Project Draft

Nischal Bhandari

06/12/2020

Load packages

```
library(tidyverse)
library(skimr)
library(datasauRus)
library(dplyr)
library(tidymodels)
```

Load data

```
heart <- read_csv("data/heart.csv")
```

Introduction

The dataset I used includes observations among potential heart disease patients. The observations are done in the following aspects of patients:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol 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
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. target (if patient have heart disease)

Heart dataset

This dataset is claimed to be frequently used by ML researchers. It is the work of Cleaveland datase that is contributed by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

I obtained this dataset from this link: https://www.kaggle.com/ronitf/heart-disease-uci

Objective:

My objective on this project is to tinker around the potential links between various bodily, metabolic conditions rather than reach a specific conclusions, so I potentially won't have much hypotheses.

Ex. 1.

Understanding our dataset

There are 14 variables and 303 observations in our dataset.

We will print out the summary of our dataset.

Table 1: Data summary

Name	Piped data
Number of rows	303
Number of columns	14
Column type frequency:	
numeric	14
Group variables	None

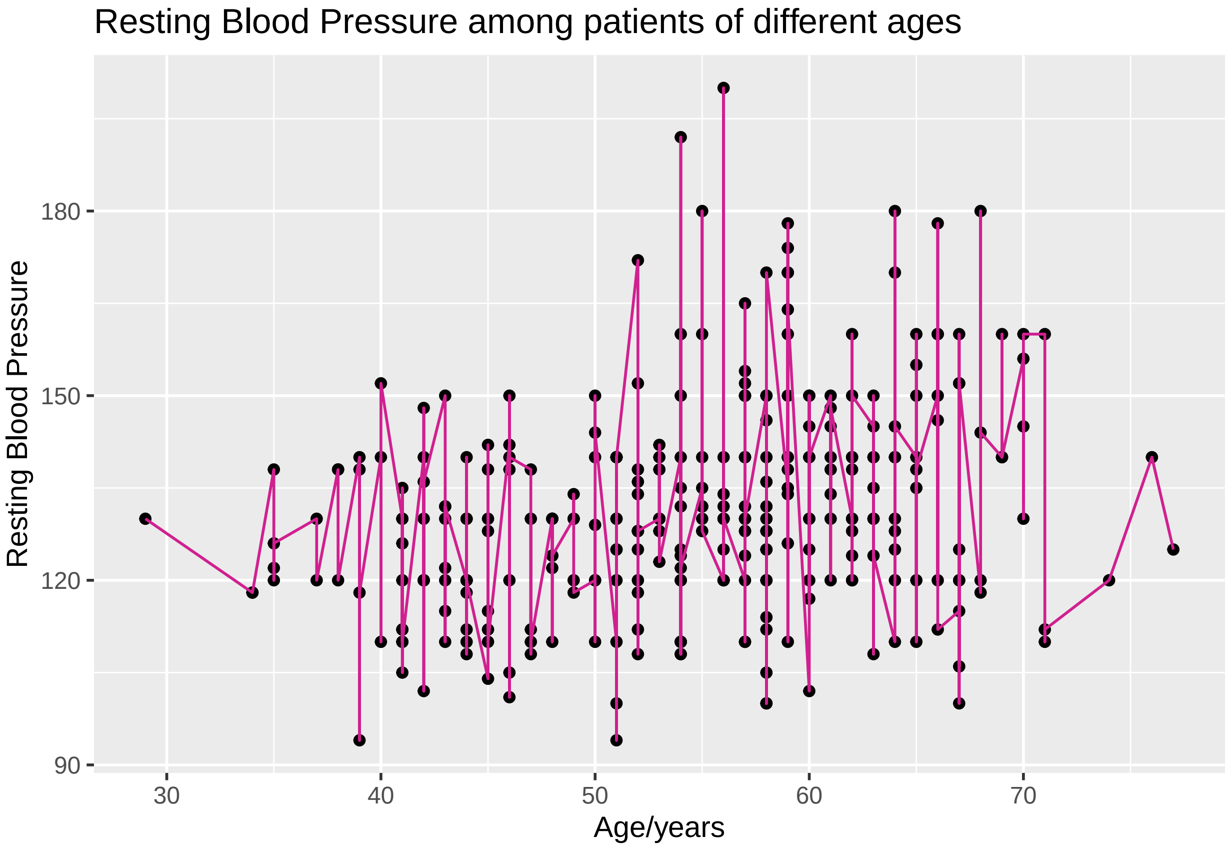
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
age	0	1	54.37	9.08	29	47.5	55.0	61.0	77.0
sex	0	1	0.68	0.47	0	0.0	1.0	1.0	1.0
cp	0	1	0.97	1.03	0	0.0	1.0	2.0	3.0
trestbps	0	1	131.62	17.54	94	120.0	130.0	140.0	200.0
chol	0	1	246.26	51.83	126	211.0	240.0	274.5	564.0
fbs	0	1	0.15	0.36	0	0.0	0.0	0.0	1.0
restecg	0	1	0.53	0.53	0	0.0	1.0	1.0	2.0
thalach	0	1	149.65	22.91	71	133.5	153.0	166.0	202.0
exang	0	1	0.33	0.47	0	0.0	0.0	1.0	1.0
oldpeak	0	1	1.04	1.16	0	0.0	0.8	1.6	6.2
slope	0	1	1.40	0.62	0	1.0	1.0	2.0	2.0
ca	0	1	0.73	1.02	0	0.0	0.0	1.0	4.0
thal	0	1	2.31	0.61	0	2.0	2.0	3.0	3.0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
target	0	1	0.54	0.50	0	0.0	1.0	1.0	1.0

