PREDICTING HEART DISEASE: HARNESSING MACHINE LEARNING FOR EARLY DETECTION

CONTENT

- Introduction
- Data exploration
- Visualization
- Machine learning

INTRODUCTION TO HEART DISEASE

The World Health Organization reports that 12 million people die from heart disease each year.

Half of these deaths are caused by cardiovascular disease.

Early diagnosis of cardiovascular disease can help people make lifestyle changes to reduce their risk of complications.

This research aims to identify the most important risk factors for heart disease and predict overall risk.

DATASET ITMES

Column Descriptions:

- `id` (Unique id for each patient)
- `age` (Age of the patient in years)
- `origin` (place of study)
- `sex` (Male/Female)
- `cp` chest pain type
    1. typical angina.
    2. atypical angina.
    3. non-anginal.
    4. asymptomatic.
- `trestbps` resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
- `chol` (serum cholesterol in mg/dl)
- `fbs` (if fasting blood sugar > 120 mg/dl)
- `restecg` (resting electrocardiographic results)
- `-- Values:` [normal, stt abnormality, lv hypertrophy]
- `thalach:` maximum heart rate achieved
- `exang:` exercise-induced angina (True/ False)
- `oldpeak:` ST depression induced by exercise relative to rest
- `slope:` the slope of the peak exercise ST segment
- `ca:` number of major vessels (0-3) colored by fluoroscopy
- `thal:` [normal; fixed defect; reversible defect]
- `num:` the predicted attribute

DATA EXPLORATION

Checking the data shape
(920, 16)

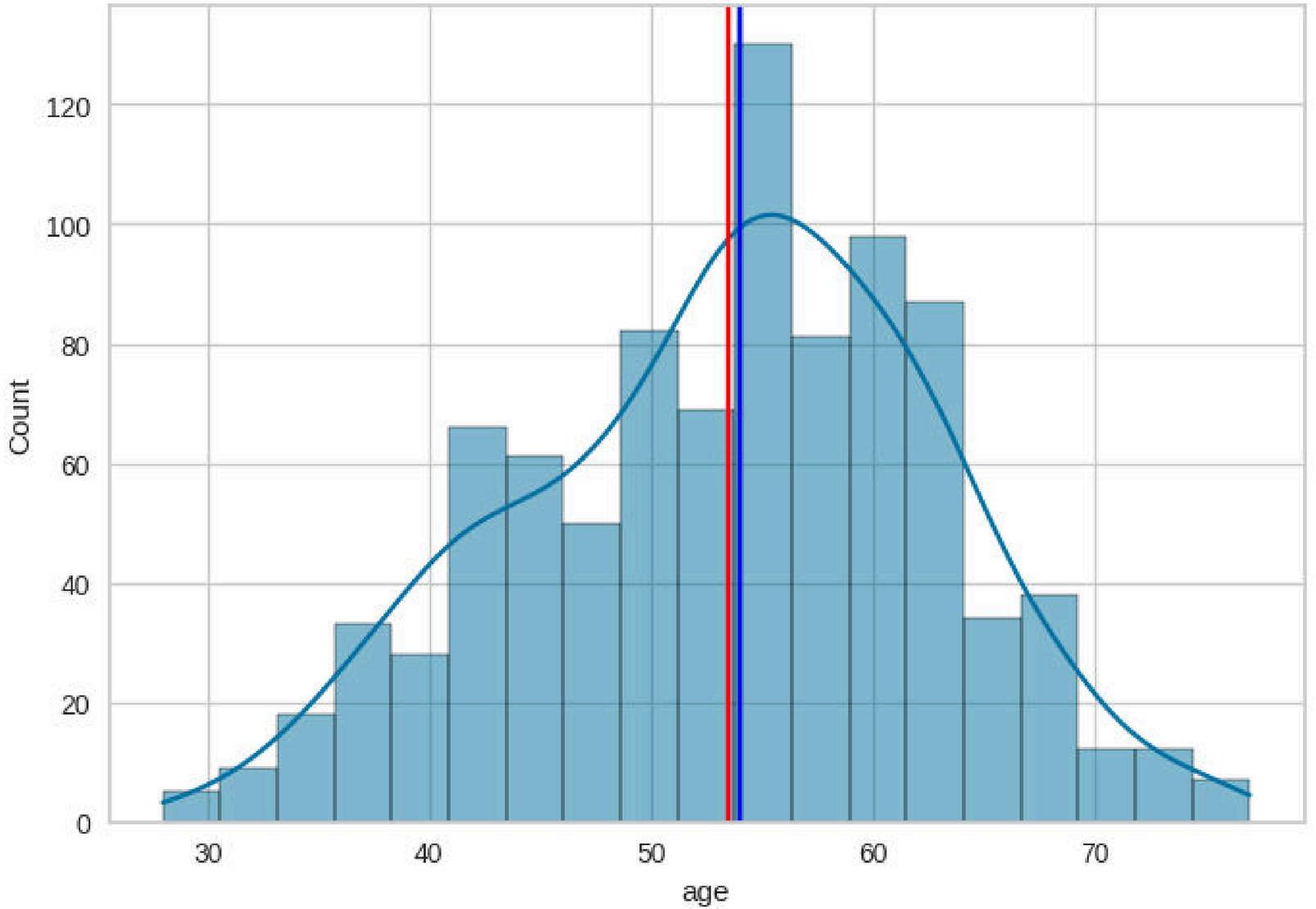
age column
(28, 77)

