Heart Disease Indicators

CPE213 – Data Model

King Mongkut's University of Technology Thonburi

Outline

01 Introduction to the problem

02 Analytic objective

03 Data description and preparation

04 Data exploration and visualization

Outline

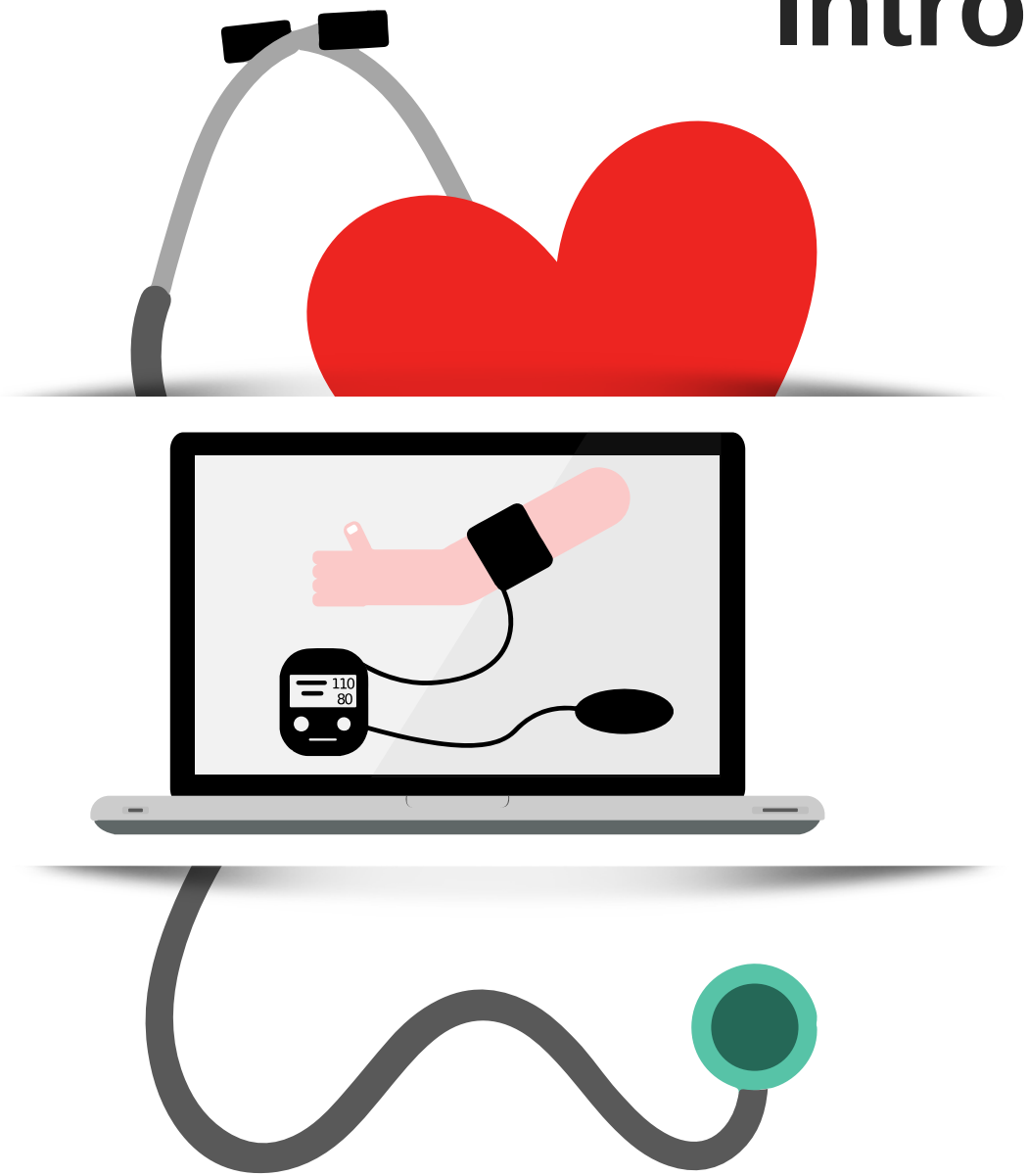
05 Model explanation

06 Modeling implementation

07 Evaluation

08 Discussion and conclusion

Introduction to the problem



สาเหตุของโรคหัวใจ

สาเหตุของโรคหัวใจขึ้นอยู่กับชนิดของโรคหัวใจนั้น ๆ สำหรับโรคหลอดเลือดหัวใจจากภาวะการเสื่อมของหลอดเลือด EN มีสาเหตุไม่ชัดเจนแต่พบว่าสัมพันธ์กับปัจจัยเสี่ยงต่าง ๆ โดยปัจจัยเสี่ยงแบ่งออกเป็นสองประเภท คือ

- ปัจจัยเสี่ยงที่ควบคุมและสามารถปรับเปลี่ยนได้
- ปัจจัยเสี่ยงที่ไม่สามารถเปลี่ยนแปลง เช่น อายุ เพศ หรือประวัติสุขภาพของคนในครอบครัว

สภาวะที่มีผลต่อโรคหัวใจ

- โรคหลอดเลือด เช่น โรคหลอดเลือดหัวใจ
- ปัญหาจังหวะการเต้นของหัวใจ (ภาวะหัวใจเต้นผิดจังหวะ)
- ข้อบกพร่องของหัวใจแต่กำเนิด (ข้อบกพร่องของหัวใจพิการแต่กำเนิด)
- โรคลิ้นหัวใจตีบหรือรั่ว
- โรคของกล้ามเนื้อหัวใจ
- การติดเชื้อที่หัวใจ
- โรคของผนังหุ้มหัวใจ

ANALYTIC
OBJECTIVE

เพื่อนำข้อมูลจากทางการแพทย์ที่ประเมิน
ไว้เบื้องต้นมาประเมินความเสี่ยงในการ
เกิดโรคหัวใจในผู้ป่วย

Data description

ข้อมูลนี้จะมีตัวแปรทั้งหมด 14 attributes

age

อายุของผู้ป่วย

ค่าที่เป็นไปได้ : ช่วงอายุ 29 ถึง 77 ปี

sex

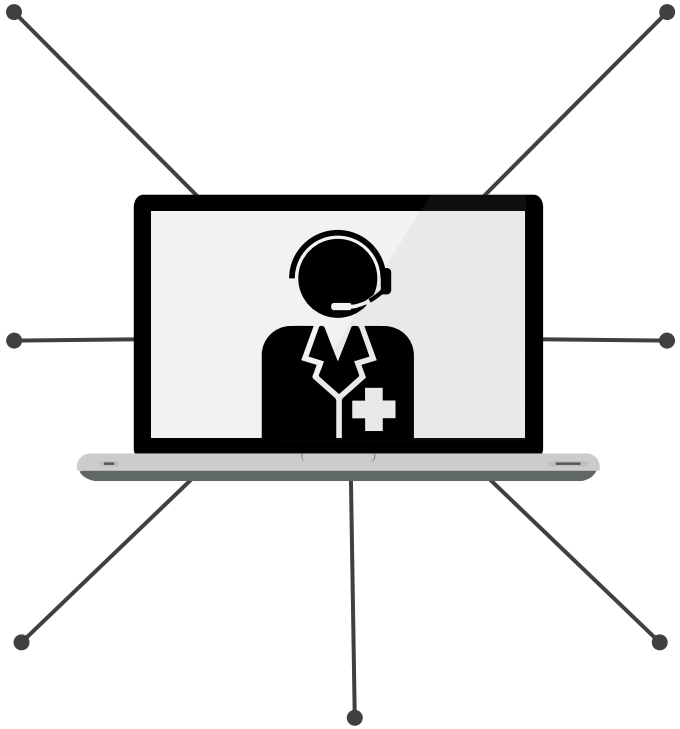
เพศของผู้ป่วย

ค่าที่เป็นไปได้ : เพศหญิง(0), เพศชาย(1)

Cp

รูปแบบการปวดหน้าอกของผู้ป่วย

ค่าที่เป็นไปได้ : หลอดเลือดหัวใจตีบ(0), หลอดเลือดหัวใจตีบผิดปกติ(1), ไม่เป็นหลอดเลือดหัวใจตีบ(2), ไม่มีอาการ(3)



restecg

ผลลัพธ์การประมวลผลการเต้นของหัวใจด้วยคลื่นไฟฟ้า

ค่าที่เป็นไปได้ : ปกติ(0), ST-T abnormal(1), hypertrophy(2)

trestbps

ช่วงการคลายตัวของหัวใจ

ค่าที่เป็นไปได้ : 94 ถึง 200

chol

ปริมาณ cholesterol ในเลือดของผู้ป่วย

ค่าที่เป็นไปได้ : 126 ถึง 564

fbs

ปริมาณน้ำตาลในเลือดของผู้ป่วยมากกว่า 120

ค่าที่เป็นไปได้ : เท็จ(0), จริง(1)

Data description (cont.)

thalach

อัตราการเต้นหัวใจสูงสุดที่วัดได้จากผู้ป่วย

ค่าที่เป็นไปได้ : 71 ถึง 202

exang

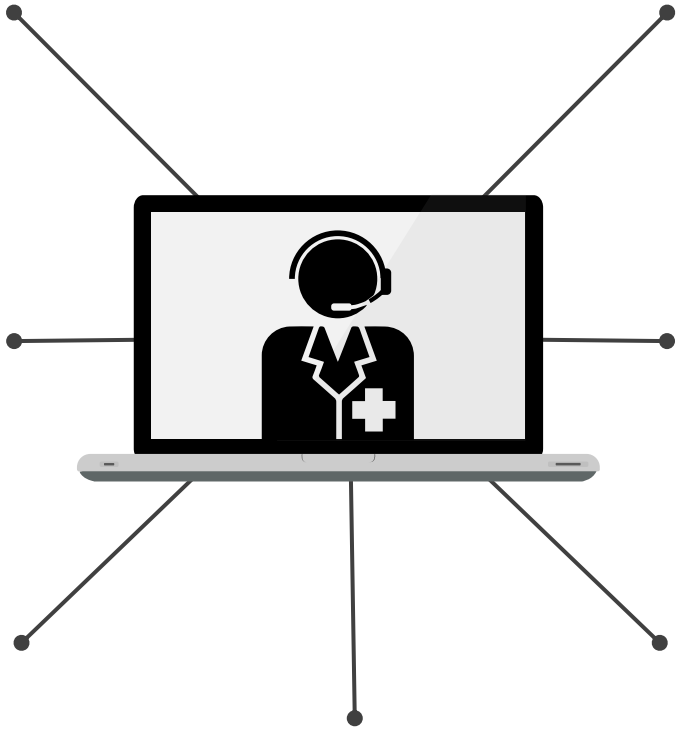
การออกกำลังกายที่ทำให้เกิดโรคหลอดเลือดหัวใจตีบ

ค่าที่เป็นไปได้ : จริง(1), เท็จ(0)

oldpeak

ระดับภาวะการซึมเศร้าจากการออกกำลังกาย
เมื่อเทียบกับสภาวะปกติของผู้ป่วย

ค่าที่เป็นไปได้ : 0 ถึง 6.2



slope

ระดับความชันของการออกกำลังกาย

ค่าที่เป็นไปได้ : ชันสูง(0), แบนราบ(1), ชันต่ำ(2)

ca

number of major vessels colored by flourosopy

ค่าที่เป็นไปได้ : 0 ถึง 3

thal

สถานะการเต้นของหัวใจของผู้ป่วย

ค่าที่เป็นไปได้ : ปกติ(3), บกพร่องคงที่(6),
บกพร่องย้อนกลับได้(7)

target

ผลสรุปของผู้ป่วย

ค่าที่เป็นไปได้ : เป็นไม่โรคหัวใจ(0), เป็นโรคหัวใจ(1)

