HS-631
STATISTICAL AND DATA ANALYSIS ON HEART DISEASE DATASET

Tejaswee Katanguri

Presentation Overview

- Introduction
- Objective
- Dataset
- Visualization
- Hypothesis testing
- ANOVA
- Regression Analysis

Objective

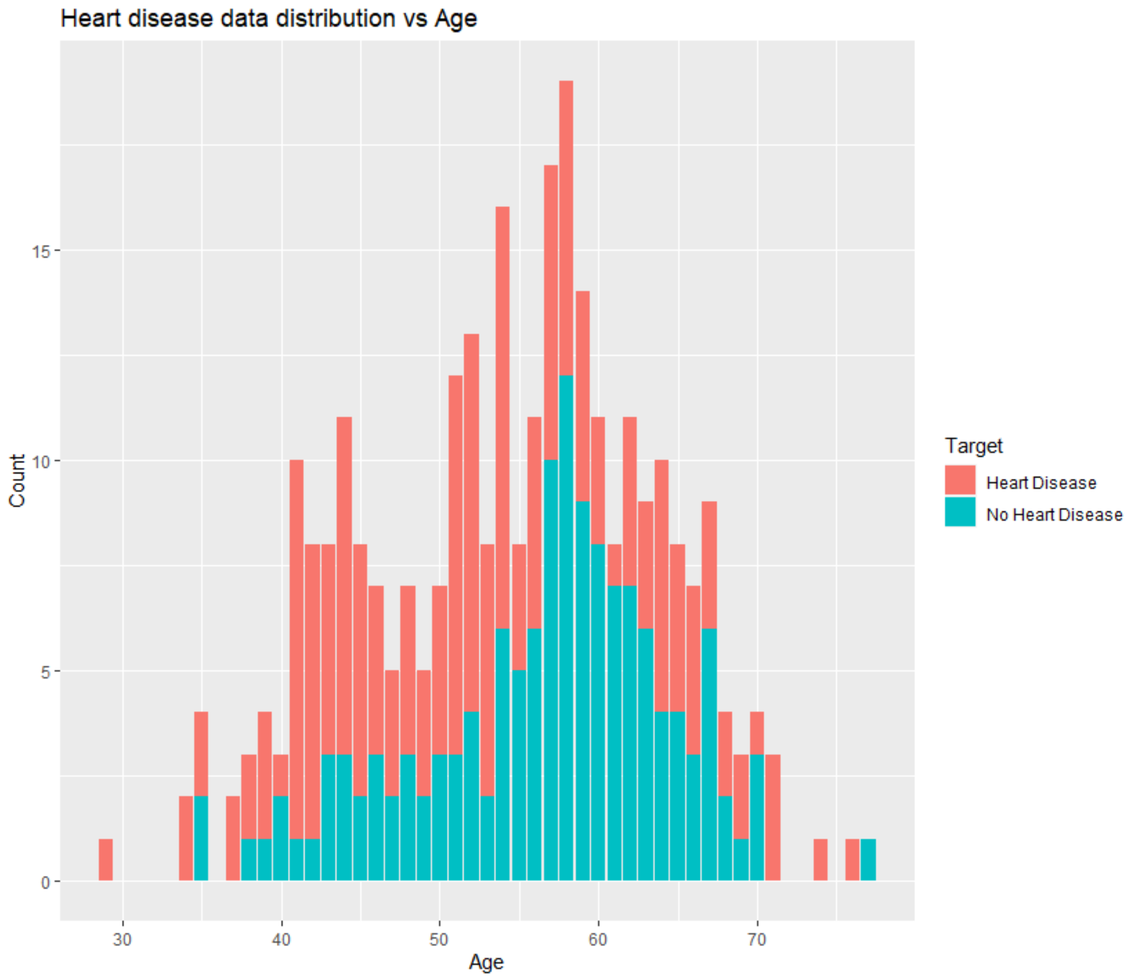
- To apply EDA techniques and statistical analysis techniques for classifying whether a person is suffering from heart disease or not, using Cleveland Heart Disease dataset from the UCI Repository.

