Heart Disease Analysis

Final project presentation,
CPE213 (Data models)

By Jidapa Thongnirun

Introduction

Introduction

- Heart disease is a fatal illness
- 17.9 million deaths each year
- Knowing analyzing data can make a better life

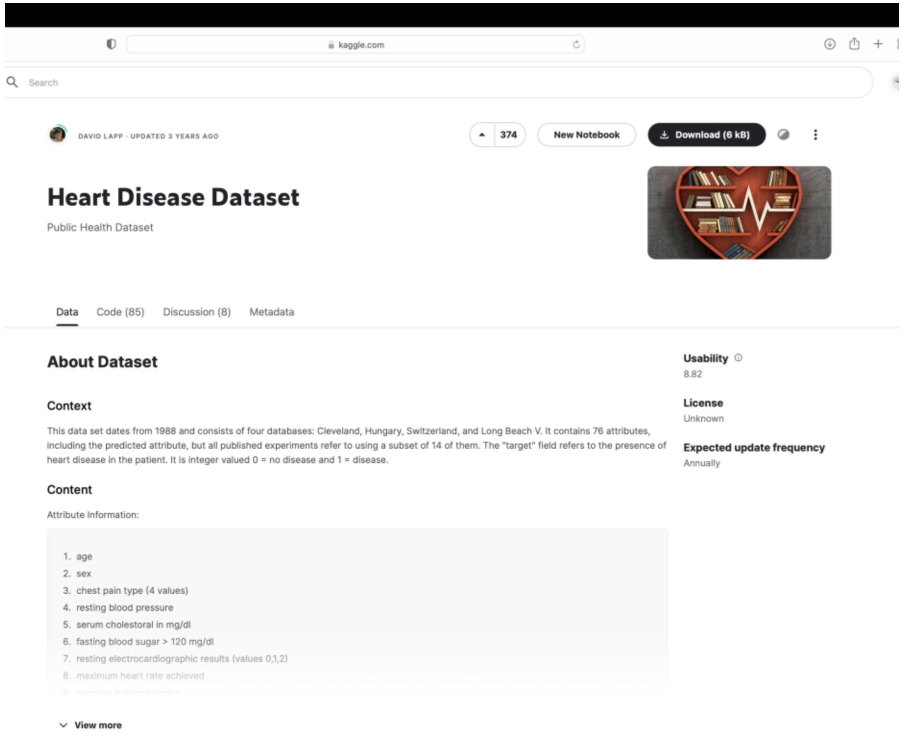
Analytic objective

Analytic objective

- Able to understand the factors that increase the rate

- Able to analyze the risk of each individual person

Data description



DAVID LAPP · UPDATED 3 YEARS AGO

Heart Disease Dataset

Public Health Dataset

About Dataset

Context

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Content

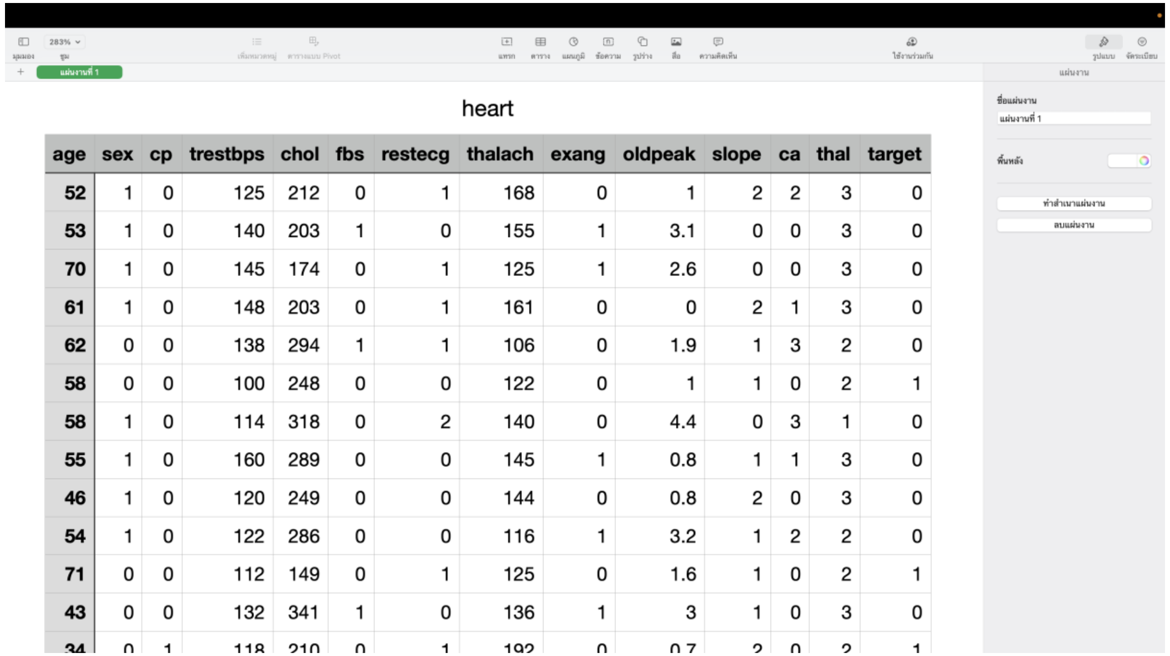
Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina

Usability
8.82

License
Unknown

Expected update frequency
Annually



แผ่นงานที่ 1

ชื่อแผ่นงาน: แผ่นงานที่ 1

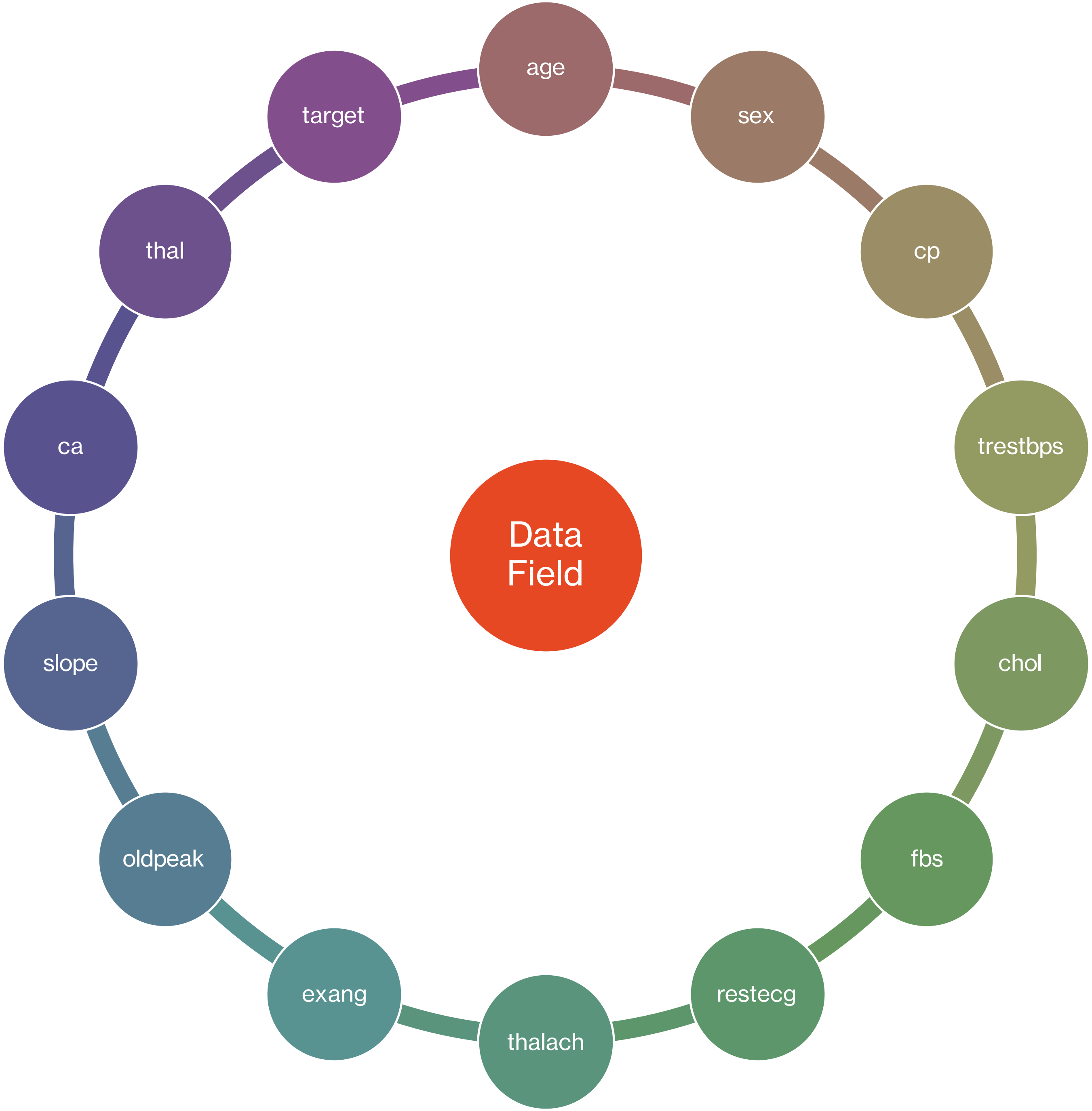
พื้นหลัง

heart

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1

Introduce to dataset

- “Heart Disease Dataset” which was uploaded by David Lapp on Kaggle.
- From 1988, contains 4 database.
- There is 14 field and 1025 row


Data Field
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target

Data description

age

- Patient's age

sex

- Patient's gender
  - Male [1]
  - Female [0]

Data description

cp

- Patient’s chest pain type
  - Asymptomatic [0]
  - Atypical angina [1]
  - Non-anginal pain [2]
  - Typical angina [3]

trestbps

- Patient’s resting blood (mmHg)

Data description

chol

- Patient’s cholesterol measurement (mg/dl)

fbs

- Patient’s fasting blood sugar > 120 mg/dl
  - True [1]
  - False [0]

Data description

restecg

- Patient’s resting electrocardiographic results
  - Showing probable or definite left ventricular hypertrophy by Estes’ criteria [0]
  - Normal [1]
  - Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)[2]

thalach

- Patient’s maximum heart rate achieved

Data description

exang

- Exercise-induced angina
  - Yes [1]
  - No [0]

oldpeak

- ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot.)

Data description

slope

- The slope of the peak exercise ST segment
  - Down sloping [0]
  - Flat [1]
  - Up sloping [2]

ca

- The number of major vessels
  - [0]
  - [1]
  - [2]
  - [3]
  - [4]

Data description

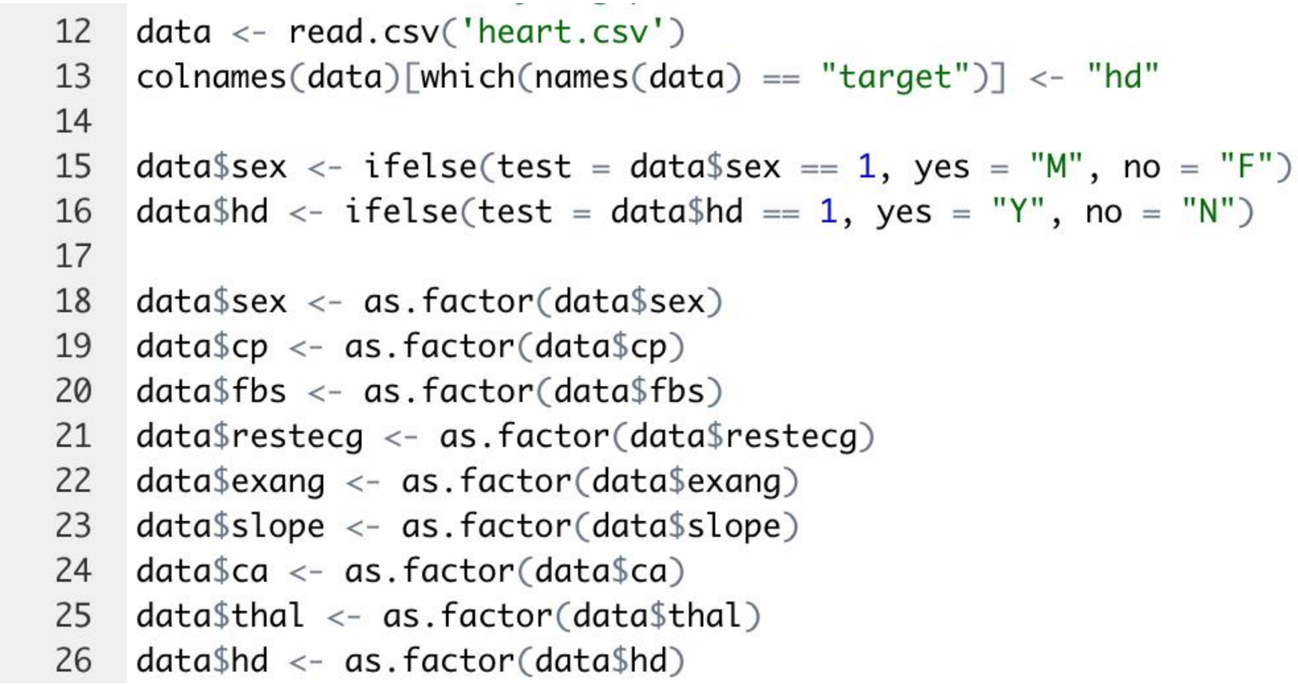
thal

- A blood disorder called thalassemia
  - [0]
  - [1]
  - [2]
  - [3]

