Data Manipulation of Heart Disease Data

Advaith Cheruvu

9/8/2021

Importance of Study

Heart disease has become the leading cause of death in the United States, causing approximately 1 in 4 deaths. The deadliest form of heart disease is coronary artery disease (CAD), responsible for approximately 16% of the world’s deaths. By analyzing symptoms of CAD and other heart diseases, doctors can make better decisions about preventative measures against heart disease and potentially save millions of lives. Thus, the data preparation done in this document is extremely important to the combating heart disease.

Angina Pectoris and Heart Disease

BioMedEd, Inc. has provided data from a study done of 304 patients that exhibited angina pectoris. Before cleaning the data, it is important to know the background behind the data. Angina pectoris is often described by patients as a squeezing or tightness in their chest. The problem for doctors is that angina is relatively common and is hard to distinguish between other types of chest pain, unrelated to heart disease. There are different types of anginas, including stable, unstable, Prinzmetal’s angina, and microvascular angina.

Stable Anginas

Stable angina develops when your heart has to work harder, usually during exercise. This type of pain can be predicted and feels similar to previous chest pains. Though the severity and duration can vary, stable anginas are usually short and dissipates quickly when resting. Stable anginas usually don’t require a medical emergency. If new symptoms occur severity/duration increases, it may signal an unstable angina or heart attack.

Unstable Anginas

Unstable anginas require a medical emergency since it may signal a heart attack. Unstable anginas are unexpected and are usually more severe than stable anginas. Occurring even at rest, unstable anginas can strike at any time and getting appropriate treatement immediately is very important.

Prinzmetal’s Angina

Prinzmetal’s angina (also called variant angina) is caused by spasms in a patient’s coronary arteries that temporarily decrease blood flow. These spasms are usually more severe and occur at rest, making them very similar to unstable anginas. Angina medication may be required to dissipate the angina, and in certain cases hospitalization may also be required.

Microvascular Angina

Microvascular Angina is caused by spasms within small coronary arteries. Much like Prinzmetal’s angina, these spasms temporarily restrict blood flow and can last longer than stable anginas. The pain associated with this type of angina may occur with shortness of breath, sleep problems, fatigue, and lethargy.

Angina in Women

Women may experience different symptoms that occur in men. Women may experience a stabbing pain instead of chest pressure and neck discomfort, both of which are not common symptoms in men. Since these symptoms are different, it may cause delays in seeking treatment.

Anginas in Relation to Heart Disease

Anginas usually signal other heart diseases and can put individuals at risk of heart attacks and coronary heart disease. All anginas are caused by decreased blood flow to the heart, but the cause of the decrease in blood flow could be different. For example, the buildup of plaque on the sides of arteries can restrict blood flow to the heart. Blood clots can also restrict blood flow to the heart and cause an angina but most of the time it is fatty buildup. This fatty buildup (called atherosclerosis) can cause a multitude of problems outside of just anginas. Atherosclerosis puts individuals at risk of a heart attack, stroke, numbness, weakness, chest pain, and transient ischemic attacks (often called mini-strokes). This shows the importance of understand anginas and the underlying causes and risk factors. By the end of this document, doctors and medical professionals can analyze the data to better understand how different variables affect the onset and risk of developing an angina.

Introduction to the Dataset

As mentioned earlier, BioMedEd, Inc. has provided data from a study done of patients that exhibited angina pectoris (a common symptom of CAD), allowing data scientists to analyze this data and gain knowledge on certain patterns related to the causes of CAD and other heart diseases. However, to use the data, the data must be cleaned and prepared for exploratory data analysis (EDA). In this document, I will go through the steps taken to prepare the heart data for EDA.

Setting up Workspace and Installing Packages

The code below shows the preparation of the workspace and cleaning up the environment, as well as installing and loading packages. “tidyverse” is an important package that includes “dplyr” which helps with data science work. Some of the organizational tools and commands from “dplyr” will be shown later. Note that the “install.packages(”tidyverse“)” line is commented because this package has already been installed on this machine.

```
# Clean up and set up
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
source("myfunctions.R")
#
# install and load libraries
# install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Loading and Looking at Heart Data

After reading the heart data into R studio, the data scientist can look at it using a variety of commands. The most important commands are “summary” which reports the quartile data, “str” which reports the type of data in the dataframe, and “names” which reports the names of the columns of data.

