**BMEN619 Heart Failure Prediction Data Exploration**

*Veronica Obodozie*

3 March 2025

**1. Introduction**

Heart Disease is one of the leading causes of death globally.

This is the data exploration portion of the Heart Disease Failure Prediction Dataset from Kaggle

**2. Data Description and Background**

**Selected data**

Heart diseases, especially heart attacks and strokes, are a leading cause of death globally.

This was a csv file gotten from Kaggle <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data> containing 11 features to predict heart disease and their labels.

* Age (years)
* Sex (M: Male or F: Female)
* ChestPainType: chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
* RestingBP: (mm Hg)
* Cholesterol (mm/dl)
* FastingBS: fasting blood sugar (1: >120 mg/dl, 0: <= 120 mg/dl)
* RestingECG (Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria)
* MaxHR (# between 60 and 202)
* ExerciseAngina: ecercise induced Angina ( Y yes, N no)
* Oldpeak: ST numeric measured in depression
* ST\_Slope: slope of peak exercise ST segment (Up upsloping, Flat, Down)
* Heartdisease: output labels (1: heart disease, 0: Normal)

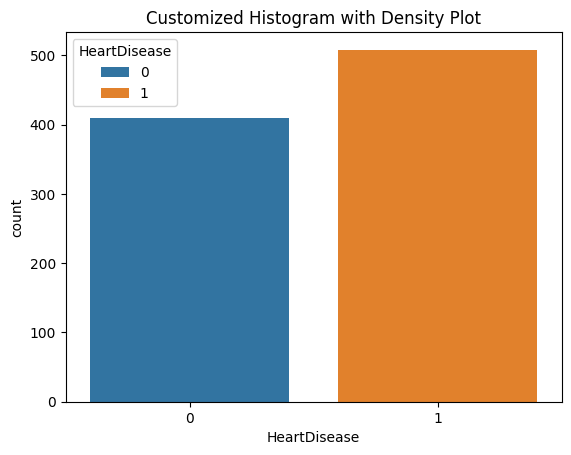
**Source:** Dataset combines 5 heart datasets over 11 common features

**Ethical Usage License**

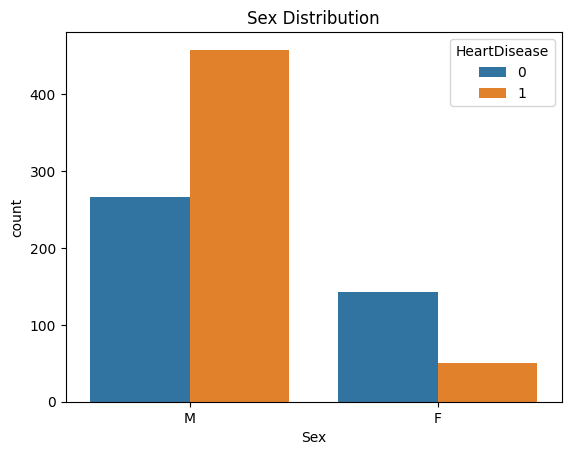
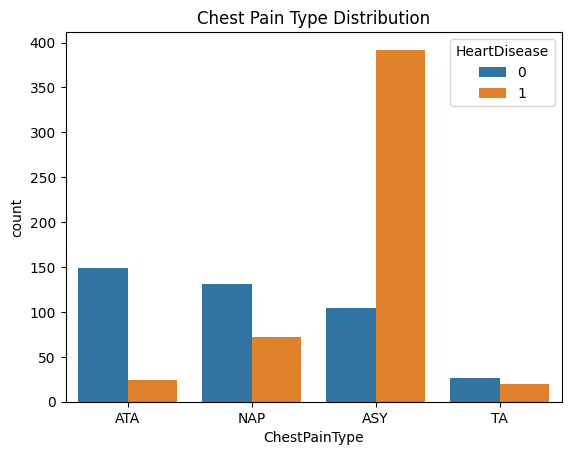
Asks for a citation when using dataset.

