

# Framingham Heart study



CPE213 : Data Model

King Mongkut's University of Technology Thonburi

# Outline

---

**Part 1:** Introduction to the Problem

---

**Part 2:** Analytic objective

---

**Part 3:** Data description

---

**Part 4:** Exploratory data analysis (EDA)

---

**Part 5:** Data Preprocessing

---

**Part 6:** Train Model

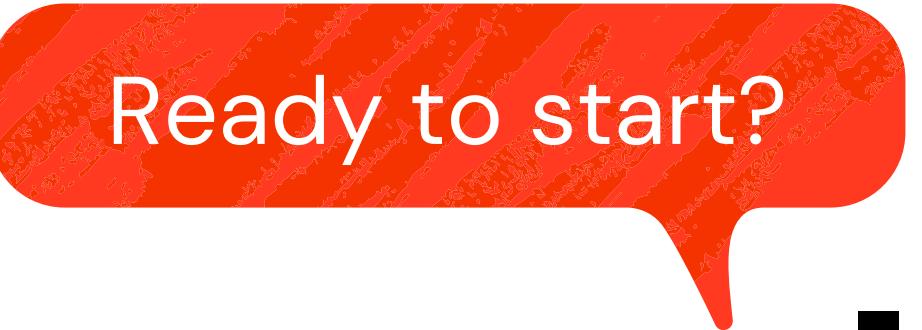
---

**Part 7:** Evaluation

---

**Part 8:** Discussion & Conclusion

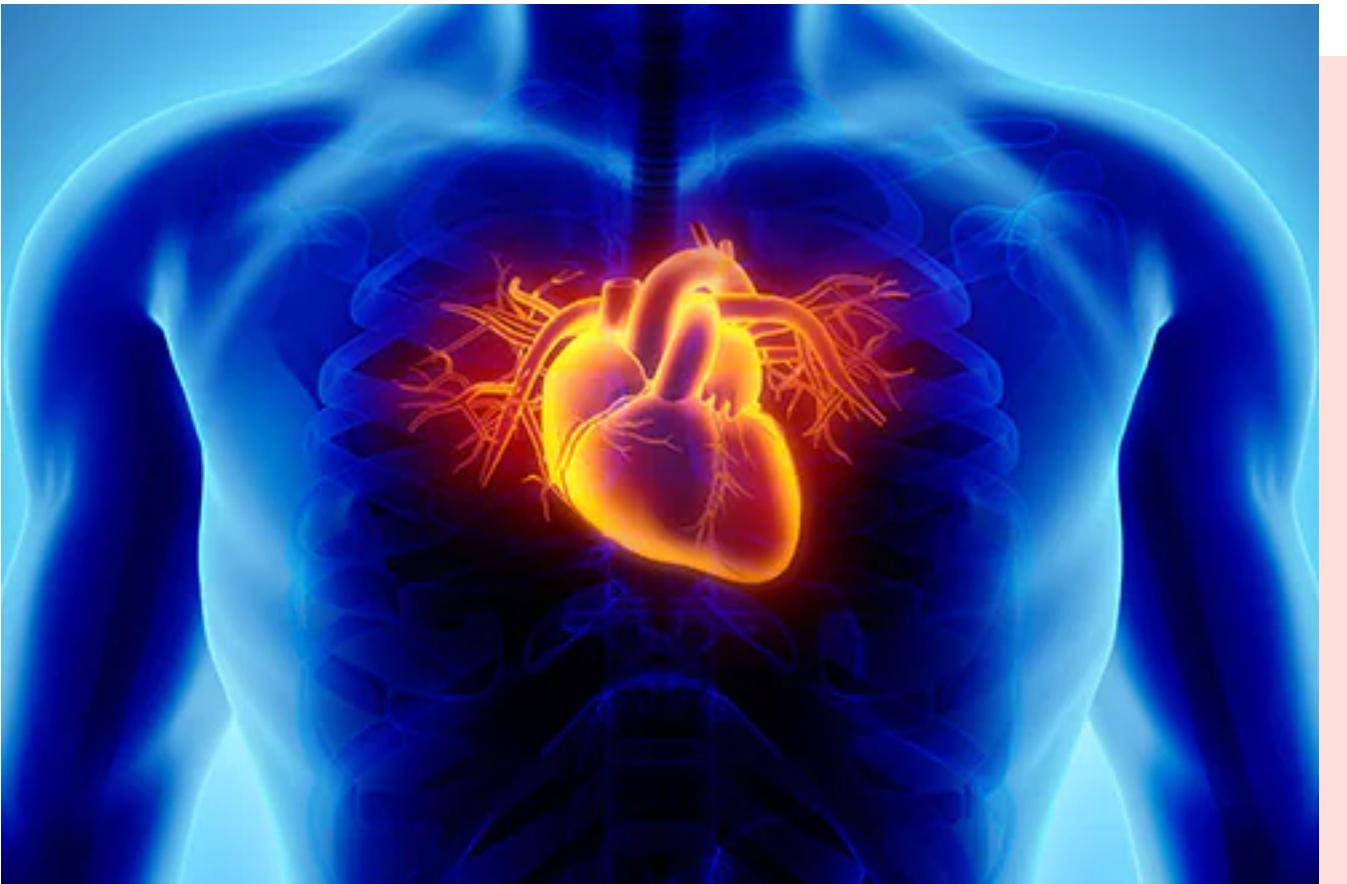
---



Ready to start?

# **Introduction to the Problem**

# Introduction to the Problem



## โรคหัวใจและหลอดเลือด

จำนวนคนเสียชีวิตจากโรคนี้

**54,530 คน**

\*ในแต่ละปีและคาดว่าจะมีแนวโน้มเพิ่มขึ้นทุกปี\*

ปัจจัยเสี่ยงที่เกี่ยวกับโรคนี้

**ปัจจัยเสี่ยงที่ปรับเปลี่ยนไม่ได้ / ได้**

\*ปัจจัยที่เปลี่ยนไม่ได้ เช่น ข้อมูลประวัติส่วนตัว ปัจจัยที่เปลี่ยนได้ เช่น ข้อมูลสุขภาพ และการสูบบุหรี่\*