Ex. 2

We will plot to visualize the relationship between `age` and `resting blood pressure` of a patient.



It can not be inferred clearly the relationship between `age` and `resting blood pressure`, but generally aged patients tend to have more `resting blood pressure`.

Summarising the `resting blood pressure`

```
##     trestbps    
##  Min.   : 94.0  
##  1st Qu.:120.0  
##  Median :130.0  
##  Mean   :131.6  
##  3rd Qu.:140.0  
##  Max.   :200.0  
```

The minimum resting blood pressure is 94.0, naximum is 200.0, and mean is 131.6.

Ex. 3.

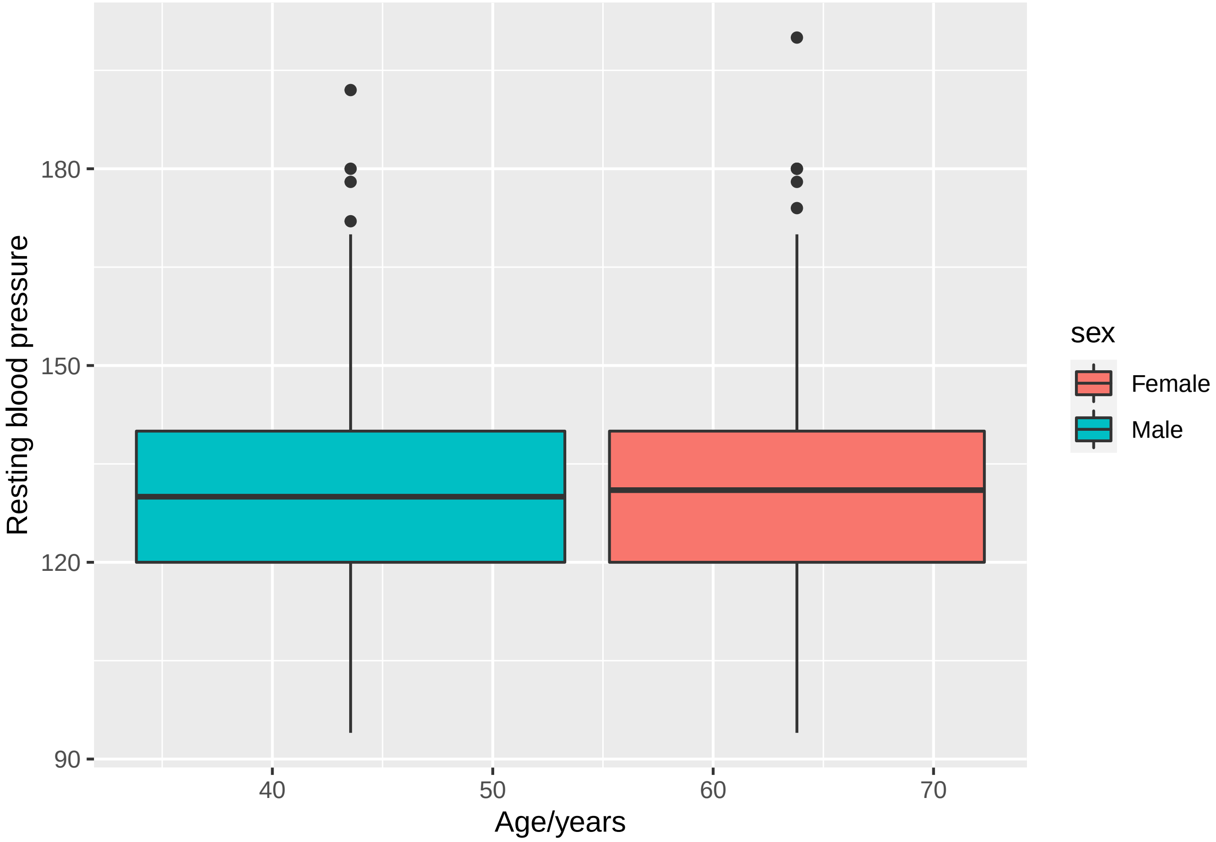
We will now calculate summary for male and female patients.

