Dirty Heart Data

Samuel Dummer

9/7/2021

Abstract

This document summarizes a dataset about the heart data in different patients. The main focus of this dataset is *angina pectoris*. This is a specific symptom of many heart diseases that causes severe pain in the chest which can also spread to other parts of the body. This pain is caused by the heart not having enough blood. This data includes age, sex, chest pain type, blood pressure, cholesterol level, electrocardiographic results, maximum heart rate, whether or not it is exercise induced angina, number of major vessels, ST depression induced by exercise relating to the rest of the patient, the slope of the peak exercise ST, and whether or not the patient has had a heart attack.

Model Setup

In this part we set the directory, imported the necessary libraries, and loaded/read in the data frame.

```
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
library(tidyverse)
dheart <- read.csv("dirtyheart.csv", header = T)
```

Checking to See How Clean the Data Is

Here we ran a few lines of code to help us observe if there was any missing data and, if so, how many lines of missing data there were.

```
clean <- ifelse(complete.cases(dheart)==TRUE,1,0)
table(clean)
```

```
paste("There are ",dim(dheart)[1]-sum(clean), " rows with missing data.")
```

```
## [1] "There are  69  rows with missing data."
```

Characterzation of Data

This part of the script includes the overall characterization of the dataset. This includes names of the columns, dimensions of the data frame, structure, head (a.k.a. first 6 rows of data), summary of the data.

```
names(dheart)
```

```
##  [1] "age"      "sex"      "cp"       "trestbps" "chol"     "fbs"     
##  [7] "restecg"  "thalach"  "exang"    "oldpeak"  "slope"    "ca"      
## [13] "thal"     "target"
```

```
dim(dheart)
```

```
## [1] 303  14
```

```
str(dheart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 NA 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: int  145 130 130 120 120 140 NA 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thal    : int  6 3 7 3 3 3 3 3 7 7 ...
##  $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(dheart)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63   1  1      145  233   1       0     150     0     2.3     0  0    6
## 2  37   1  4      130  250   0       1     187     0     3.5     0  0    3
## 3  41   0  4      130  204   0       0     172     0     1.4     2  0    7
## 4  NA   1  3      120  236   0       1     178     0     0.8     2  0    3
## 5  57   0  2      120  354   0       1     163     1     0.6     2  0    3
## 6  57   1  2      140  192   0       1     148     0     0.4     1  0    3
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

```
summary(dheart)
```

```
##       age             sex               cp           trestbps    
##  Min.   : 0.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0  
##  1st Qu.:46.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :3.000   Median :130.0  
##  Mean   :53.08   Mean   :0.6768   Mean   :3.158   Mean   :131.7  
##  3rd Qu.:60.50   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##  NA's   :12      NA's   :6                        NA's   :13
##      chol          fbs            restecg          thalach
##  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.:132.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :245.2   Mean   :0.1515   Mean   :0.5217   Mean   :149.4
##  3rd Qu.:274.0   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##  NA's   :13      NA's   :6        NA's   :4        NA's   :8
##     exang           oldpeak         slope             ca
##  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   :1.05   Mean   :1.403   Mean   :0.7322
##  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
##  NA's   :3        NA's   :5      NA's   :5       NA's   :8
##      thal           target
##  Min.   :3.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:0.000
##  Median :3.000   Median :1.000
##  Mean   :4.734   Mean   :0.539
##  3rd Qu.:7.000   3rd Qu.:1.000
##  Max.   :7.000   Max.   :1.000
##  NA's   :2       NA's   :8
```

Removing/Replacing rows with missing values

In this part of the script, we ran some code from the "tidyverse" library to help remove data with missing values. There were two ways this was done. Either the whole row was removed or the data was replaced with the average. We also removed the data with age "0" since this age was such an outlier in comparison to the rest of the data that the data must have been input incorrectly.