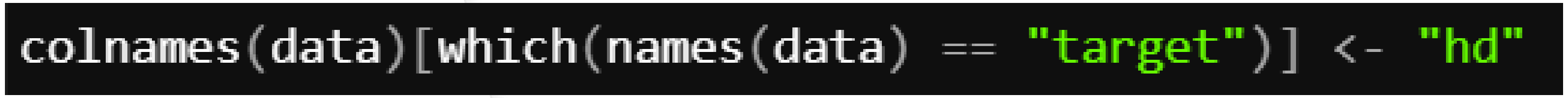
target

- Diagnosis of heart disease
  - ○ Yes / disease [1]
  - ○ No / No disease [0]

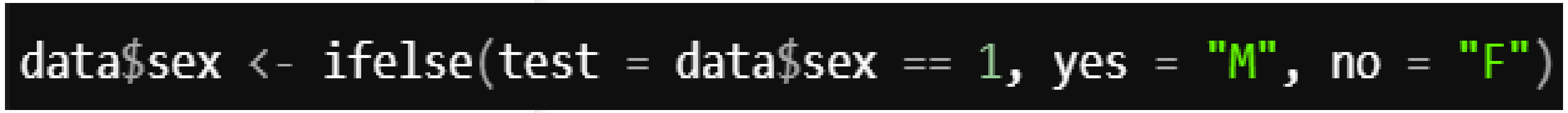
Data preparation



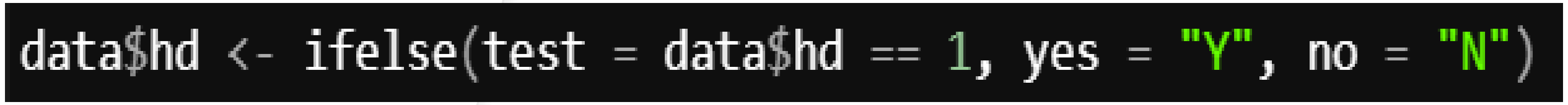
```r
data <- read.csv('heart.csv')
colnames(data)[which(names(data) == "target")] <- "hd"

data$sex <- ifelse(test = data$sex == 1, yes = "M", no = "F")
data$hd <- ifelse(test = data$hd == 1, yes = "Y", no = "N")

data$sex <- as.factor(data$sex)
data$cp <- as.factor(data$cp)
data$fbs <- as.factor(data$fbs)
data$restecg <- as.factor(data$restecg)
data$exang <- as.factor(data$exang)
data$slope <- as.factor(data$slope)
data$ca <- as.factor(data$ca)
data$thal <- as.factor(data$thal)
data$hd <- as.factor(data$hd)
```



```
colnames(data)[which(names(data) == "target")] <- "hd"
```

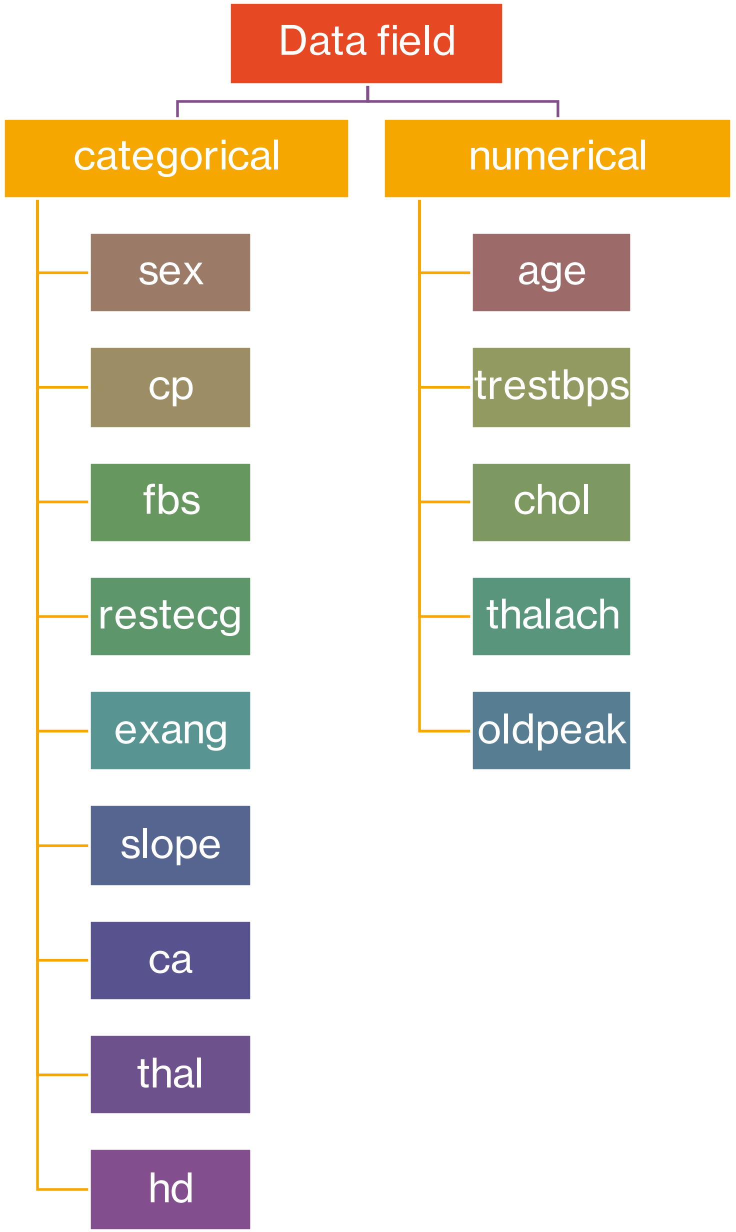


```
data$sex <- ifelse(test = data$sex == 1, yes = "M", no = "F")
```

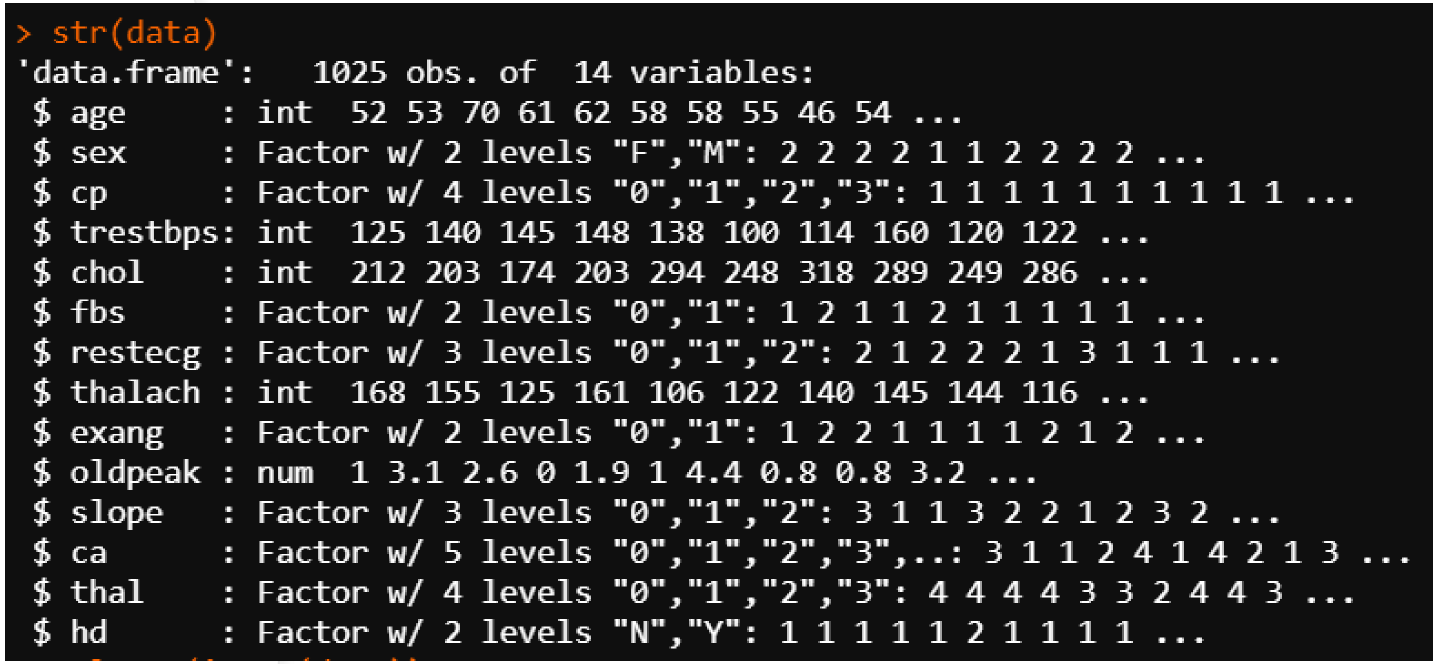


```
data$hd <- ifelse(test = data$hd == 1, yes = "Y", no = "N")
```

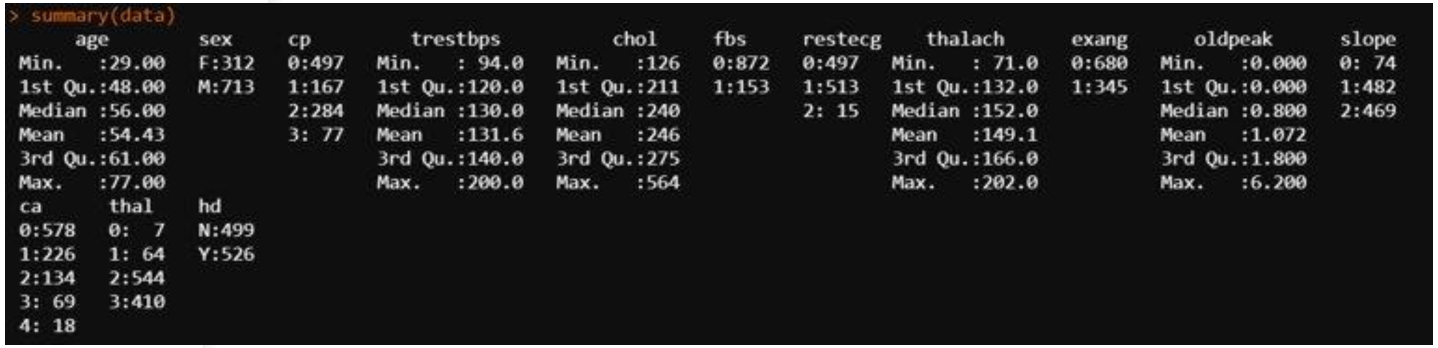
Data exploration



Data visualization



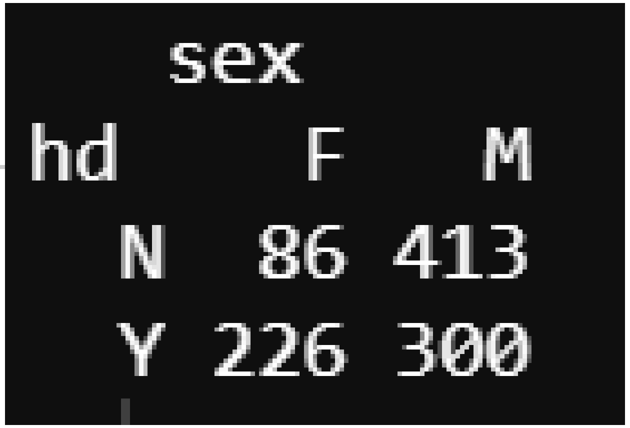
```
> str(data)
'data.frame':	1025 obs. of  14 variables:
 $ age     : int  52 53 70 61 62 58 58 55 46 54 ...
 $ sex     : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 2 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ trestbps: int  125 140 145 148 138 100 114 160 120 122 ...
 $ chol    : int  212 203 174 203 294 248 318 289 249 286 ...
 $ fbs     : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 1 1 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 2 1 2 2 2 1 3 1 1 1 ...
 $ thalach : int  168 155 125 161 106 122 140 145 144 116 ...
 $ exang   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
 $ oldpeak : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
 $ slope   : Factor w/ 3 levels "0","1","2": 3 1 1 3 2 2 1 2 3 2 ...
 $ ca      : Factor w/ 5 levels "0","1","2","3",..: 3 1 1 2 4 1 4 2 1 3 ...
 $ thal    : Factor w/ 4 levels "0","1","2","3": 4 4 4 4 3 3 2 4 4 3 ...
 $ hd      : Factor w/ 2 levels "N","Y": 1 1 1 1 1 2 1 1 1 1 ...
```



