TDS10 Final Project

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

Last compiled on: 25 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.

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...,K-1.

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

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

The probability of the baseline category is:

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

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 categorial 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
##   thalch    exang  oldpeak      hdc
##        0        0        0        0
```

```
##   age sex     place             cp trestbps chol fbs        restecg thalch
## 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.

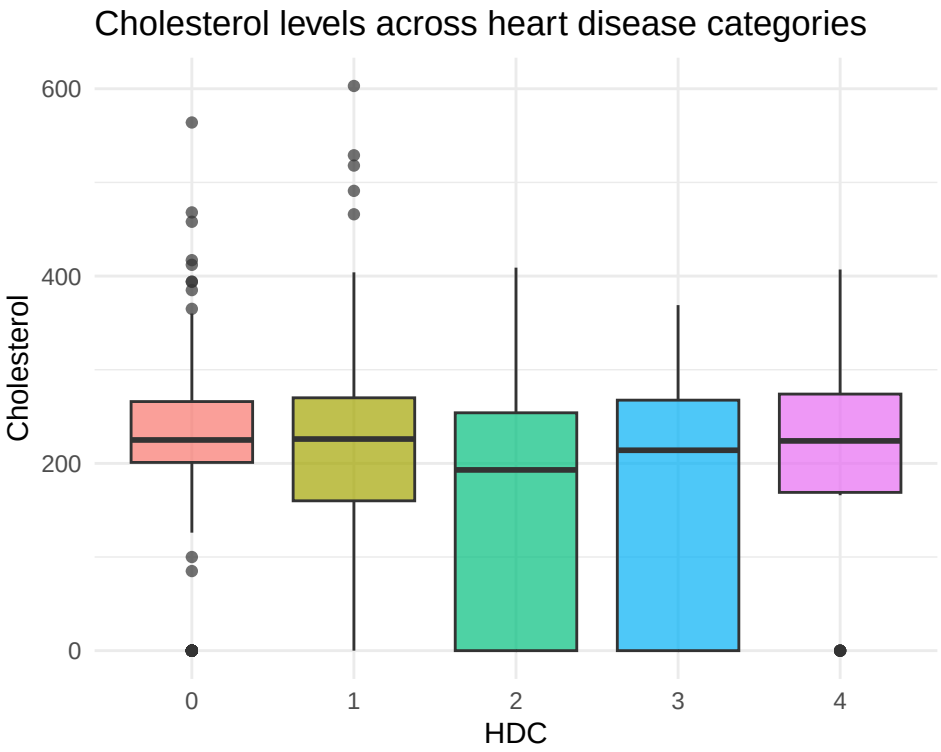
To investigate numerical factors like cholesterol, resting blood pressure, maximal heart rate, and ST depression, we employed boxplots. These charts illustrate the variations in these variables' distributions among the various categories of heart disease.

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.

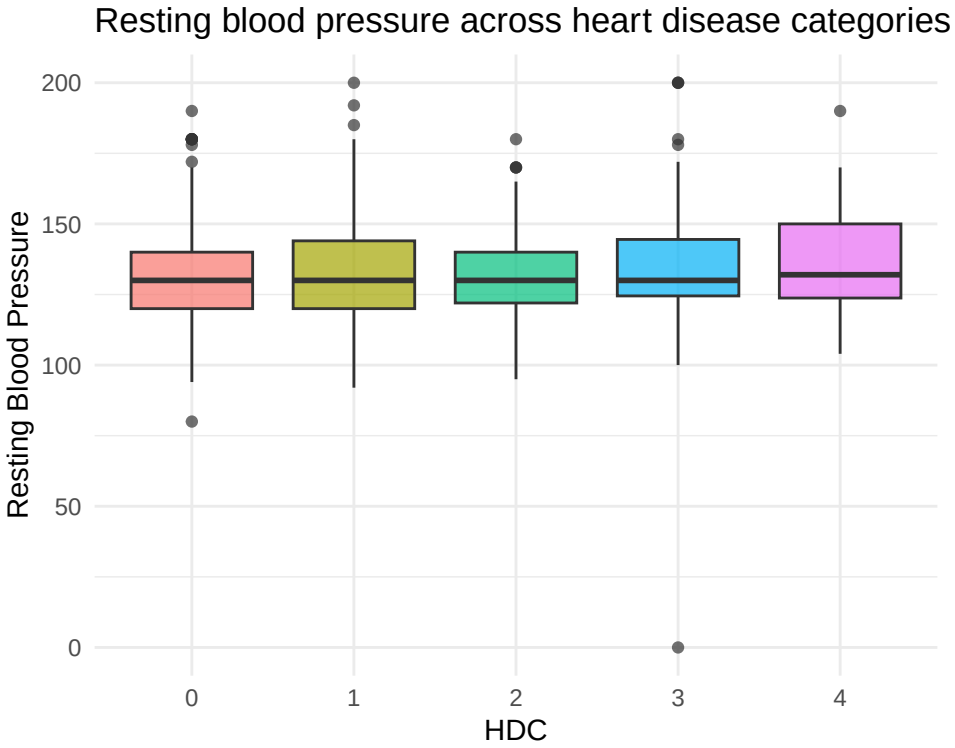
An introductory comprehension of the connections between predictors and the response variable is provided by this visual investigation.

