MAS 646 Initial Report:

Ana Raquel Chacin, Oliver Mazariegos, Shannon Land, Mariana Gomez Del Nogal

April 07, 2025

1 Introduction

2 Dataset Description

2.1 Categorical Variables

2.2 Numerical Variables

3 Exploratory Data Analysis

3.1 Structure of Dataset

```
## 'data.frame':    1888 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "Typical angina",..: 4 3 2 2 1 1 2 2 3 3 ...
##  $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : Factor w/ 2 levels "False","True": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "Normal","ST-T wave abnormality",..: 1 2 1 2 2 2 1 2
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : Factor w/ 3 levels "Upsloping","Flat",..: 1 1 3 3 3 2 2 3 3 3 ...
##  $ ca      : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ thal    : Factor w/ 3 levels "Normal","Fixed defect",..: 1 2 2 2 2 1 2 3 3 2 ...
##  $ target  : Factor w/ 2 levels "No heart attack",..: 2 2 2 2 2 2 2 2 2 2 ...
```

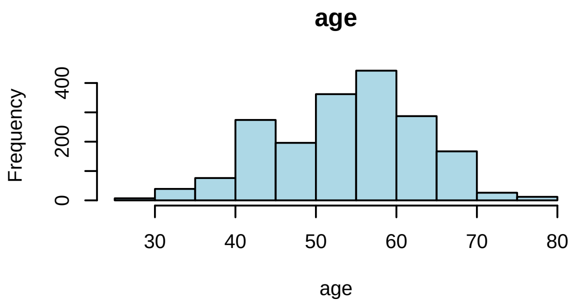
3.2 Checking for Null Values

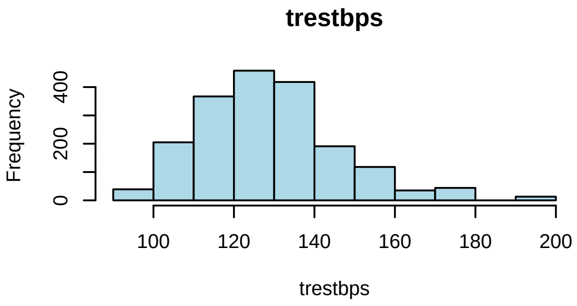
```
##      age      sex       cp trestbps     chol      fbs  restecg thalachh
##        0        0      130        0        0        0        0        0
##    exang  oldpeak    slope       ca     thal   target
##        0        0       18       28      130        0
```

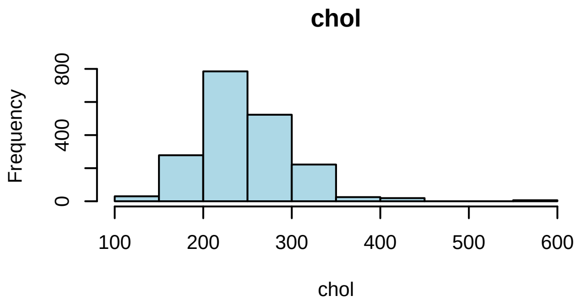
3.3 Statistical Summary

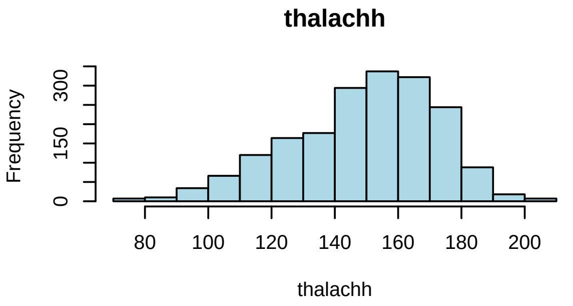
```
##       age            sex                        cp         trestbps
##  Min.   :29.00   Female: 588   Typical angina  :766   Min.   : 94.0
##  1st Qu.:47.75   Male  :1300   Atypical angina :291   1st Qu.:120.0
##  Median :55.00                 Non-anginal pain:499   Median :130.0
##  Mean   :54.35                 Asymptomatic    :202   Mean   :131.5
##  3rd Qu.:61.00                 NA’s            :130   3rd Qu.:140.0
##  Max.   :77.00                                        Max.   :200.0
##       chol          fbs                               restecg
##  Min.   :126.0   False:1608   Normal                      :918
##  1st Qu.:211.0   True : 280   ST-T wave abnormality       :812
##  Median :241.0                Left ventricular hypertrophy:158
##  Mean   :246.9
##  3rd Qu.:276.0
##  Max.   :564.0
##     thalachh     exang         oldpeak              slope        ca
##  Min.   : 71.0   No :1262   Min.   :0.000   Upsloping  :114   0   :1084
##  1st Qu.:133.0   Yes: 626   1st Qu.:0.000   Flat       :882   1   : 410
##  Median :152.0              Median :0.800   Downsloping:874   2   : 239
##  Mean   :149.4              Mean   :1.054   NA’s       : 18   3   : 127
##  3rd Qu.:166.0              3rd Qu.:1.600                     NA’s:  28
##  Max.   :202.0              Max.   :6.200
##                  thal                 target
##  Normal           : 96   No heart attack:911
##  Fixed defect     :874   Heart attack   :977
##  Reversible defect:788
##  NA’s             :130
##
##
```