3 ใน 4 ของการเสียชีวิตจากโรคนี้

**โรคหลอดเลือดสมอง  
และ โรคหัวใจขาดเลือด**

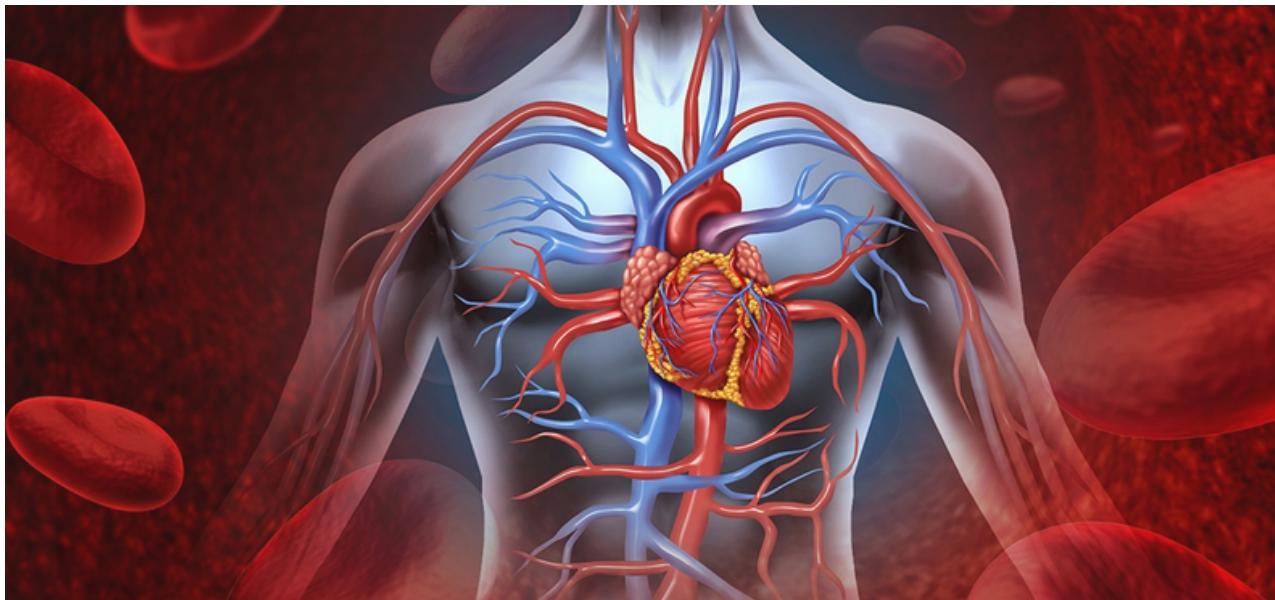
โรคอื่น ๆ ที่มีผลต่อการเกิดโรคนี้

**โรคไตเรื้อรัง โรคหัวใจเต้นผิดจังหวะ ชนิดโรคข้ออักเสบรูมาตอยด์ และ ภาวะหัวใจห้องล่างช้ายโต**

Part 2

# Analytic objective

# Analytic objective



เพื่อ นำข้อมูลปัจจัยเสี่ยง ของโรคหัวใจ และ<sup>ก</sup>  
หลอดเลือดมา หาความเสี่ยง ในการเกิดโรคนี้  
ภายในระยะเวลา 10 ปี



Part 3

# Data description

# Data description

Dataset เป็นข้อมูลที่เปิดเผยแพร่บน Kaggle  
เป็นข้อมูลที่ใช้ในการศึกษา ระบบหัวใจ และ<sup>1</sup>  
หลอดเลือดของผู้อยู่อาศัยในเมือง  
Framingham, Massachusetts  
โดยมีข้อมูลกว่า 4,000 records

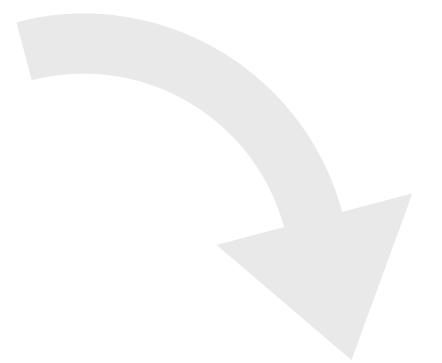


## Framingham Heart study dataset

Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.

[k kaggledatasets](#)

\*Ref Link in project\*



## Ten Year CHD

ในระยะเวลา 10 ปี กี่จะเสี่ยง  
เป็นโรคหลอดเลือดหัวใจ

1 = ใช่,  
0 = ไม่ใช่

\*target\*