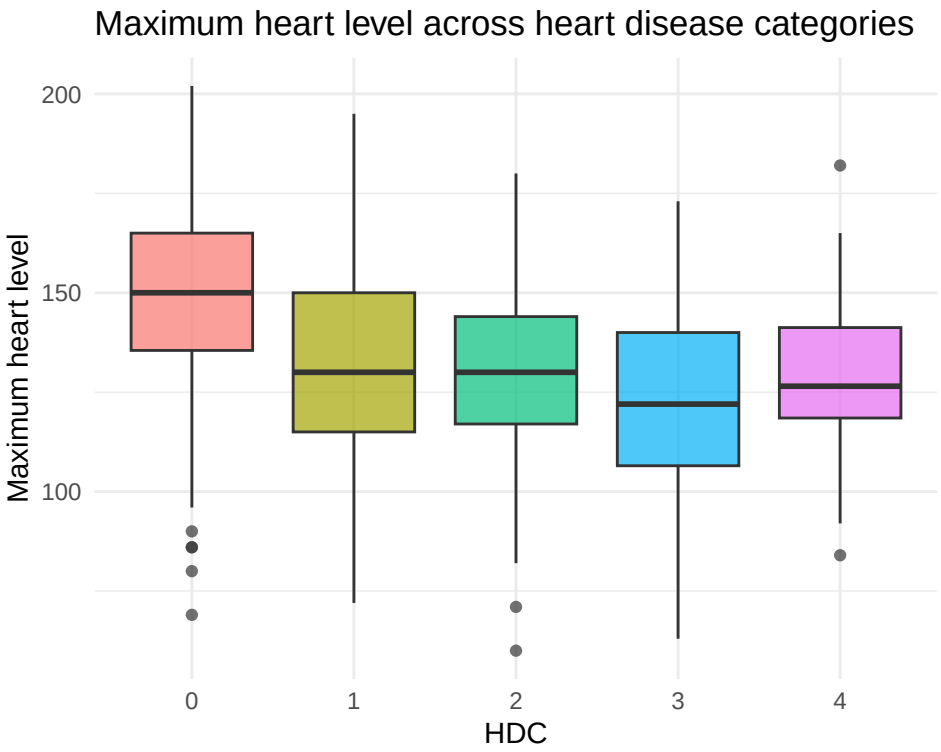
After cholesterol boxplot: From the boxplot, cholesterol tends to be higher in some of the higher heart disease categories, although the variability overlaps across groups.