```
#
# load heart data
heart <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\dirtyHeart.csv",header=T)
#
# look at heart data
names(heart)
```

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

```
summary(heart)
```

```
##       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
```

```
str(heart)
```

```
## '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 ...
```

```
dim(heart)
```

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

```
class(heart)
```

```
## [1] "data.frame"
```

```
glimpse(heart)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, NA, 57, 57, 56, 44, 52, 57, 54, 48, NA, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 1, 4, 4, 3, 2, 2, 4, 4, 4, 4, 4, 2, 3, 2, 3, 3, 2, 4, 3, 2, 1~
## $ trestbps <int> 145, 130, 130, 120, 120, 140, NA, 120, 172, 150, 140, 130, 13~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, NA, 0, ~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, NA, ~
## $ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, NA, 0, 0, 0, ~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, NA, 0.~
## $ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, N~
## $ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
## $ thal     <int> 6, 3, 7, 3, 3, 3, 3, 3, 7, 7, 6, 3, 6, 7, 7, 3, 7, 3, 3, 3, 3~
## $ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
head(heart)
```

```
##   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
```

As you can see, these commands provide useful information about the quality of the data as well as what the data is.

Replacing Column Names

Based on the “names” command from earlier, the column names of the data in this dataset is very ambiguous. By using a “dplyr” command “rename” we can rename the columns so that it is easier to read.

```
#
# replacing column names
heart <- rename(heart, "Age" = "age", "Sex" = "sex", "Chest Pain Type" = "cp", "Resting Blood Pr
essure" = "trestbps", "Cholesterol" = "chol", "Fasting Blood Sugar" = "fbs", "Resting EEG Result
s" = "restecg", "Maximum Heart Rate" = "thalach", "Angina by Exercise" = "exang", "ST Depression
by Exerise" = "oldpeak", "ST Segment Slope" = "slope", "Major Vessels" = "ca", "Thalassemia" =
"thal", "Heart Attack" = "target")
```

Though this looks complicated, all this is doing is converting the old ambiguous column names to something more understandable.

```
#
# renamed columns
names(heart)
```

```
##  [1] "Age"                       "Sex"
##  [3] "Chest Pain Type"           "Resting Blood Pressure"
##  [5] "Cholesterol"               "Fasting Blood Sugar"
##  [7] "Resting EEG Results"       "Maximum Heart Rate"
##  [9] "Angina by Exercise"        "ST Depression by Exerise"
## [11] "ST Segment Slope"          "Major Vessels"
## [13] "Thalassemia"               "Heart Attack"
```

Replacing missing data

Most columns have missing data, either as “NA” or data that just doesn’t make sense. We can check this by running the “summary” command which reports the number of NA’s in a column.

```
summary(heart)
```

```
##       Age             Sex         Chest Pain Type Resting Blood Pressure
##  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
##   Cholesterol    Fasting Blood Sugar Resting EEG Results Maximum Heart Rate
##  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
##  Angina by Exercise ST Depression by Exerise ST Segment Slope Major Vessels
##  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
##   Thalassemia     Heart Attack
##  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
```

To fix this with numerical data, we can replace the NA’s with the average of the whole column (excluding NA’s).

```
#
# replace missing age data
heart$Age <- ifelse(is.na(heart$Age), round(mean(heart$Age, na.rm=TRUE),0), heart$Age)
heart$Age <- ifelse(heart$Age == 0, round(mean(heart$Age, na.rm=TRUE),0), heart$Age)
#
# replace Resting Blood Pressure data
heart$`Resting Blood Pressure` <- ifelse(is.na(heart$`Resting Blood Pressure`), round(mean(heart
$`Resting Blood Pressure`, na.rm=TRUE),0), heart$`Resting Blood Pressure`)
#
# replace cholesterol data
heart$Cholesterol <- ifelse(is.na(heart$Cholesterol), round(mean(heart$Cholesterol, na.rm=TRUE),
0), heart$Cholesterol)
#
# replace Maximum Heart Rate data
heart$`Maximum Heart Rate` <- ifelse(is.na(heart$`Maximum Heart Rate`), round(mean(heart$`Maximu
m Heart Rate`, na.rm=TRUE),0), heart$`Maximum Heart Rate`)
#
# replace ST Depression by Exercise data
heart$`ST Depression by Exerise` <- ifelse(is.na(heart$`ST Depression by Exerise`), round(mean(h
eart$`ST Depression by Exerise`, na.rm=TRUE),0), heart$`ST Depression by Exerise`)
#
# replace Major Vessels data
heart$`Major Vessels` <- ifelse(is.na(heart$`Major Vessels`), round(mean(heart$`Major Vessels`,
 na.rm=TRUE),0), heart$`Major Vessels`)
```

It should be noted that for age, there were a few observations with an age of 0 which does not make sense in the context of the collected data, so we must replace those values with the mean as well. Some of the non-numerical variables had only 2 levels, which means that I could input the most common level in place of the NA's. This can be done for "Sex" data, "Fasting Blood Sugar" data, "Angina by Exercise" data, and "Heart Attack" data.