Data exploration and visualization

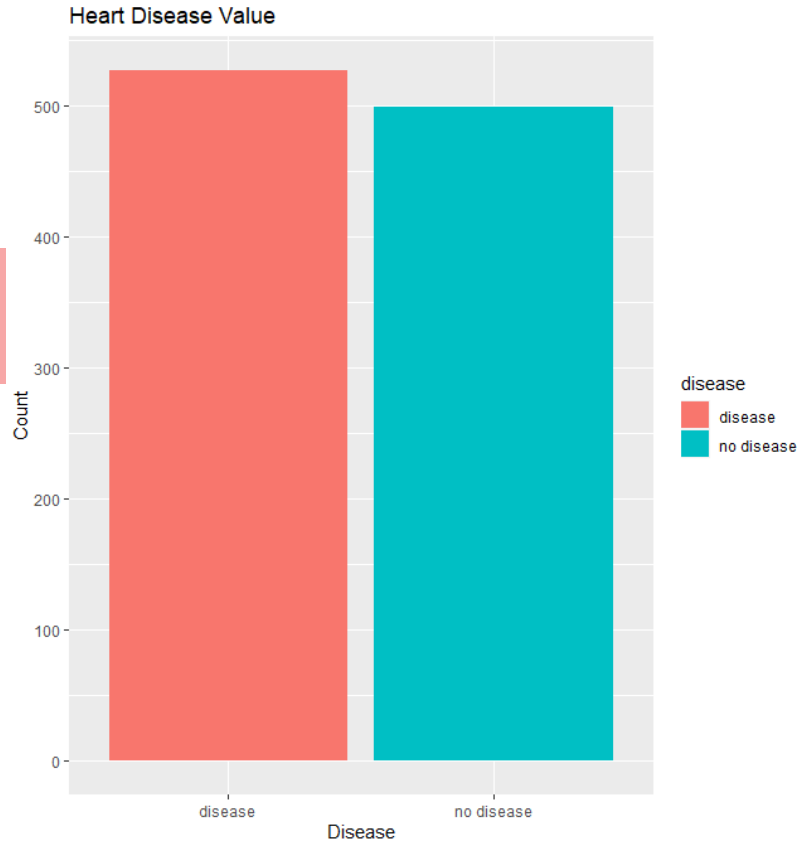
Data table

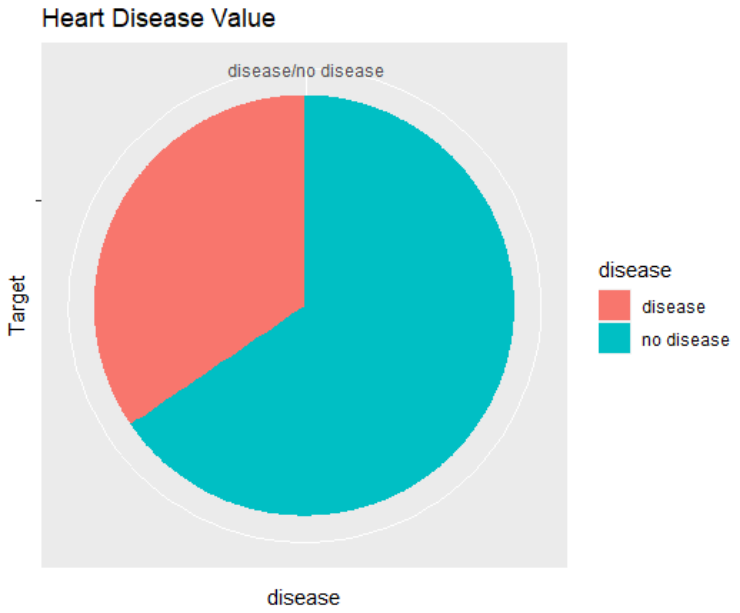
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
2	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
3	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
4	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
5	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
6	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
7	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
8	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
9	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
10	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
11	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
12	43	0	0	132	341	1	0	136	1	3.0	1	0	3	0
13	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
14	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
15	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0
16	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
17	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1

Data structure

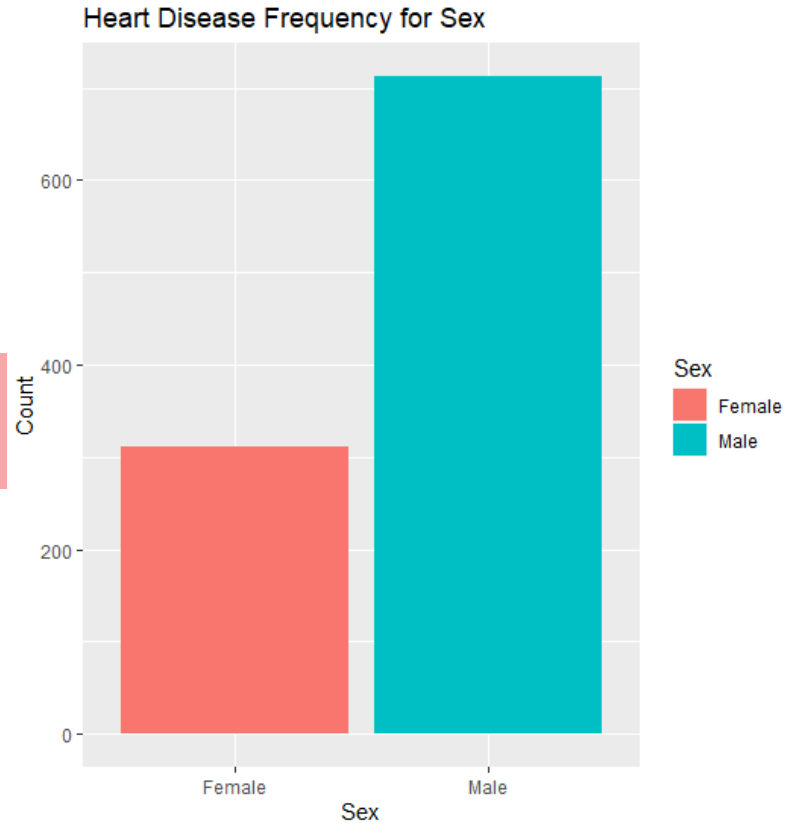
```
'data.frame':   1025 obs. of  14 variables:
 $ age     : int  52 53 70 61 62 58 58 55 46 54 ...
 $ sex     : int  1 1 1 1 0 0 1 1 1 1 ...
 $ cp      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ trestbps: int  125 140 145 148 138 100 114 160 120 122 ...
 $ chol    : int  212 203 174 203 294 248 318 289 249 286 ...
 $ fbs     : int  0 1 0 0 1 0 0 0 0 0 ...
 $ restecg : int  1 0 1 1 1 0 2 0 0 0 ...
 $ thalach : int  168 155 125 161 106 122 140 145 144 116 ...
 $ exang   : int  0 1 1 0 0 0 0 1 0 1 ...
 $ oldpeak : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
 $ slope   : int  2 0 0 2 1 1 0 1 2 1 ...
 $ ca      : int  2 0 0 1 3 0 3 1 0 2 ...
 $ thal    : int  3 3 3 3 2 2 1 3 3 2 ...
 $ target  : int  0 0 0 0 0 1 0 0 0 0 ...
```