First of all, we will mutate `sex` variable.

```
## # A tibble: 303 x 14
##      age sex      cp trestbps  chol   fbs restecg thalach exang oldpeak slope
##    <dbl> <chr> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>
##  1    63 Male      3      145   233     1       0     150     0     2.3     0
##  2    37 Male      2      130   250     0       1     187     0     3.5     0
##  3    41 Fema~     1      130   204     0       0     172     0     1.4     2
##  4    56 Male      1      120   236     0       1     178     0     0.8     2
##  5    57 Fema~     0      120   354     0       1     163     1     0.6     2
##  6    57 Male      0      140   192     0       1     148     0     0.4     1
##  7    56 Fema~     1      140   294     0       0     153     0     1.3     1
##  8    44 Male      1      120   263     0       1     173     0     0       2
##  9    52 Male      2      172   199     1       1     162     0     0.5     2
## 10    57 Male      2      150   168     0       1     174     0     1.6     2
## # ... with 293 more rows, and 3 more variables: ca <dbl>, thal <dbl>,
## #   target <dbl>
```

Now, we have categorized gender of patients as a character variable.

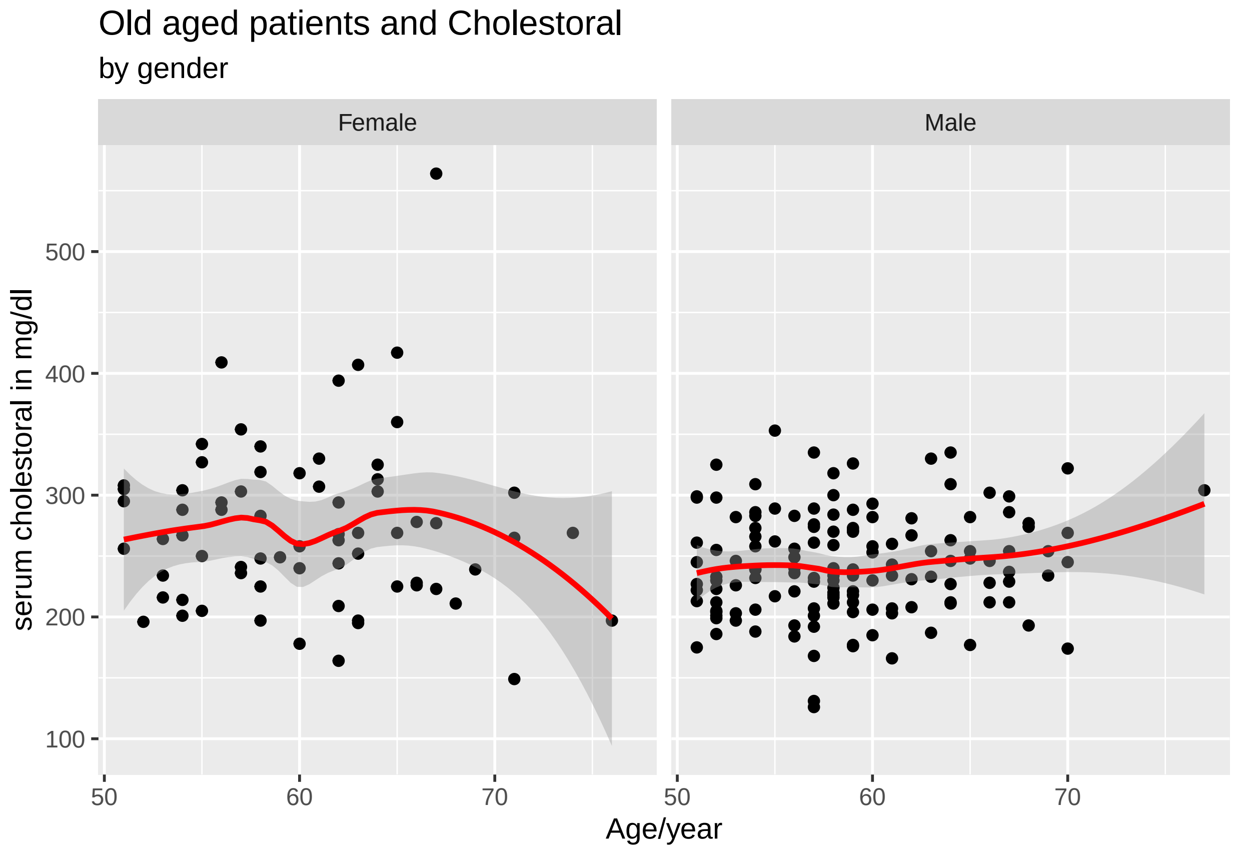
We will create boxplot for `age`, `resting blood pressure`, `gender` variables.