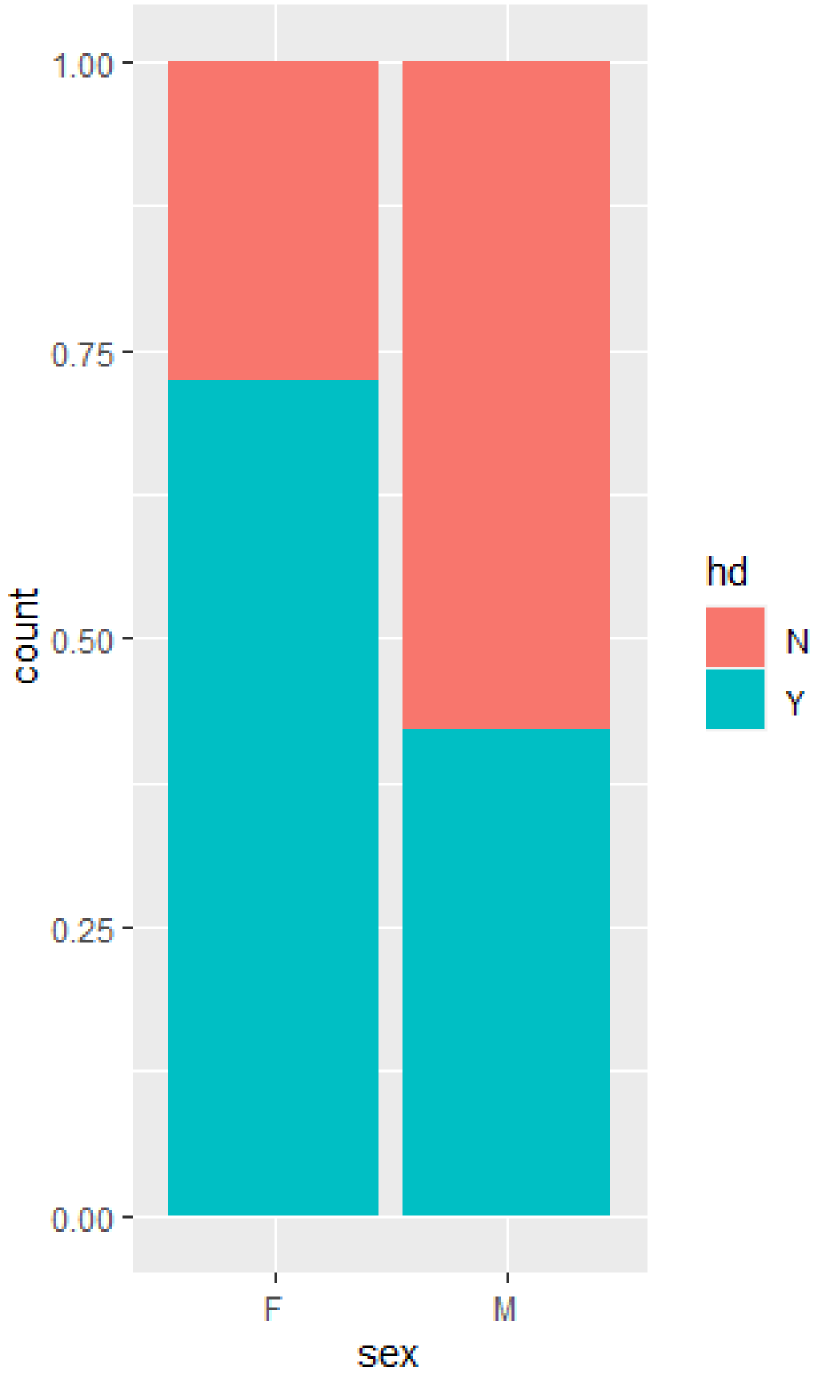
```
> summary(data)
      age         sex     cp         trestbps          chol        fbs     restecg     thalach       exang       oldpeak       slope  
 Min.   :29.00   F:312   0:497   Min.   : 94.0   Min.   :126   0:872   0:497   Min.   : 71.0   0:680   Min.   :0.000   0: 74  
 1st Qu.:48.00   M:713   1:167   1st Qu.:120.0   1st Qu.:211   1:153   1:513   1st Qu.:132.0   1:345   1st Qu.:0.000   1:482  
 Median :56.00           2:284   Median :130.0   Median :240           2: 15   Median :152.0           Median :0.800   2:469  
 Mean   :54.43           3: 77   Mean   :131.6   Mean   :246                   Mean   :149.1           Mean   :1.072          
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:275                   3rd Qu.:166.0           3rd Qu.:1.800          
 Max.   :77.00                   Max.   :200.0   Max.   :564                   Max.   :202.0           Max.   :6.200          
 ca       thal     hd     
 0:578   0:  7   N:499  
 1:226   1: 64   Y:526  
 2:134   2:544          
 3: 69   3:410          
 4: 18                  
```

Categorical Variables

hd and sex

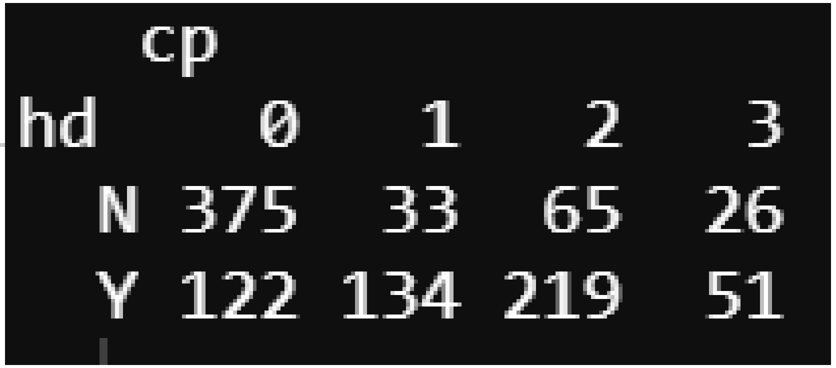


```
    sex
hd     F   M
  N   86 413
  Y  226 300
```

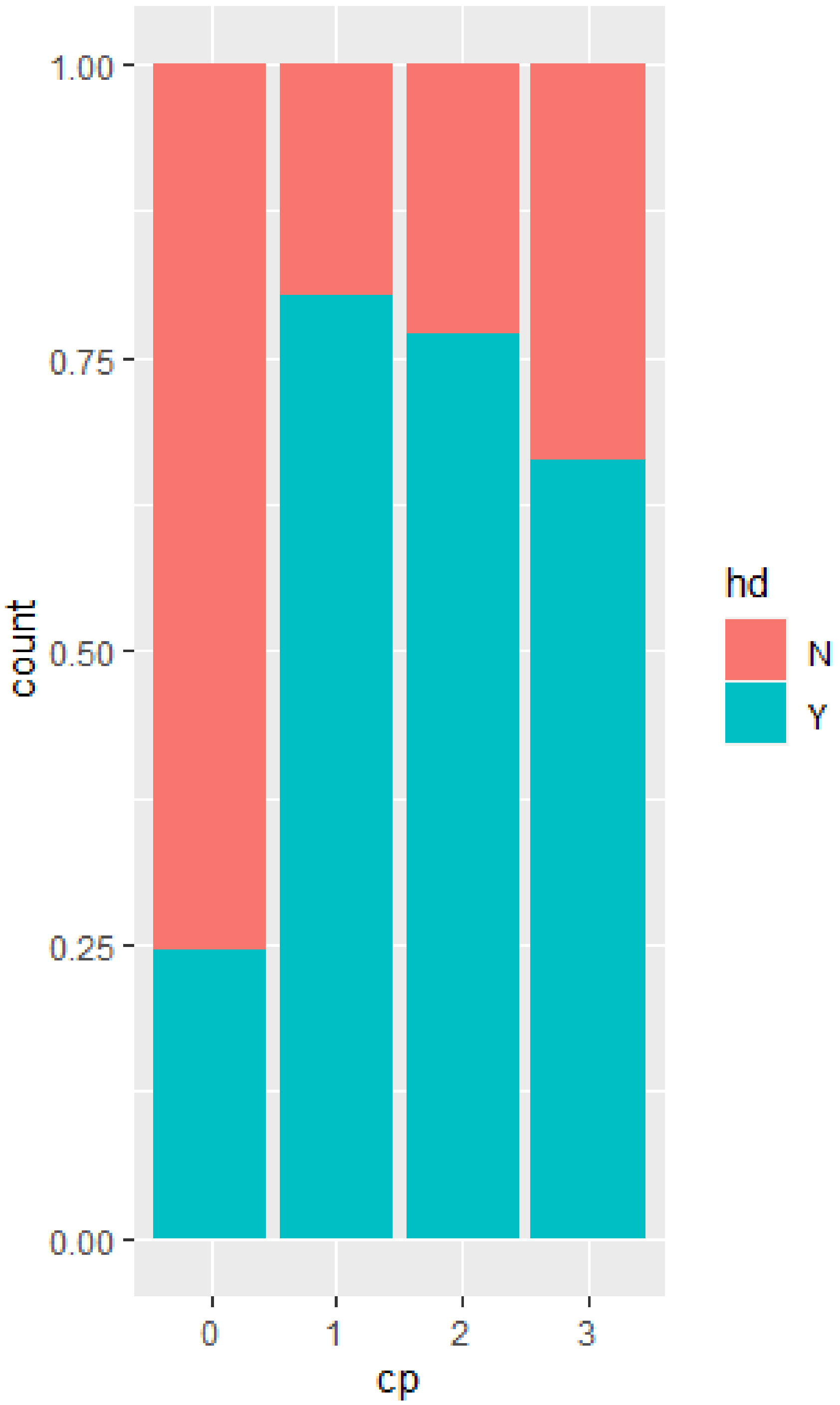


Categorical Variables

hd and cp

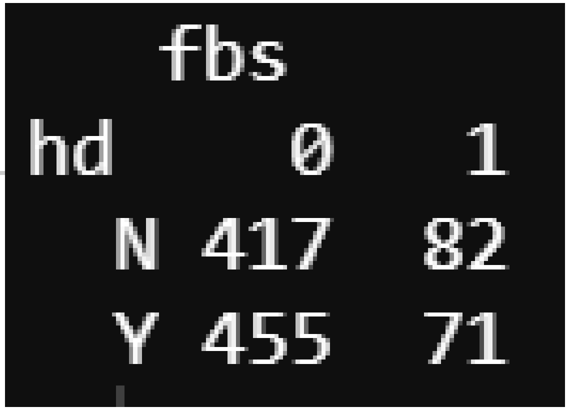


```
    cp
hd     0   1   2   3
  N  375  33  65  26
  Y  122 134 219  51
```

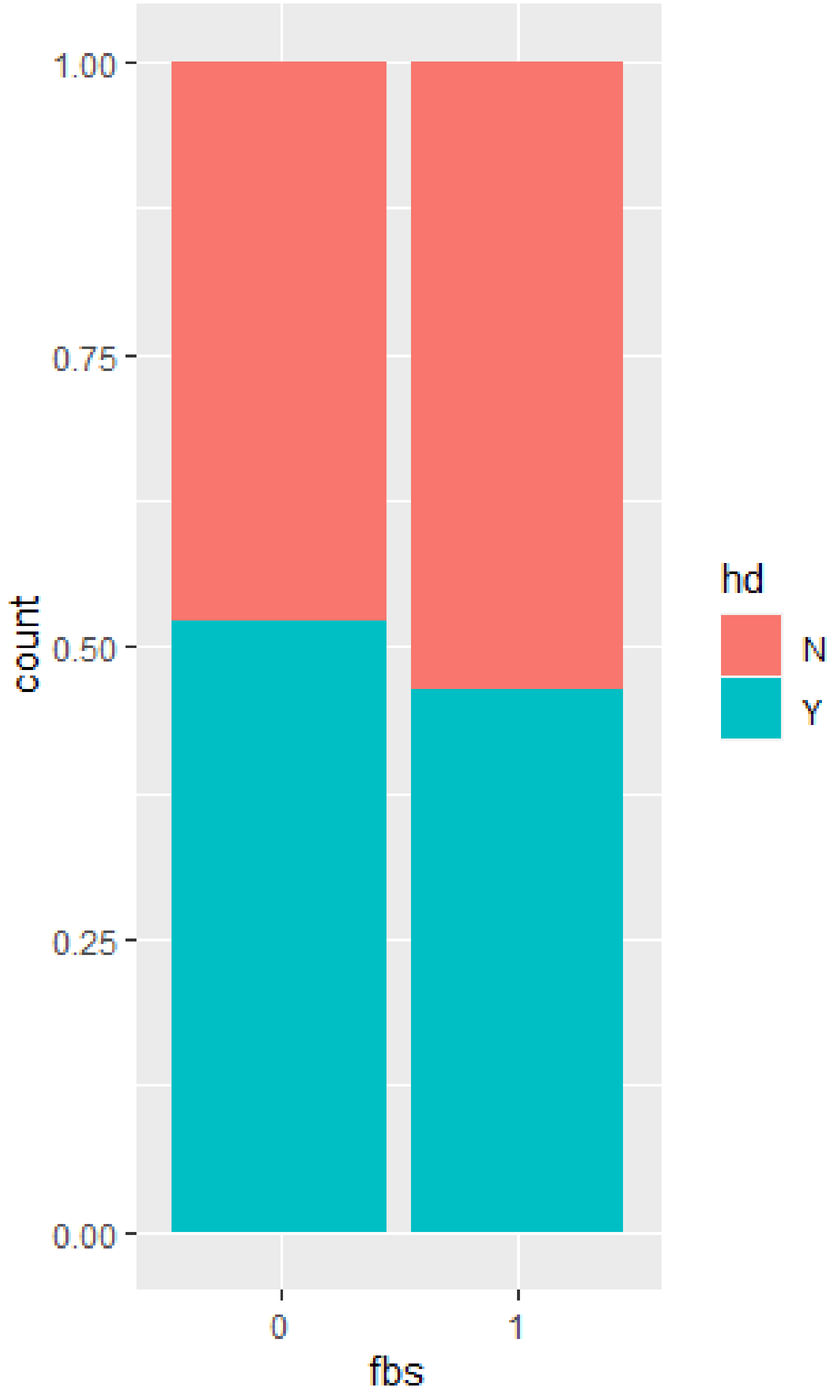


Categorical Variables

hd and fbs

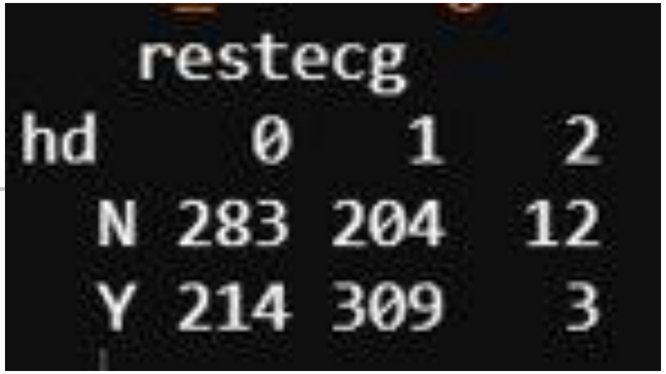


```
    fbs
hd     0   1
  N 417  82
  Y 455  71
```