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

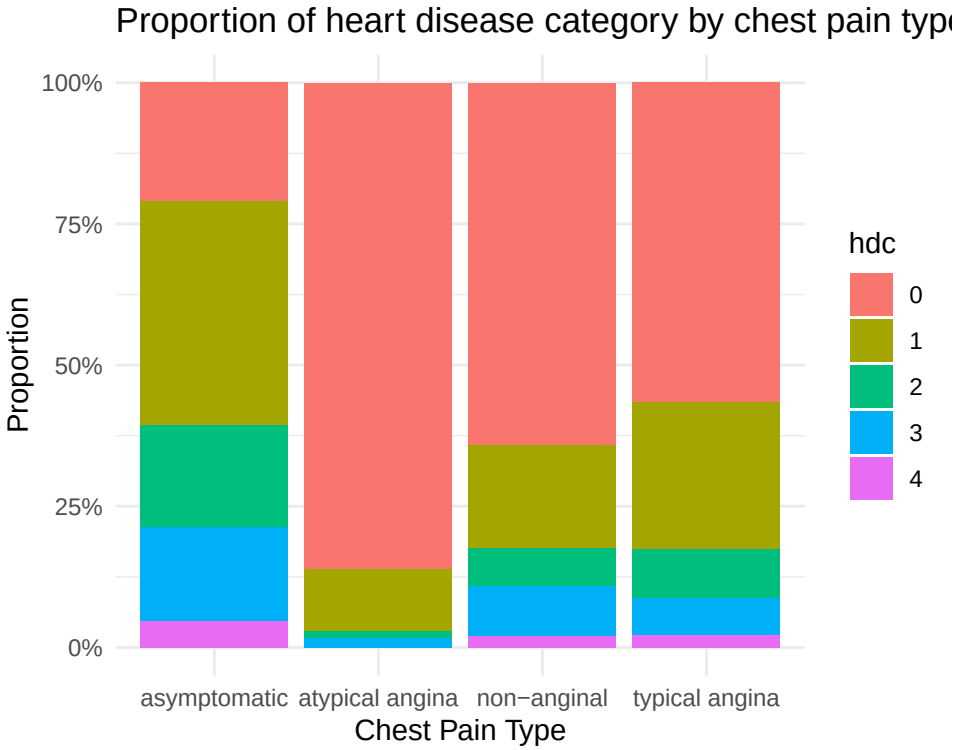


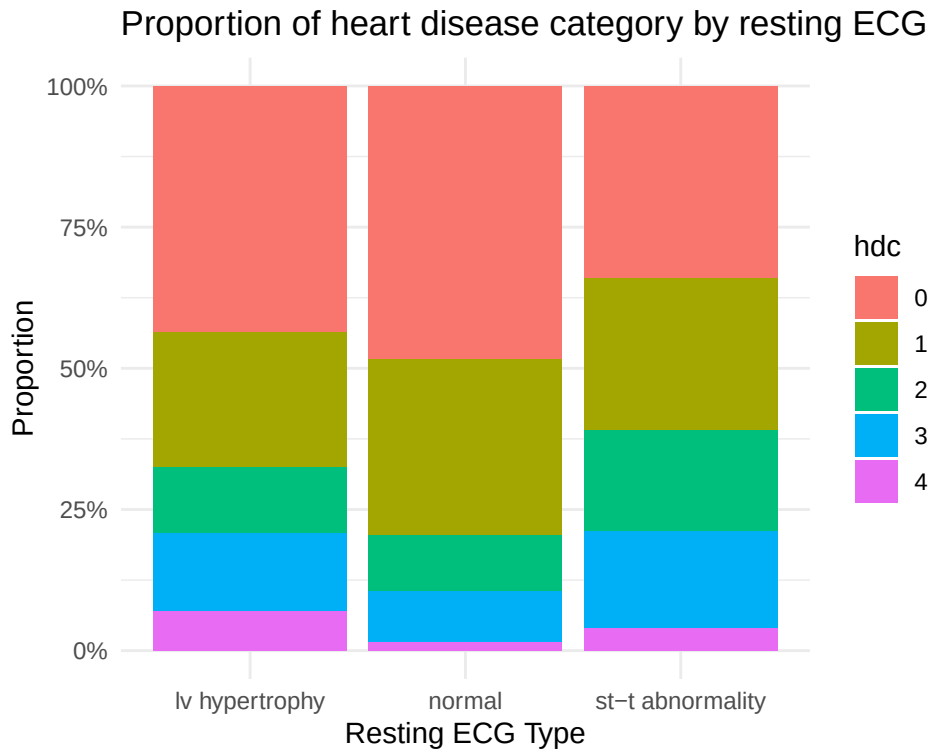
After Resting Blood Pressure boxplot: Blood pressure appears to increase slightly for patients with higher heart disease categories, but the differences are not very clear.

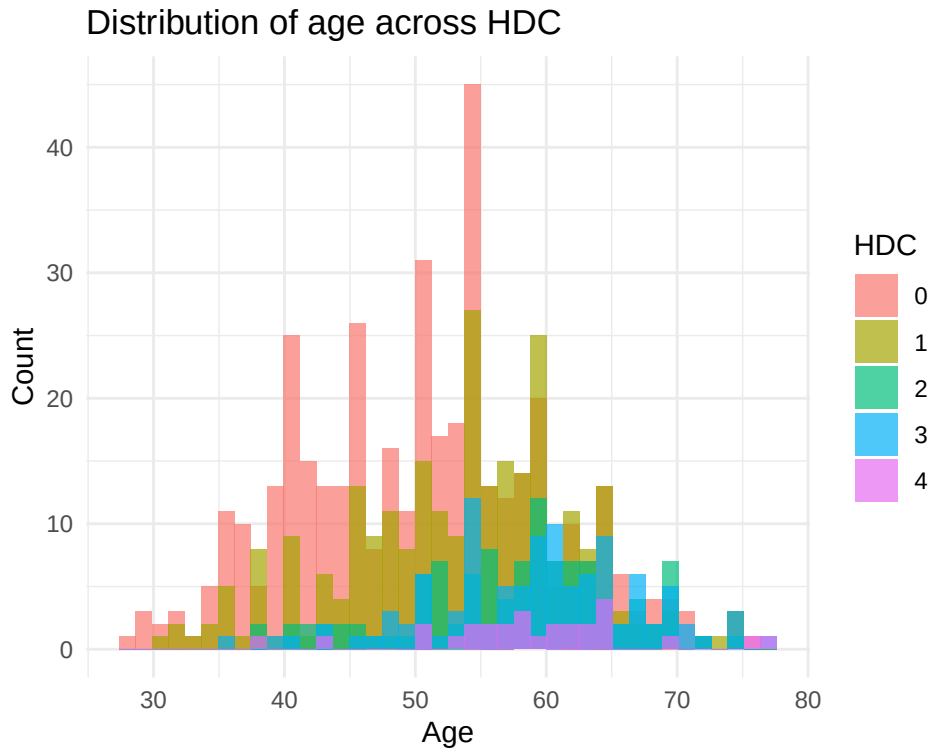




After cp vs hdc barplot: Certain chest pain types appear more often in higher heart disease categories, suggesting that chest pain type may be a useful predictor.




Proportion of heart disease category by resting ECG
100%
75%
50%
25%
0%
Proportion
lv hypertrophy
normal
st−t abnormality
Resting ECG Type
hdc
0
1
2
3
4


Distribution of age across HDC
40
30
20
10
0
Count
30
40
50
60
70
80
Age
HDC
0
1
2
3
4

Multinomial Logistic Regression

(Fit model + interpretation)

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)