Ex. 4

We will filter patients that are above 50 to visualize the relationship between their `age` and `blood cholestoral`.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Generally, the ’blood cholesterol‘ in female is low as their age increases, but in case of male, their cholesterol level generally rises with the increase in their age.

Ex. 5

We’ll count the cases of excercise-induced angina among heart patients.

Mutating the variable first.

```
## # A tibble: 303 x 14
##      age sex      cp trestbps  chol   fbs restecg thalach exang oldpeak slope
##    <dbl> <chr> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <chr>   <dbl> <dbl>
##  1    63 Male      3      145   233     1       0     150 NO        2.3     0
##  2    37 Male      2      130   250     0       1     187 NO        3.5     0
##  3    41 Fema~     1      130   204     0       0     172 NO        1.4     2
##  4    56 Male      1      120   236     0       1     178 NO        0.8     2
##  5    57 Fema~     0      120   354     0       1     163 Yes       0.6     2
##  6    57 Male      0      140   192     0       1     148 NO        0.4     1
##  7    56 Fema~     1      140   294     0       0     153 NO        1.3     1
##  8    44 Male      1      120   263     0       1     173 NO        0         2
##  9    52 Male      2      172   199     1       1     162 NO        0.5     2
## 10    57 Male      2      150   168     0       1     174 NO        1.6     2
## # ... with 293 more rows, and 3 more variables: ca <dbl>, thal <dbl>,
## #   target <dbl>
```

Now, we will count the number of exercise-induced angina.

```
heart %>%
  count(exang, sort = TRUE)
```