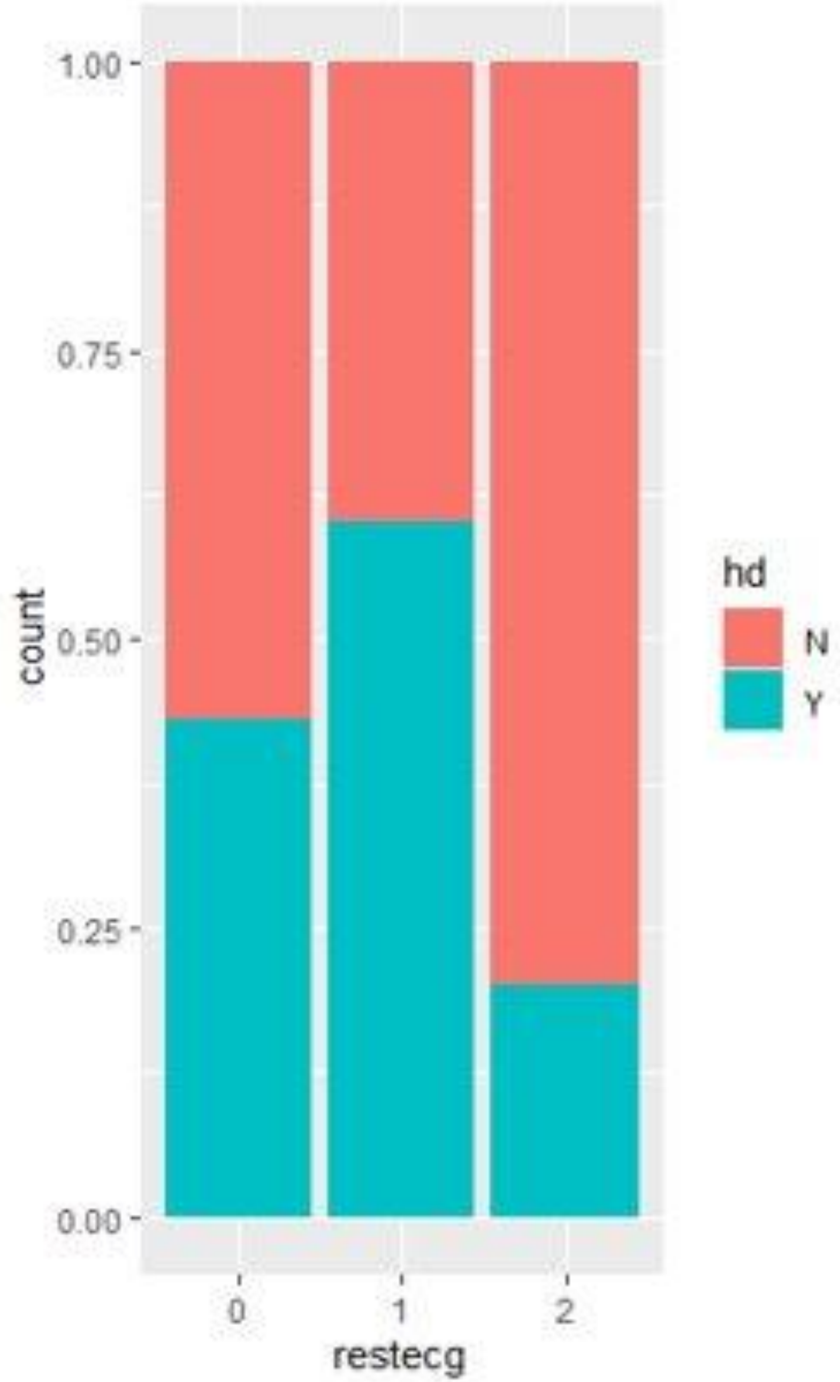


Categorical Variables

hd and restecg

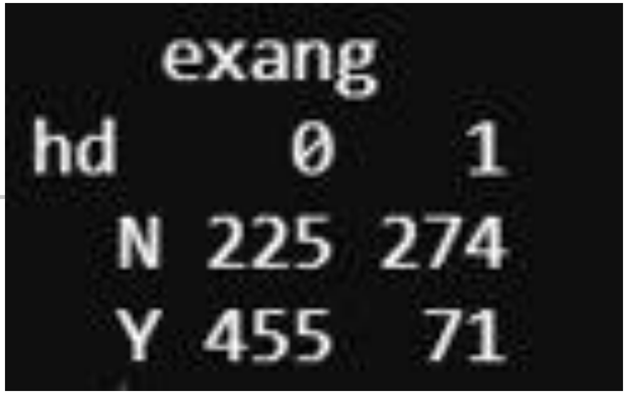


```
       restecg
hd      0   1   2
  N   283 204  12
  Y   214 309   3
```

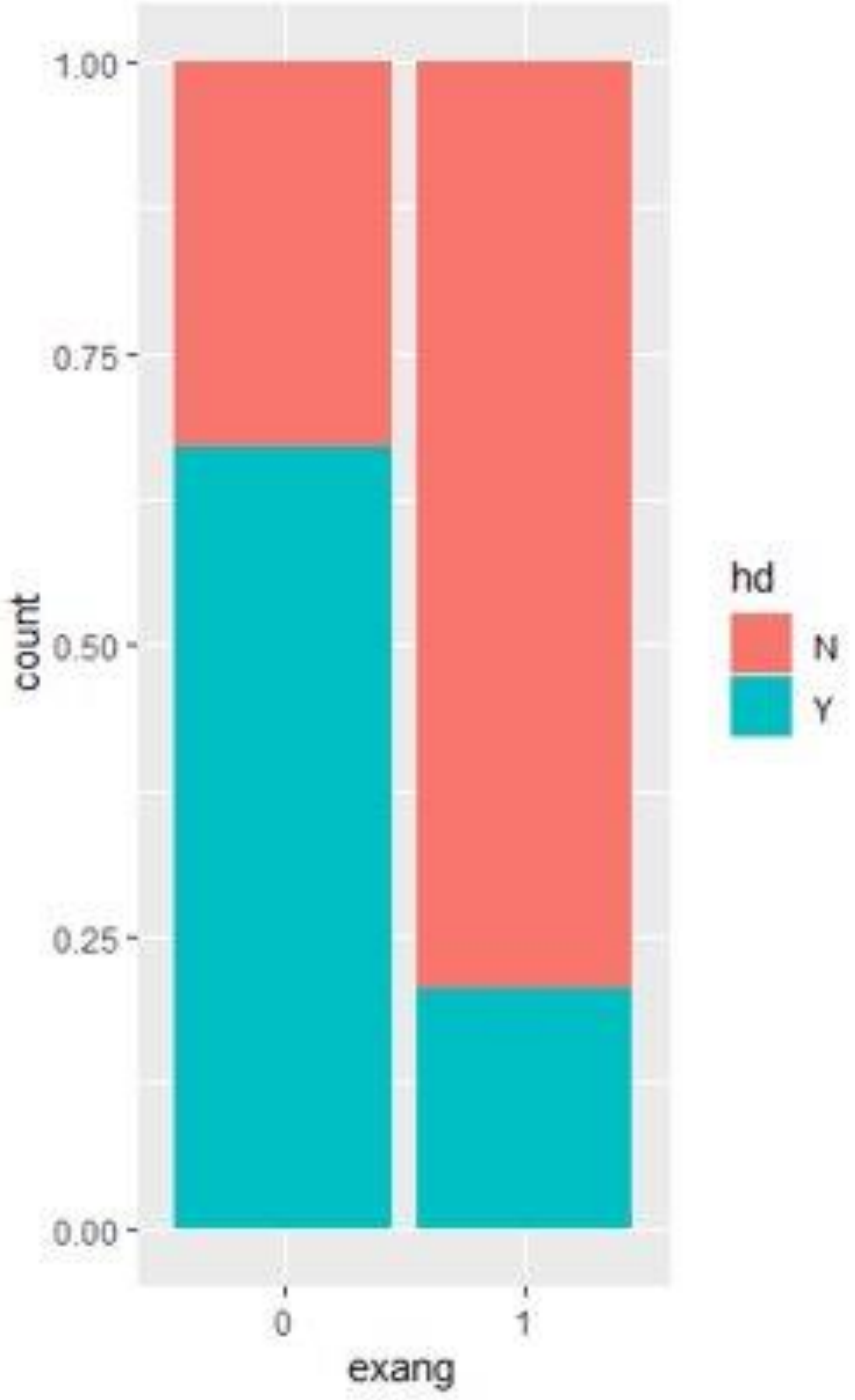


Categorical Variables

hd and exang

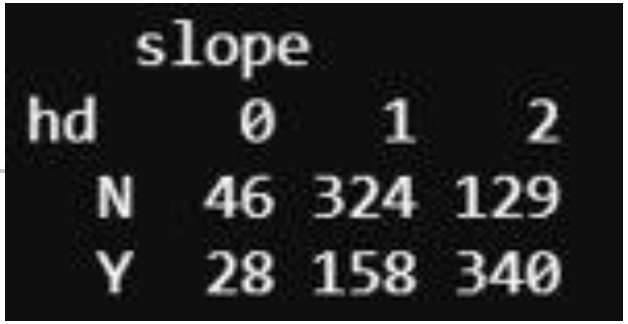


```
     exang
hd      0   1
   N 225 274
   Y 455  71
```

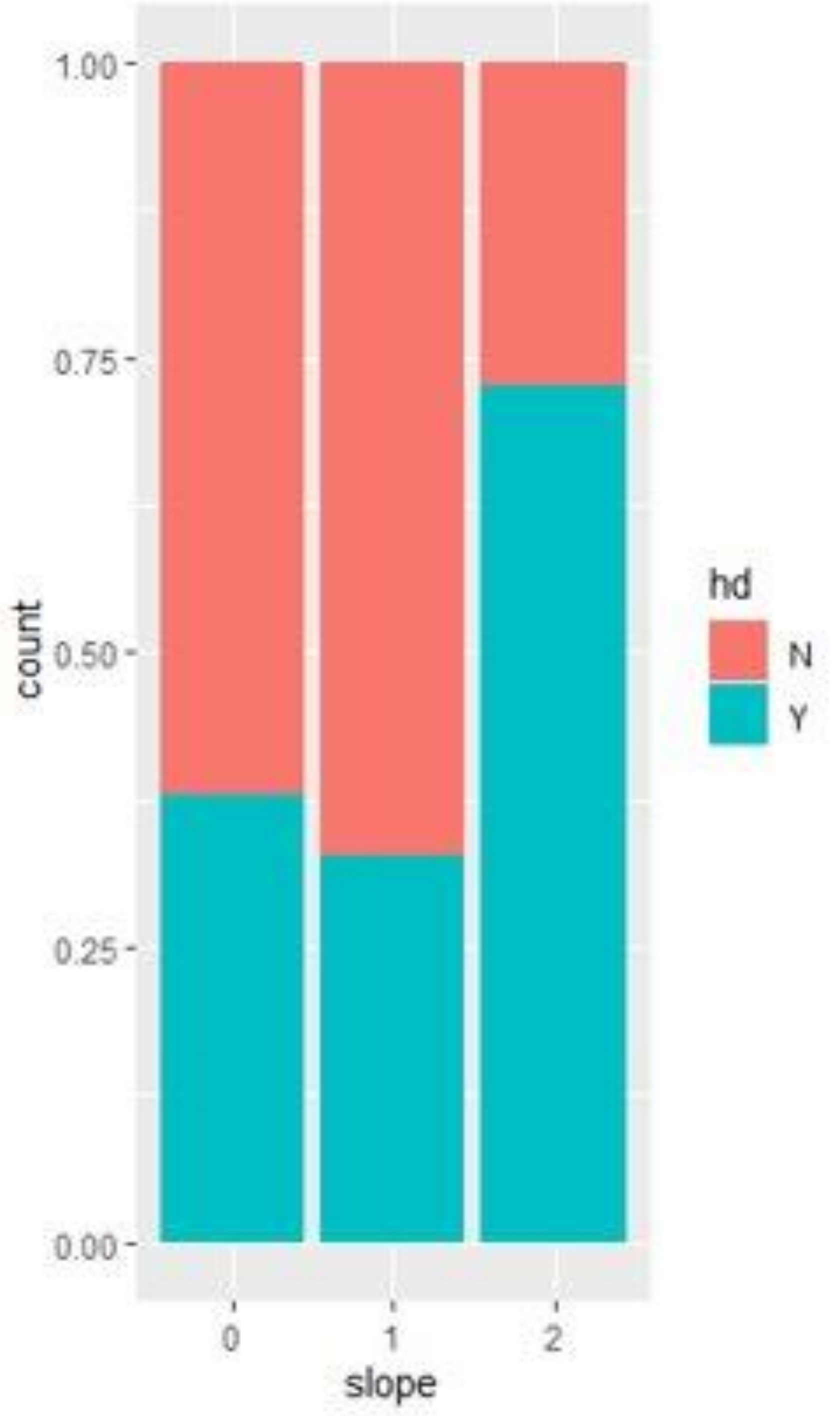


Categorical Variables

hd and slope

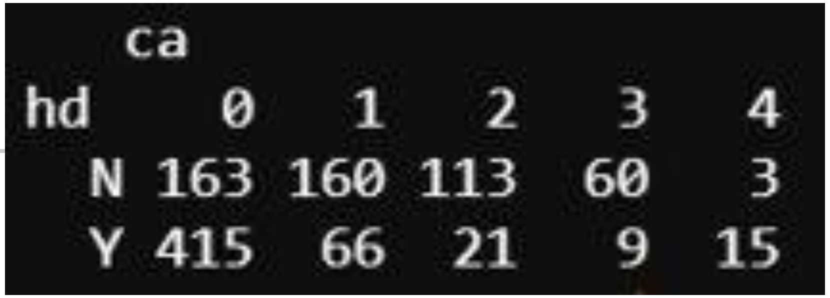


```
    slope
hd    0   1   2
  N  46 324 129
  Y  28 158 340
```