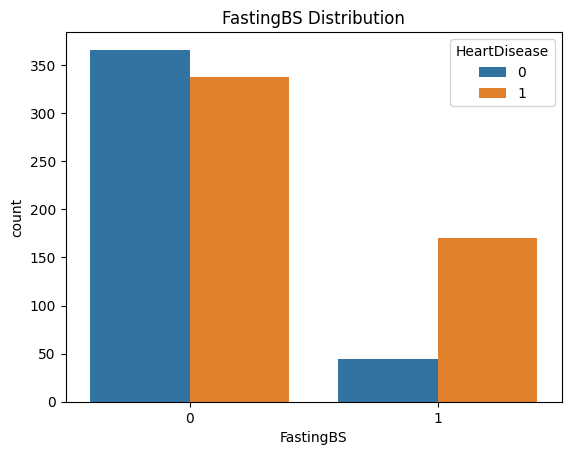
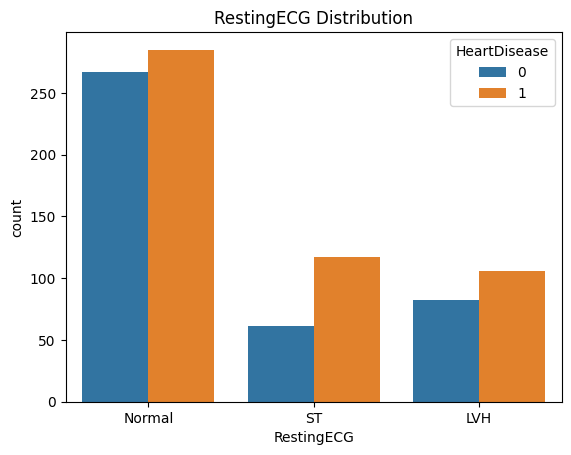
**3. Data Exploration**

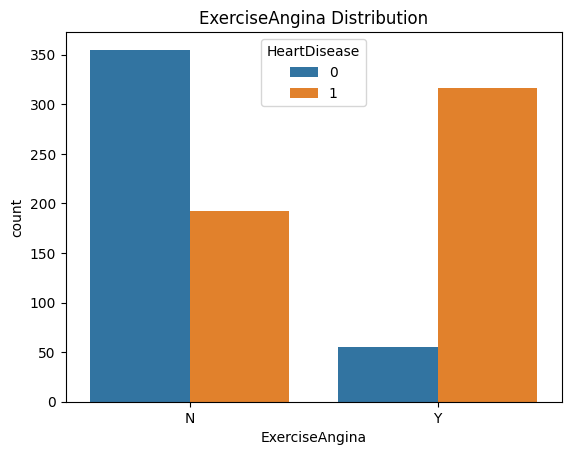
1. **How many samples:** 918
2. Patterns
3. How it looks
4. **Any missing data**: No missing data
5. Visualization
   1. Visual and categorization



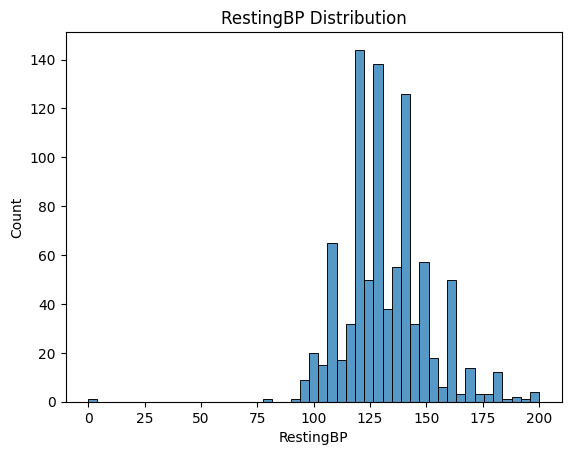
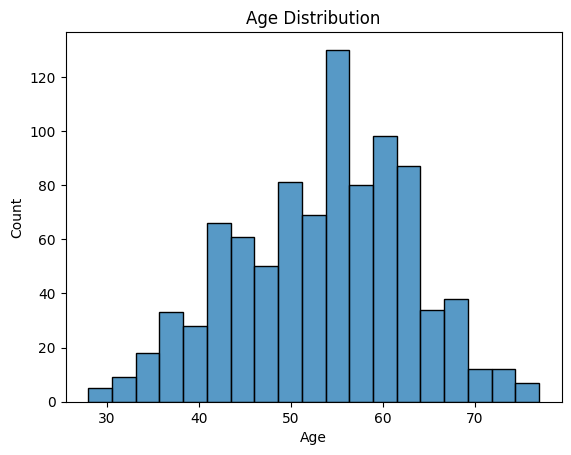
* 1. Patterns