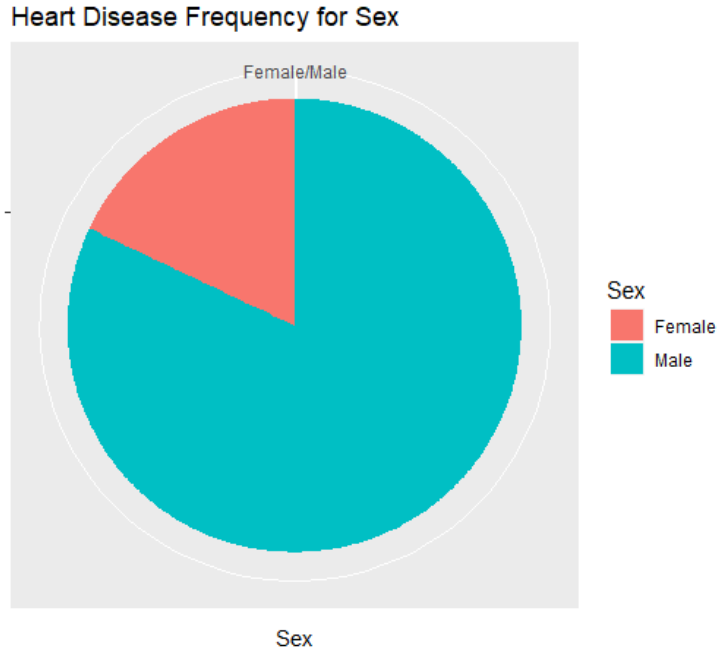
Data Analysis



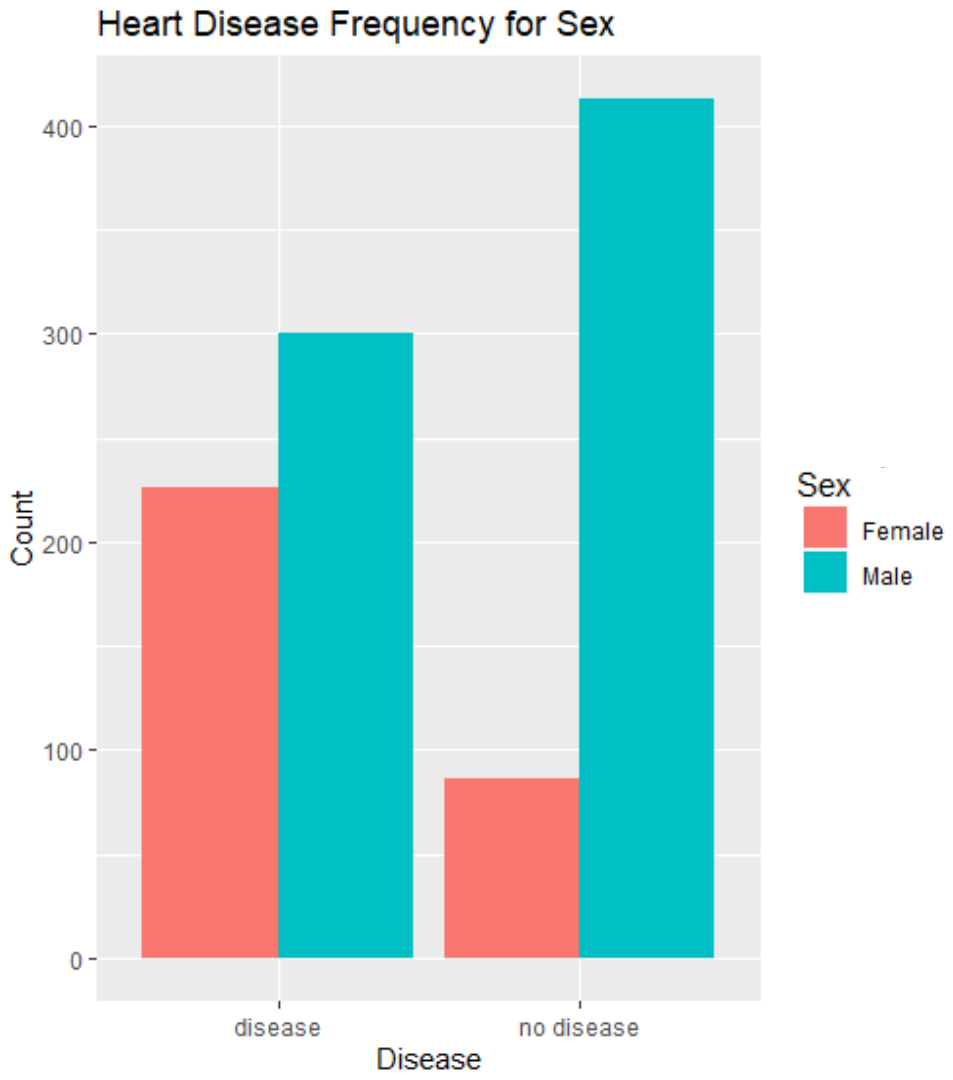


Graph
Sex_char

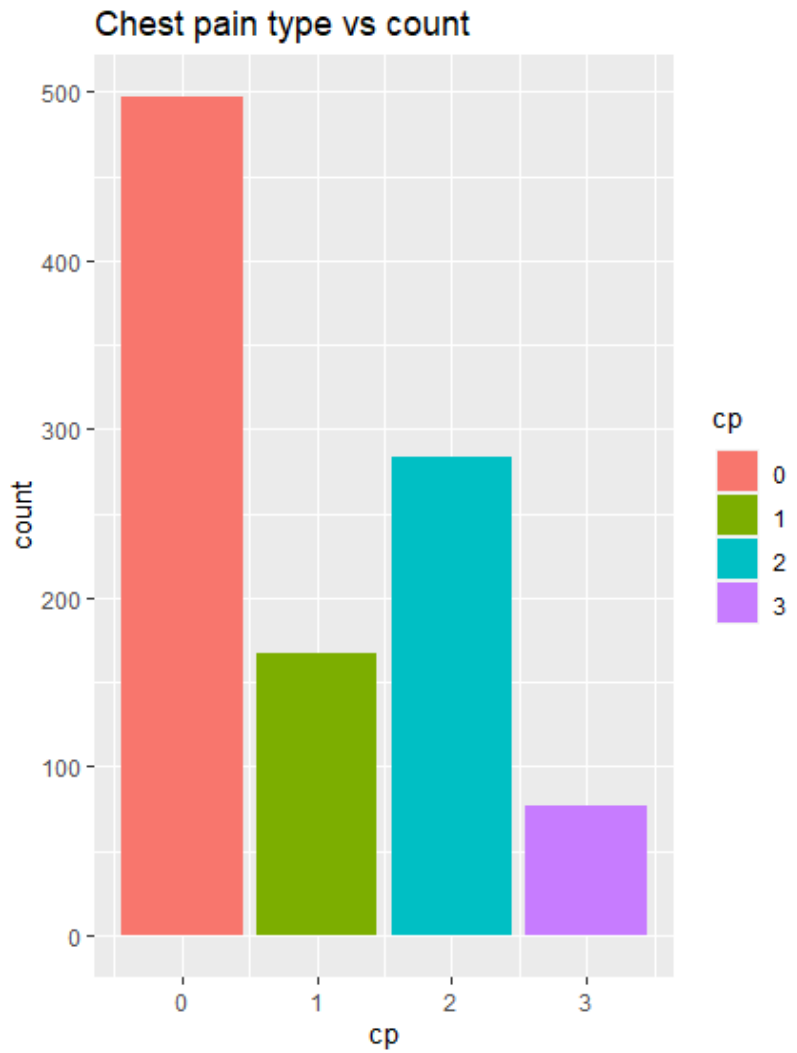




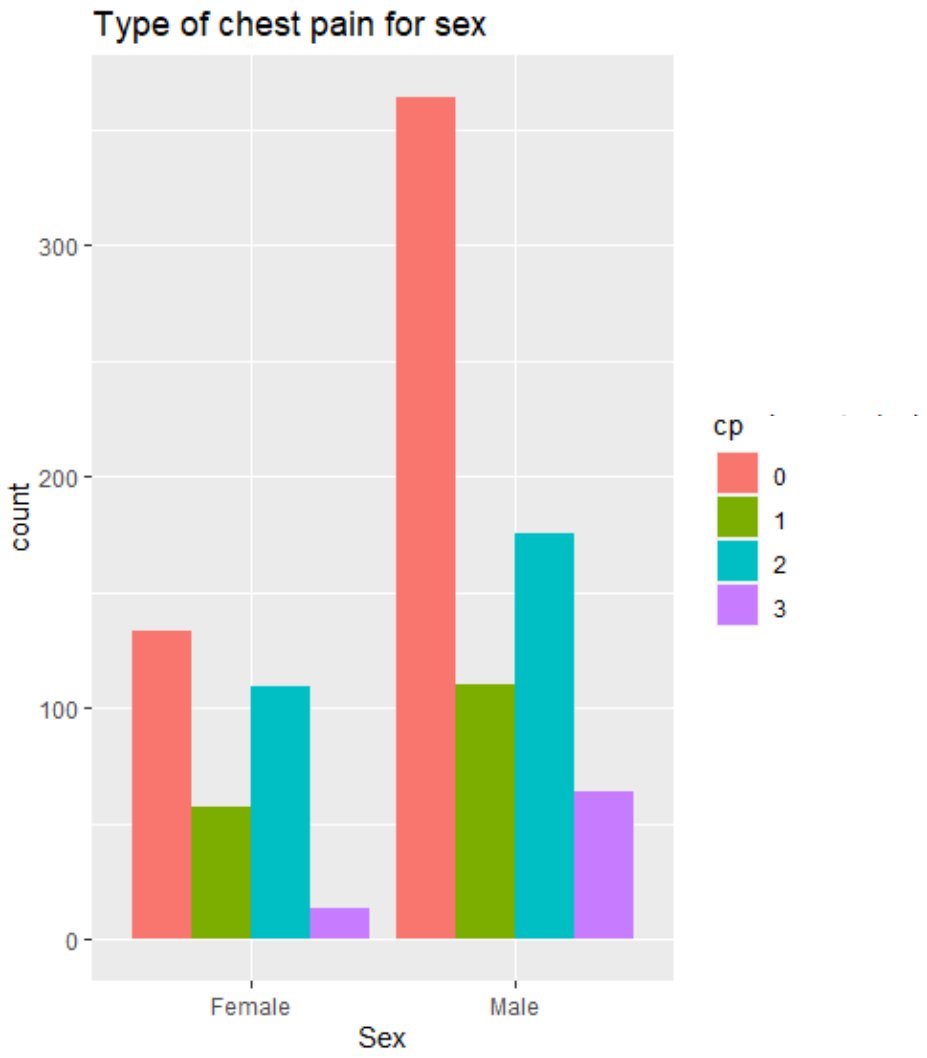
	disease	no disease
Female	226	86
Male	300	413


Heart Disease Frequency for Sex
400
300
200
100
0
Count
disease
no disease
Disease
Sex
Female
Male

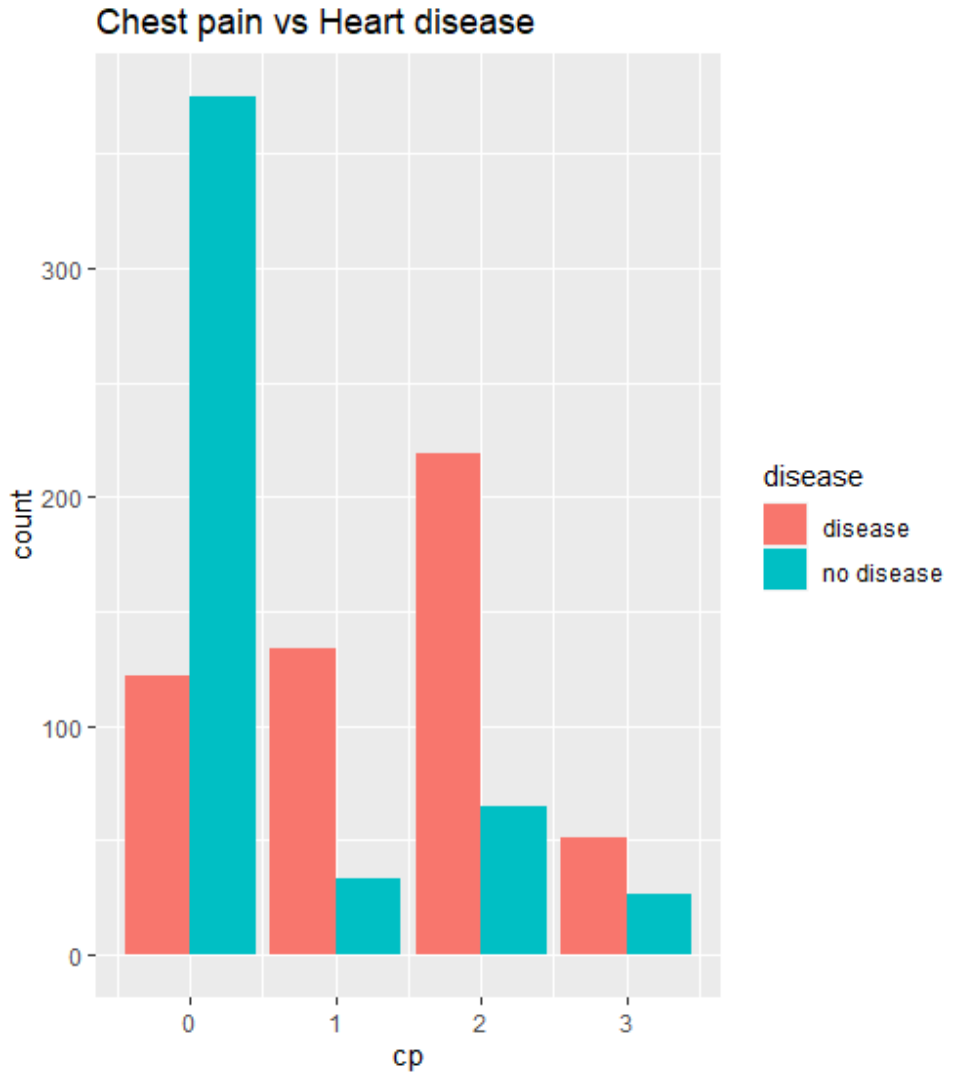
Graph
CP


Chest pain type vs count
500
400
300
200
100
0
count
0
1
2
3
cp
cp
0
1
2
3

	0	1	2	3
Female	133	57	109	13
Male	364	110	175	64

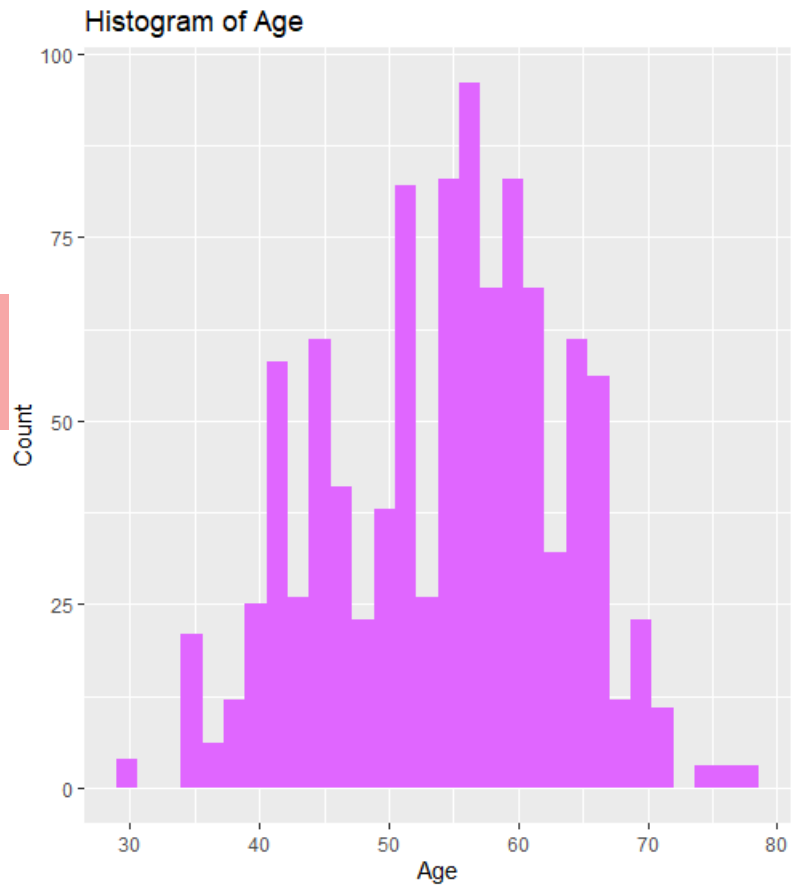


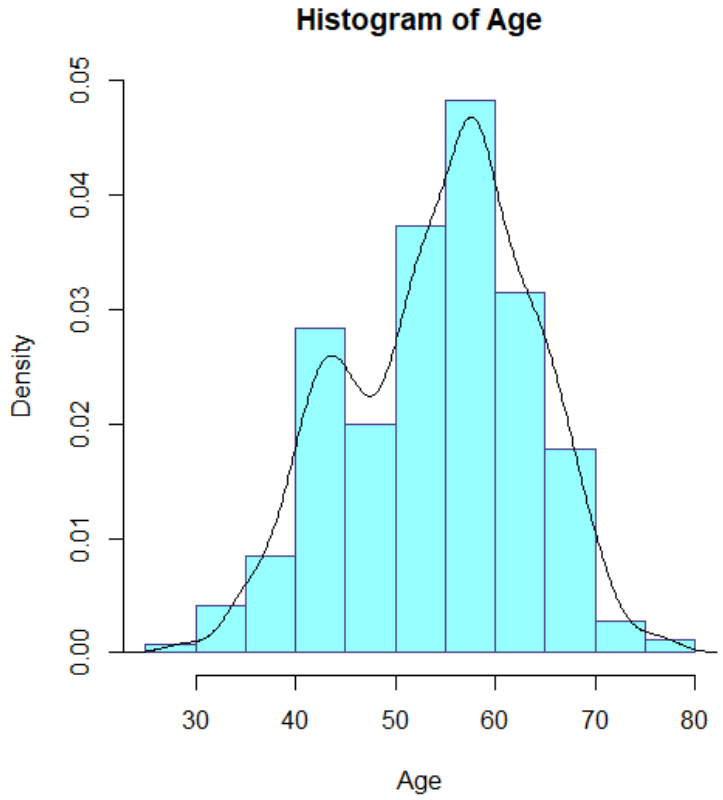
	0	1	2	3
disease	122	134	219	51
no disease	375	33	65	26