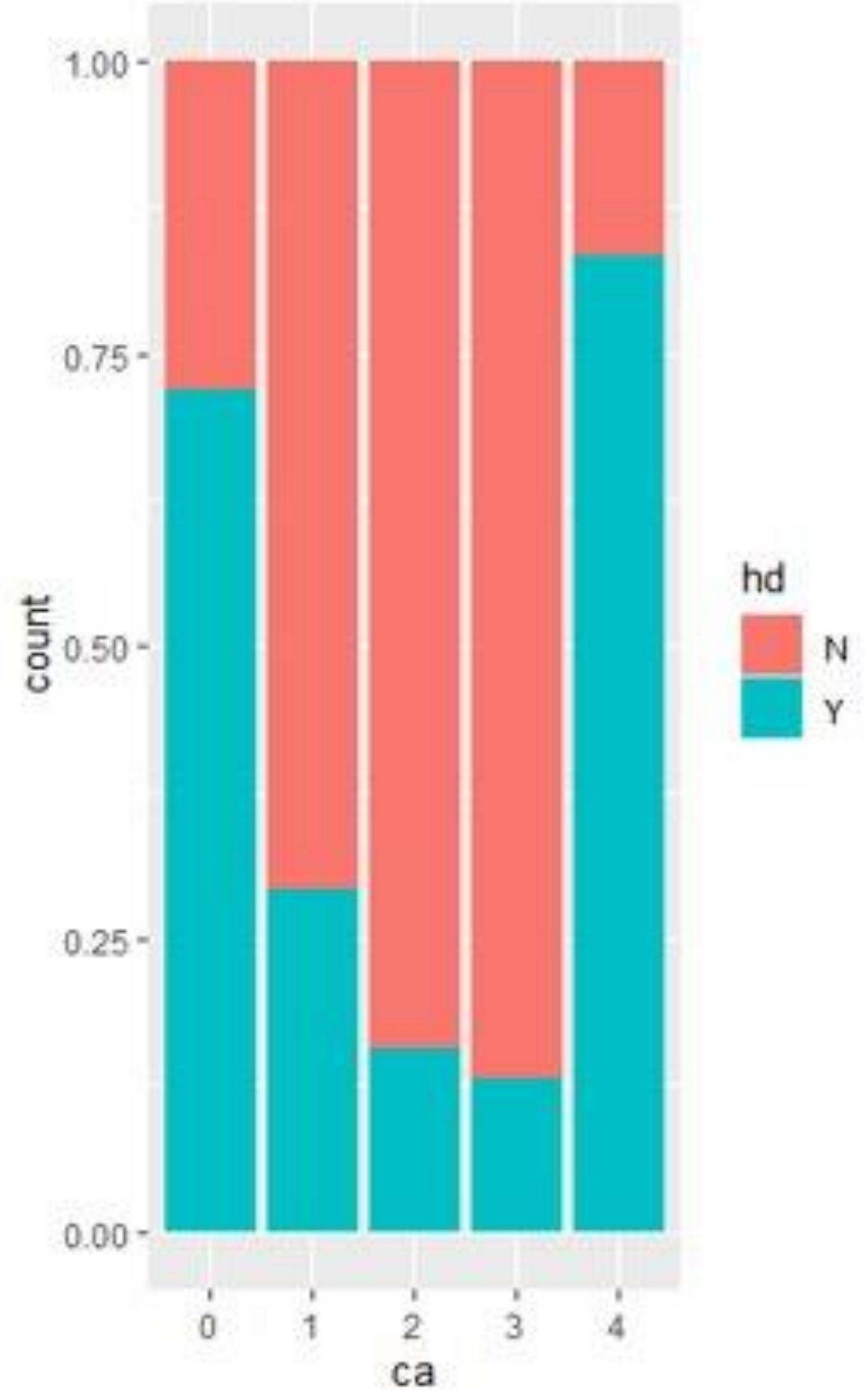


Categorical Variables

hd and ca

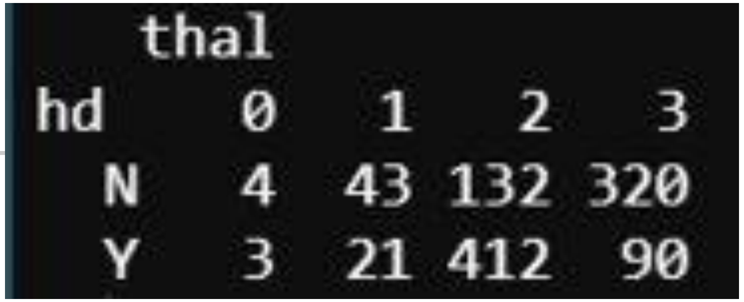


```
    ca
hd     0   1   2   3   4
  N 163 160 113  60   3
  Y 415  66  21   9  15
```

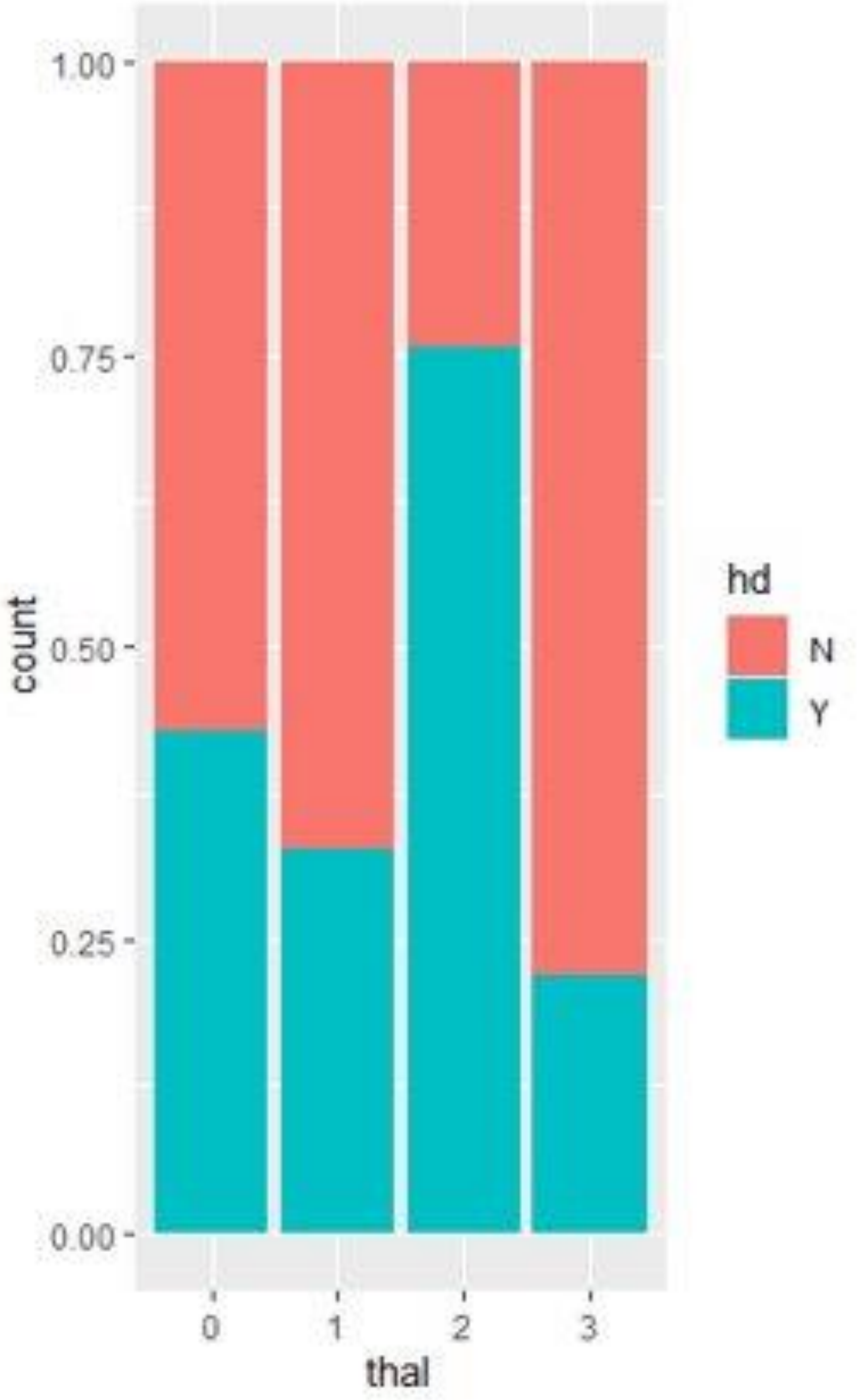


Categorical Variables

hd and thal

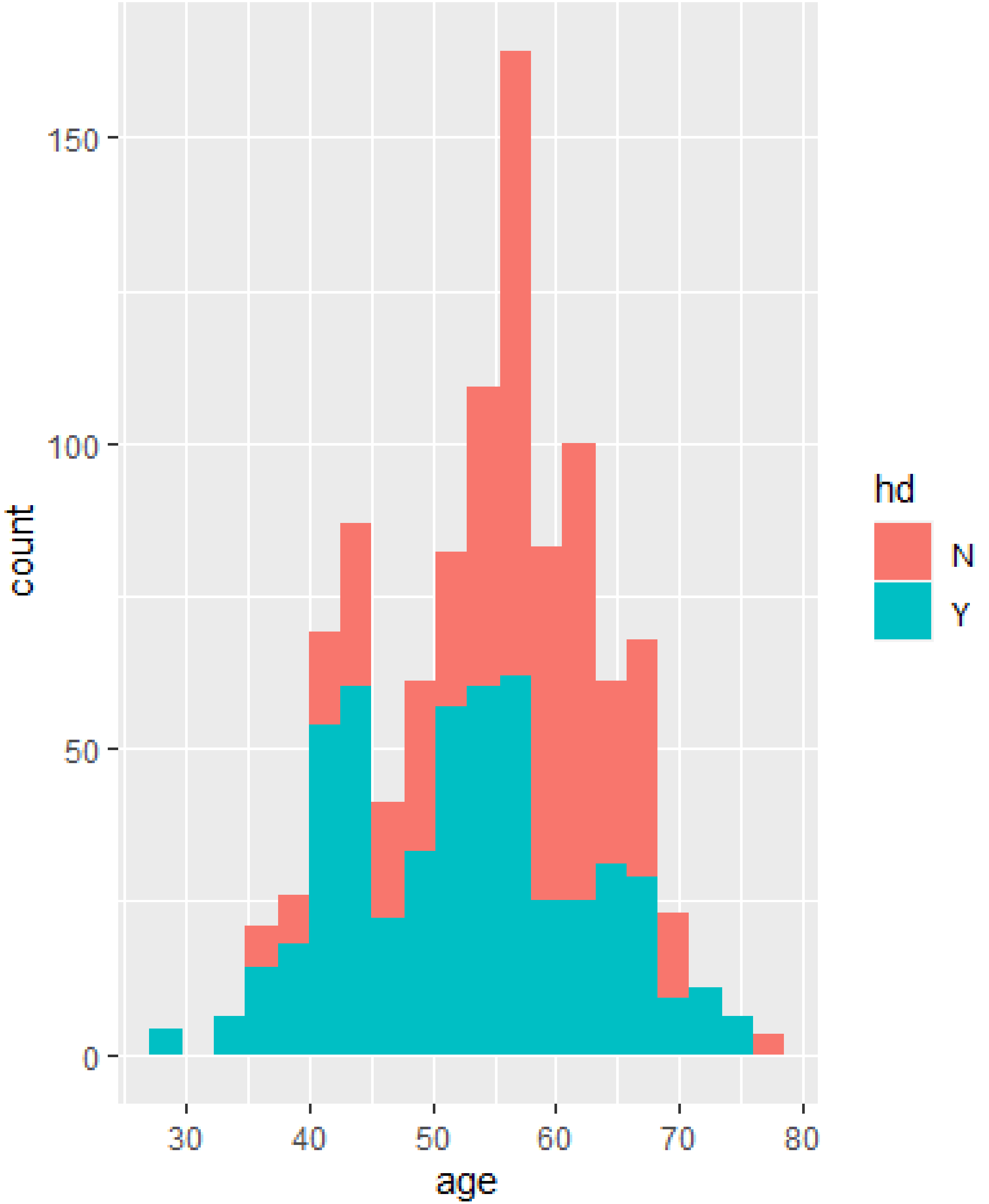


```
       thal
hd        0    1    2    3
   N      4   43  132  320
   Y      3   21  412   90
```