Dataset Description

- The dataset used is the Cleveland Heart Disease dataset taken from the UCI repository.
- The dataset contains 14 attributes from 303 patients/samples.
- Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Max heart rate achieved, Exercise induced angina, ST depression induced by exercise relative to rest, Peak exercise ST segment, Number of major vessels (0–3) colored by fluoroscopy, Thal, Diagnosis of heart disease.

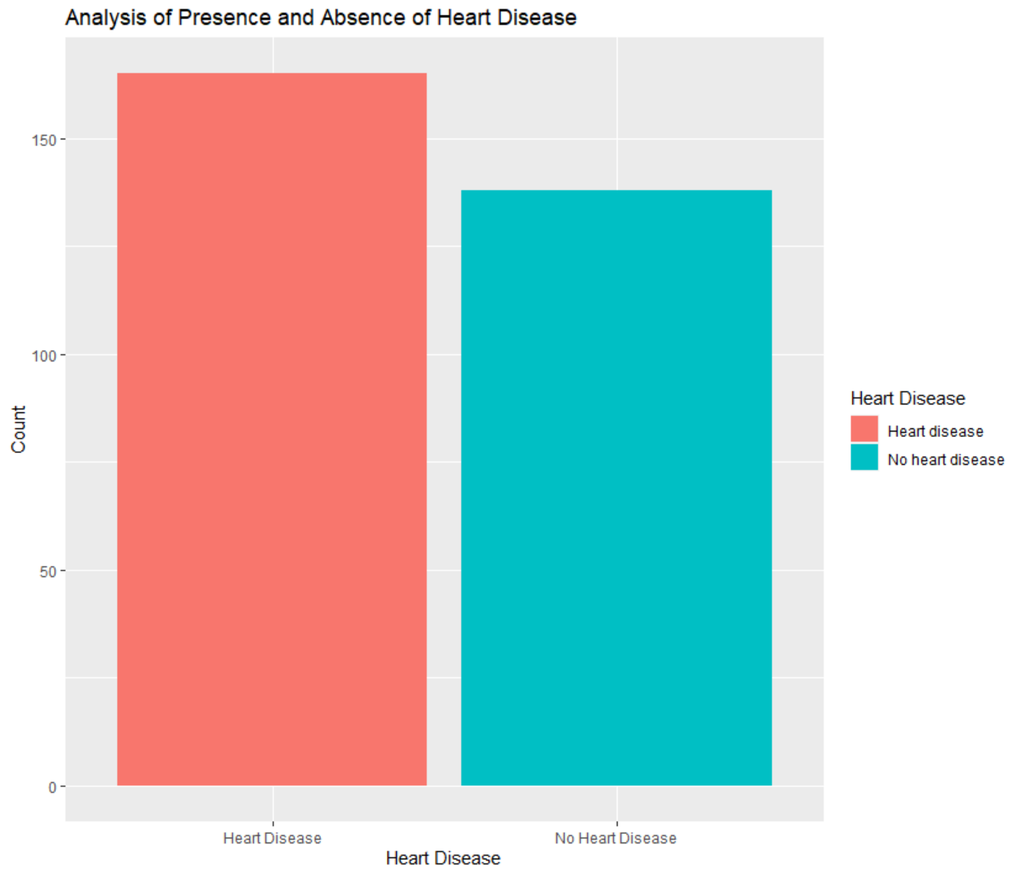
Data Analysis

- The graph shows data distribution between presence of heart disease vs Age.
- Age < 55 there are lot more samples with No heart disease vs heart disease.
- Highest number of samples are between ages 55 - 65.



