

TDS10 Final Project

Aleksandrov Boyan, 320961, b.aleksandrov@studenti.luiss.it
Nishtelkova Maria

Last compiled on: 26 November, 2025

Abstract

IF you wish, you may add here a short abstract of 100 words max.

Introduction

“adsadsds”

9999999

After restecg vs hdc barplot: Some Rest ECG categories show different frequencies across hdc levels, indicating a possible relationship.

hist Histograms show the distribution of the numeric variables. Some variables are skewed (especially oldpeak), and the plots help us understand the data before modeling.

Dataset Description

(Write about heart.csv here)

Multinomial Logistic Regression — Theory

Question(1.1):

We are using multinomial Logistic Regression because the response variable can take more than 2 categories. For these categories there is a separate set of coefficients and we choose one as the baseline. The coefficients describe how the predictors(age, sex, chol etc.) affect the probability of belonging to each outcome category.

In this regression the response variable Y can take K -number of different categories. We have to pick one of the categories to be the baseline - category 0, for every other category - $k = 1, 2, \dots, K-1$.

The model shows the probability of an observation belonging to category K using the multinomial logistic regression function:

$$P(Y = k | X) = \frac{\exp(\beta_{0k} + \beta_k^T X)}{1 + \sum_{j=1}^{K-1} \exp(\beta_{0j} + \beta_j^T X)}$$

The probability of the baseline category is:

$$P(Y = 0 | X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\beta_{0j} + \beta_j^T X)}$$

Data Preparation

There are several variables in the dataset containing missing values. In order to prepare the data for the multinomial logistic model, we observed how many missing values each variable had and saw that the variables ca, thal, and slope had extremely high numbers of missing values. Because these variables are categorical, and because imputing such a large amount of missing data would introduce strong bias, we decided to remove them from the dataset. For the remaining numeric variables with less missing values- (trestbps, chol, thalach, oldpeak) we applied median imputation.

We chose this method because replacing each missing value with the median of the corresponding variable is a robust measure that is not affected by extreme values.

For the categorical variables with very few missing values, we replaced missing entries with the most frequent category (the mode).

This approach ensures that there are no missing values before using the multinomial logistic regression model.

```
##      age      sex      place      cp trestbps      chol      fbs      restecg
##         0         0         0         0         0         0         0         0
##  thalach    exang  oldpeak      hdc
##         0         0         0         0
```

```
##  age sex      place      cp trestbps chol fbs      restecg thalach
## 1  63  1 Cleveland  typical angina    145  233  1 lv hypertrophy    150
## 2  67  1 Cleveland  asymptomatic    160  286  0 lv hypertrophy    108
## 3  67  1 Cleveland  asymptomatic    120  229  0 lv hypertrophy    129
## 4  37  1 Cleveland  non-anginal    130  250  0          normal    187
## 5  41  0 Cleveland  atypical angina    130  204  0 lv hypertrophy    172
## 6  56  1 Cleveland  atypical angina    120  236  0          normal    178
##  exang oldpeak hdc
## 1     0     2.3  0
## 2     1     1.5  2
```

```
## 3      1      2.6    1
## 4      0      3.5    0
## 5      0      1.4    0
## 6      0      0.8    0
```

Exploratory Data Analysis

Before establishing the multinomial logistic regression model, we looked at some of the predictors' visual variations over the response variable `hdc`'s various levels. This enables us to recognise potential patterns and comprehend which factors might be helpful in predicting the seriousness of heart disease.