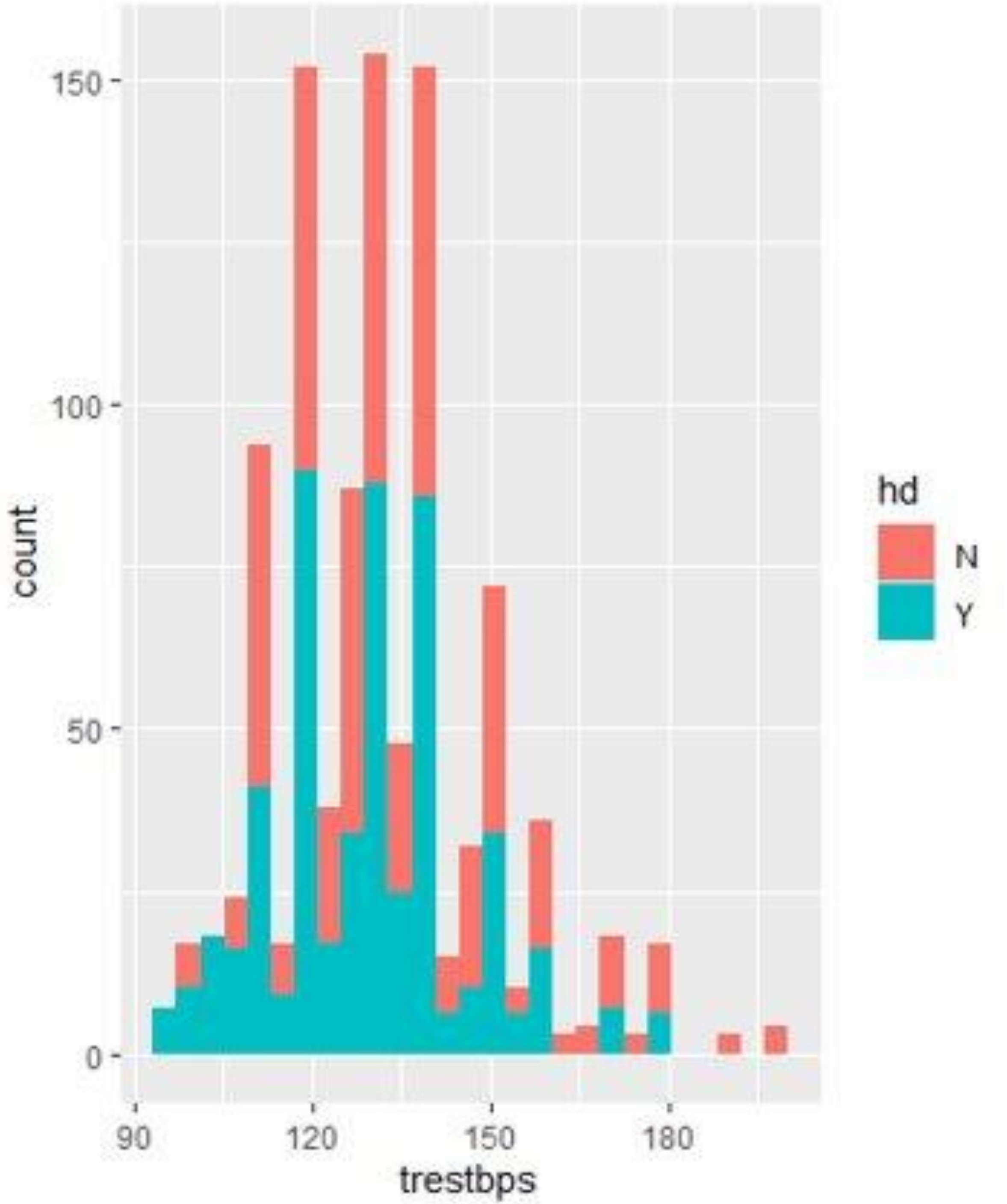
Numerical Variables

hd and age



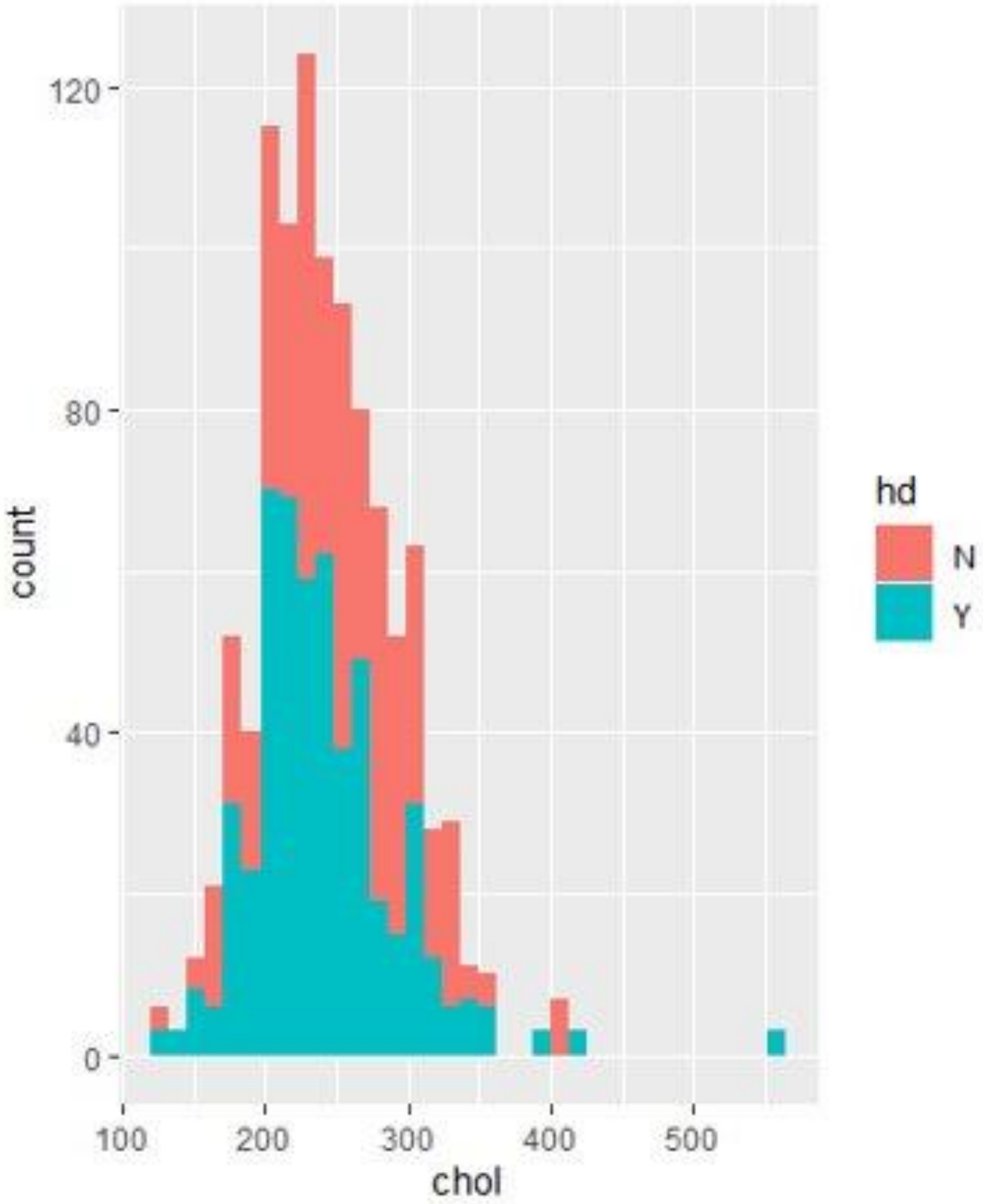
Numerical Variables

hd and trestbps



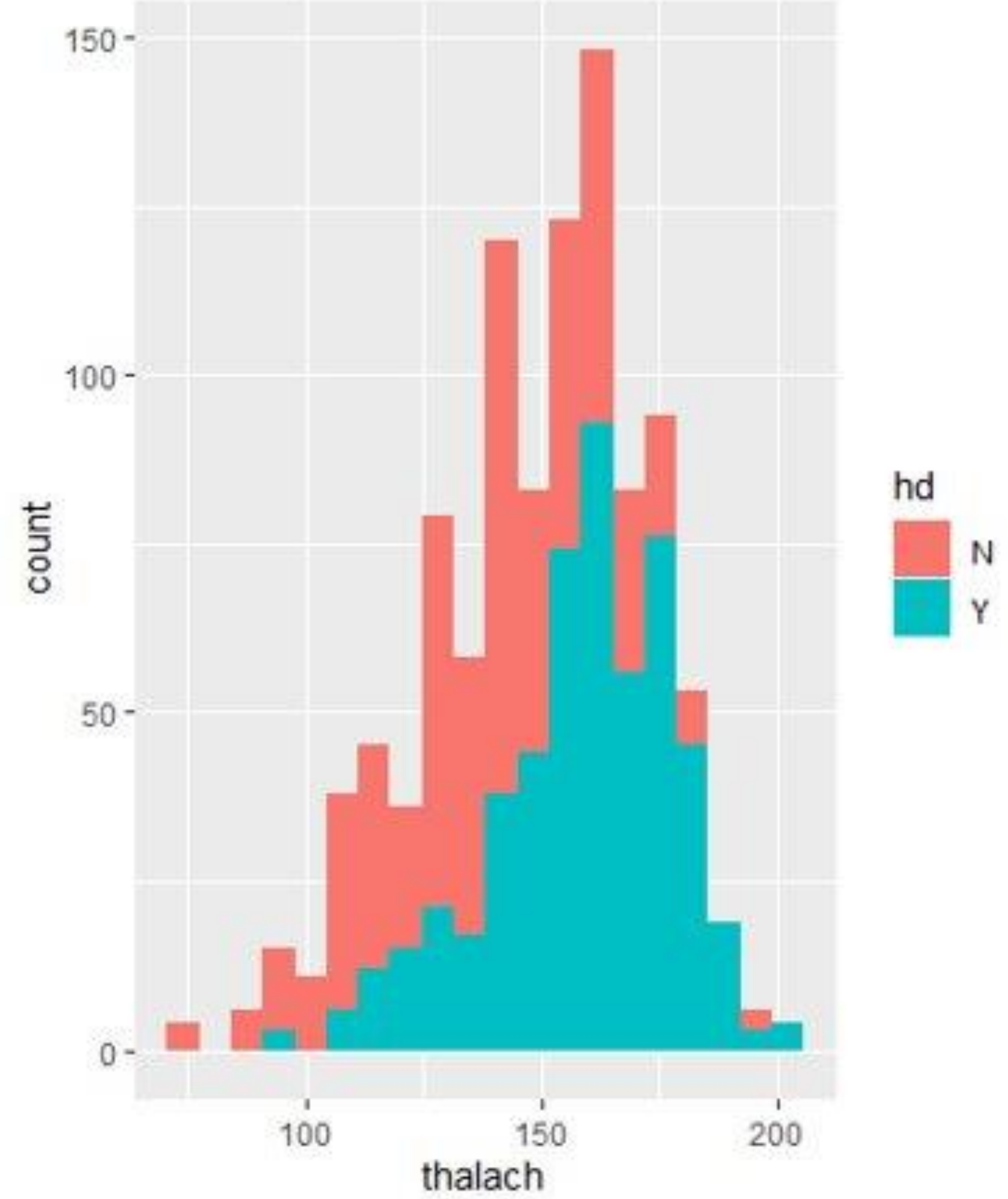
Numerical Variables

hd and chol



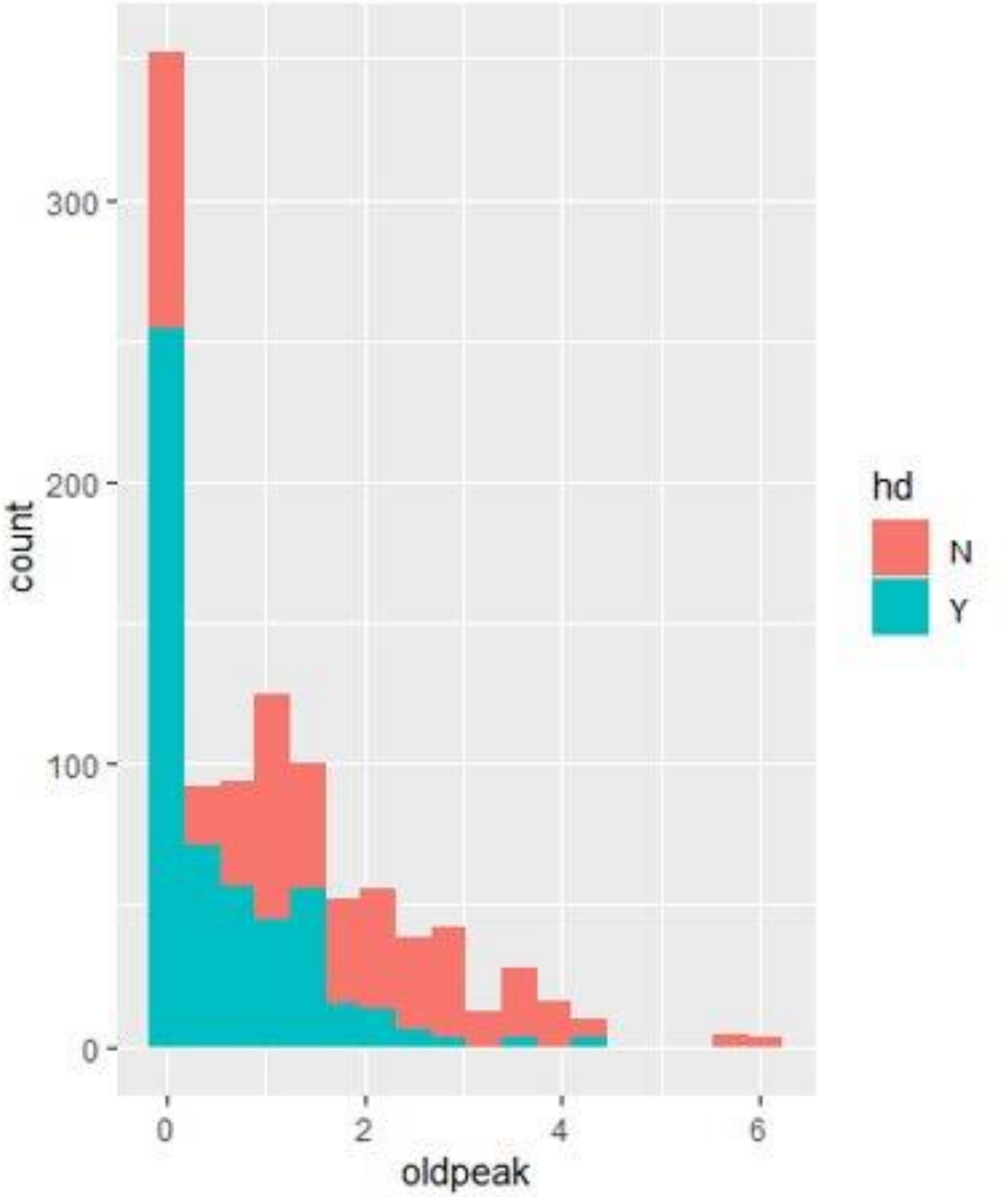
Numerical Variables

hd and thalach



Numerical Variables

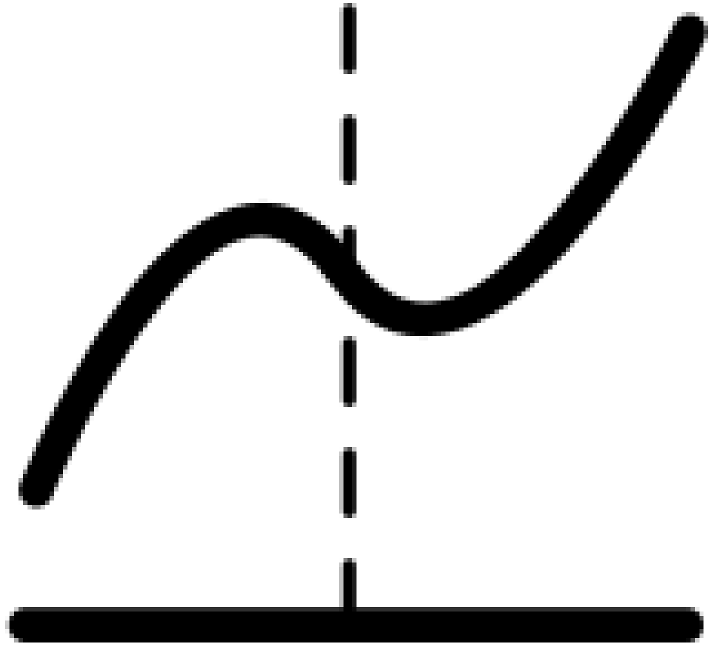
hd and oldpeak



Model explanation

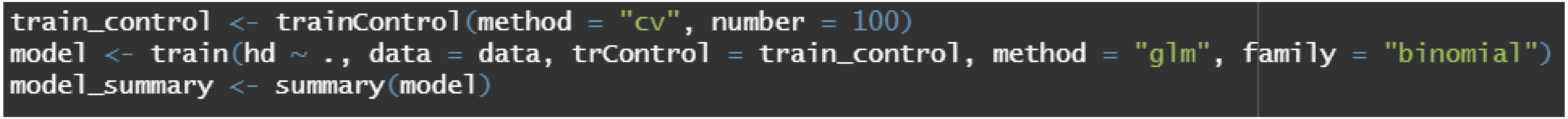
Why logistic regression?

- Because it's a process of modeling the probability of a discrete outcome given an input variable.
- And our output from data is a categorical which is either having disease or not.



Modeling implementation

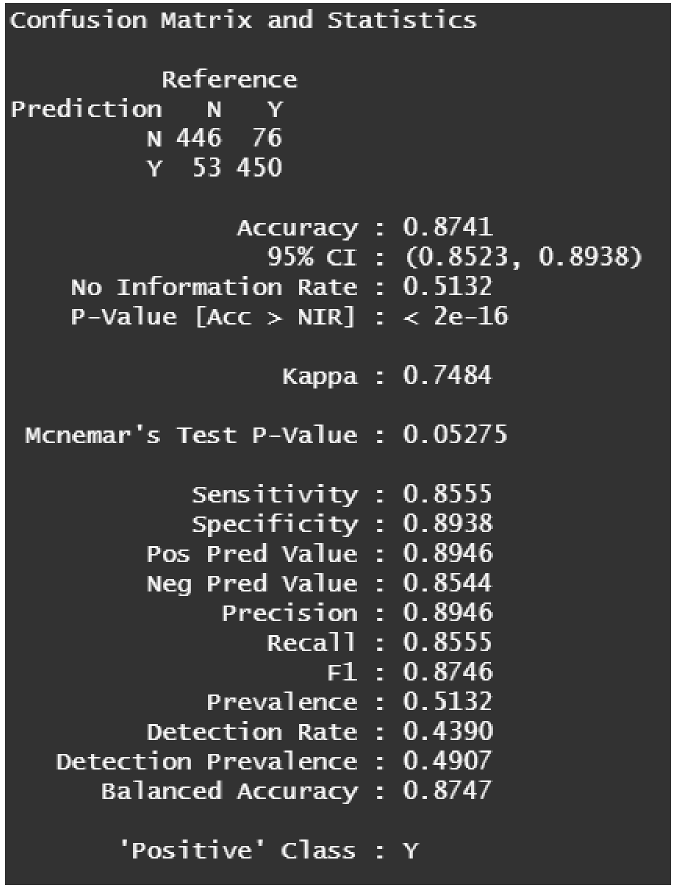
R Code



```
train_control <- trainControl(method = "cv", number = 100)
model <- train(hd ~ ., data = data, trControl = train_control, method = "glm", family = "binomial")
model_summary <- summary(model)
```

Evaluation

Confusion Matrix



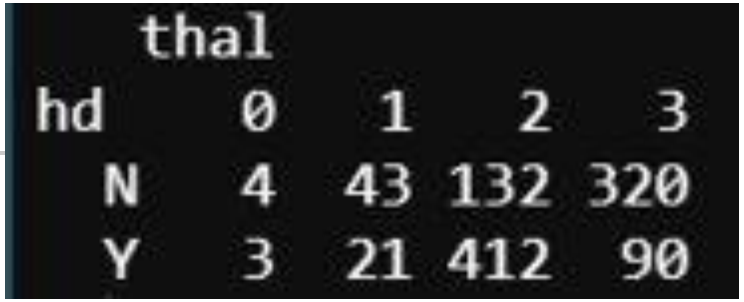