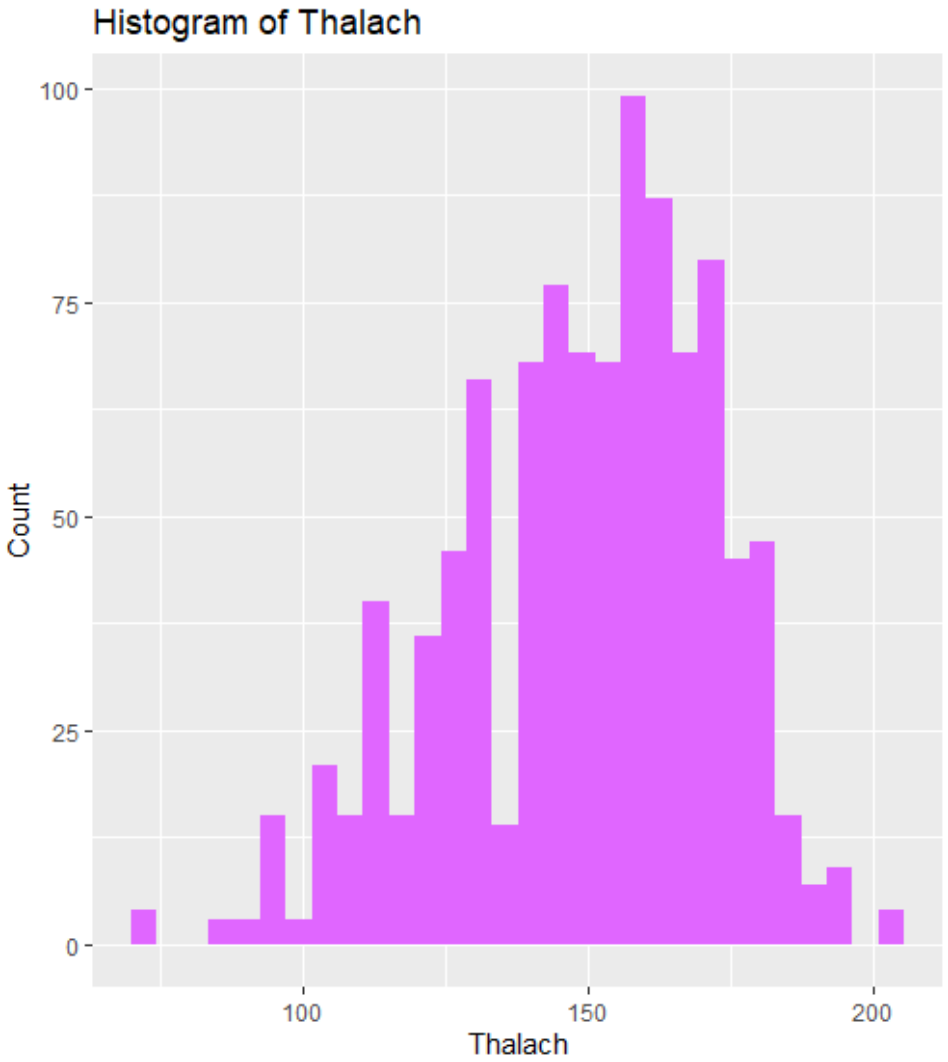


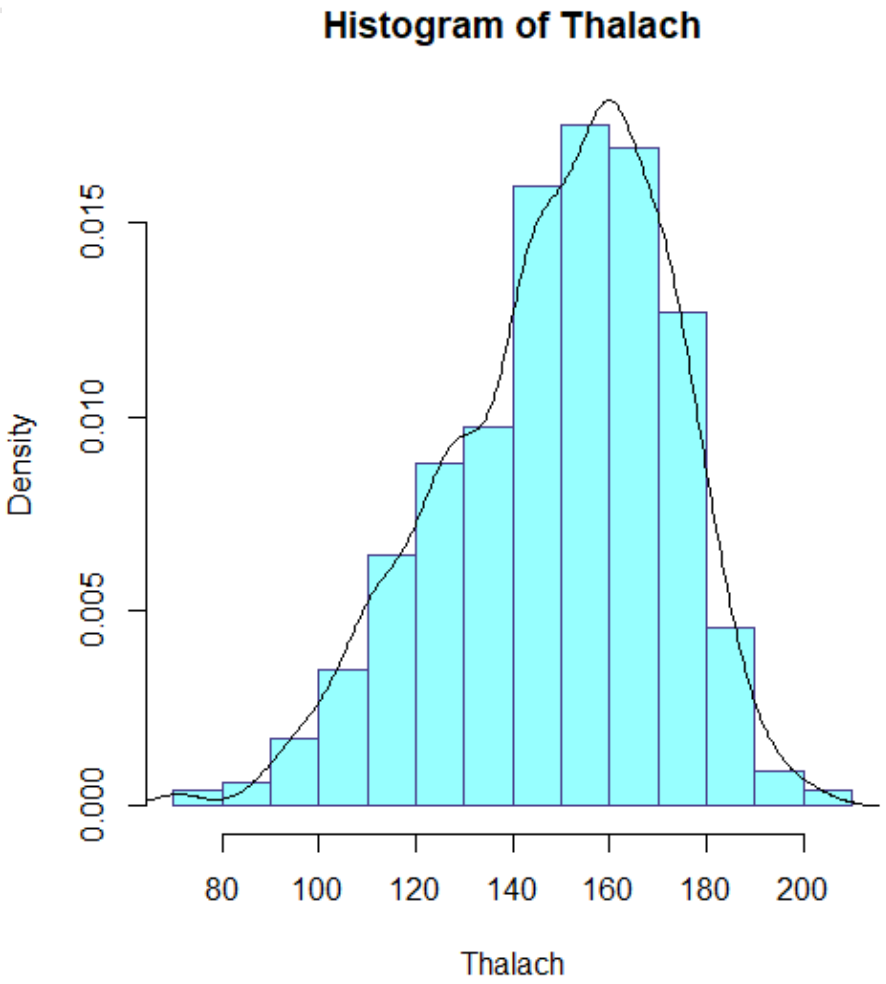
Graph

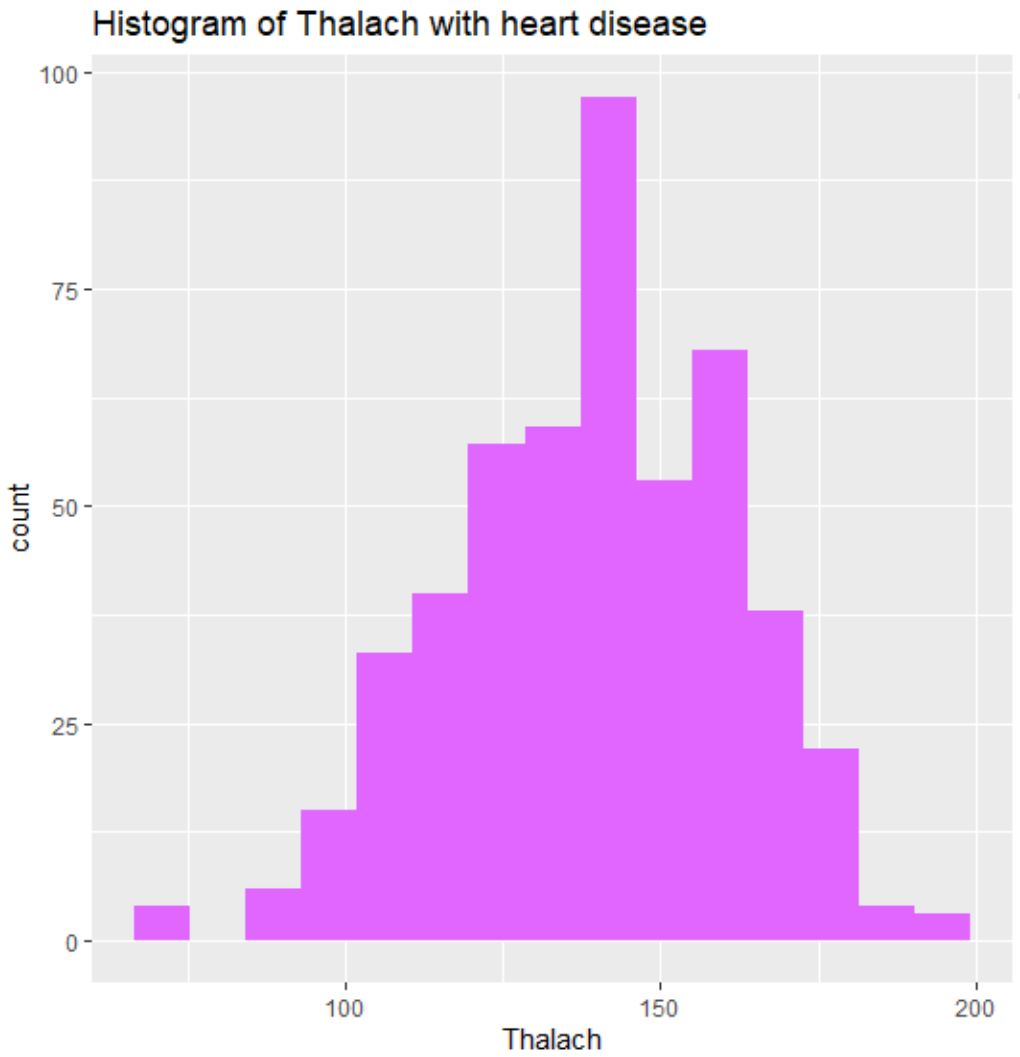
histogram

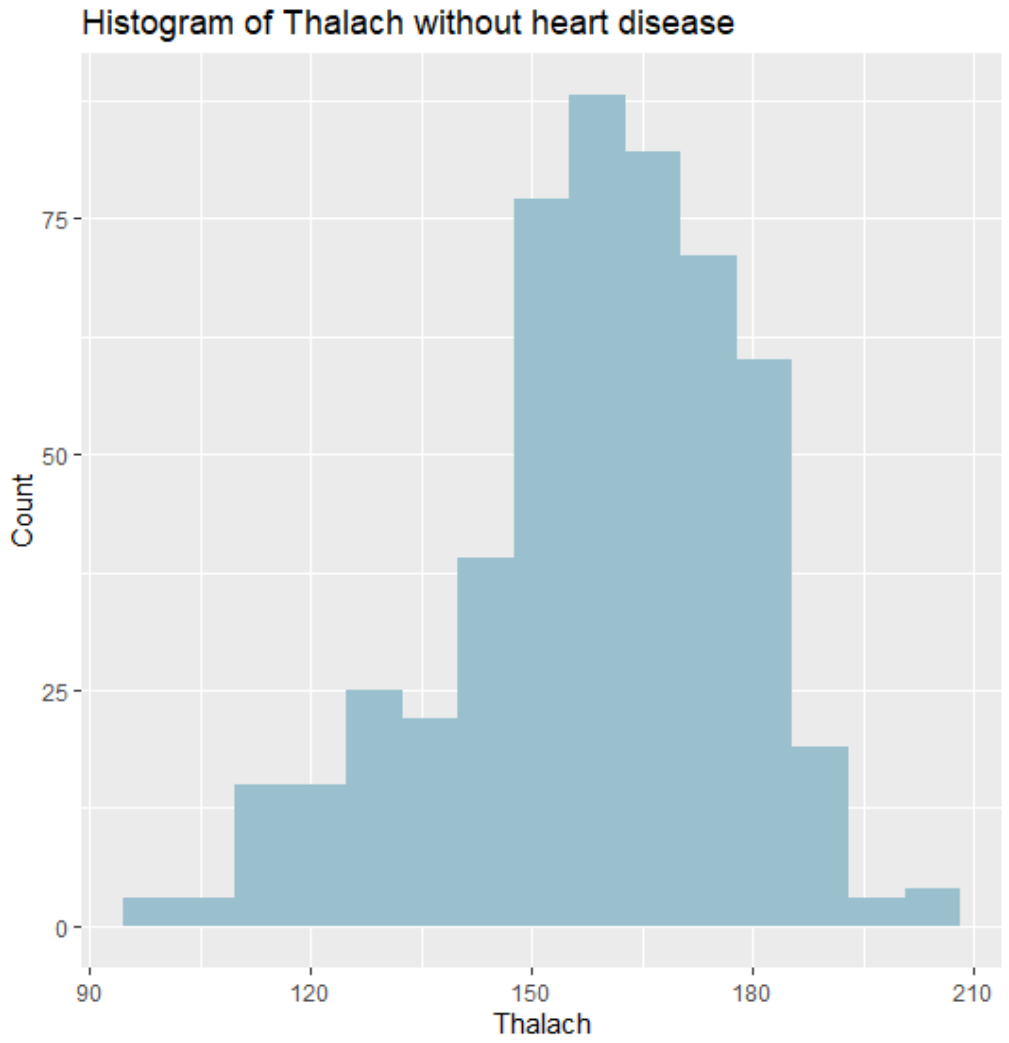


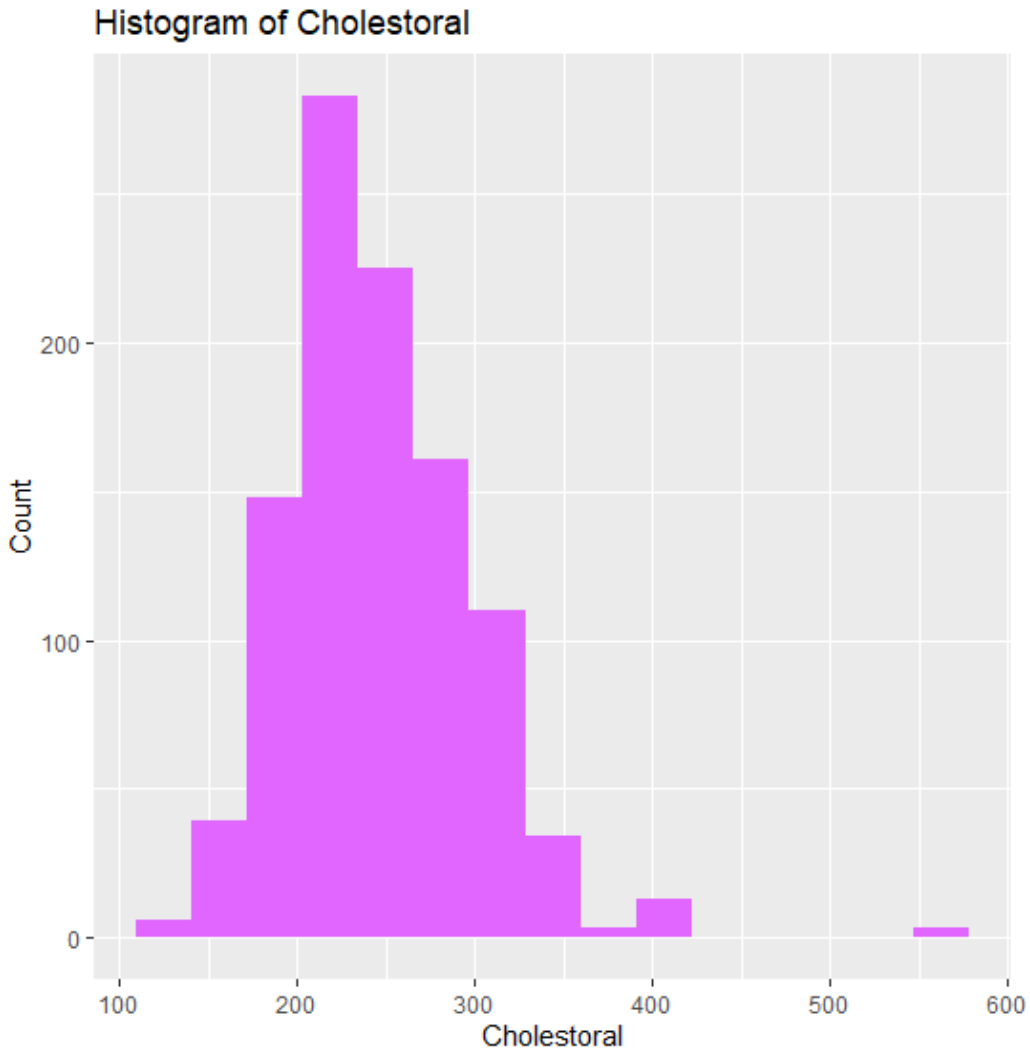


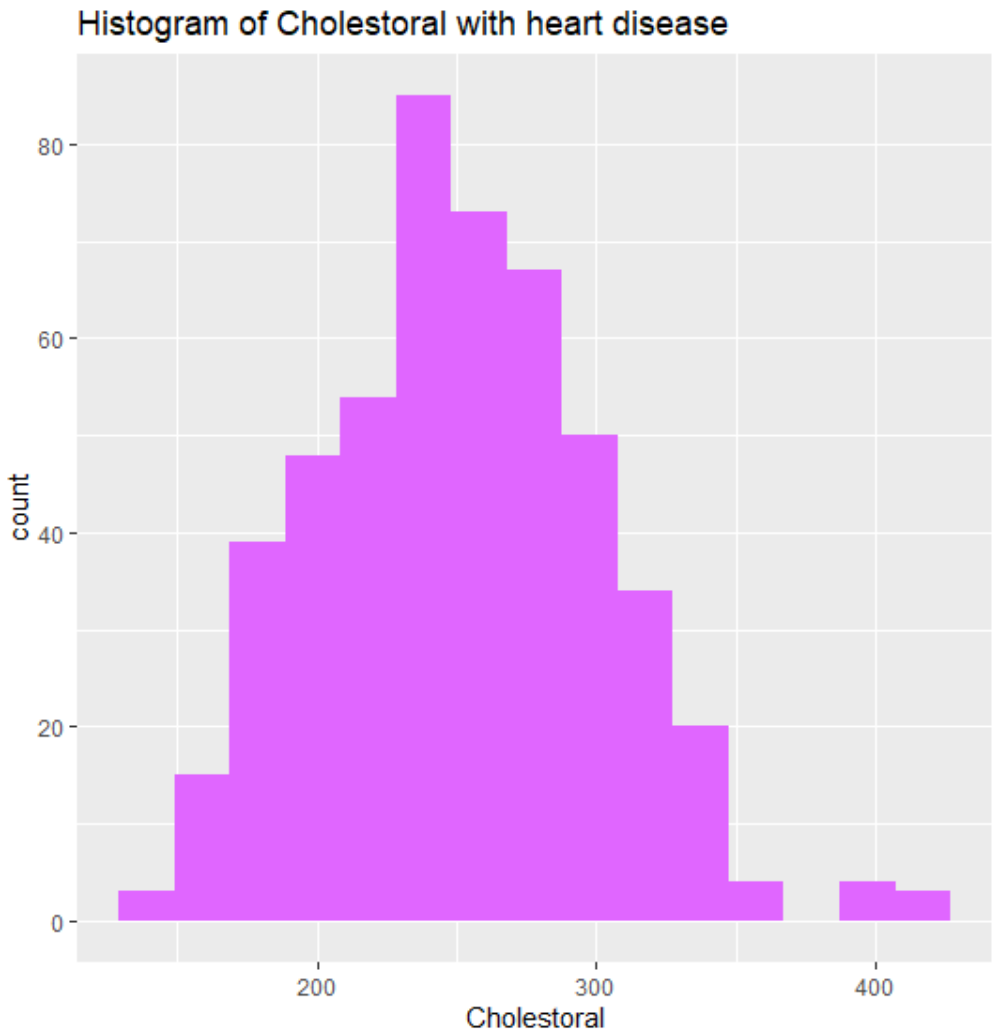

Histogram of Thalach
Count
Thalach
0
25
50
75