```r
dheart <- filter(dheart, sex == 0 | sex == 1)
dheart <- filter(dheart, fbs == 0 | fbs == 1)
dheart <- filter(dheart, exang == 0 | exang == 1)
dheart <- filter(dheart, restecg == 0 | restecg == 1 | restecg == 2)
dheart <- filter(dheart, slope == 0 | slope == 1 | slope == 2)
dheart <- filter(dheart, ca == 0 | ca == 1 | ca == 2 | ca == 3)
dheart <- filter(dheart, thal == 3 | thal == 6 | thal == 7)
dheart <- filter(dheart, target == 0 | target == 1)
dheart$age <- ifelse(is.na(dheart$age), mean(dheart$age, na.rm=TRUE), dheart$age)
dheart$age <- ifelse(dheart$age == 0, mean(dheart$age, na.rm=TRUE), dheart$age)
dheart$trestbps <- ifelse(is.na(dheart$trestbps), mean(dheart$trestbps, na.rm=TRUE), dheart$trestbps)
dheart$chol <- ifelse(is.na(dheart$chol), mean(dheart$chol, na.rm=TRUE), dheart$chol)
dheart$thalach <- ifelse(is.na(dheart$thalach), mean(dheart$thalach, na.rm=TRUE), dheart$thalach)
dheart$oldpeak <- ifelse(is.na(dheart$oldpeak), mean(dheart$oldpeak, na.rm=TRUE), dheart$oldpeak)
clean <- ifelse(complete.cases(dheart)==TRUE,1,0)
table(clean)
```

```r
paste("There are ",dim(dheart)[1]-sum(clean), " rows with missing data.")
```

```
## [1] "There are  0  rows with missing data."
```

Changing Names of Labels’

In this section, we changed the labels of all the columns that used numbers as labels. For example, 0 and 1 in the sex column was changed to “male” and “female.” This helps a lot more with comprehension of the dataset.

```
dheart$sex  <- factor(dheart$sex, levels=c(0,1), labels = c("male", "female"))
dheart$cp  <- factor(dheart$cp, levels=c(1,2,3,4), labels = c("typical angina", "atypical angina", "non
dheart$fbs  <- factor(dheart$fbs, levels=c(0,1), labels = c("false", "true"))
dheart$restecg  <- factor(dheart$restecg, levels=c(0,1,2), labels = c("normal", "wave abnomality", "lef
dheart$exang  <- factor(dheart$exang, levels=c(0,1), labels = c("no", "yes"))
dheart$slope  <- factor(dheart$slope, levels=c(0,1,2), labels = c("upsloping", "flat", " downsloping"))
dheart$thal  <- factor(dheart$thal, levels=c(3,6,7), labels = c("normal", "fixed defect", "reversible d
dheart$target  <- factor(dheart$target, levels=c(0,1), labels = c("no", "yes"))
head(dheart)
```

```
##        age    sex               cp trestbps chol   fbs         restecg thalach
## 1 63.00000 female   typical angina      145  233  true          normal     150
## 2 37.00000 female     asymptomatic      130  250 false wave abnomality     187
## 3 41.00000   male     asymptomatic      130  204 false          normal     172
## 4 53.35857 female non-anginal pain      120  236 false wave abnomality     178
## 5 57.00000   male  atypical angina      120  354 false wave abnomality     163
## 6 57.00000 female  atypical angina      140  192 false wave abnomality     148
##   exang oldpeak        slope ca              thal target
## 1    no     2.3    upsloping  0      fixed defect    yes
## 2    no     3.5    upsloping  0            normal    yes
## 3    no     1.4  downsloping  0 reversible defect    yes
## 4    no     0.8  downsloping  0            normal    yes
## 5   yes     0.6  downsloping  0            normal    yes
## 6    no     0.4         flat  0            normal    yes
```

Arranging the Dataset

This part includes some code that rearranges the data to be sorted from youngest to oldest.