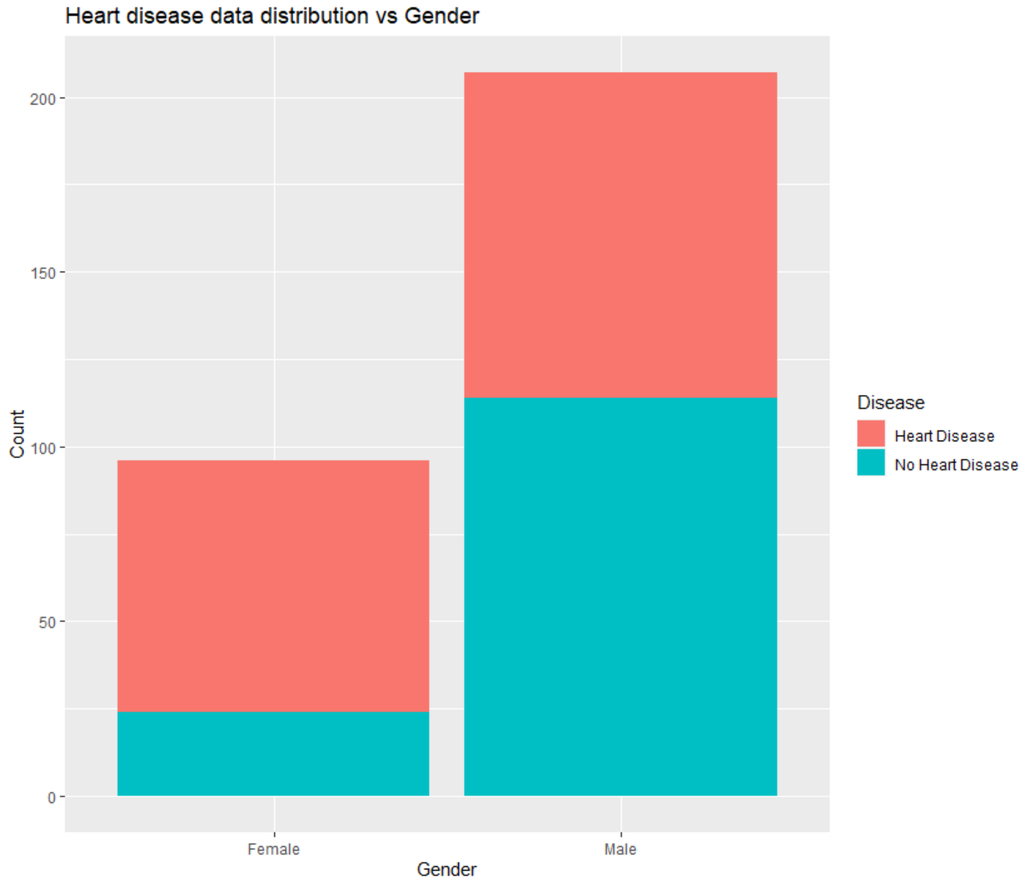
Data Analysis

- There are ~20 more samples with heart disease vs no heart disease.