100
100
150
200

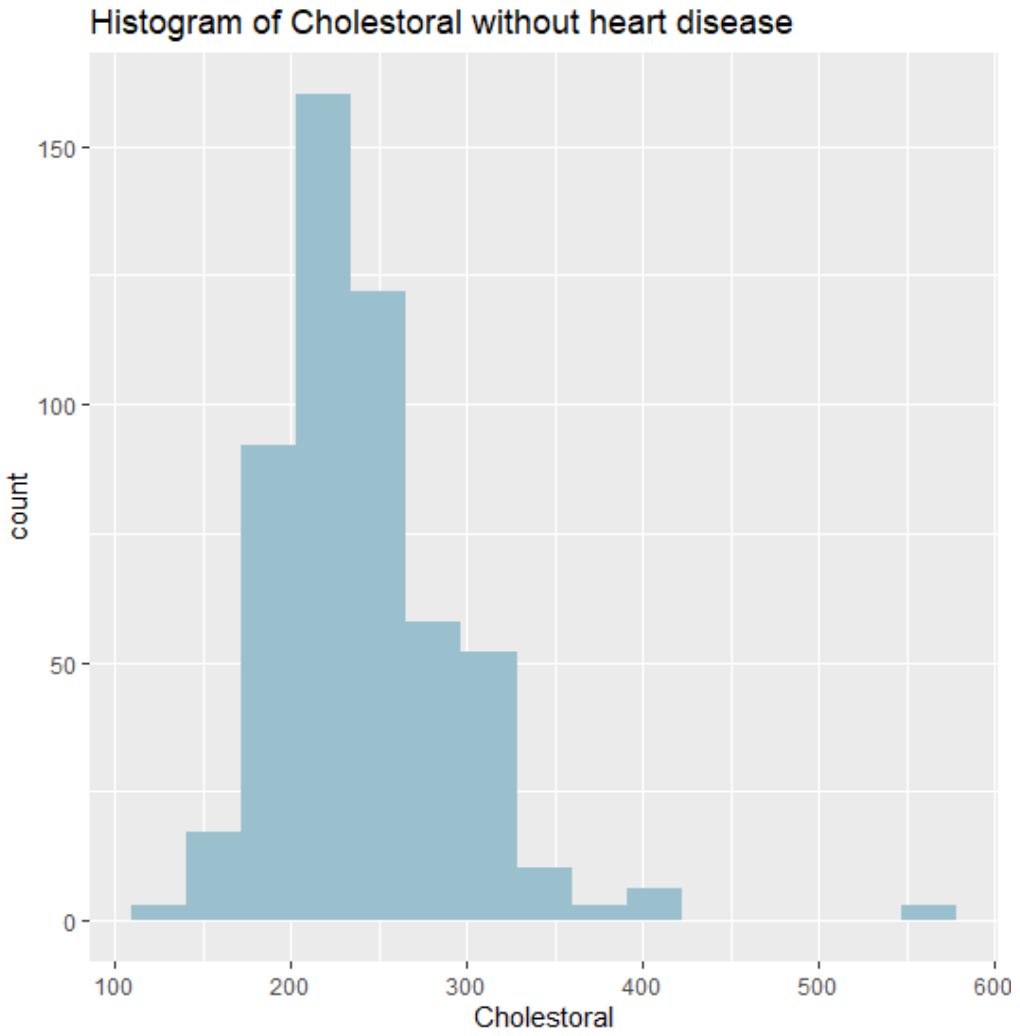

Histogram of Thalach
Density
Thalach
0.000
0.005
0.010
0.015
80
100
120
140
160
180
200


Histogram of Thalach with heart disease
count
100
75
50
25
0
100
150
200
Thalach


Histogram of Thalach without heart disease
Count
75
50
25
0
90
120
150
180
210
Thalach