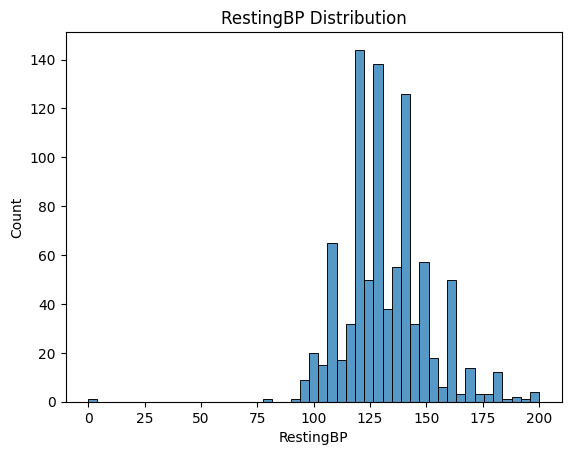
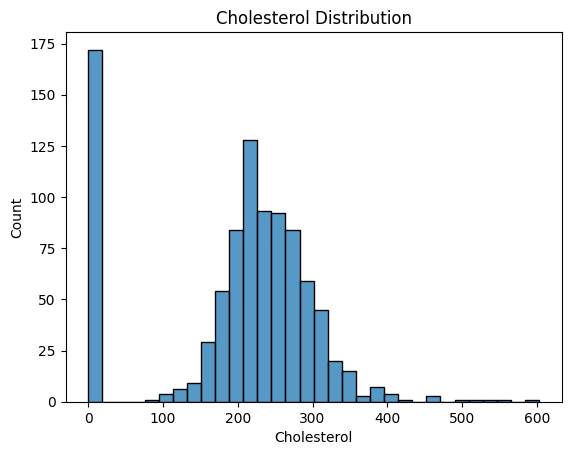
 

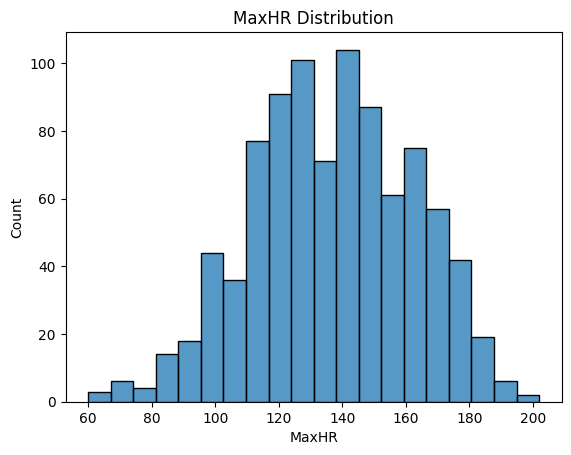
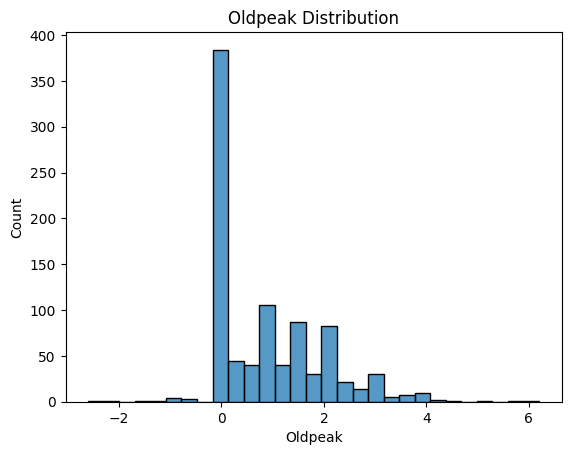
 