```
dheart <- arrange(dheart, age, sex)
head(dheart)
```

```
##   age    sex               cp trestbps chol   fbs         restecg thalach exang
## 1  29 female     asymptomatic  130.000  204 false          normal     202    no
## 2  34   male  atypical angina  118.000  210 false wave abnomality     192    no
## 3  34 female non-anginal pain  118.000  182 false          normal     174    no
## 4  35   male     asymptomatic  138.000  183 false wave abnomality     182    no
## 5  35 female     asymptomatic  130.869  192 false wave abnomality     174    no
## 6  35 female non-anginal pain  120.000  198 false wave abnomality     130   yes
##   oldpeak        slope ca              thal target
## 1     0.0  downsloping  0 reversible defect    yes
## 2     0.7  downsloping  0            normal    yes
## 3     0.0  downsloping  0            normal    yes
## 4     1.4  downsloping  0 reversible defect    yes
## 5     0.0  downsloping  0 reversible defect    yes
## 6     1.6         flat  0            normal     no
```

Summarizing Heart rate, Cholesterol, and Blood Pressure

Here, we used the summarize command from the tidyverse package to quickly give us the mean of the Cholesterol, Maximum Heart rate, and Resting Blood Pressure.

```
numsum <- summarize(dheart, meanchol = mean(chol), meanbp = mean(trestbps), meanhr = mean(thalach))
numsum
```

```
##   meanchol  meanbp   meanhr
## 1 246.1216 130.869 148.5769
```

Creating a Table that Compares Cholesterol and Whether or not the Patient had a Heart Attack

In this section of the script, we quickly created a new table that could compare the cholesterol levels to the occurrence of a heart attack. This was to see whether or not there was a correlation between the two. We used the select function from tidyverse to get the columns then arranged it from the largest to smallest cholesterol levels. In the end, there didn't seem to be a correlation.

```
cholcomp <- dheart %>% select("chol", "target") %>% arrange(-chol)
head(cholcomp)
```

```
##   chol target
## 1  564    yes
## 2  417    yes
## 3  407     no
## 4  360    yes
## 5  354    yes
## 6  353     no
```

Creating a Summary of the Heart rate, Cholesterol, and Blood Pressure Levels Grouped by Age

Here, we specifically showed the mean of the Heart rate, Cholesterol, and Blood Pressure levels of each age specifically. This was done using the group_by and summarize functions in the tidyverse library.

```
agesum <- dheart %>% group_by(age) %>% summarize(meanchol = mean(chol), meanbp = mean(trestbps), meanhr
head(agesum)
```

```
## # A tibble: 6 x 4
##     age meanchol meanbp meanhr
##   <dbl>    <dbl>  <dbl>  <dbl>
## 1    29      204    130    202
## 2    34      196    118    183
## 3    35      214.   129.   160.
## 4    37      232.   125    178.
## 5    38      231    120    182
## 6    39      246.   117.   167
```

Normalizing Data

In this part, we normalized the cholesterol, heart rate, and blood sugar levels. This is done by dividing each column by the largest value. We used the mutate function from the tidyverse library to help us succeed in doing this.

```
dheart <- mutate(dheart, chol = chol/max(chol))
dheart <- mutate(dheart, trestbps = trestbps/max(trestbps))
dheart <- mutate(dheart, thalach = thalach/max(thalach))
head(dheart[c(4, 5, 8)])
```

```
##    trestbps      chol   thalach
## 1 0.6500000 0.3617021 1.0000000
## 2 0.5900000 0.3723404 0.9504950
## 3 0.5900000 0.3226950 0.8613861
## 4 0.6900000 0.3244681 0.9009901
## 5 0.6543452 0.3404255 0.8613861
## 6 0.6000000 0.3510638 0.6435644
```

Renaming Columns

Renaming the columns was done to help make the dataset more legible to other readers. This renaming was done using the rename function in tidyverse.

```
dheart <- rename(dheart, "chest pain type" = "cp", "resting blood pressure" = "trestbps", "cholesterol"
names(dheart)
```

```
##  [1] "age"                           "sex"
##  [3] "chest pain type"               "resting blood pressure"
##  [5] "cholesterol"                   "fasting blood sugar"
##  [7] "resting electrocarfiographic" "max heart rate"
##  [9] "excercise induced angina"     "st depression"
## [11] "slope"                         "number of vessels"
## [13] "thalassemia"                   "heart attack"
```