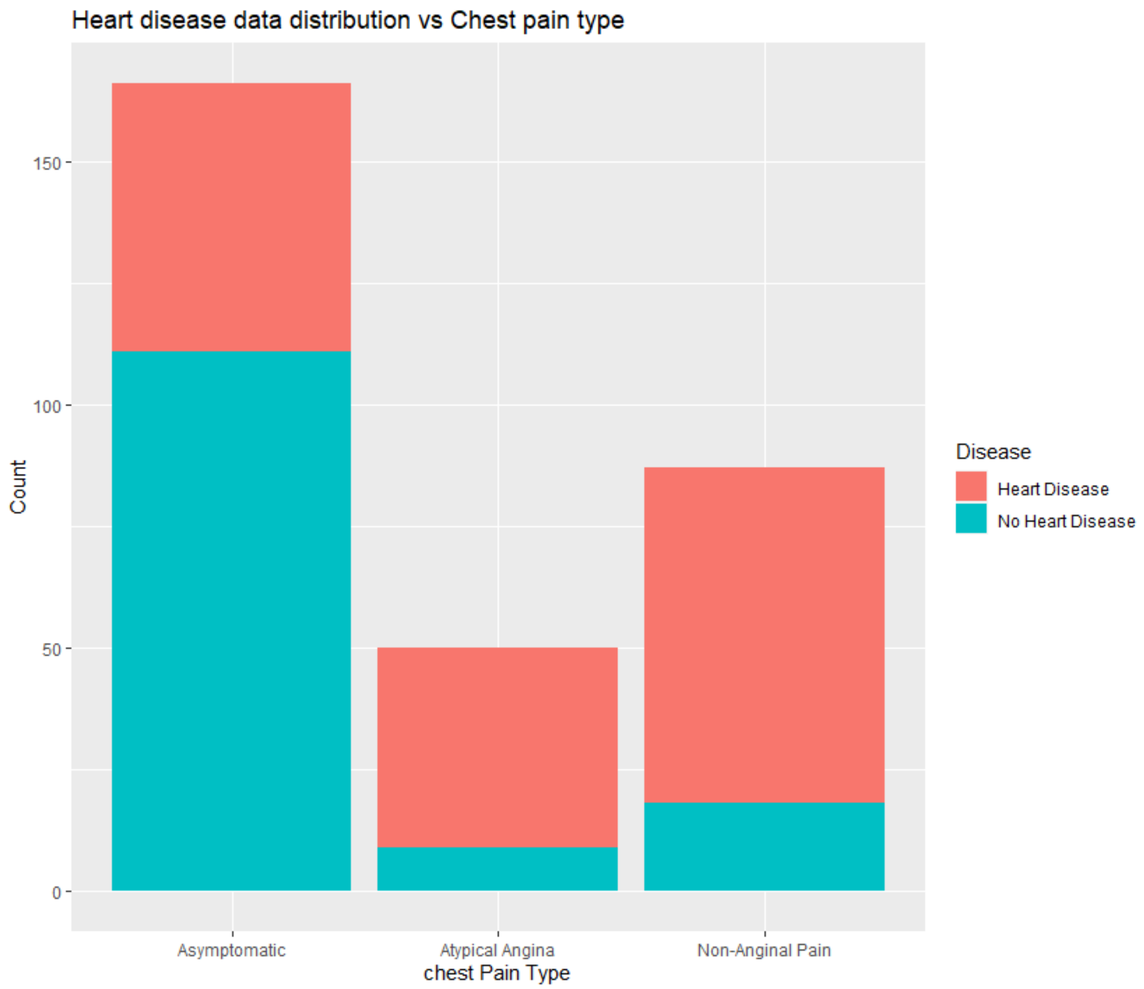
Data Analysis

- There are more samples with Males than Females. Male samples with heart disease vs no heart disease is evenly split.
- There are more Female samples with heart disease.