```
#
# replace sex data
heart$Sex <- ifelse(is.na(heart$Sex), round(mean(heart$Sex, na.rm=TRUE),0), heart$Sex)
#
# replace Fasting blood sugar data
heart$`Fasting Blood Sugar` <- ifelse(is.na(heart$`Fasting Blood Sugar`), round(mean(heart$`Fast
ing Blood Sugar`, na.rm=TRUE),0), heart$`Fasting Blood Sugar`)
#
# replace Angina by Exerise data
heart$`Angina by Exercise` <- ifelse(is.na(heart$`Angina by Exercise`), round(mean(heart$`Angina
by Exercise`, na.rm=TRUE),0), heart$`Angina by Exercise`)
#
# replace Heart Attack data
heart$`Heart Attack` <- ifelse(is.na(heart$`Heart Attack`), round(mean(heart$`Heart Attack`, na.
rm=TRUE),0), heart$`Heart Attack`)
```

For the remaining variables ("Chest Pain Type", "Resting EEG Results", "ST Segment Slope", and "Thalessemia") a different approach was taken to fix the missing data. With more than 2 levels, the NAs were just omitted from the dataframe.

```
#
# removing remaining data
heart <- na.omit(heart)
```

With more than 2 possible options for the missing data, it would be reasonably unfair to assign the missing data to the most common option. Now we can check the dimensions of the dataset.

```
dim(heart)
```

```
## [1] 292  14
```

Changing Values Within Columns

The data found in each column is also ambiguous and must be dealt with. For example, changing the “Sex” values to “male” and “female” instead of “1” and “0”.

```
#
# Changing Sex values
heart$Sex<-ifelse(heart$Sex== 1,"Male","Female")
#
# Changing Chest Pain values
heart$`Chest Pain Type`<- ifelse(heart$`Chest Pain Type`== 1,"typical angina",
    ifelse(heart$`Chest Pain Type`== 2, "atypical angina",
    ifelse(heart$`Chest Pain Type`== 3, "non-anginal pain",
    ifelse(heart$`Chest Pain Type`== 4, "asymptomatic", heart$`Chest Pain Type`))))
#
# Changing fasting blood sugar values
heart$`Fasting Blood Sugar`<-ifelse(heart$`Fasting Blood Sugar`== 1,"High","Low")
#
# Changing resting EEG values
heart$`Resting EEG Results`<- ifelse(heart$`Resting EEG Results`== 0,"normal",
    ifelse(heart$`Resting EEG Results`== 1, "ST-T wave abnormality",
    ifelse(heart$`Resting EEG Results`== 2, "ventricular hypertrophy", heart$`Resting EEG Result
s`)))
#
# Changing exercise induced angina values
heart$`Angina by Exercise`<-ifelse(heart$`Angina by Exercise`== 0,"no","yes")
#
# Changing ST Segment Slope values
heart$`ST Segment Slope`<- ifelse(heart$`ST Segment Slope`== 0,"upsloping",
                                     ifelse(heart$`ST Segment Slope`== 1, "flat",
    ifelse(heart$`ST Segment Slope`== 2, "downsloping", heart$`ST Segment Slope`)))
#
# Changing Thalassemia values
heart$Thalassemia <- ifelse(heart$`Thalassemia`== 3,"normal",
                           ifelse(heart$`Thalassemia`== 6, "fixed defect",
                            ifelse(heart$Thalassemia ==7, "reversible defect", heart$Thalassemi
a)))
#
# Changing Heart Attack values
heart$`Heart Attack`<- ifelse(heart$`Heart Attack`== 1, "yes", "no")
```

Now we can check if the data is readable