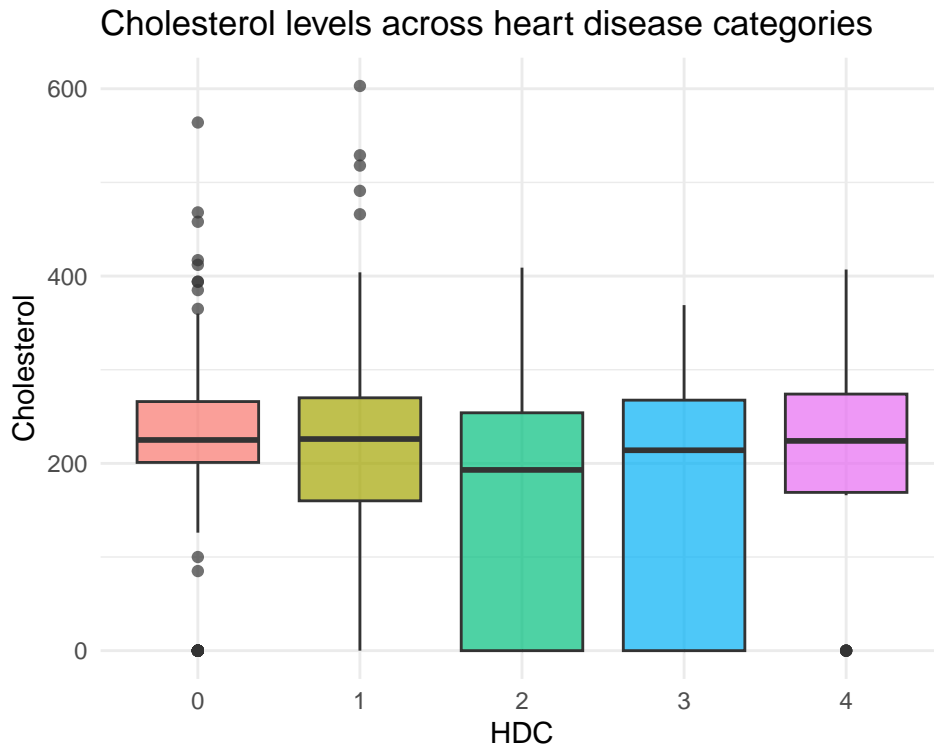
To investigate numerical factors like cholesterol, resting blood pressure, maximal heart rate, and ST depression, we employed boxplots. These charts illustrate the variations in these variables' distributions among the various categories of heart disease.

To determine how the frequency of each category varies throughout the `hdc` levels, we made barplots for categorical variables including the type of chest discomfort and the resting ECG result.

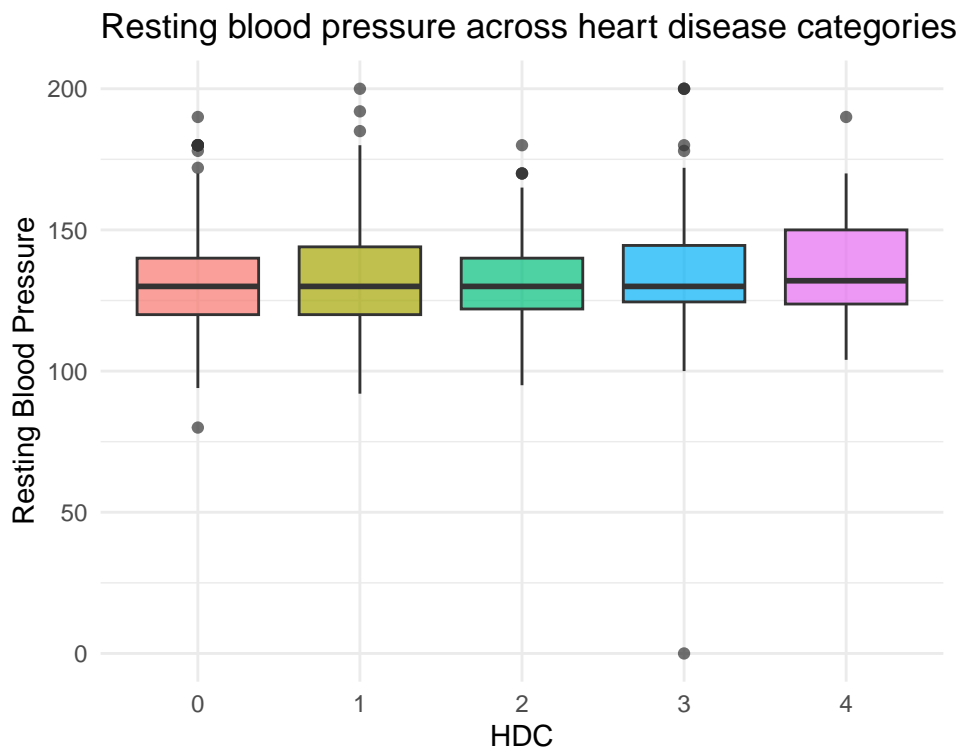
An introductory comprehension of the connections between predictors and the response variable is provided by this visual investigation.