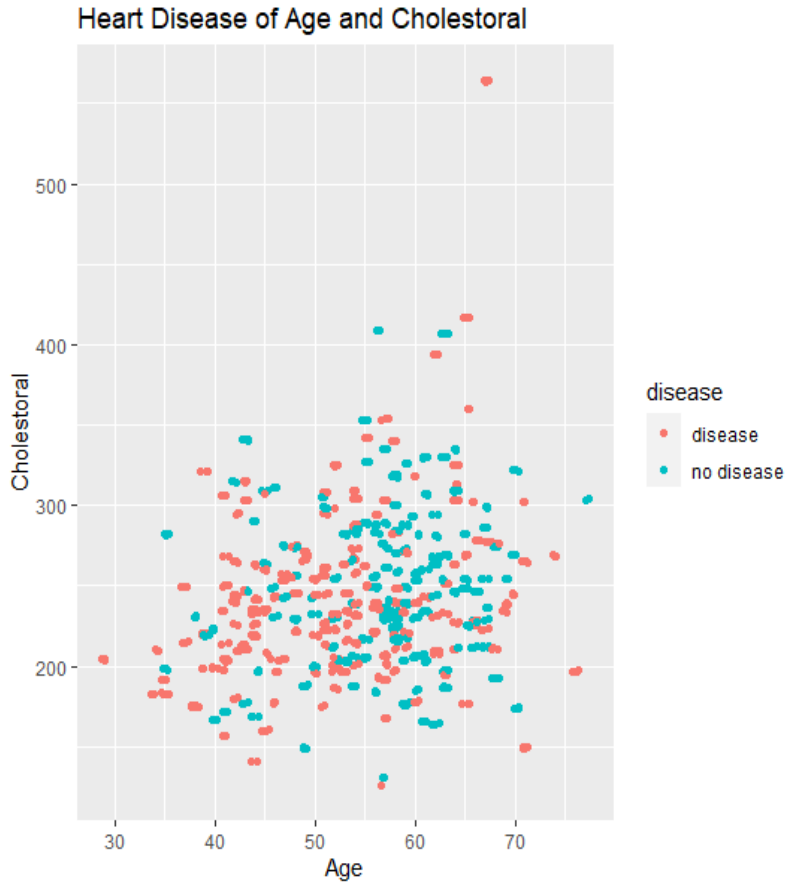

Histogram of Cholestoral
Count
200
100
0
100
200
300
400
500
600
Cholestoral





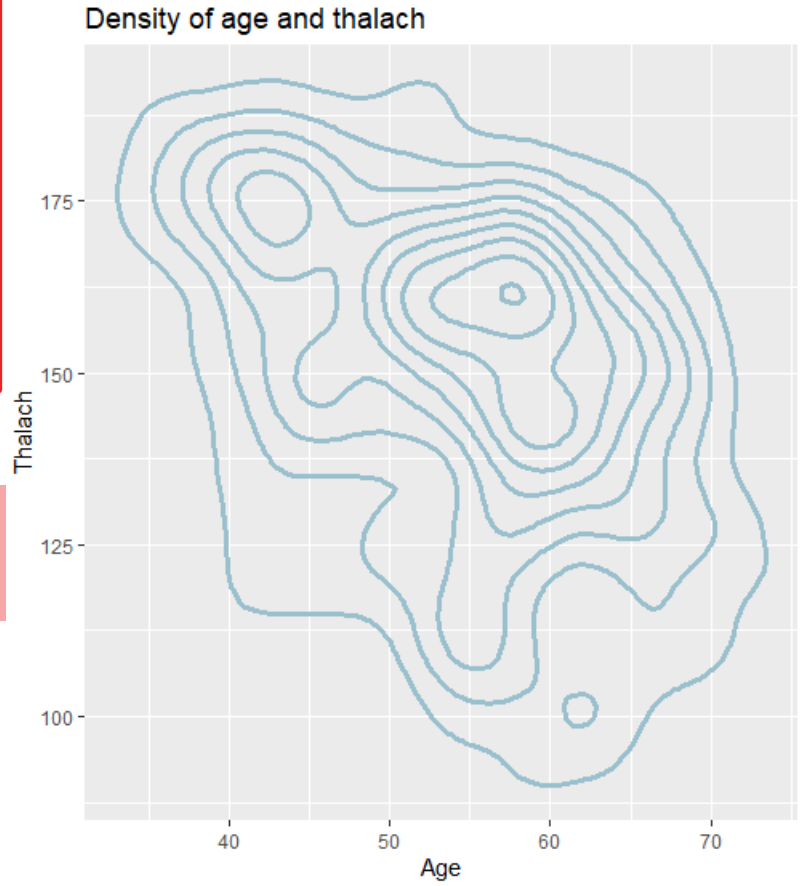
	0	1
Female	238	74
Male	442	271

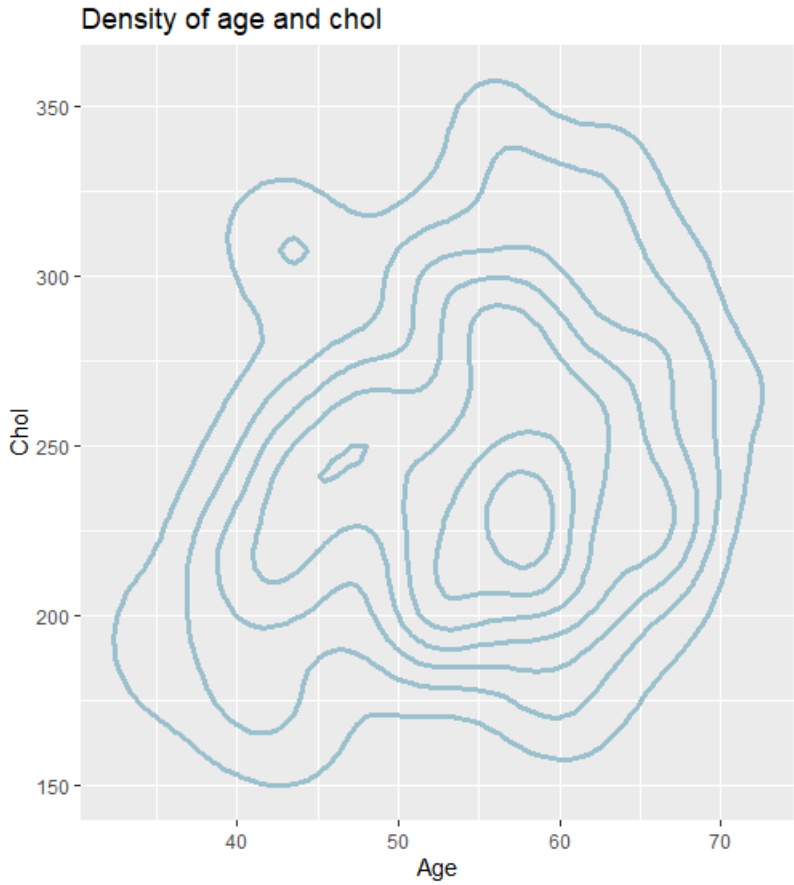
Graph
Scatter



Graph

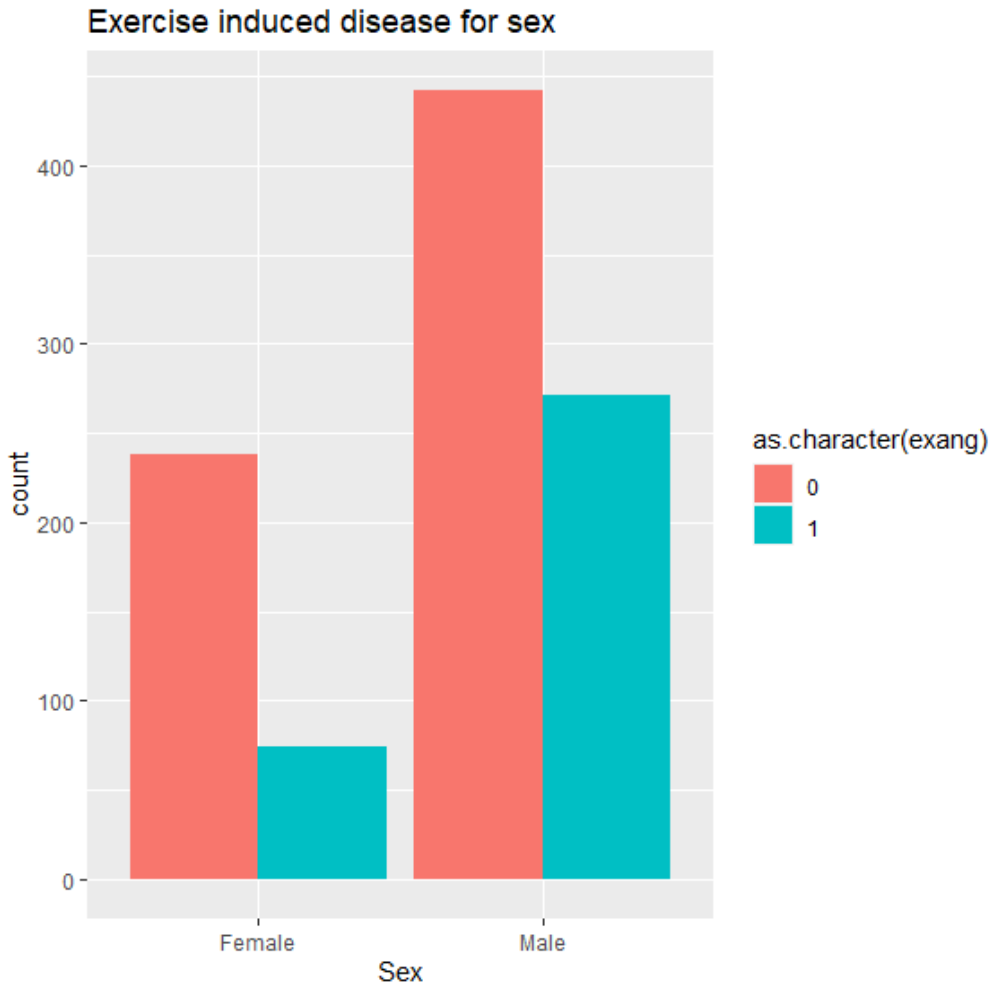
2D density



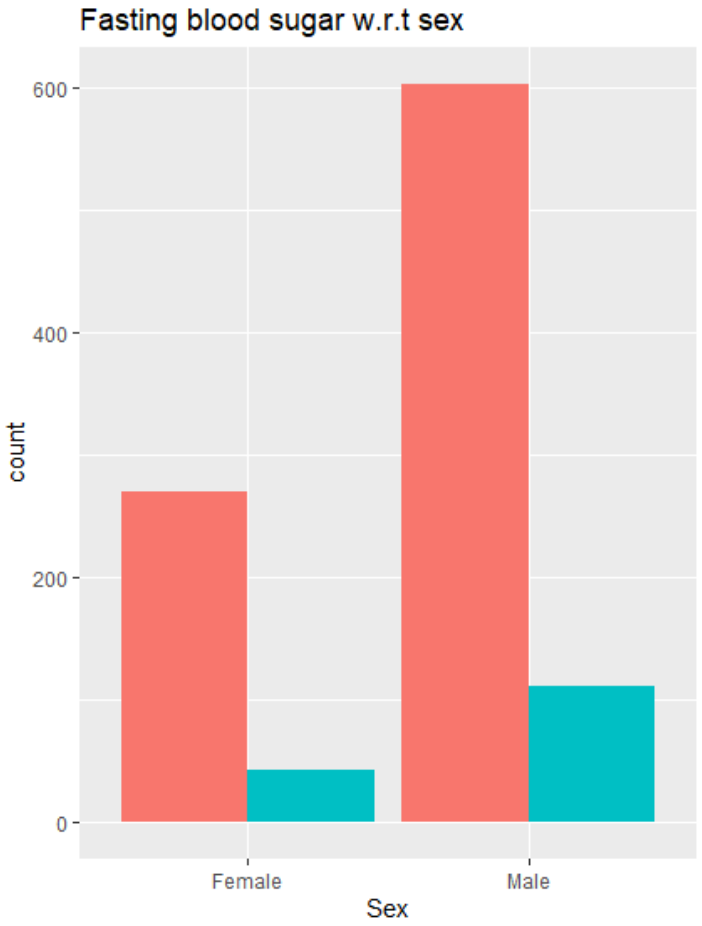


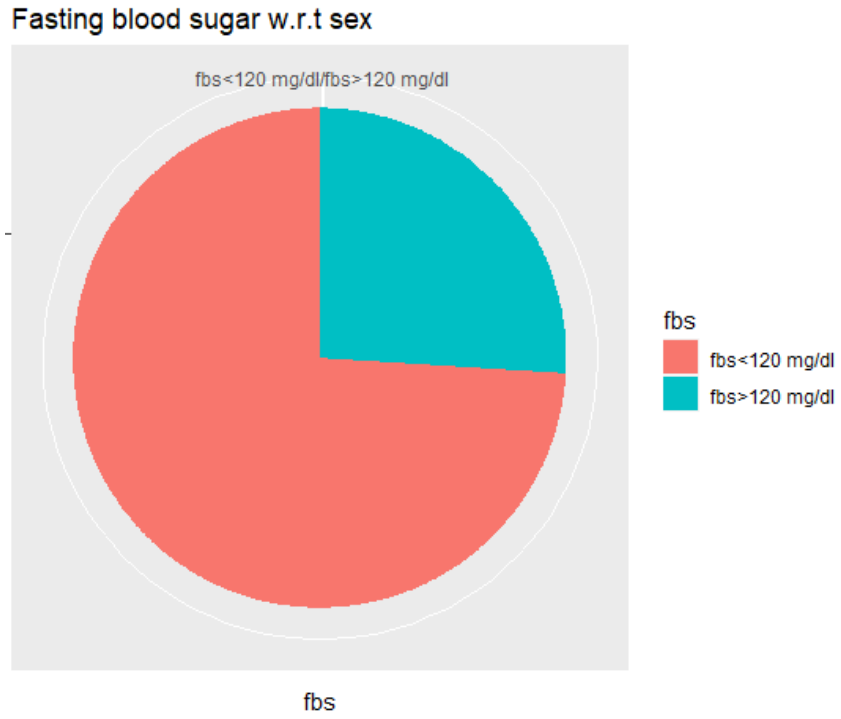
Graph
exang

	0	1
Female	238	74
Male	442	271



Graph
fbs





	0	1
Female	270	42
Male	602	111

Model explanation

Logistic regression

Logistic regression

เป็นเทคนิคการวิเคราะห์สถิติ เชิงคุณภาพ (qualitative statistical techniques)