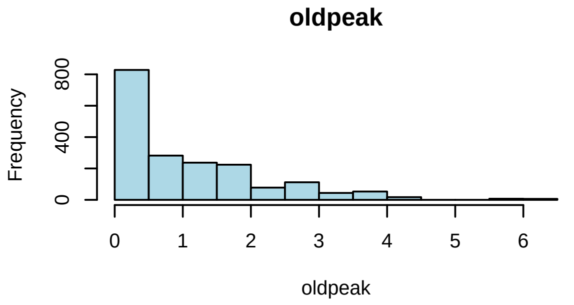
3.4 Frequency Distributions



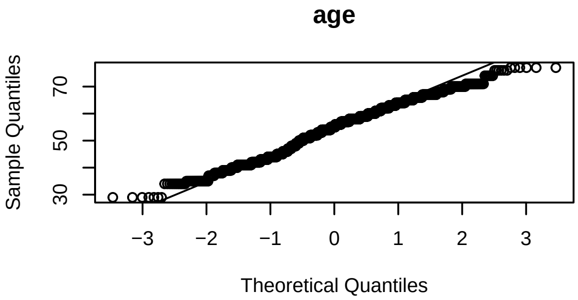


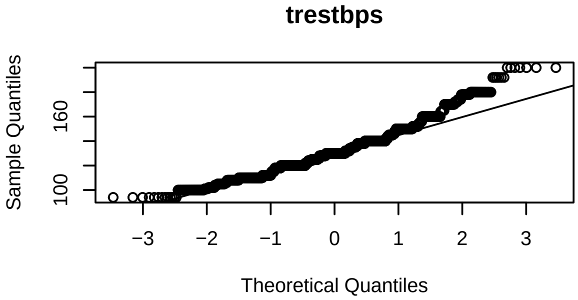


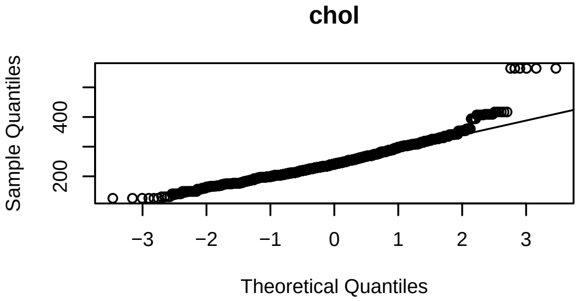


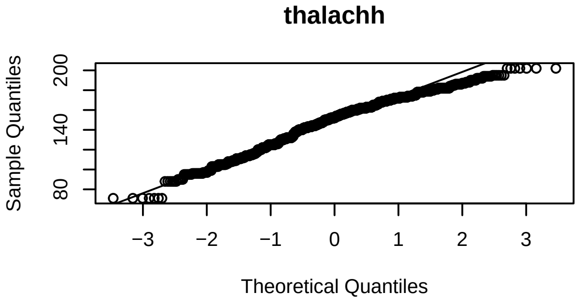


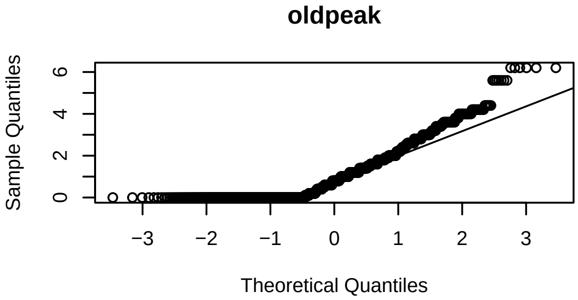