Cholesterol levels across heart disease categories Cholesterol levels appear broadly similar across categories, though higher HDC groups show slightly more variability.

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

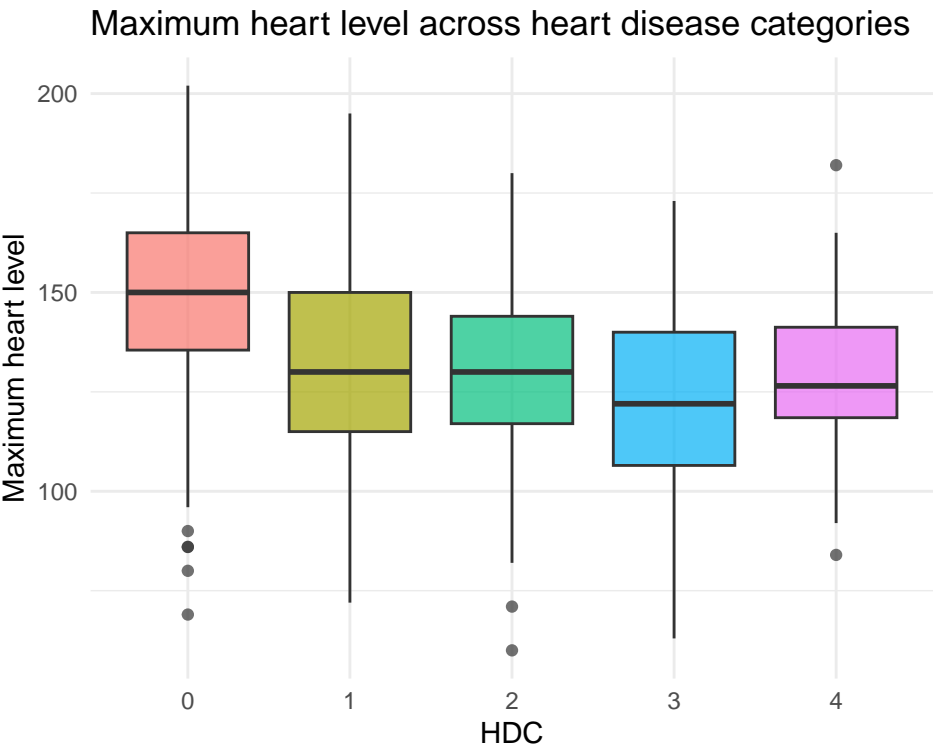


Resting blood pressure across heart disease categories Resting blood pressure shows a mild increasing trend with higher HDC, but overall differences between groups are modest.