* 1. Show distribution



1. **Statistical Analysis**

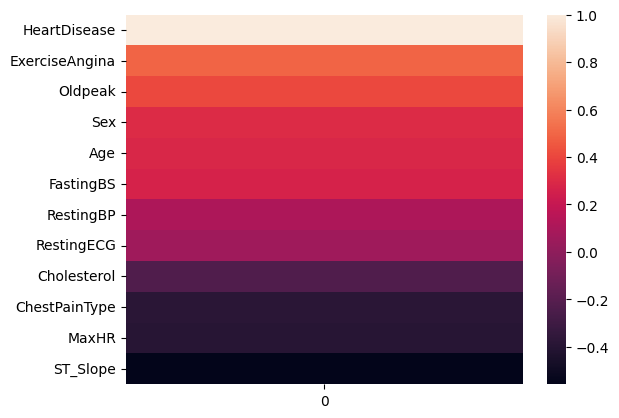
|  | **Age** | **RestingBP** | **Cholesterol** | **FastingBS** | **MaxHR** | **Oldpeak** | **HeartDisease** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 |
| mean | 53.510893 | 132.396514 | 198.799564 | 0.233115 | 136.809368 | 0.887364 | 0.553377 |
| std | 9.432617 | 18.514154 | 109.384145 | 0.423046 | 25.460334 | 1.066570 | 0.497414 |
| min | 28.000000 | 0.000000 | 0.000000 | 0.000000 | 60.000000 | -2.600000 | 0.000000 |
| 25% | 47.000000 | 120.000000 | 173.250000 | 0.000000 | 120.000000 | 0.000000 | 0.000000 |
| 50% | 54.000000 | 130.000000 | 223.000000 | 0.000000 | 138.000000 | 0.600000 | 1.000000 |
| 75% | 60.000000 | 140.000000 | 267.000000 | 0.000000 | 156.000000 | 1.500000 | 1.000000 |
| max | 77.000000 | 200.000000 | 603.000000 | 1.000000 | 202.000000 | 6.200000 | 1.00000 |

1. **Outlier detection**
   1. 3std of mean z score

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| position | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST\_Slope | HeartDisease |
| 76 | 32 | M | ASY | 118 | 529 | 0 | Normal | 130 | N | 0 | Flat | 1 |
| 109 | 39 | M | ATA | 190 | 241 | 0 | Normal | 106 | N | 0 | Up | 0 |
| 149 | 54 | M | ASY | 130 | 603 | 1 | Normal | 125 | Y | 1 | Flat | 1 |
| 166 | 50 | M | ASY | 140 | 231 | 0 | ST | 140 | Y | 5 | Flat | 1 |
| 241 | 54 | M | ASY | 200 | 198 | 0 | Normal | 142 | Y | 2 | Flat | 1 |
| 324 | 46 | M | ASY | 100 | 0 | 1 | ST | 133 | N | -2.6 | Flat | 1 |
| 365 | 64 | F | ASY | 200 | 0 | 0 | Normal | 140 | Y | 1 | Flat | 1 |
| 390 | 51 | M | ASY | 140 | 0 | 0 | Normal | 60 | N | 0 | Flat | 1 |
| 399 | 61 | M | NAP | 200 | 0 | 1 | ST | 70 | N | 0 | Flat | 1 |
| 449 | 55 | M | NAP | 0 | 0 | 0 | Normal | 155 | N | 1.5 | Flat | 1 |
| 592 | 61 | M | ASY | 190 | 287 | 1 | LVH | 150 | Y | 2 | Down | 1 |
| 616 | 67 | F | NAP | 115 | 564 | 0 | LVH | 160 | N | 1.6 | Flat | 0 |
| 702 | 59 | M | TA | 178 | 270 | 0 | LVH | 145 | N | 4.2 | Down | 0 |
| 732 | 56 | F | ASY | 200 | 288 | 1 | LVH | 133 | Y | 4 | Down | 1 |
| 759 | 54 | M | ATA | 192 | 283 | 0 | LVH | 195 | N | 0 | Up | 1 |
| 771 | 55 | M | ASY | 140 | 217 | 0 | Normal | 111 | Y | 5.6 | Down | 1 |
| 791 | 51 | M | ASY | 140 | 298 | 0 | Normal | 122 | Y | 4.2 | Flat | 1 |
| 850 | 62 | F | ASY | 160 | 164 | 0 | LVH | 145 | N | 6.2 | Down | 1 |
| 900 | 58 | M | ASY | 114 | 318 | 0 | ST | 140 | N | 4.4 | Down | 1 |

1. Balance
   1. Balance
   2. Can cause bias to models balancing improves fairness and generalizability
   3. Proportional representation of categories
2. Feature Correlation.

Correlation between heart disease and features:



**4. References**

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.