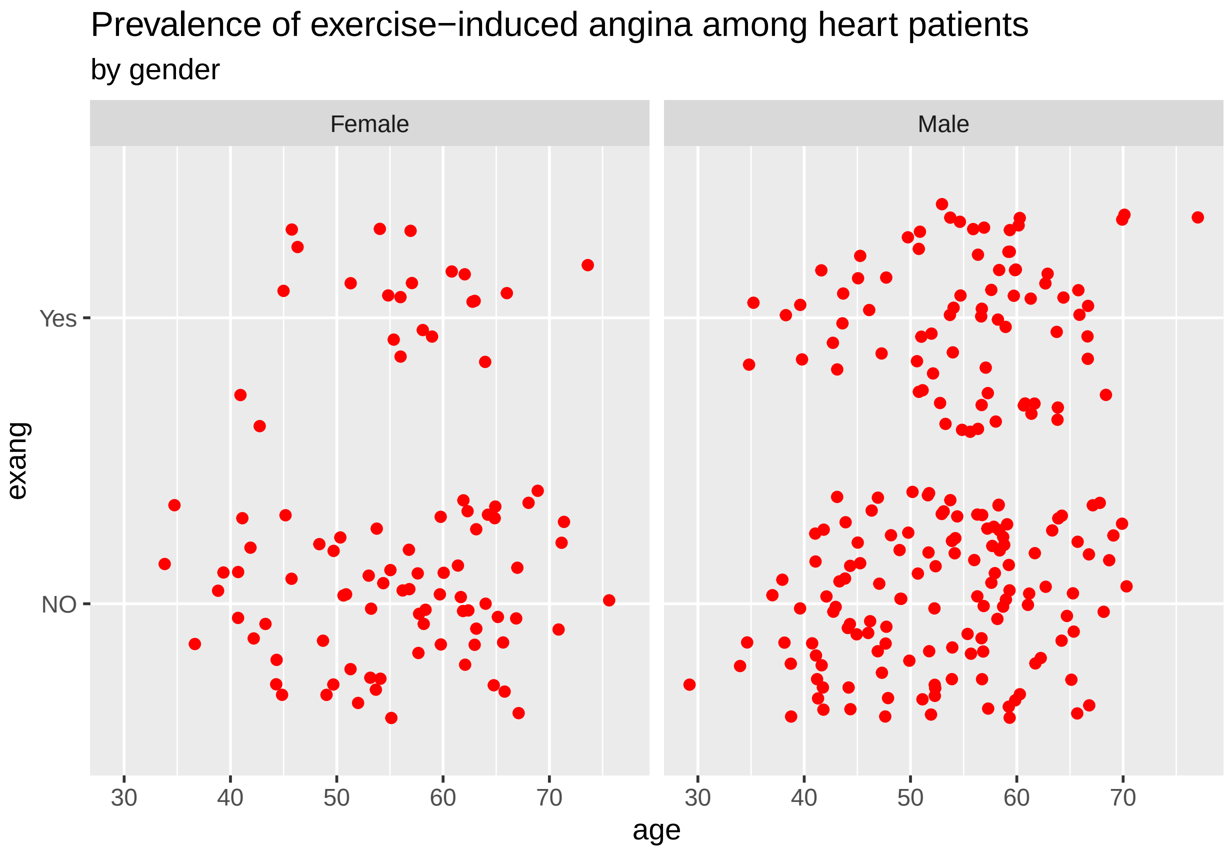
```
## # A tibble: 2 x 2
##   exang     n
##   <chr> <int>
## 1 NO      204
## 2 Yes      99
```

So 204 patients didn't have exercise-induced angina while 99 had.

Ex. 6

We will love to see if such induced angina was more prevalent among male.

```
ggplot(heart, mapping = aes(x = age, y = exang))+
geom_jitter(color="red")+
facet_wrap(.~sex)+
labs(
title = "Prevalence of exercise-induced angina among heart patients",
subtitle = "by gender"
)
```



Comparatively, less number of women with heart diseases have exercise-induced angina than man.

Ex. 7

Now, we will counnt the distribution of `chest pain` among our patients.

```
## # A tibble: 4 x 2
##       cp     n
##   <dbl> <int>
## 1     0   143
## 2     2    87
## 3     1    50
## 4     3    23
```

Majority i.e. 143 patients had `type-0` chest pain while 87 had `type-2`, 50 `type-1` and 23 `type-3` which represent least number of patients.

Ex. 8.

We will try to model `thalach` which is maximum heart rate achieved as dependent on `age`, `the resting blood pressure`, and `Cholesterol level`.

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 190.       11.0         17.2  1.84e-46
## 2 age          -1.09      0.141      -7.74 1.58e-13
## 3 chol          0.0329    0.0239      1.38 1.70e- 1
## 4 trestbps      0.0849    0.0719      1.18 2.39e- 1
```

The model can be expressed in a formula as:

$predicted_maximum_heart_rate = 190 + 0.0329 * chol - 1.09 * age + 0.0849 * trestbps$

Ex. 9

We will check if our model represent an accurate interaction between variables.

```
## [1] 0.1686218
```

```
## [1] 0.1602802
```

The R squared value is 0.1686218 while adjusted R squared value is 0.1602802. These values are pretty low, so this model doesn't represent good interaction between the response and independent variables.