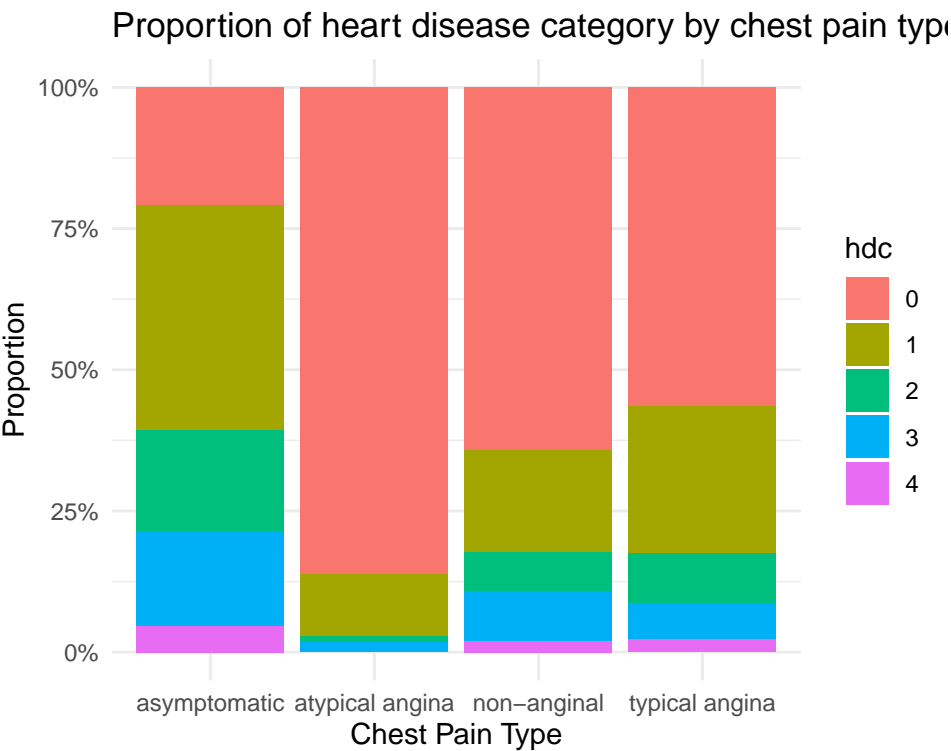


Maximum heart rate across heart disease categories Patients with higher HDC values gen-

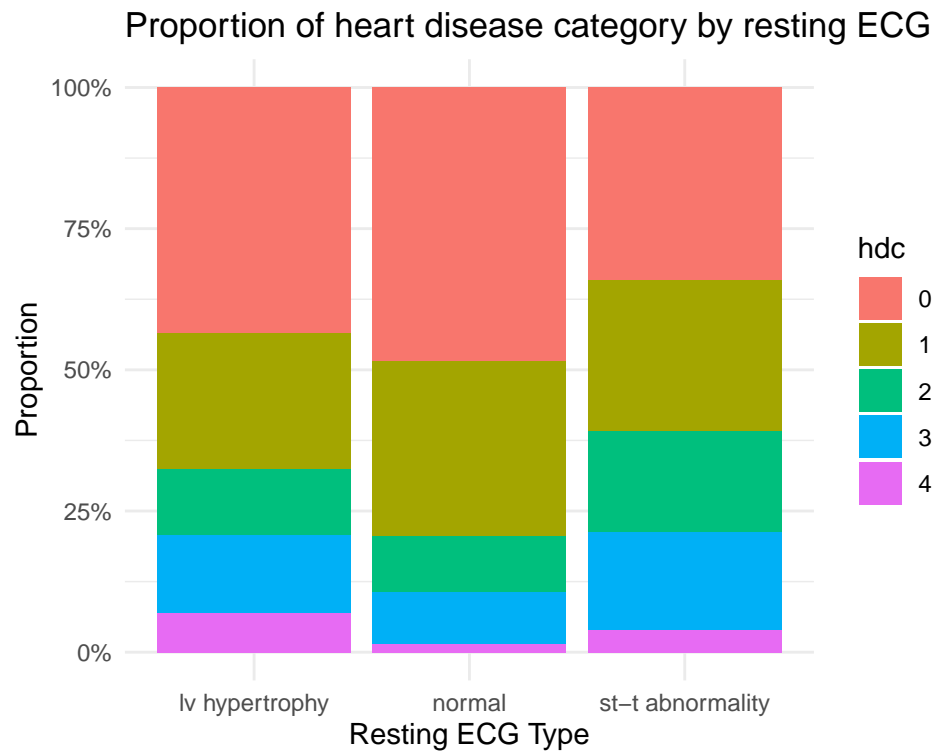
erally demonstrate lower maximum heart rates compared to those without heart disease.