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        920 non-null    int64
 1   age       920 non-null    int64
 2   sex       920 non-null    object
 3   dataset   920 non-null    object
 4   cp        920 non-null    object
 5   trestbps  861 non-null    float64
 6   chol      890 non-null    float64
 7   fbs       830 non-null    object
 8   restecg   918 non-null    object
 9   thalch    865 non-null    float64
 10  exang     865 non-null    object
 11  oldpeak   858 non-null    float64
 12  slope     611 non-null    object
 13  ca        309 non-null    float64
 14  thal      434 non-null    object
 15  num       920 non-null    int64
dtypes: float64(5), int64(3), object(8)
```

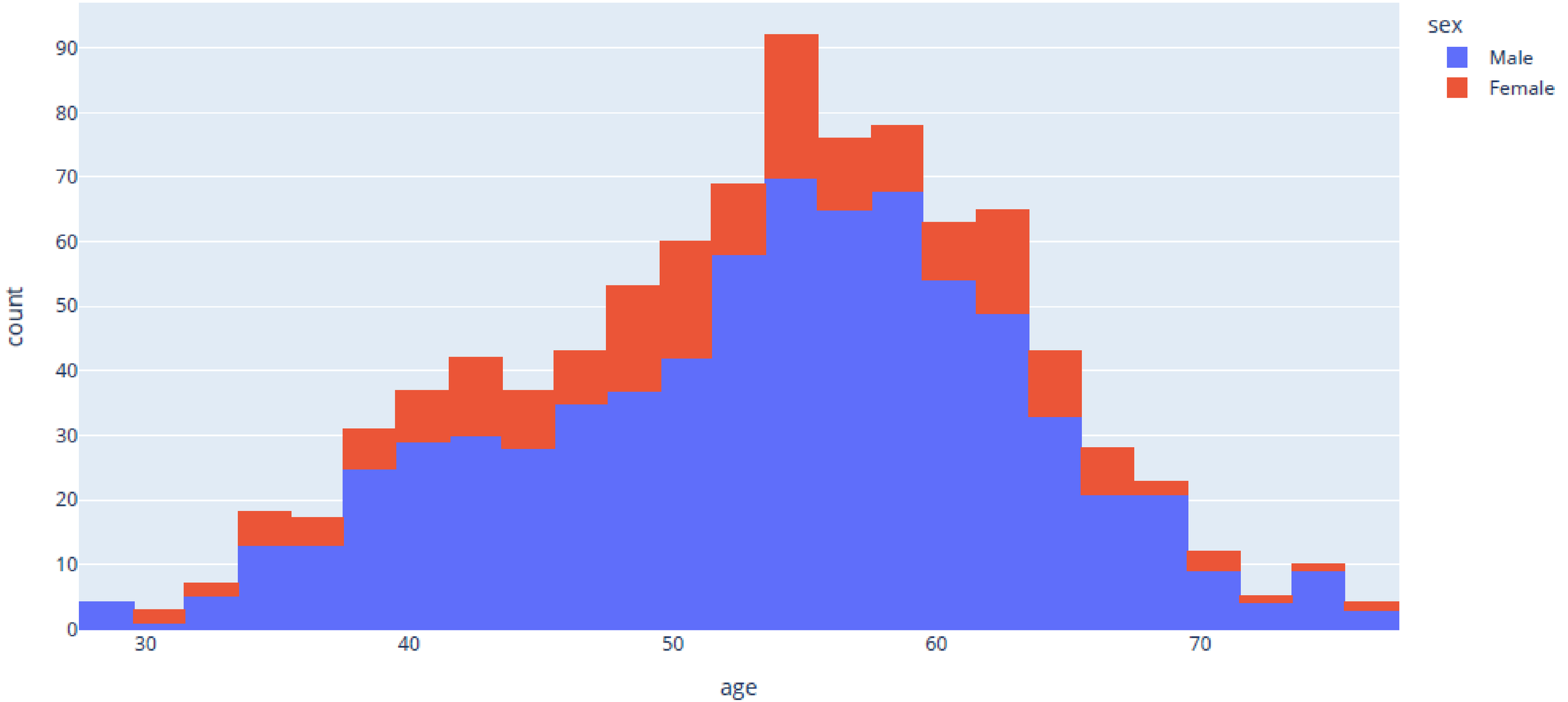
VISUALIZATIONS

Mean 53.51086956521739
Median 54.0
Mode 0 54

AGE COLUMN DISTRIBUTION



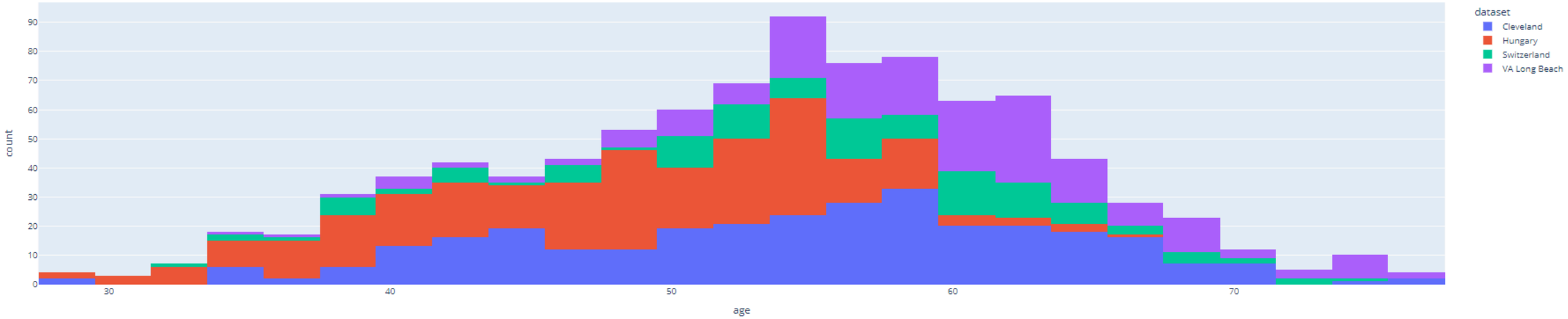
GENDER BASE DISTRIBUTION



Male percentage in the data: 78.91%

Female percentage in the data : 21.09%

EXPLORING AGE BY DATASET

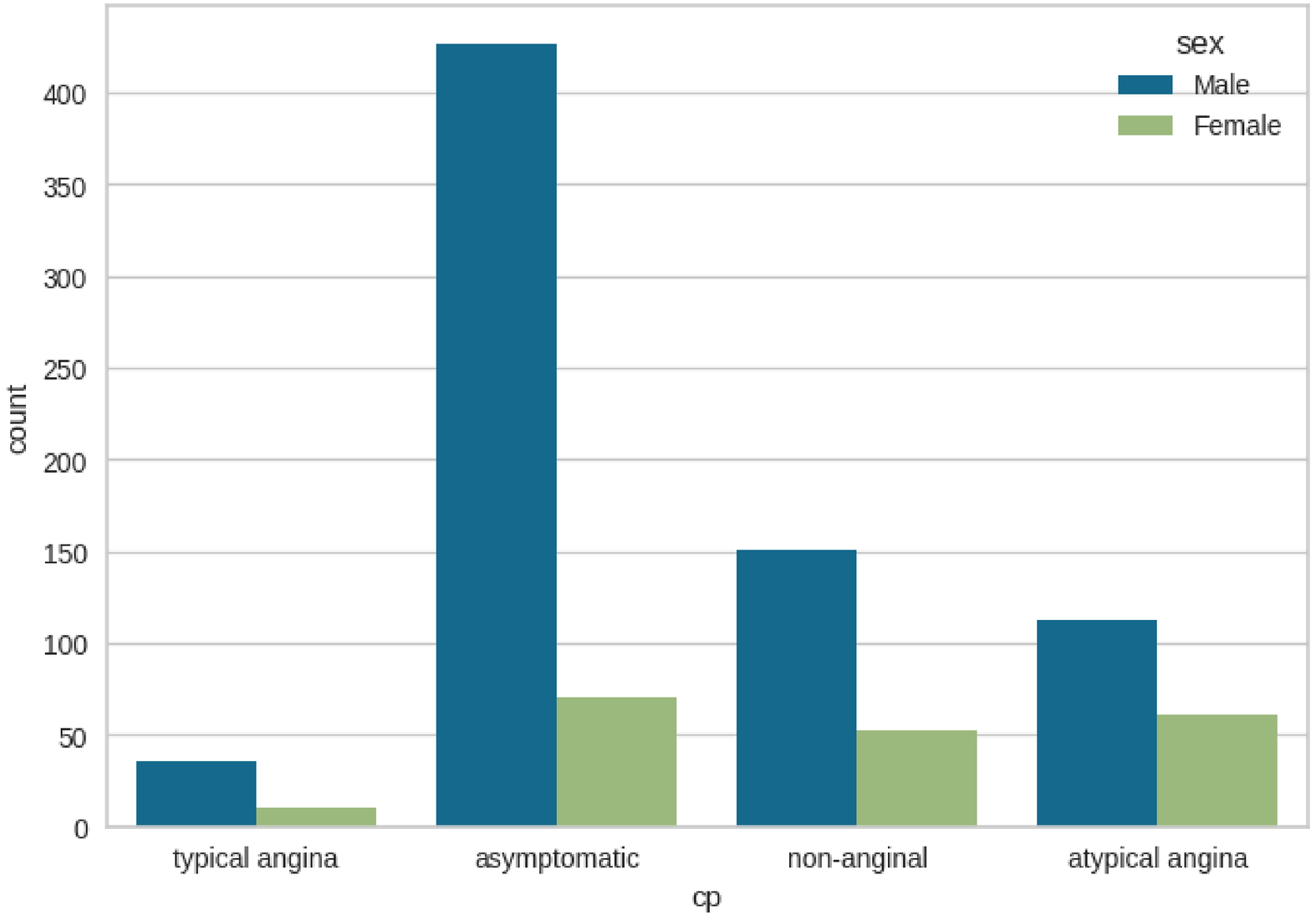