```
head(heart)
```

```
##   Age    Sex  Chest Pain Type Resting Blood Pressure Cholesterol
## 1  63   Male   typical angina                    145         233
## 2  37   Male     asymptomatic                    130         250
## 3  41 Female     asymptomatic                    130         204
## 4  53   Male non-anginal pain                    120         236
## 5  57 Female  atypical angina                    120         354
## 6  57   Male  atypical angina                    140         192
##   Fasting Blood Sugar   Resting EEG Results Maximum Heart Rate
## 1                High                normal                150
## 2                 Low ST-T wave abnormality                187
## 3                 Low                normal                172
## 4                 Low ST-T wave abnormality                178
## 5                 Low ST-T wave abnormality                163
## 6                 Low ST-T wave abnormality                148
##   Angina by Exercise ST Depression by Exerise ST Segment Slope Major Vessels
## 1                 no                      2.3        upsloping             0
## 2                 no                      3.5        upsloping             0
## 3                 no                      1.4      downsloping             0
## 4                 no                      0.8      downsloping             0
## 5                yes                      0.6      downsloping             0
## 6                 no                      0.4             flat             0
##         Thalassemia Heart Attack
## 1      fixed defect          yes
## 2            normal          yes
## 3 reversible defect          yes
## 4            normal          yes
## 5            normal          yes
## 6            normal          yes
```

Grouping By Hypertension

Using the completed dataset we can manipulate the data using “dplyr” commands. For example, we can add a variable for hypertension. This is useful for doctors and medical professionals because hypertension is a risk factor for many heart diseases and can work in conjunction with angina pectoris to help doctors diagnose severe heart diseases. Also, since hypertension only requires a person’s blood pressure, we can easily determine whether someone has hypertension based upon the “Resting Blood Pressure” values. I will be using 140 mm Hg as the cutoff value for hypertension. Those with a resting blood pressure of greater than 140 mm Hg will be marked as “yes” for having hypertension and those with a resting blood pressure of less than 140 mm Hg will be marked as “no” for not having hypertension.

```
#
# adding a variable for hypertension based on resting blood pressure
heart <- mutate(heart, Hypertension = factor(1 * (heart$`Resting Blood Pressure` >= 140), labels
= c("no", "yes")))
```

We can now check to make sure this new variable has been implemented by using the “names” command

```
names(heart)
```

```
##  [1] "Age"                     "Sex"                    
##  [3] "Chest Pain Type"         "Resting Blood Pressure" 
##  [5] "Cholesterol"             "Fasting Blood Sugar"    
##  [7] "Resting EEG Results"     "Maximum Heart Rate"     
##  [9] "Angina by Exercise"      "ST Depression by Exerise"
## [11] "ST Segment Slope"        "Major Vessels"          
## [13] "Thalassemia"             "Heart Attack"           
## [15] "Hypertension"
```

Using this new variable, we can group the data based on hypertension to observe some basic patterns related to those with hypertension. This code below tells us the mean age, mean cholesterol, and mean maximum heart rate of those with or without hypertension.

```
#
# grouping data based upon hypertension
hypertensionGroup <- group_by(heart, Hypertension)
summarize(hypertensionGroup, Age = mean(Age), Cholesterol = mean(Cholesterol))
```

```
## # A tibble: 2 x 3
##   Hypertension   Age Cholesterol
##   <fct>        <dbl>       <dbl>
## 1 no            52.6        244.
## 2 yes           57.8        249.
```

This is especially useful for doctors and medical professionals because we can see that those with hypertension tend to be slightly older and have higher cholesterol. This puts those with hypertension at a high risk for heart disease, more so than with previously thought because of the other trends associated with those with hypertension.

Arranging By Age

It may be useful to arrange the dataset by ascending cholesterol since angina pectoris is primarily caused by cholesterol blocking the coronary arteries. This way, we can look for patterns in the data as the cholesterol increases.

```
#
# arranging data by increasing Cholesterol
heart <- arrange(heart, Cholesterol)
```

This can be done for other variables that may prove to be useful. For example, we can order the dataset by maximum heart rate, age, or resting blood pressure to see different patterns related to how prevalent angina is in different types of patients and what kind.

Filtering and Selecting Based on Different Variables

The data that was given contains many people that didn't have typical angina pain. To reveal the patterns of why this is the case, we must filter out the people with typical angina pain to work with those who are different.

