DataCleaningHeartAttack

Maria Blanco González-Mohíno y Eloisa Baquero Candau

2024-01-06

Contents

1 Descripción del dataset

Fuente de este dataset: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion/248773 Fuente de Kaggle: https://archive.ics.uci.edu/dataset/45/heart+disease

1.1 Variables del dataset

- **Age**: Edad del paciente.
- **Sex**: Sexo del paciente.
- **cp**: Dolor torácico tipo (1 = angina típica; 2 = angina atípica; 3 = dolor no anginoso; 4 = asintomático).
- **ca**: Número de grandes vasos sanguíneos (0-3).
- **trtbps**: Presión arterial en reposo (en mm Hg).
- **chol**: colesterol en mg/dl obtenido a través del sensor de IMC.
- **fbs**: (glucemia en ayunas > 120 mg/dl) (1 = verdadero; 0 = falso).
- **rest_ecg**: resultados electrocardiográficos en reposo (0 = normal; 1 = presenta anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST de > 0,05 mV); 2 = muestra hipertrofia ventricular izquierda probable o definida según los criterios de Estes).
- **thalach**: frecuencia cardiaca máxima alcanzada
- **exng**: Angina inducida por el ejercicio (1 = si; 0 = no).
- **oldpeak**: Depresión del ST inducida por el ejercicio en relación con el reposo.
- **slp**: Slope / Bajada.
- **caa**: Número de vasos mayores (0-3) coloreados por flouroscopia.
- **thall**: No hay información al respecto.
- **output**: 0 = menor probabilidad de infarto; 1 = mayor probabilidad de infarto

1.2 ¿Por qué es importante? y ¿qué pregunta responde?

2 Exploración de los datos

```
heart <- read_csv("../data/heart.csv", show_col_types = FALSE)
head(heart, 5)
```

```
## # A tibble: 5 x 14
##     age   sex    cp trtbps  chol   fbs restecg thalachh  exng oldpeak   slp
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>   <dbl> <dbl>
## 1    63     1     3    145   233     1       0      150     0     2.3     0
## 2    37     1     2    130   250     0       1      187     0     3.5     0
## 3    41     0     1    130   204     0       0      172     0     1.4     2
## 4    56     1     1    120   236     0       1      178     0     0.8     2
## 5    57     0     0    120   354     0       1      163     1     0.6     2
## # i 3 more variables: caa <dbl>, thall <dbl>, output <dbl>
```

```
str(heart)
```

```
## spc_tbl_ [303 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : num [1:303] 1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_double(),
##   ..   cp = col_double(),
##   ..   trtbps = col_double(),
##   ..   chol = col_double(),
##   ..   fbs = col_double(),
##   ..   restecg = col_double(),
##   ..   thalachh = col_double(),
##   ..   exng = col_double(),
##   ..   oldpeak = col_double(),
##   ..   slp = col_double(),
##   ..   caa = col_double(),
##   ..   thall = col_double(),
##   ..   output = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
summary(heart)
```

```
##       age             sex               cp            trtbps     
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs             restecg          thalachh     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##       exng           oldpeak          slp              caa        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##      thall           output     
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000  
##  Mean   :2.314   Mean   :0.5446  
##  3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.0000
```

2.1 Limpieza de los datos

Búsqueda de valores nulos por columnas.

```r
sapply(heart, function(x) sum(is.na(x)))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh 
##        0        0        0        0        0        0        0        0 
##     exng  oldpeak      slp      caa    thall   output 
##        0        0        0        0        0        0
```

Enfocandonos en la naturaleza del dataset, hay algunos casos especiales que también se consideraran valores nulos

```r
unique(heart$sex)
```

```
## [1] 1 0
```

2.2 Visualización de variables de interes

```
hist(heart$age)
```