```
Mean of the dataset:  dataset
Cleveland          54.351974
Hungary            47.894198
Switzerland        55.317073
VA Long Beach      59.350000
Name: age, dtype: float64
```

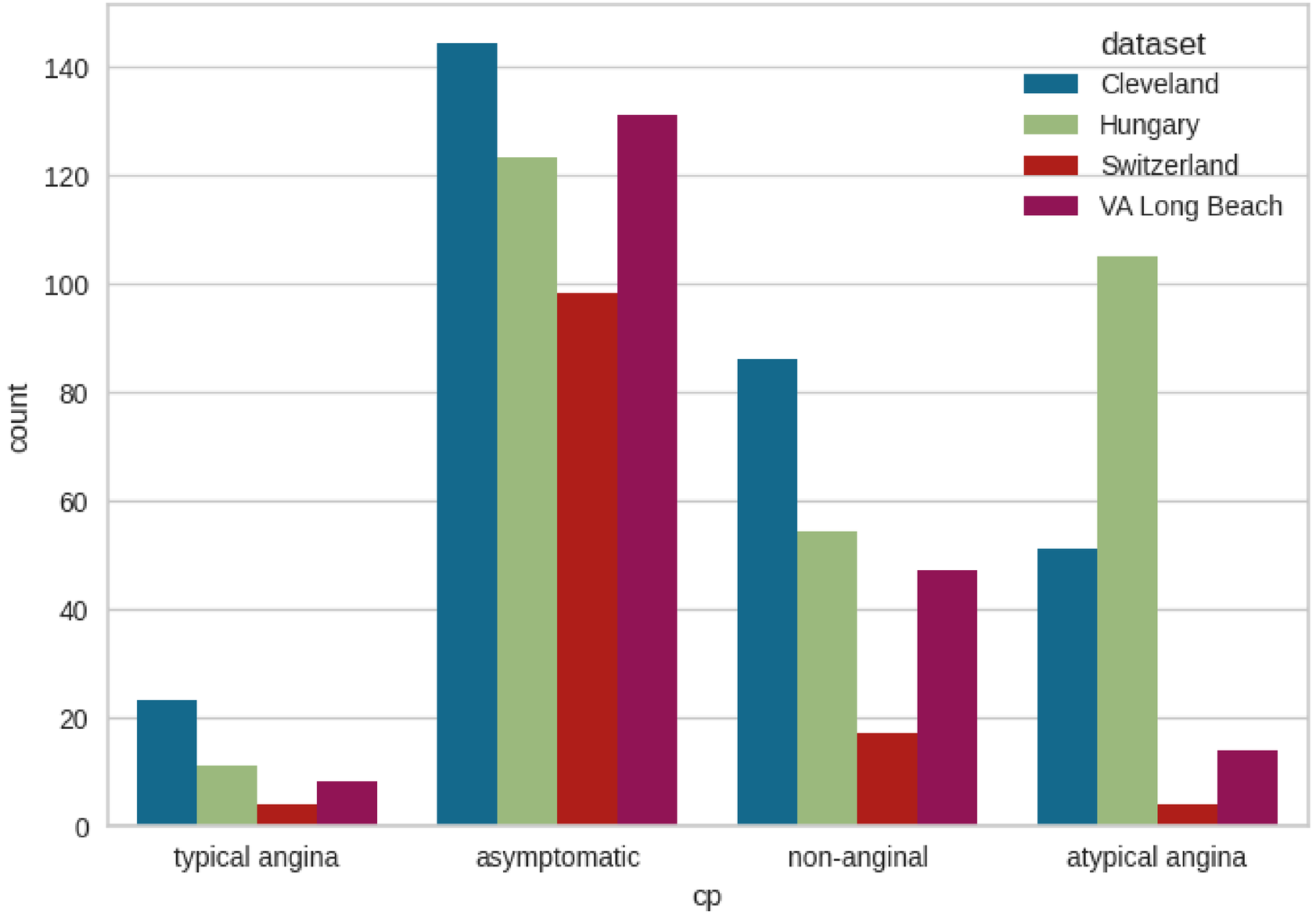
```
Median of the dataset:  dataset
Cleveland          55.5
Hungary            49.0
Switzerland        56.0
VA Long Beach      60.0
Name: age, dtype: float64
```