3.5 Check for Normality



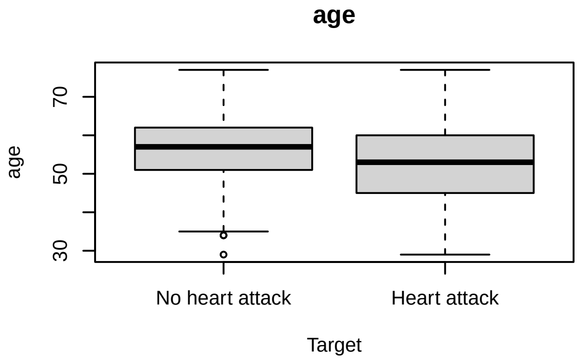


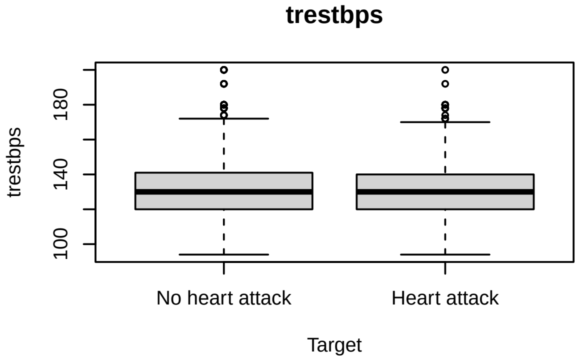


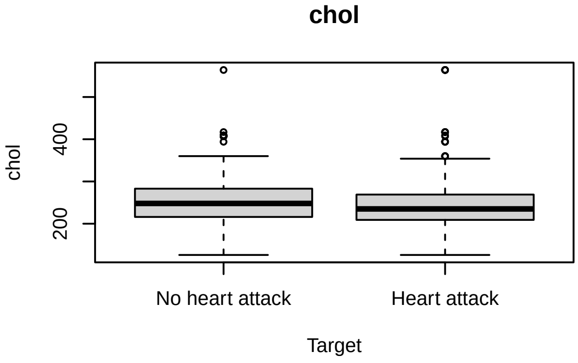


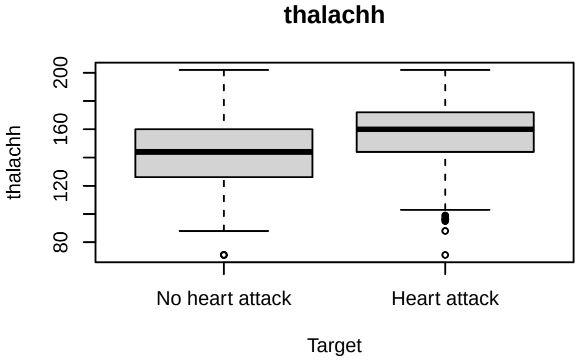


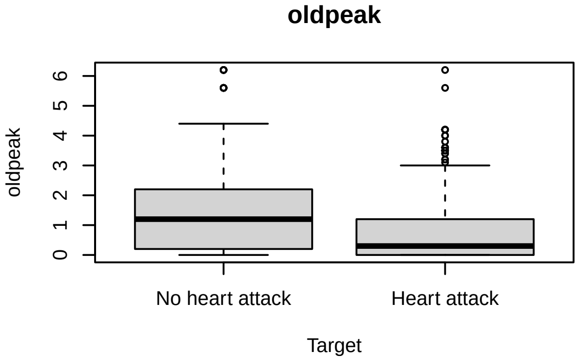
3.6 Pairwise Relationship: Numerical Predictors