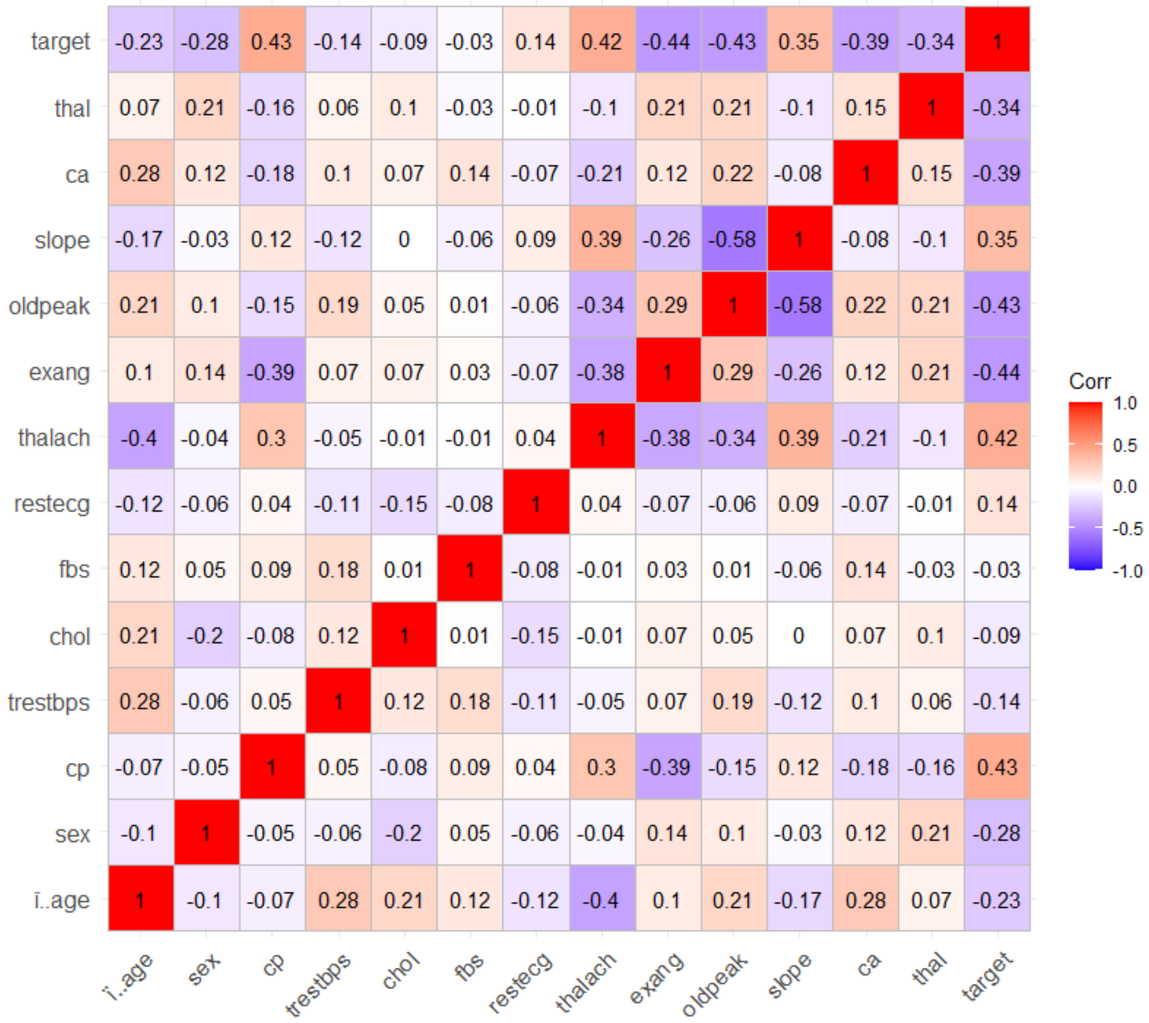
Data Analysis

- There are 3 pain types in the dataset.
- There are more heart disease samples in “Atypical Angina” and “Non-Anginal Pain” vs Asymptomatic.



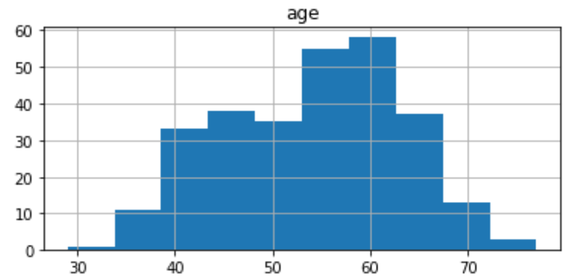
Correlation Plot

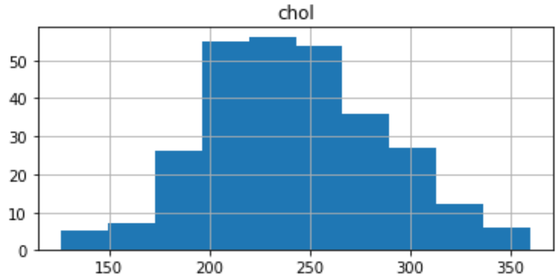
- Correlation plot to find the most significant variables with respect to target.
- We can see slope, thalach(Maximum heart rate) and cp(chest pain) are most related to target.