```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 446  76
         Y  53 450

               Accuracy : 0.8741
                 95% CI : (0.8523, 0.8938)
    No Information Rate : 0.5132
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7484

 Mcnemar's Test P-Value : 0.05275

            Sensitivity : 0.8555
            Specificity : 0.8938
         Pos Pred Value : 0.8946
         Neg Pred Value : 0.8544
              Precision : 0.8946
                 Recall : 0.8555
                     F1 : 0.8746
             Prevalence : 0.5132
         Detection Rate : 0.4390
   Detection Prevalence : 0.4907
      Balanced Accuracy : 0.8747

       'Positive' Class : Y
```

		Actual Class	
		Positive	**Negative**
Prediction	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

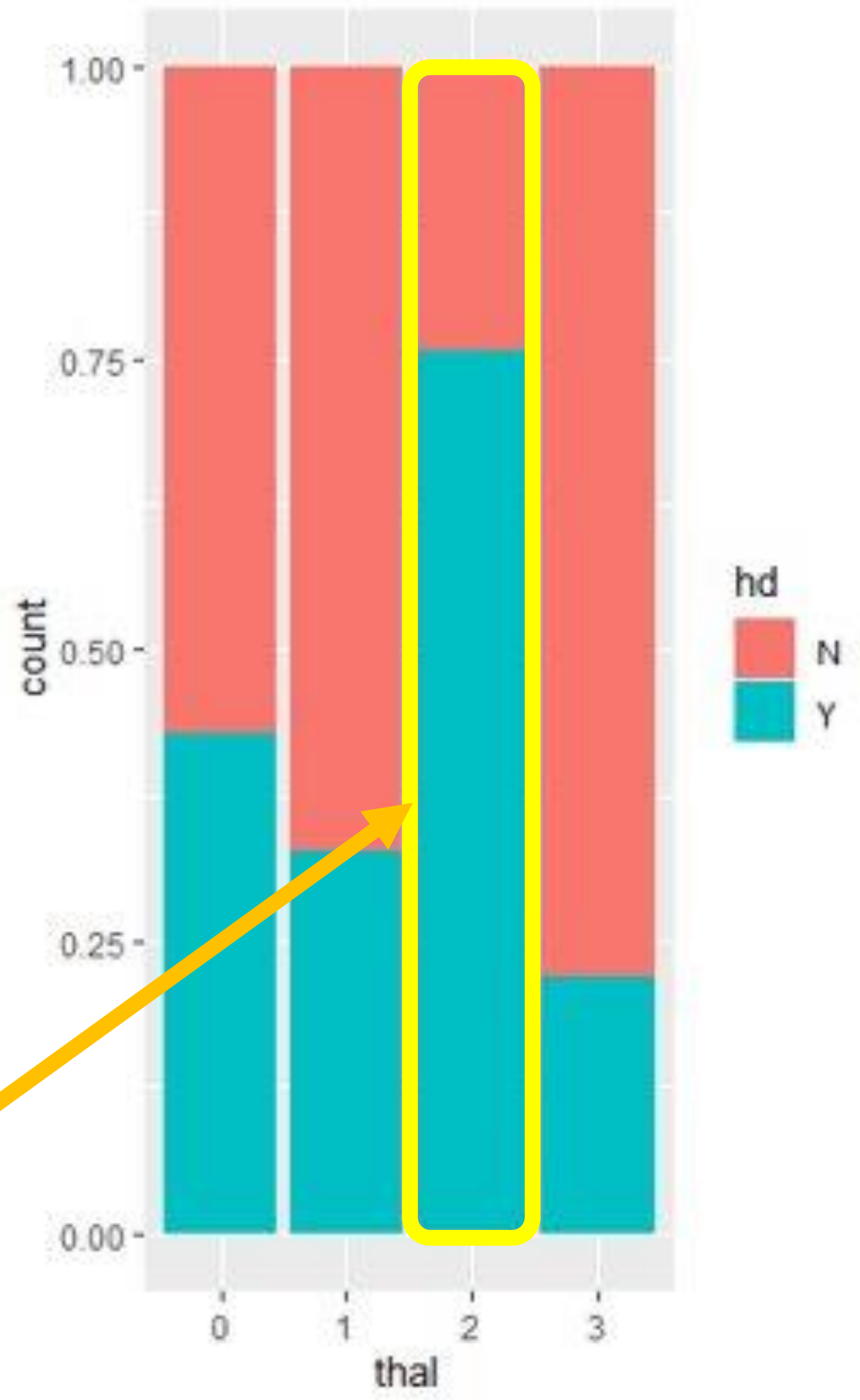
Discussion and Conclusion

Categorical Variables

hd and thal



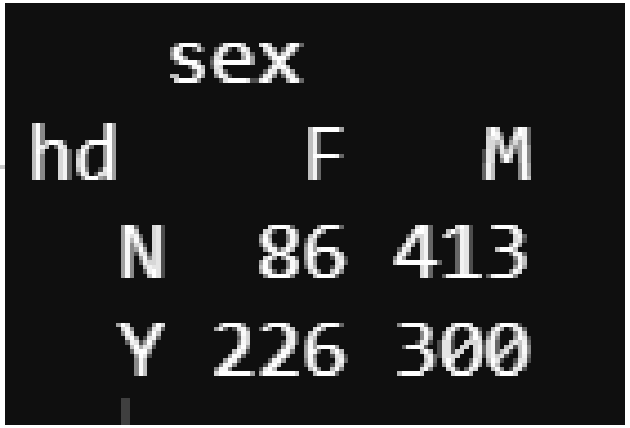
```
     thal
hd     0   1   2   3
  N    4  43 132 320
  Y    3  21 412  90
```