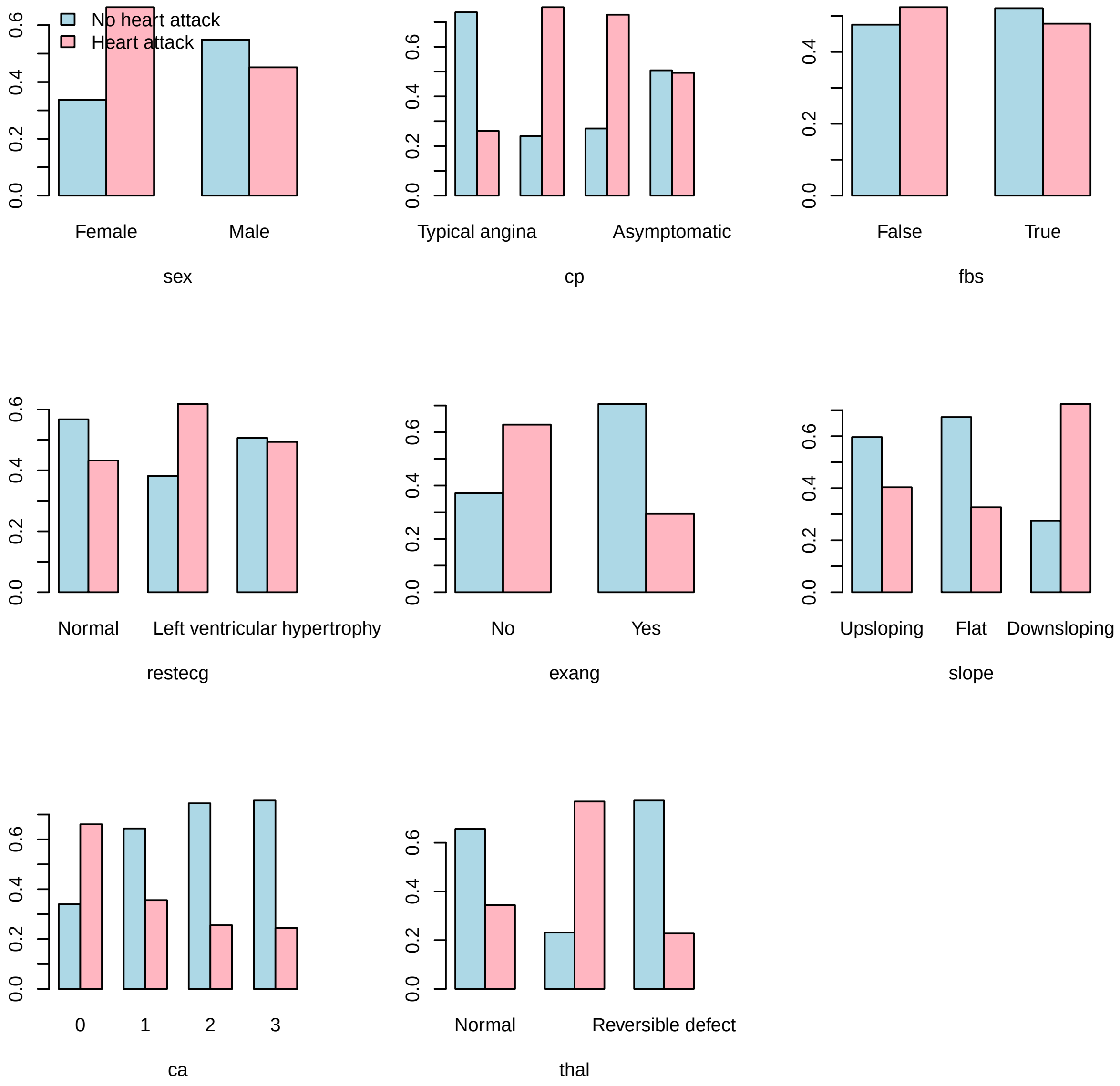








3.7 Pairwise Relationship: Categorical Predictors



4 Model Building

4.1 Simple Logistic Regression Model

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = df)
##
```

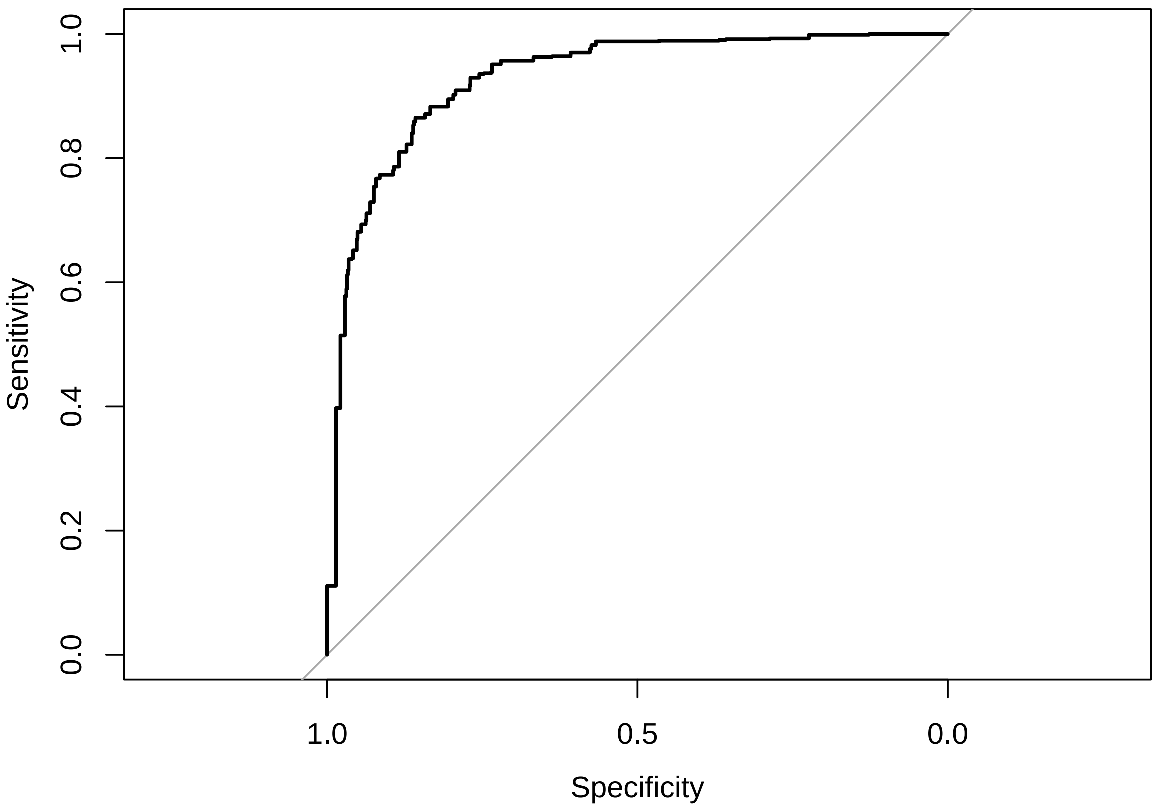
```
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -0.437032   1.138619  -0.384 0.701107
## age                                  0.031159   0.010011   3.113 0.001855 **
## sexMale                             -0.791800   0.193163  -4.099 4.15e-05 ***
## cpAtypical angina                    0.884477   0.235066   3.763 0.000168 ***
## cpNon-anginal pain                   1.511737   0.205525   7.355 1.90e-13 ***
## cpAsymptomatic                       1.191119   0.247181   4.819 1.44e-06 ***
## trestbps                            -0.013597   0.004643  -2.929 0.003403 **
## chol                                -0.002063   0.001581  -1.304 0.192077
## fbsTrue                              0.472182   0.229404   2.058 0.039562 *
## restecgST-T wave abnormality         0.750704   0.161851   4.638 3.51e-06 ***
## restecgLeft ventricular hypertrophy -1.573527   0.429201  -3.666 0.000246 ***
## thalachh                             0.015176   0.004519   3.358 0.000785 ***
## exangYes                            -0.592068   0.183573  -3.225 0.001259 **
## oldpeak                             -0.256745   0.094531  -2.716 0.006608 **
## slopeFlat                           -0.597720   0.342591  -1.745 0.081036 .
## slopeDownsloping                     0.595699   0.364304   1.635 0.102013
## ca1                                 -1.973069   0.204303  -9.658  < 2e-16 ***
## ca2                                 -2.944669   0.307642  -9.572  < 2e-16 ***
## ca3                                 -2.140256   0.381295  -5.613 1.99e-08 ***
## thalFixed defect                     0.420311   0.317641   1.323 0.185760
## thalReversible defect               -1.650432   0.314411  -5.249 1.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2320.7  on 1673  degrees of freedom
## Residual deviance: 1140.0  on 1653  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 1182
##
## Number of Fisher Scoring iterations: 6
```

4.2 Model Evaluation

4.2.1 Confusion Matrix

```
##                  Reference
## Prediction        No heart attack Heart attack
##   No heart attack             697          108
##   Heart attack                139          730
```

4.2.2 ROC & AUC



```
## Area under the curve: 0.9283
```