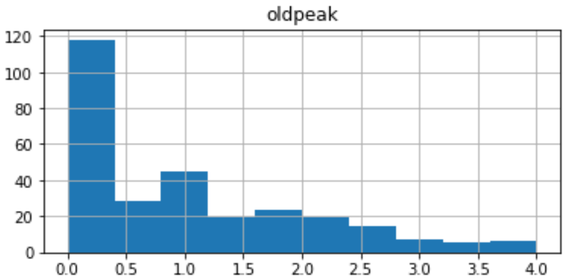


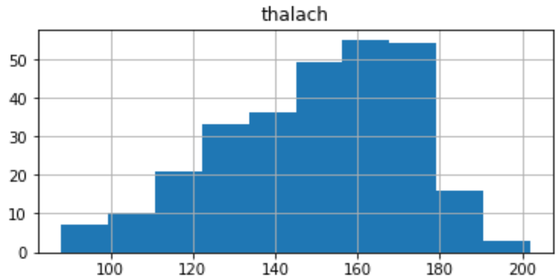
Data distribution

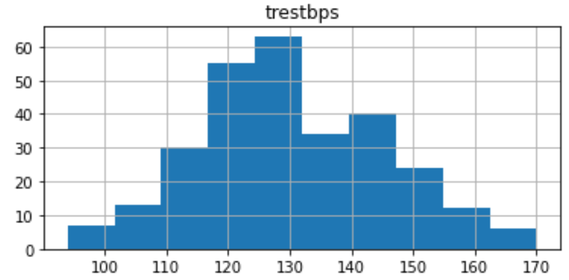
- Age, Chol(cholesterol) and trestbps (resting blood pressure) are normally distributed, whereas oldpeak and thalach(maximum heart rate) are skewed to left and right.











Hypothesis Testing

- Claim 1: Average mean of BP value of people who have heart disease is more than 120.

- Hypothesis:
  - H_0(Null) : mean(BP) <=120
  - H_1(Alternate): mean(BP) > 120

- P-value is less than 0.05, hence we reject the null hypothesis.

- Claim 2: Average mean of BP of a person with heart disease is greater than the average mean BP of a person without heart disease.

- Hypothesis:
  - H_0(Null) : mean(A) - mean(B) <= 0
  - H_1(Alternate): mean(A) - mean(B) > 0

- P-value is greater than 0.05, we fail to reject the null hypothesis

Hypothesis Testing - ANOVA

- Hypothesis Testing
- H_0: All means are equal
- H_1: at least two means differ
- Performed one-way ANOVA to determine whether three types of chest pain means differ.
- P-value is < 0.05, which tells us that the means of three types of chest pain are different.
- P-value = 2e-16

Regression Analysis

- Chose significant variables to feed to the logistic regression model.
- Used glm function with binomial type.
- P-values for the significant variables were lower than 0.05
- F1 - Score for the model 80%

Attributes	Pr(>chi) values
CP(chest pain type)	4.257E-15
CHOL(cholesterol)	0.147
FBS(fasting blood sugar)	0.176
THALACH(Maximum heart rate)	5.745E-09
SLOPE	0.0000115

Thank You

Regression Analysis - Backup

- Chose significant variables to feed to the logistic regression model.
- Used glm function with binomial type.
- P-values for the significant variables were lower than 0.05
- F1 - Score for the model 80%

Heart Disease	0	1
0	0.32	0.135
1	0.089	0.455