```
Mode of the dataset:  dataset
Cleveland                58
Hungary                  54
Switzerland              61
VA Long Beach    [62, 63]
Name: age, dtype: object
```

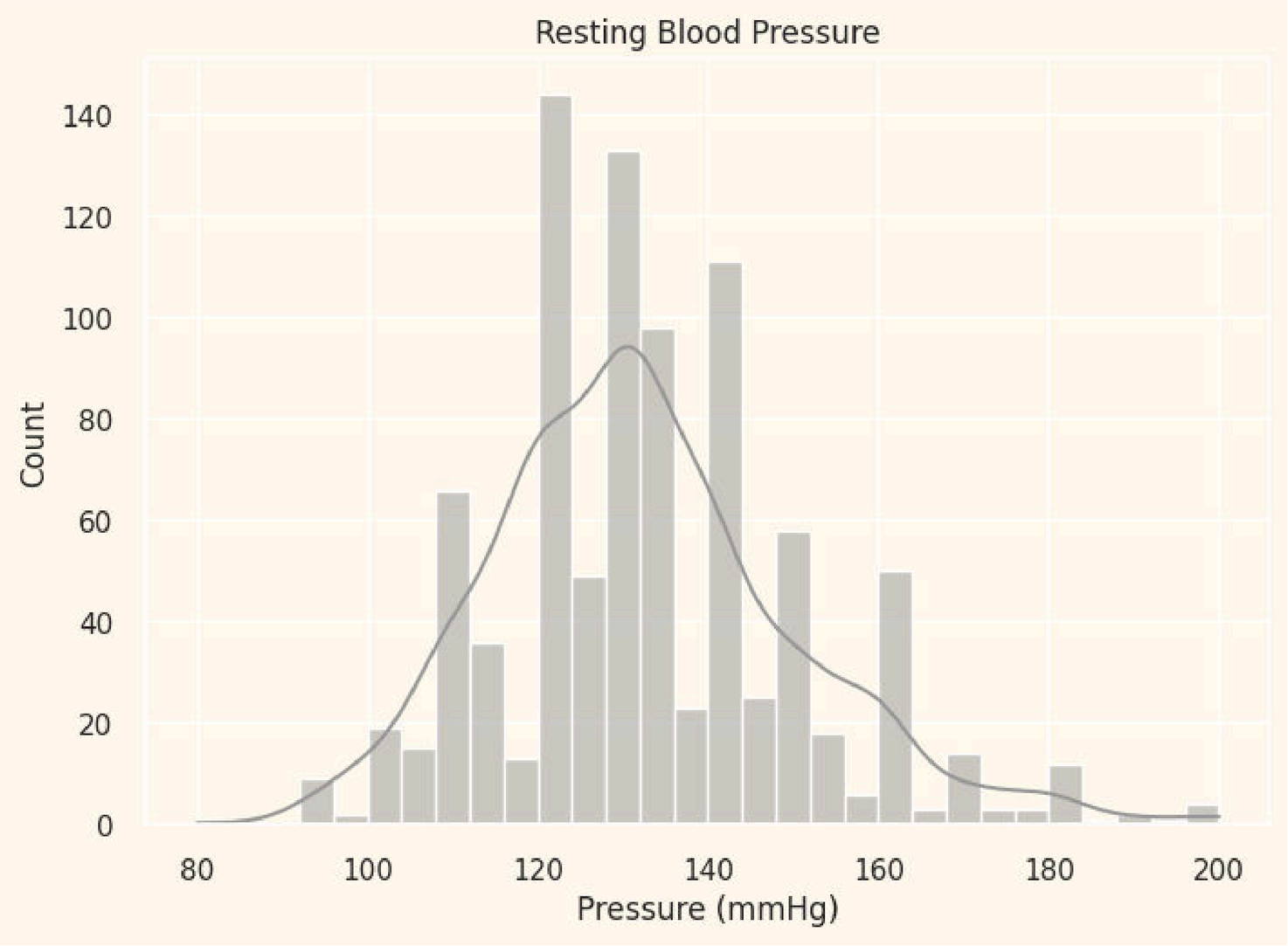
EXPLORING CP (CHEST PAIN)



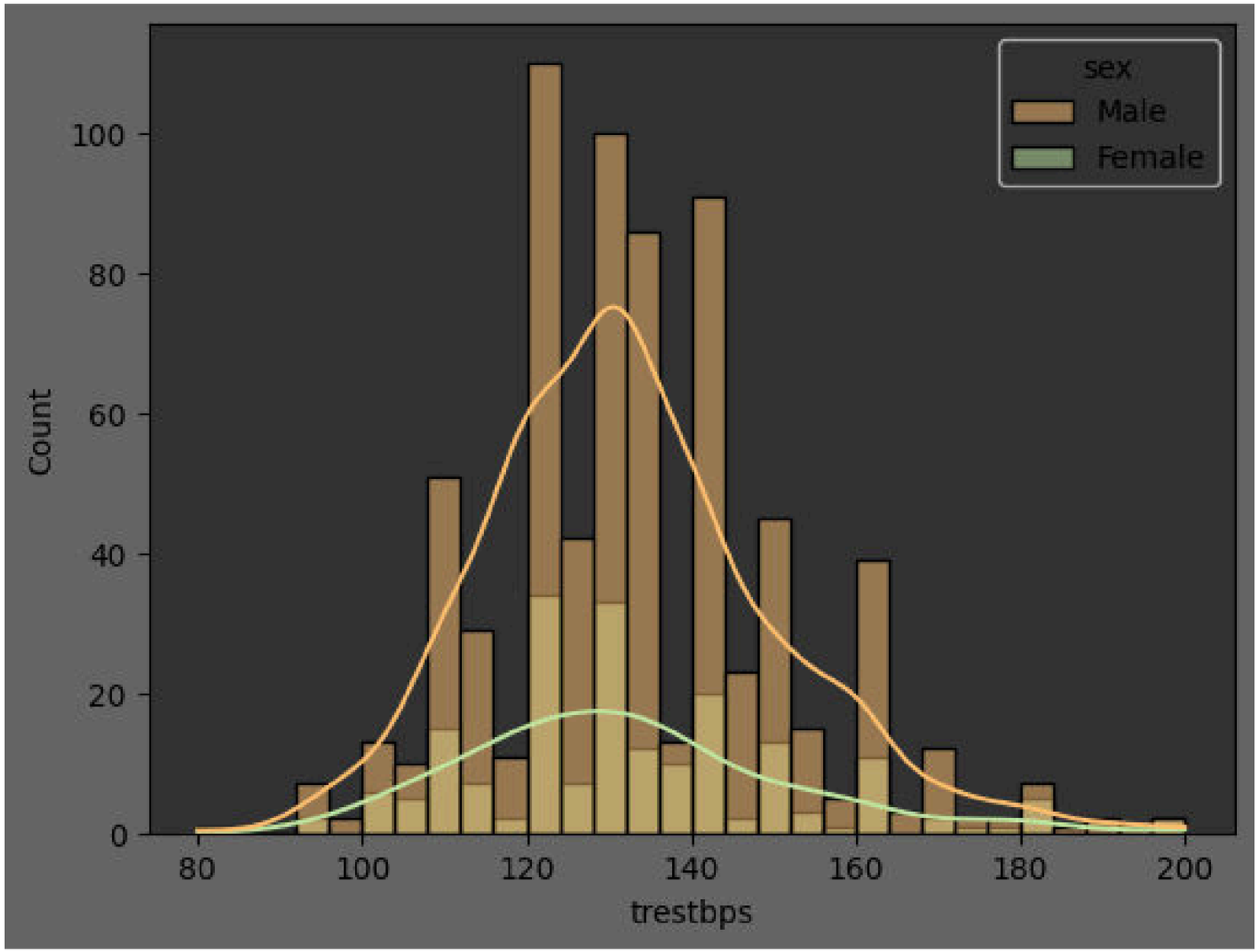
EXPLORING CP BY DATASET



RESTING BLOOD PRESSURE



TRESTBPS


sex
Male
Female
Count
0
20
40
60
80
100
80
100
120
140
160
180
200
trestbps

MACHINE LEARINING
MODEL EVALUATION

Model Logistic Regression
Accuracy: 0.4945652173913043

Model Gradient Boosting
Accuracy: 0.6358695652173914

Model KNeighbors Classifier
Accuracy: 0.592391304347826

Model Decision Tree Classifier
Accuracy: 0.6141304347826086

Model AdaBoost Classifier
Accuracy: 0.5760869565217391

Model Random Forest
Test Accuracy: 0.6086956521739131

Model XGboost Classifier
Accuracy: 0.625

Model Support Vector Machine
Accuracy: 0.5815217391304348

Model Naye base Classifier
Accuracy: 0.5217391304347826

Best Model: GradientBoosting

CONCULSION

The minimum age to have a heart disease start from 28 years old

Most of the people get heart disease at the age of 53 to 54 years.

We have the highest number of people from Clveland(304) and lowest from Switzerland (123)

The highest number of female in this dataset are from Cleveland(97) and lowest are from VA Long Beach(6)

The highest number of male are from Hungary(212) and lowest from Switzerland(113)

The age in which highest number of Typical Angina happened is 62 to 63 years.

The age in which highest number of Asymtomatic Angina happened is 56 to 57 years.

The age in which highest number of Non Anginal happened is 54 to 55 years.

The age in which highest number of Atypical Angina happened is 54 to 55 years.

Thanks!

Abdallah Mohamed Shokry
a.shokry811@email.com