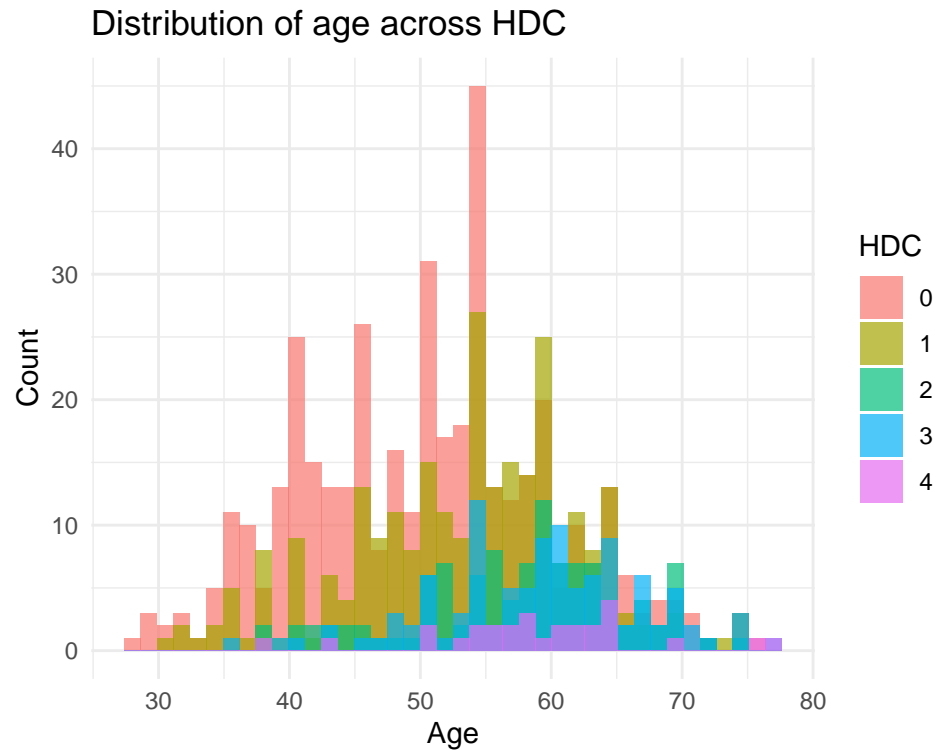
Proportion of heart disease category by chest pain type Atypical and non-anginal chest pain types contain a higher proportion of severe heart disease cases compared to typical angina.



Proportion of heart disease category by resting ECG type ST-T abnormalities and LV hypertrophy are associated with a greater share of higher heart disease categories compared to normal ECG results.