```
#
# filtering Chest Pain Type data
nonTypicalAngina <-filter(heart, 'Chest Pain Type' != "typical angina")
#
# viewing nonTypicalAngina
head(nonTypicalAngina)
```

```
##   Age    Sex  Chest Pain Type Resting Blood Pressure Cholesterol
## 1  57   Male     asymptomatic                    150         126
## 2  53   Male  atypical angina                    130         131
## 3  44 Female  atypical angina                    108         141
## 4  71 Female     asymptomatic                    112         149
## 5  49   Male non-anginal pain                    118         149
## 6  45 Female non-anginal pain                    112         160
##   Fasting Blood Sugar   Resting EEG Results Maximum Heart Rate
## 1                High ST-T wave abnormality                173
## 2                 Low ST-T wave abnormality                115
## 3                 Low ST-T wave abnormality                175
## 4                 Low ST-T wave abnormality                125
## 5                 Low                normal                126
## 6                 Low ST-T wave abnormality                138
##   Angina by Exercise ST Depression by Exerise ST Segment Slope Major Vessels
## 1                 no                      0.2      downsloping             1
## 2                yes                      1.2             flat             1
## 3                 no                      0.6             flat             0
## 4                 no                      1.6             flat             0
## 5                 no                      0.8      downsloping             3
## 6                 no                      0.0             flat             0
##    Thalassemia Heart Attack Hypertension
## 1       normal          yes          yes
## 2       normal           no           no
## 3       normal          yes           no
## 4       normal          yes           no
## 5 fixed defect           no           no
## 6       normal          yes           no
```

Doctors and medical professionals can also analyze the EEG results in conjunction with the pain type. To prepare the data for such analysis, we can isolate those two variables using this code.

```
#
# Selecting chest pain type and EEG results data
painTypeEEGResults <- select(heart, `Chest Pain Type`, `Resting EEG Results`)
```

We can also isolate people based on the values of variables. For example, it would be useful for doctors and medical professionals if we were to isolate a group of highly vulnerable individuals. so that we could analyze their medical information. The code below isolates those who have a cholesterol higher than 272 and a resting blood pressure higher than 140. These numbers came from the 3rd quartile of variables, which can be seen from the summary function.

```
summary(heart$`Resting Blood Pressure`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    94.0   120.0   130.0   131.5   140.0   200.0
```

```
summary(heart$Cholesterol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   126.0   211.8   243.0   245.6   273.2   564.0
```

```
#
# filtering cholesterol and resting blood pressure data
highCholBP <-filter(heart, Cholesterol >= 273 & heart$'Resting Blood Pressure' >= 140)
dim(highCholBP)
```