Binary logistic regression analysis

ใช้กับตัวแปรตามที่แบ่งออกเป็น 2 กลุ่มย่อย โดยมีค่าเป็น 0 กับ 1

Multinomial logistic regression analysis

ตัวแปรตามมีหลายค่ามากกว่า 2 กลุ่ม เช่น การมีมาตรฐานสูง ปานกลาง ต่ำ

Purpose

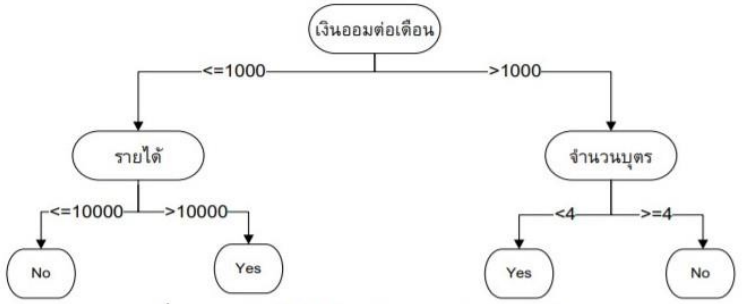
1. เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการโลจิสติคที่สร้างขึ้น
2. ตัวแปรอิสระใดบ้างที่สามารถใช้อธิบาย โอกาสการเกิดเหตุการณ์หรือการไม่เกิดเหตุการณ์ ที่สนใจตามตัวแปรตาม

Decision tree

Decision tree

แบบจำลองทางคณิตศาสตร์ เพื่อการหาทางเลือกที่ดีที่สุด
โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้

มีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning)
สามารถสร้างแบบจำลองการจัดหมวดหมู่ (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่
กำหนดไว้ล่วงหน้า (Training set) ได้โดยอัตโนมัติ และสามารถพยากรณ์กลุ่มของ
รายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย



ภาพที่ 2-4 ตัวอย่างต้นไม้ตัดสินใจ (สุภชัย ประคองศิลป์, 2551)

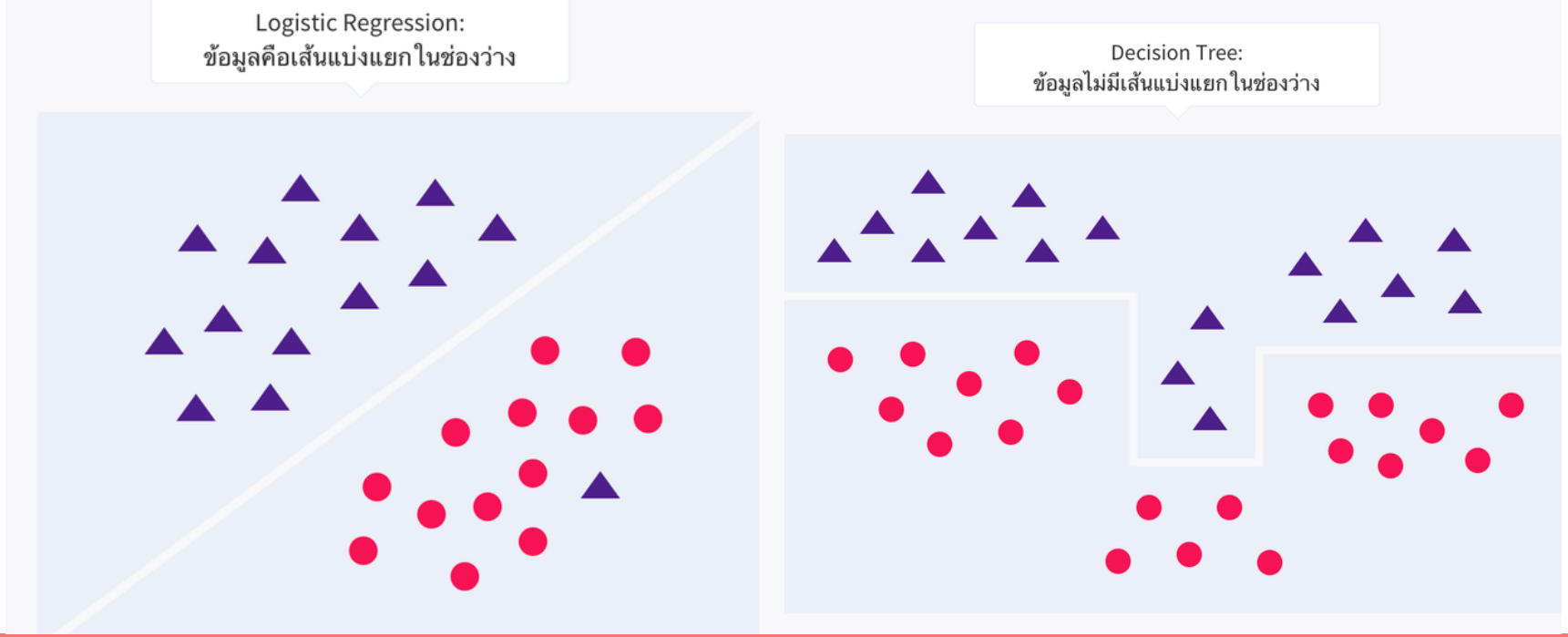
Complement

1. โหนด (Node) คือคุณสมบัติต่างๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทาง
ใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า
โหนดราก(Root Node)

2. กิ่ง (Branch) คือ คุณสมบัติของคุณสมบัติในโหนดที่แตกออกมา โดยจำนวน
ของกิ่งจะเท่ากับคุณสมบัติของโหนด

3. ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล

Logistic regression vs Decision tree



“ เราเลือกใช้ทั้ง logistic regression และ decision tree เพื่อเปรียบเทียบตัว model ว่าตัวไหนทำงานได้ดีกว่ากันกับข้อมูลที่มีอยู่เพื่อการทำนายผลว่าผู้ป่วยเป็นโรคหัวใจหรือไม่จากข้อมูล โดยแต่ละ model ก็จะมีเกณฑ์ที่แตกต่างกันไป ในการสร้าง เราจะอธิบายต่อไปในหัวข้อ Modeling implementation ”

Modeling
implementation

Logistic regression

```
data$target<-factor(ifelse(data$target==1,"yes","no"))
data <- select(data,-disease)

ind <- sample(nrow(data),0.3*nrow(data))
training <- data[-ind,]
testing <- data[ind,]

model1 <- glm(target~.,data=training,family = binomial)
result1 <- predict(model1,testing)
result1
res_1 <- factor(ifelse(result1>0,"yes","no"))
confusionMatrix(res_1,testing$target,mode = "prec_recall",positive = "yes")

model2 <- glm(target~thalach*oldpeak,data=training,family = binomial)
result2 <- predict(model2,testing)
result2
res_2 <- factor(ifelse(result2>0,"yes","no"))
confusionMatrix(res_2,testing$target,mode = "prec_recall",positive = "yes")

model3 <- glm(target~thalach*oldpeak*slope,data=training,family = binomial)
result3 <- predict(model3,testing)
result3
res_3 <- factor(ifelse(result3>0,"yes","no"))
confusionMatrix(res_3,testing$target,mode = "prec_recall",positive = "yes")

model4 <- glm(target~thalach*oldpeak*slope*thal,data=training,family = binomial)
result4 <- predict(model4,testing)
result4
res_4 <- factor(ifelse(result4>0,"yes","no"))
confusionMatrix(res_4,testing$target,mode = "prec_recall",positive = "yes")
```

เมื่อลองใช้ variable หลากหลายค่าในการสร้าง model ขึ้นมา
เพื่อเปรียบเทียบประสิทธิภาพของการทำงาน
Logistic regression model สามารถใช้ glm function
ในการสร้างโมเดลขึ้นมาได้ภายในโปรแกรม R ดังภาพที่แสดง

Modeling

IMPLEMENTATION

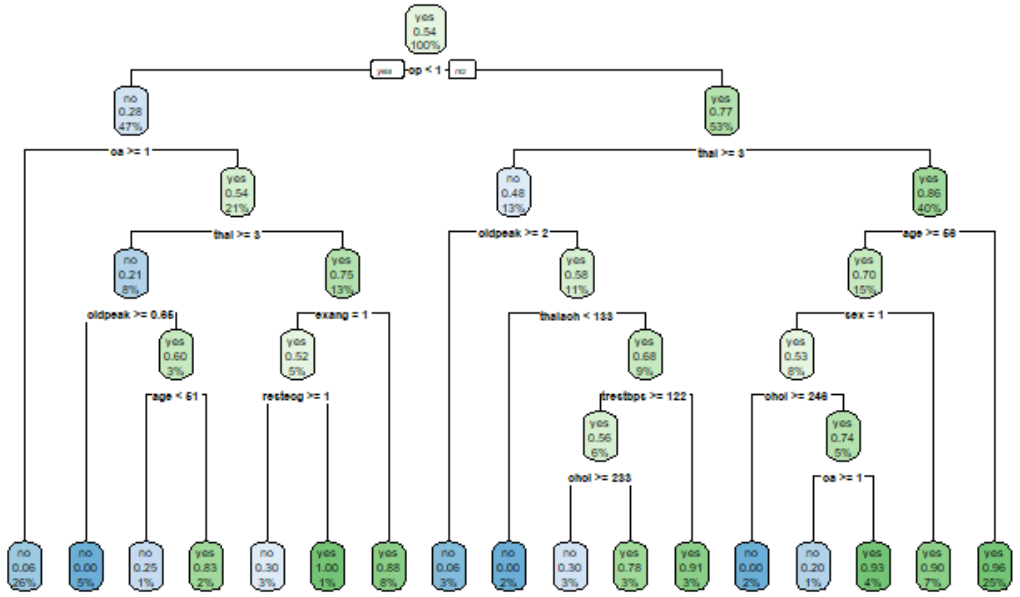
Decision tree

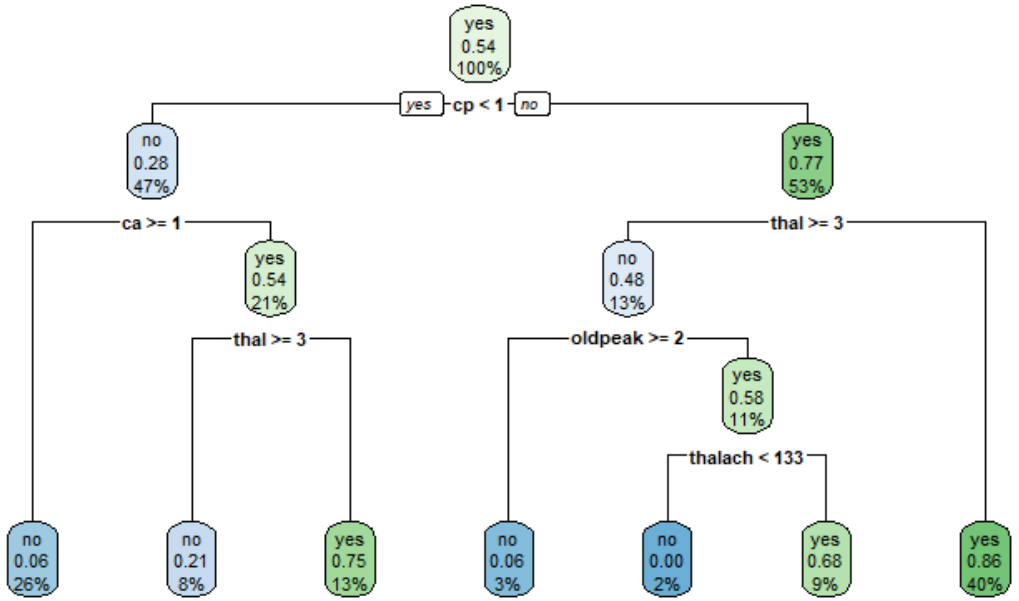
```
library(rpart)
library(rpart.plot)

tree <- rpart(target~.,data=training)
rpart.plot(tree)
tree$variable.importance
res <- predict(tree,testing,type="class")
confusionMatrix(res,testing$target,positive="yes",mode="prec_recall")

control <- trainControl(method = "cv",number = 100)
metric <- "Accuracy"
model_rand <- train(target~.,data=training,method="rpart",metric=metric,trControl=control)
model_rand
#cp = 0.02416918#
tree <- rpart(target~.,data=training,control = rpart.control(cp = 0.02416918))
rpart.plot(tree)
tree$variable.importance
res <- predict(tree,testing,type="class")
confusionMatrix(res,testing$target,positive="yes",mode="prec_recall")
```

ใช้ library rpart ในการสร้าง decision tree ขึ้นมา
โดยเราได้มีการลองใช้ cross validation เพื่อหาค่า cp ในการลด
Complexity ของ tree ออกมาและเปรียบเทียบเพื่อดูประสิทธิภาพในการทำนายผลว่าทำงาน
ได้ดีกว่ากันมากน้อยแค่ไหนและโมเดลใดที่เหมาะสมสำหรับการใช้งานในครั้งนี้มากกว่ากัน

Modeling

Decision tree with cp value





> Cp value สามารถกำหนดเพื่อลด complexity ของ tree ได้ เราได้ค่าของ cp จากการทำ cross validation ออกมาจาก จากนั้นจึงทำการ apply กับ decision tree ของ เพื่อนำ model มาเปรียบเทียบกันว่า model ใดที่ทำงานได้ดีกว่ากันกับข้อมูลที่มีอยู่ของเรา จากการใช้ confusion matrix

