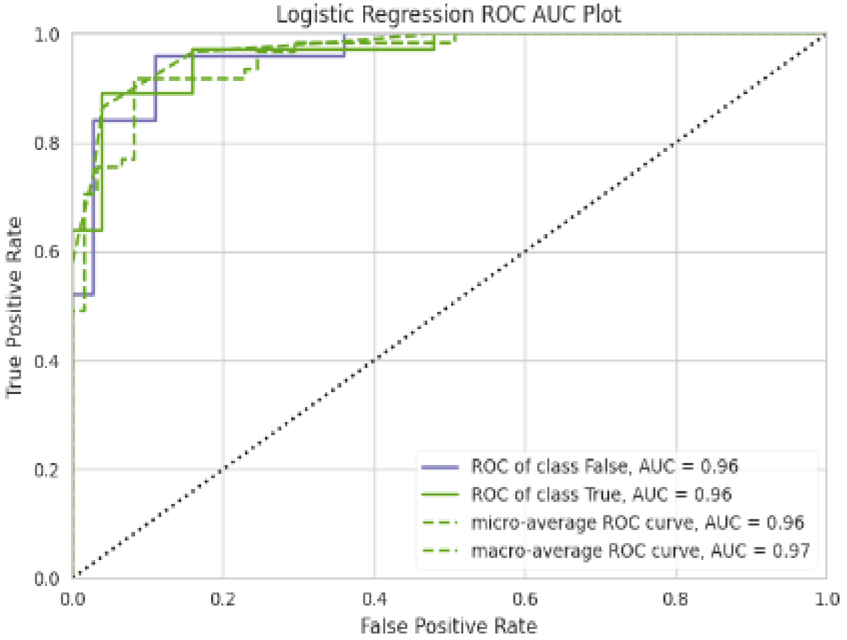
Classification report

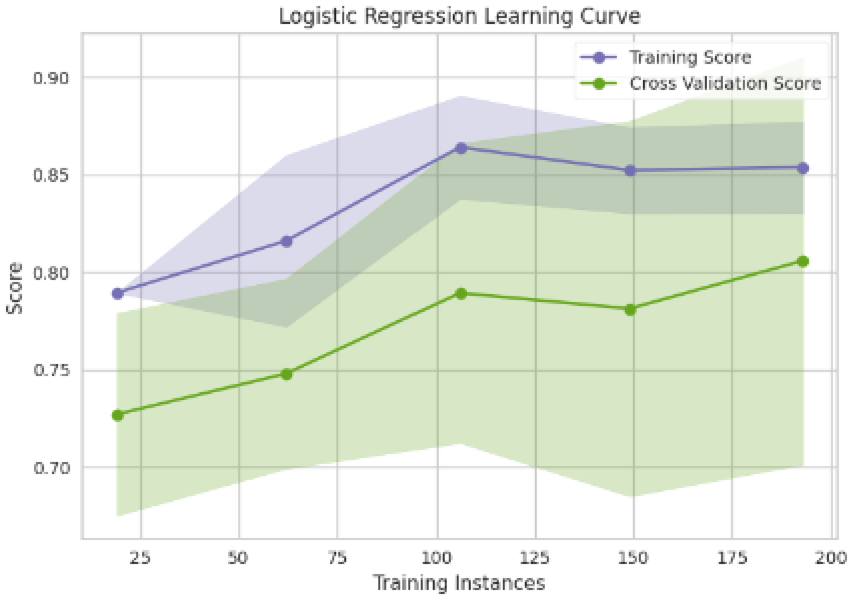


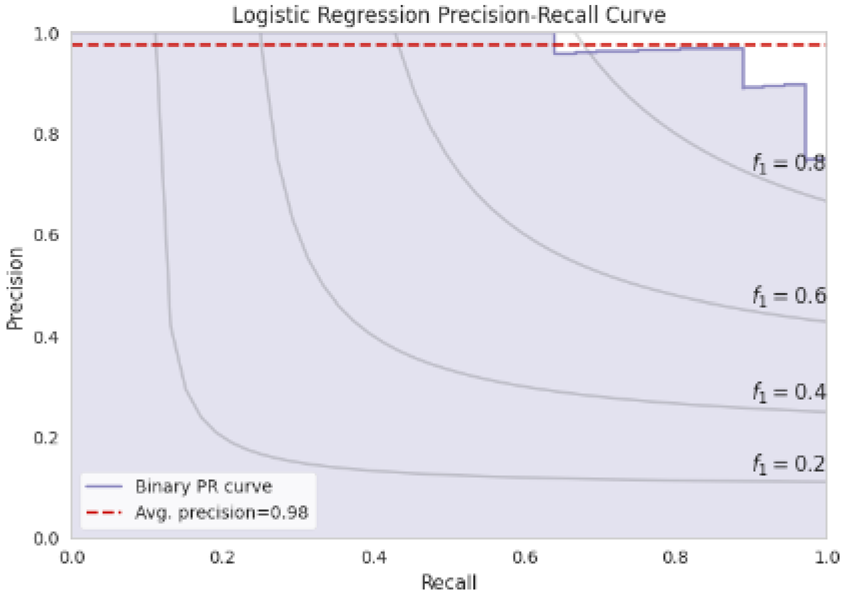
```
.:. Logistic Regression Accuracy: 91.80% .:.

.: Classification Report
*************************
              precision    recall  f1-score   support

           0       0.86      0.96      0.91        25
           1       0.97      0.89      0.93        36

    accuracy                           0.92        61
   macro avg       0.91      0.92      0.92        61
weighted avg       0.92      0.92      0.92        61
```

Performance evolution

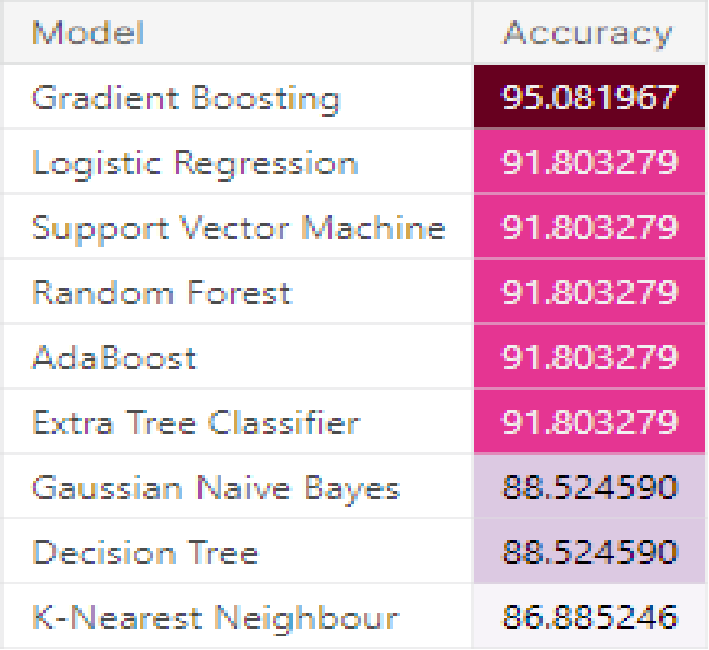





Logistic Regression Learning Curve
Training Score
Cross Validation Score
Score
0.90
0.85
0.80
0.75
0.70
25
50
75
100
125
150
175
200
Training Instances


Logistic Regression Precision-Recall Curve
1.0
0.8
0.6
0.4
0.2
0.0
Precision
$f_1 = 0.8$
$f_1 = 0.6$
$f_1 = 0.4$
$f_1 = 0.2$
Binary PR curve
Avg. precision=0.98
0.0
0.2
0.4
0.6
0.8
1.0
Recall

Model comparision



Model	Accuracy
Gradient Boosting	95.081967
Logistic Regression	91.803279
Support Vector Machine	91.803279
Random Forest	91.803279
AdaBoost	91.803279
Extra Tree Classifier	91.803279
Gaussian Naive Bayes	88.524590
Decision Tree	88.524590
K-Nearest Neighbour	86.885246

THANK YOU
FOR
REVIEWING
MY WORK