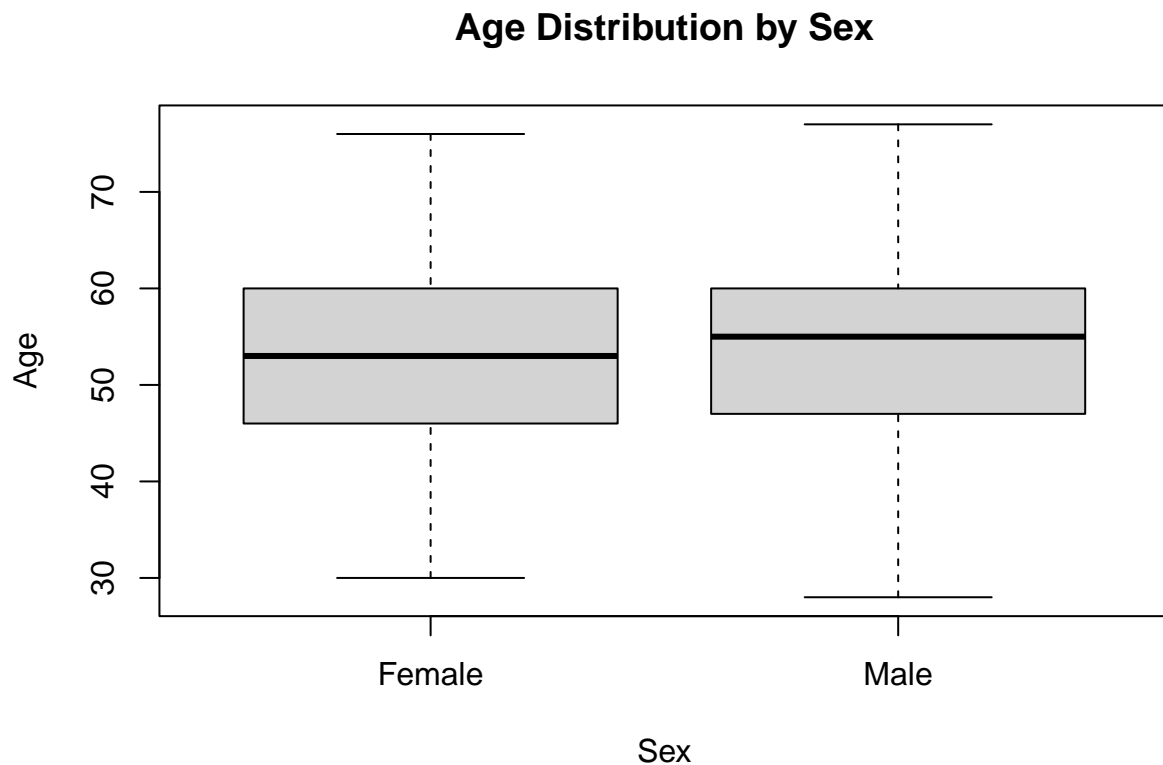
Distribution of age across HDC Patients with higher heart disease categories tend to be older, with the age distribution shifting toward later decades as HDC increases.



By looking at the box plot and p-test results we can conclude that we do not have statistically significant evidence that males and females have different mean ages.

With the p-test being below 0.05, we get the confirmation that the difference is not strong enough to be called significant.

```
boxplot(age ~ sex, data = heartData,  
        names = c("Female", "Male"),  
        main = "Age Distribution by Sex",  
        xlab = "Sex",  
        ylab = "Age")
```



```
t.test(age ~ sex, data = heartData)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by sex
## t = -1.7155, df = 301.61, p-value = 0.08727
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
## 95 percent confidence interval:
##  -2.8205168  0.1932128
## sample estimates:
## mean in group 0 mean in group 1
##      52.47423      53.78788
```

Multinomial Logistic Regression

We decided to fit the regression model using `multinom()` from the `nnet`, with `hdc` as the target variable and we included all of the predictors.

The baseline category is chosen to be $hdc = 0$ and for each other category $hdc = 1-4$ the model estimates how much each predictor affects the odds of belonging to a category 1-4 versus the baseline.

Multiple predictors appear as strong. Sex has a positive coefficient, indicating that males have higher chance of heart disease compared to females. Another one is `exang`. It increases the odds of being in any of the disease categories 1-4. In addition `cp` has a large negative indicating that certain chest pains reduces the risk of higher disease categories.

Overall the model selects `oldpeak`, `exang`, `cp` and maximum heart rate as key predictors

```
library(nnet)
multinomial_regression <- multinom(
  hdc ~ age + sex + place + chol + cp + trestbps
  + fbs + restecg + thalch + exang + oldpeak,
  data = heartData
)

summary(multinomial_regression)
```

Model Evaluation

(Cross-validation)

Model Improvement

(Stepwise model / alternative model)

Binary Logistic Regression

(Create `hdc01` + logistic model)

Model Comparison

(Compare multinomial vs binary)

Conclusion

(Brief summary)