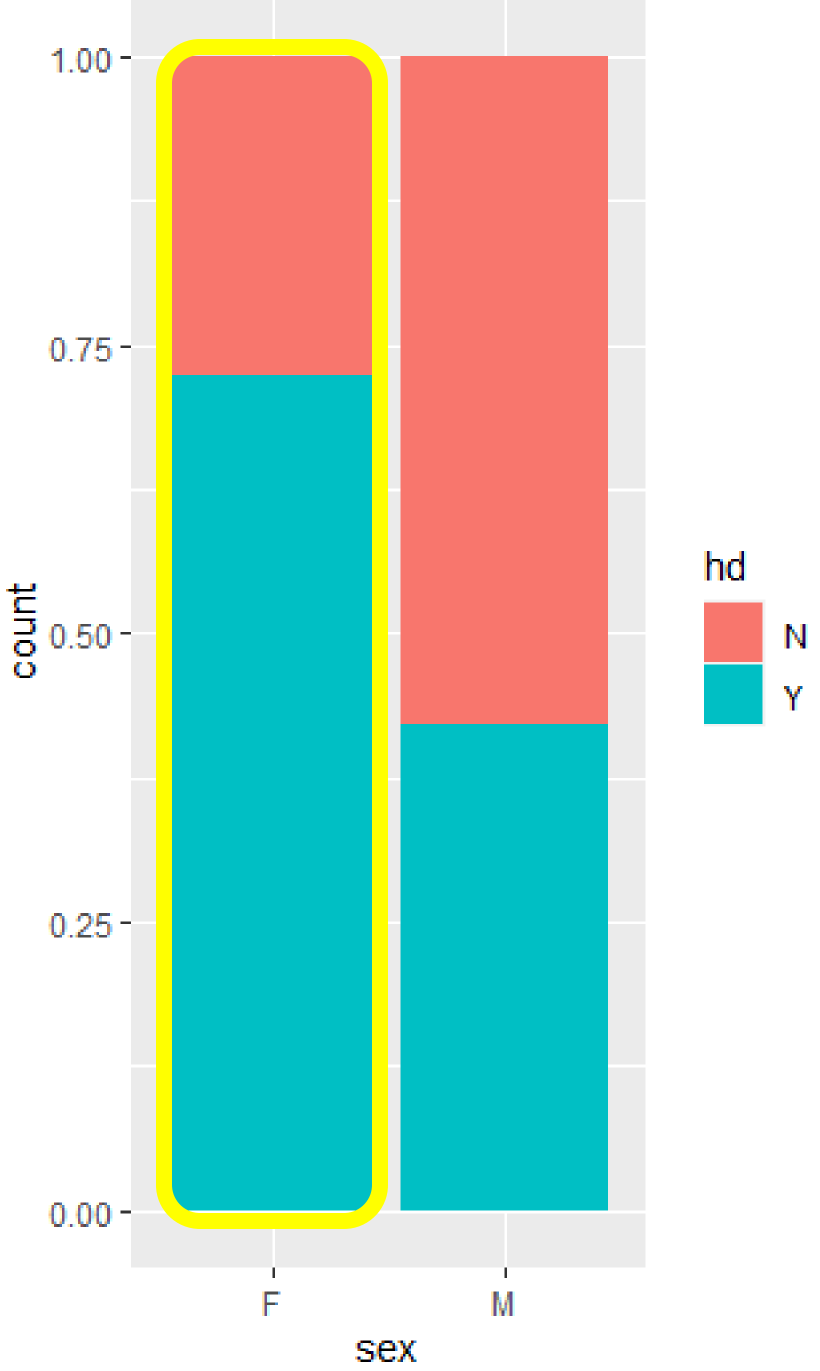
(2) Fixed defect

Categorical Variables

hd and sex

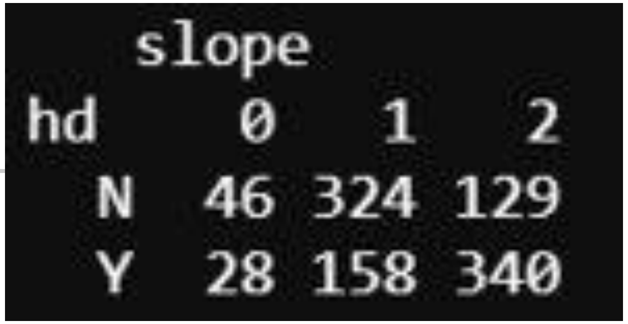


```
    sex
hd     F   M
  N   86 413
  Y  226 300
```

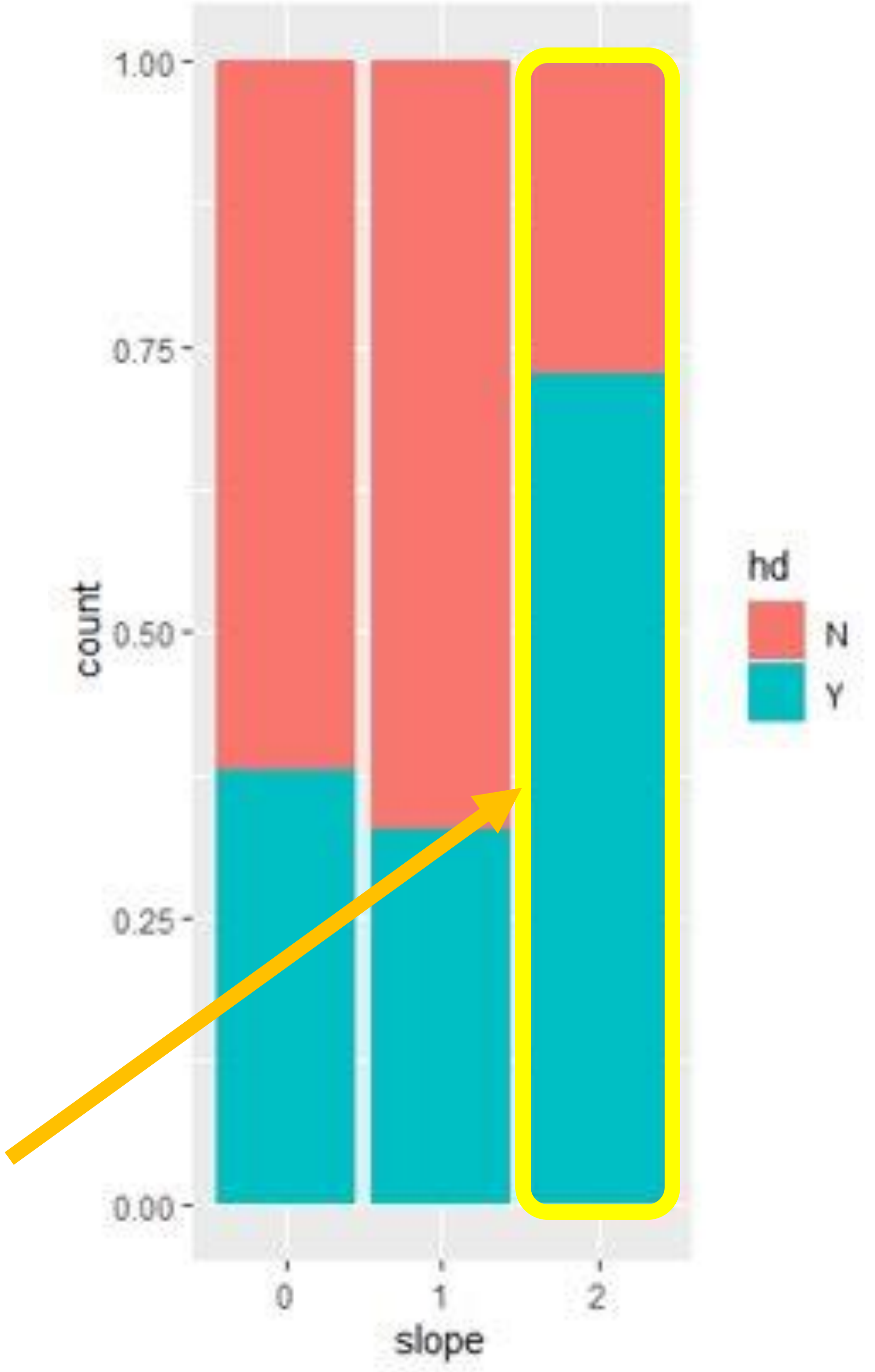


Categorical Variables

hd and slope



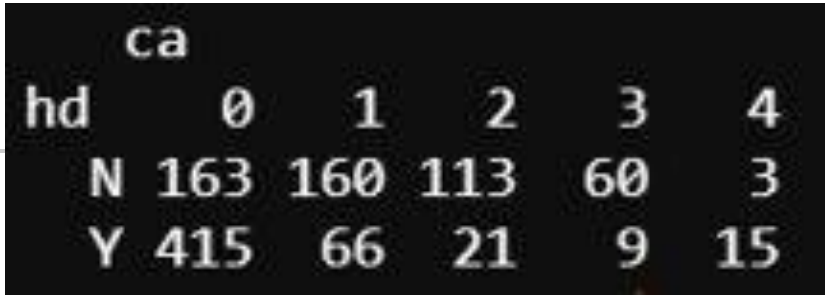
```
    slope
hd    0   1   2
  N  46 324 129
  Y  28 158 340
```



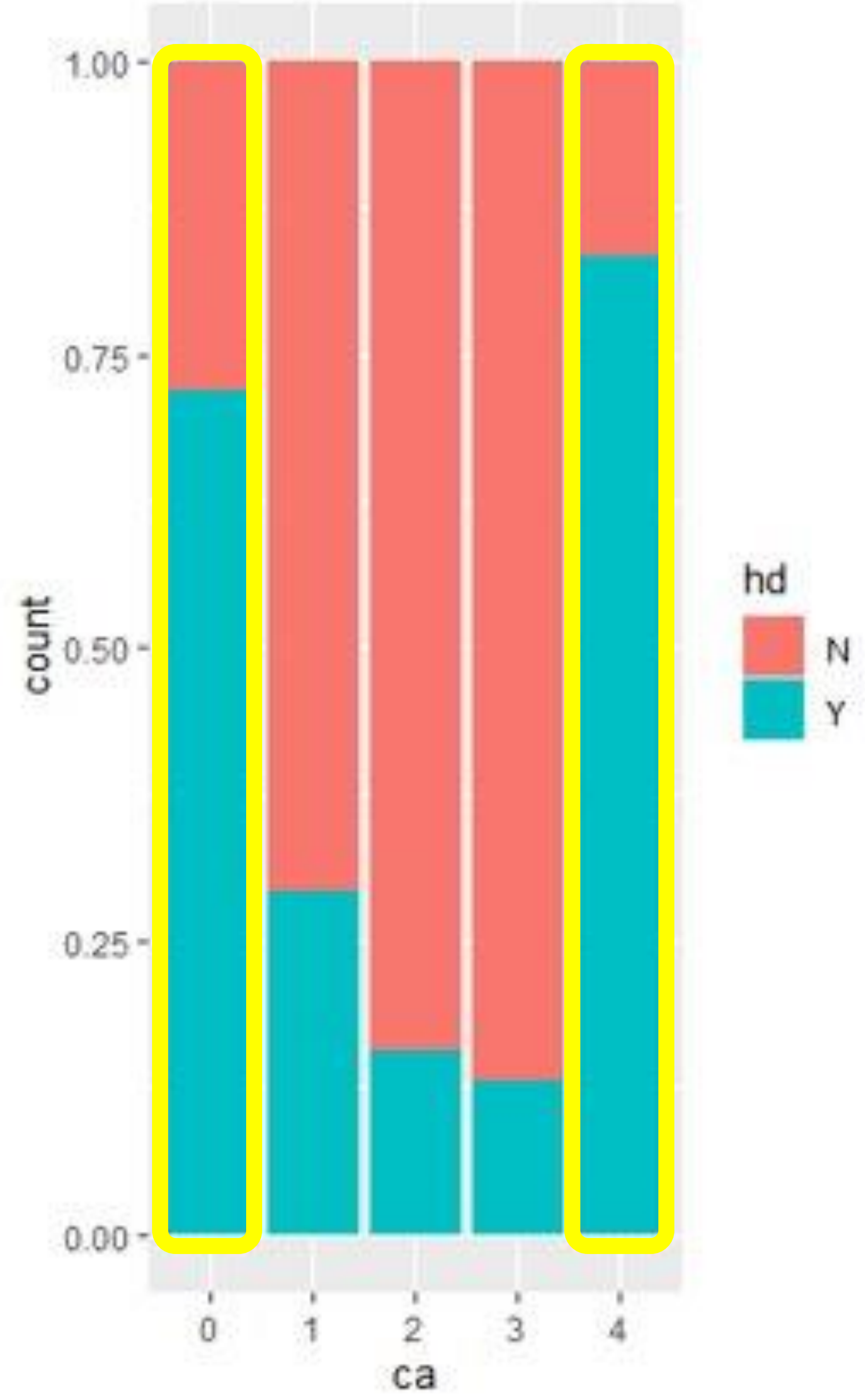
(2) Upsloping

Categorical Variables

hd and ca



```
    ca
hd     0   1   2   3   4
  N 163 160 113  60   3
  Y 415  66  21   9  15
```



count
ca
hd
N
Y