```
## [1] 28 15
```

```
highCholBP
```

```
##    Age    Sex  Chest Pain Type Resting Blood Pressure Cholesterol
## 1   59   Male     asymptomatic                    160         273
## 2   68   Male non-anginal pain                    180         274
## 3   57   Male   typical angina                    152         274
## 4   57   Male non-anginal pain                    150         276
## 5   67 Female     asymptomatic                    152         277
## 6   60   Male     asymptomatic                    145         282
## 7   58 Female non-anginal pain                    150         283
## 8   54   Male     asymptomatic                    192         283
## 9   67   Male     asymptomatic                    160         286
## 10  56 Female     asymptomatic                    200         288
## 11  59   Male     asymptomatic                    170         288
## 12  55   Male non-anginal pain                    160         289
## 13  60   Male     asymptomatic                    140         293
## 14  51   Male     asymptomatic                    140         298
## 15  51   Male     asymptomatic                    140         299
## 16  71 Female non-anginal pain                    160         302
## 17  61 Female     asymptomatic                    145         307
## 18  51 Female     asymptomatic                    140         308
## 19  45   Male non-anginal pain                    142         309
## 20  46   Male     asymptomatic                    140         311
## 21  64 Female   typical angina                    140         313
## 22  39   Male     asymptomatic                    140         321
## 23  64 Female     asymptomatic                    180         325
## 24  59   Male     asymptomatic                    170         326
## 25  64   Male     asymptomatic                    140         335
## 26  65 Female non-anginal pain                    160         360
## 27  63 Female non-anginal pain                    150         407
## 28  65 Female     asymptomatic                    140         417
##    Fasting Blood Sugar   Resting EEG Results Maximum Heart Rate
## 1                  Low                normal                125
## 2                 High                normal                150
## 3                  Low ST-T wave abnormality                 88
## 4                  Low                normal                112
## 5                  Low ST-T wave abnormality                172
## 6                  Low                normal                142
## 7                 High                normal                162
## 8                  Low                normal                149
## 9                  Low                normal                108
## 10                High                normal                133
## 11                 Low                normal                159
## 12                 Low                normal                145
## 13                 Low                normal                170
## 14                 Low ST-T wave abnormality                122
## 15                 Low ST-T wave abnormality                173
## 16                 Low ST-T wave abnormality                162
## 17                 Low                normal                146
## 18                 Low                normal                142
## 19                 Low                normal                147
## 20                 Low ST-T wave abnormality                120
## 21                 Low ST-T wave abnormality                133
## 22                 Low                normal                182
## 23                 Low ST-T wave abnormality                154
## 24                 Low                normal                140
## 25                 Low ST-T wave abnormality                158
## 26                 Low                normal                151
## 27                 Low                normal                154
## 28                High                normal                157
##    Angina by Exercise ST Depression by Exerise ST Segment Slope Major Vessels
## 1                  no                      0.0      downsloping             0
## 2                 yes                      1.6             flat             0
## 3                 yes                      1.2             flat             1
## 4                 yes                      0.6             flat             1
## 5                  no                      0.0      downsloping             1
## 6                 yes                      2.8             flat             2
## 7                  no                      1.0      downsloping             0
## 8                  no                      0.0      downsloping             1
## 9                 yes                      1.5             flat             3
## 10                yes                      4.0        upsloping             2
## 11                 no                      0.2             flat             0
## 12                yes                      0.8             flat             1
## 13                 no                      1.2             flat             2
## 14                yes                      4.2             flat             3
## 15                yes                      1.6      downsloping             0
## 16                 no                      0.4      downsloping             2
## 17                yes                      1.0             flat             0
## 18                 no                      1.5      downsloping             1
## 19                yes                      0.0             flat             3
## 20                yes                      1.8             flat             2
## 21                 no                      0.2      downsloping             0
## 22                 no                      0.0      downsloping             0
## 23                yes                      0.0      downsloping             0
## 24                yes                      3.4        upsloping             0
## 25                 no                      0.0      downsloping             0
## 26                 no                      0.8      downsloping             0
## 27                 no                      4.0             flat             3
## 28                 no                      0.8      downsloping             1
##          Thalassemia Heart Attack Hypertension
## 1             normal           no          yes
## 2  reversible defect           no          yes
## 3             normal           no          yes
## 4             normal           no          yes
## 5  reversible defect          yes          yes
## 6             normal           no          yes
## 7  reversible defect          yes          yes
## 8  reversible defect           no          yes
## 9  reversible defect           no          yes
## 10 reversible defect           no          yes
## 11            normal           no          yes
## 12            normal           no          yes
## 13 reversible defect           no          yes
## 14      fixed defect           no          yes
## 15            normal           no          yes
## 16            normal          yes          yes
## 17 reversible defect           no          yes
## 18 reversible defect          yes          yes
## 19            normal           no          yes
## 20      fixed defect          no          yes
## 21      fixed defect         yes          yes
## 22            normal         yes          yes
## 23 reversible defect         yes          yes
## 24            normal          no          yes
## 25      fixed defect          no          yes
## 26            normal         yes          yes
## 27            normal          no          yes
## 28            normal         yes          yes
```

As you can see, there are 28 high risk individuals with higher cholesterol and blood pressure. By isolating these people, we can analyze some other shared variables to determine potential causes of angina pectoris and assess the risk of those who don’t have a higher cholesterol and blood pressure. By grouping, filtering, and selecting certain data, we prepare the data for analysis.

Work Cited

“Angina (Chest PAIN).” Www.heart.org, American Heart Association, 2021, www.heart.org/en/health-topics/heart-attack/angina-chest-pain.

“Angina.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 12 June 2020, www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373.

“Atherosclerosis.” National Heart Lung and Blood Institute, U.S. Department of Health and Human Services, 2021, www.nhlbi.nih.gov/health-topics/atherosclerosis.

“Facts about Hypertension.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 July 2021, www.cdc.gov/bloodpressure/facts.htm.

“Heart Disease.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Jan. 2021, www.cdc.gov/heartdisease/index.htm.

“The Top 10 Causes of Death.” World Health Organization, World Health Organization, 9 Dec. 2020, www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.