Evaluation

Precision : เป็นจริงตามผลที่ทำนายออกมาหรือไม่มากน้อยเพียงใด

Recall : model นี้เมื่อใช้กับ testing data แล้ว คัดกรองคนที่เป็นโรคจริงได้มากน้อยเพียงใด

F1 : ค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall นักวิจัยสร้าง **F1** ขึ้นมาเพื่อเป็น single metric ที่วัดความสามารถของโมเดล (ไม่ต้องเลือกระหว่าง precision, recall เพราะเฉลี่ยให้แล้ว)

Accuracy : มีความแม่นยำในการทำนายผลมากน้อยเพียงใด

Evaluation

เปรียบเทียบ confusion matrix เพื่อดูประสิทธิภาพการทำงานของ model

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  135  13
       yes  33 126

               Accuracy : 0.8502
                 95% CI : (0.8052, 0.8882)
    No Information Rate : 0.5472
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7013

 Mcnemar's Test P-Value : 0.005088

            Precision : 0.7925
               Recall : 0.9065
                   F1 : 0.8456
           Prevalence : 0.4528
       Detection Rate : 0.4104
 Detection Prevalence : 0.5179
    Balanced Accuracy : 0.8550

     'Positive' Class : yes
```

รูปภาพที่ 1 : Confusion matrix of logistic regression model ที่มีประสิทธิภาพสูงสุดคือการใช้ variable ทุกตัว

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  155   9
       yes  13 130

               Accuracy : 0.9283
                 95% CI : (0.8935, 0.9545)
    No Information Rate : 0.5472
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8557

 Mcnemar's Test P-Value : 0.5224

            Precision : 0.9091
               Recall : 0.9353
                   F1 : 0.9220
           Prevalence : 0.4528
       Detection Rate : 0.4235
 Detection Prevalence : 0.4658
    Balanced Accuracy : 0.9289

     'Positive' Class : yes
```

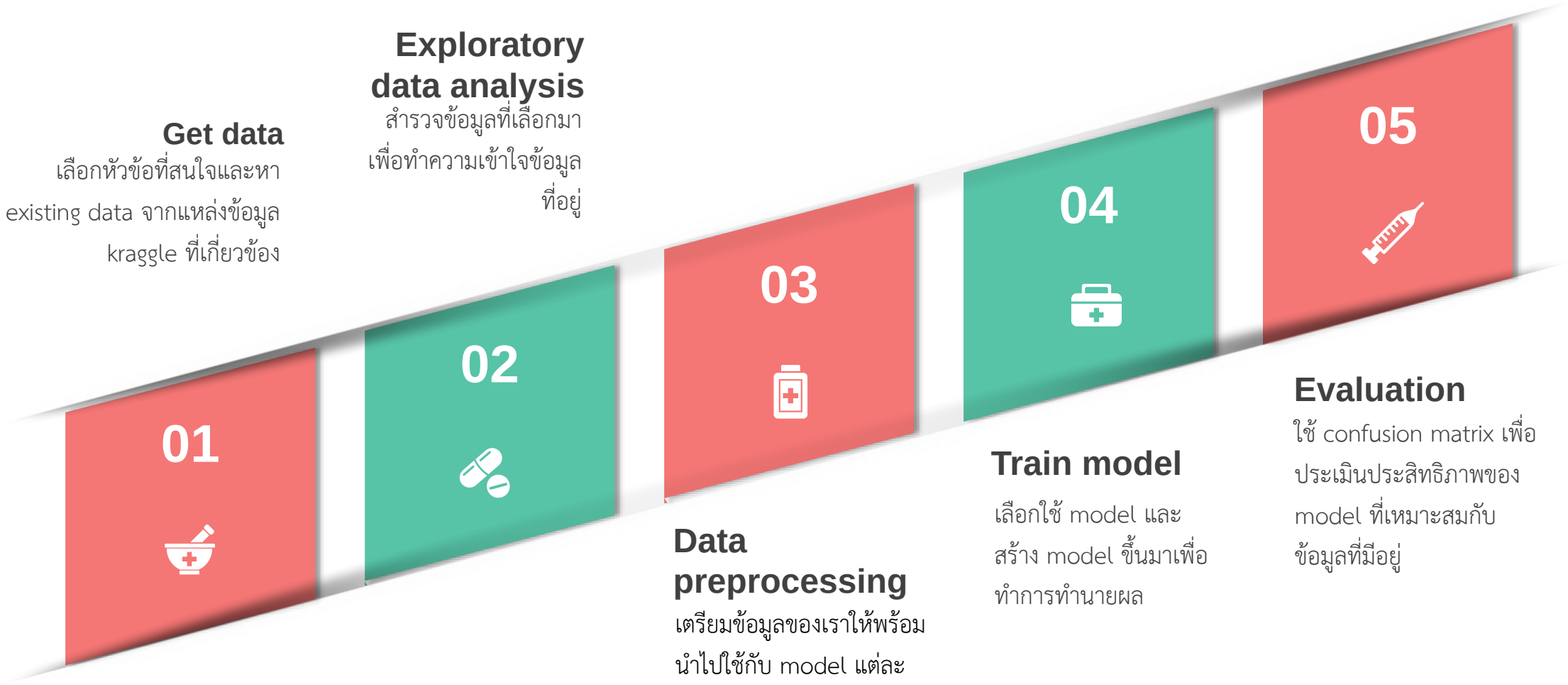
รูปภาพที่ 2 : Confusion matrix of decision tree model

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  130   8
       yes  38 131

               Accuracy : 0.8502
                 95% CI : (0.8052, 0.8882)
    No Information Rate : 0.5472
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7032

 Mcnemar's Test P-Value : 1.904e-05

            Precision : 0.7751
               Recall : 0.9424
                   F1 : 0.8506
           Prevalence : 0.4528
       Detection Rate : 0.4267
 Detection Prevalence : 0.5505
    Balanced Accuracy : 0.8581

     'Positive' Class : yes
```

รูปภาพที่ 3 : Confusion matrix of decision tree model that fixed cp value

Discussion and conclusion

Summary



Get data

เลือกหัวข้อที่สนใจและหา existing data จากแหล่งข้อมูล kraggle ที่เกี่ยวข้อง

Exploratory data analysis

สำรวจข้อมูลที่เลือกมาเพื่อทำความเข้าใจข้อมูลที่อยู่

Data preprocessing

เตรียมข้อมูลของเราให้พร้อมนำไปใช้กับ model แต่ละชนิดของเรา

Train model

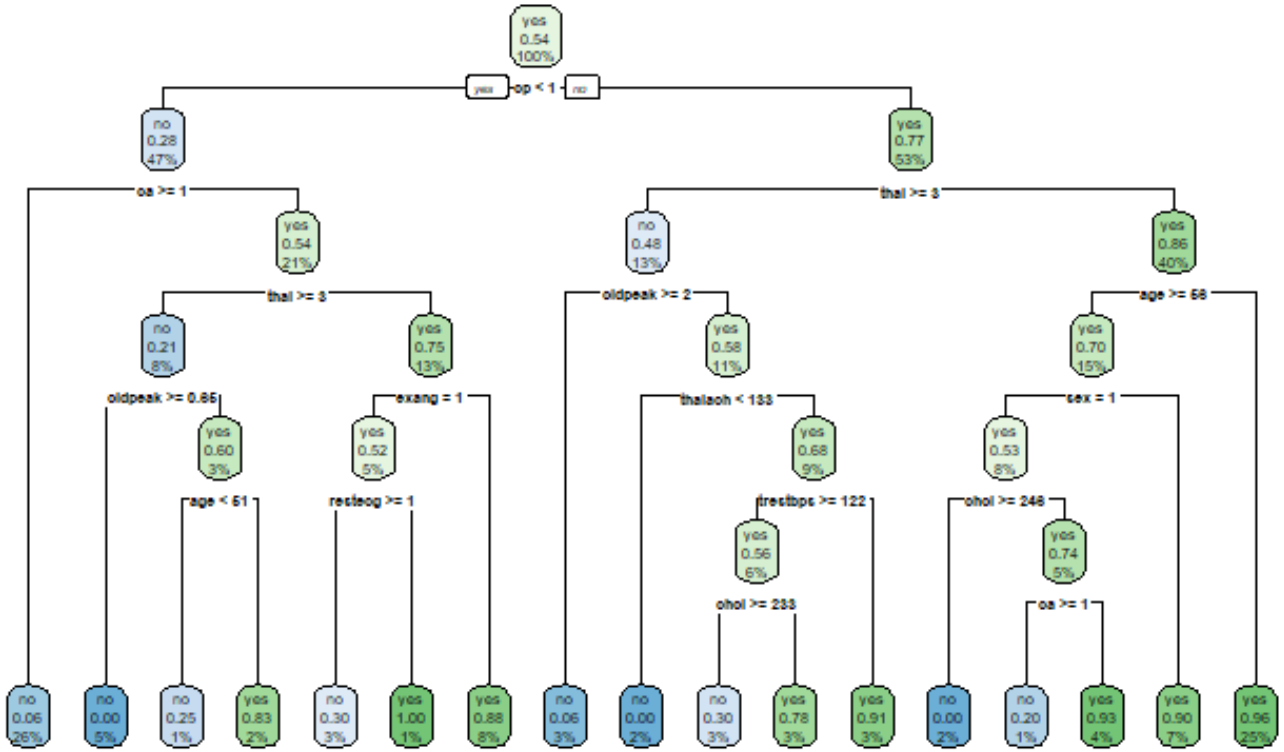
เลือกใช้ model และสร้าง model ขึ้นมาเพื่อทำการทำนายผล

Evaluation

ใช้ confusion matrix เพื่อประเมินประสิทธิภาพของ model ที่เหมาะสมกับข้อมูลที่มีอยู่

Discussion

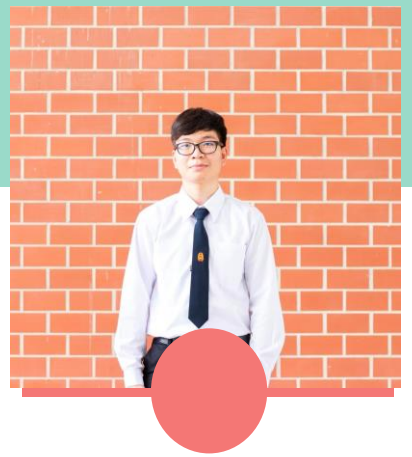
Decision tree model



จากการ train model & evaluate ทำให้เราเห็นว่า decision tree model ทำให้เกิดประสิทธิภาพในการทำนายผลสูงกว่า model ตัวอื่นที่ทำการ train ขึ้นมา สังเกตุได้จาก F1 & Accuracy value

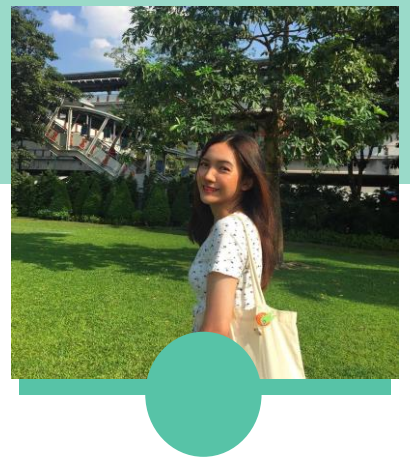
แต่ทั้งนี้ทั้งนั้นตัว model สำหรับการทำนายผลโรคนั้นเป็นเพียงแค่การบ่งบอกถึงความเสี่ยง ไม่สามารถใช้แทนการวินิจฉัยของแพทย์ได้ เพื่อความปลอดภัยผู้ใช้งานควรเข้ารับการตรวจวินิจฉัยที่โรงพยาบาลเพื่อความมั่นใจและความปลอดภัย

Our Team



Silamat Fankaew

61070507221



Warisara Sangsakulrungrueng

61070507220

Napat Lertjarad

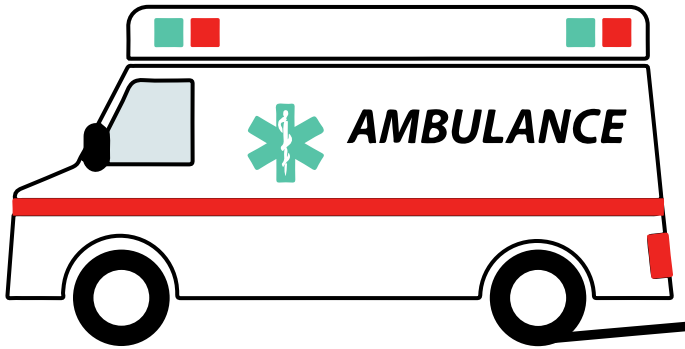
61070507203

Rapeepan Masatitsup

61070507218

Thanawat Onbut

61070207210



Thank You