# Data description



ในชุดข้อมูลนี้มีตัวแปรกั้งหนด

## 15 attributes

### Male

เพศหญิงหรือชาย

1 = ชาย,  
0 = หญิง

### Age

อายุ

### Education

ระดับการศึกษา

1 = Some High School,  
2 = High School or GED,  
3 = Some College or Vocational School,  
4 = college

### Current Smoker

สูบบุหรี่หรือไม่

1 = สูบบุหรี่,  
0 = ไม่สูบบุหรี่

### Cigs Per Day

จำนวนการสูบบุหรี่ต่อวัน

### BP Meds

ใช้ยาลดความดัน  
เลือดหรือไม่

1 = ใช้ยาลดความดันเลือด,  
0 = ไม่ใช้ยาลดความดันเลือด

### Prevalent Stroke

เป็นโรคหลอดเลือดใน  
สมองหรือไม่

1 = เป็น,  
0 = ไม่เป็น

### Prevalent Hyp

เป็นโรคโรคความดัน  
โลหิตสูงหรือไม่

1 = เป็น,  
0 = ไม่เป็น

### Diabetes

เป็นโรคเบาหวาน  
หรือไม่

1 = เป็น,  
0 = ไม่เป็น

### Tot Chol

ระดับคอเลสเตอรอลกั้งหนด  
(มิลลิกรัม/เดซิลิตร)

### Sys BP

ความดันโลหิตช่วงบน

### Dia BP

ความดันโลหิตช่วงล่าง

### BMI

ดัชนีมวลกาย

### Heart Rate

อัตราการเต้นของหัวใจ

### Glucose

ระดับน้ำตาล



Part 4

# Exploratory Data Analysis (EDA)

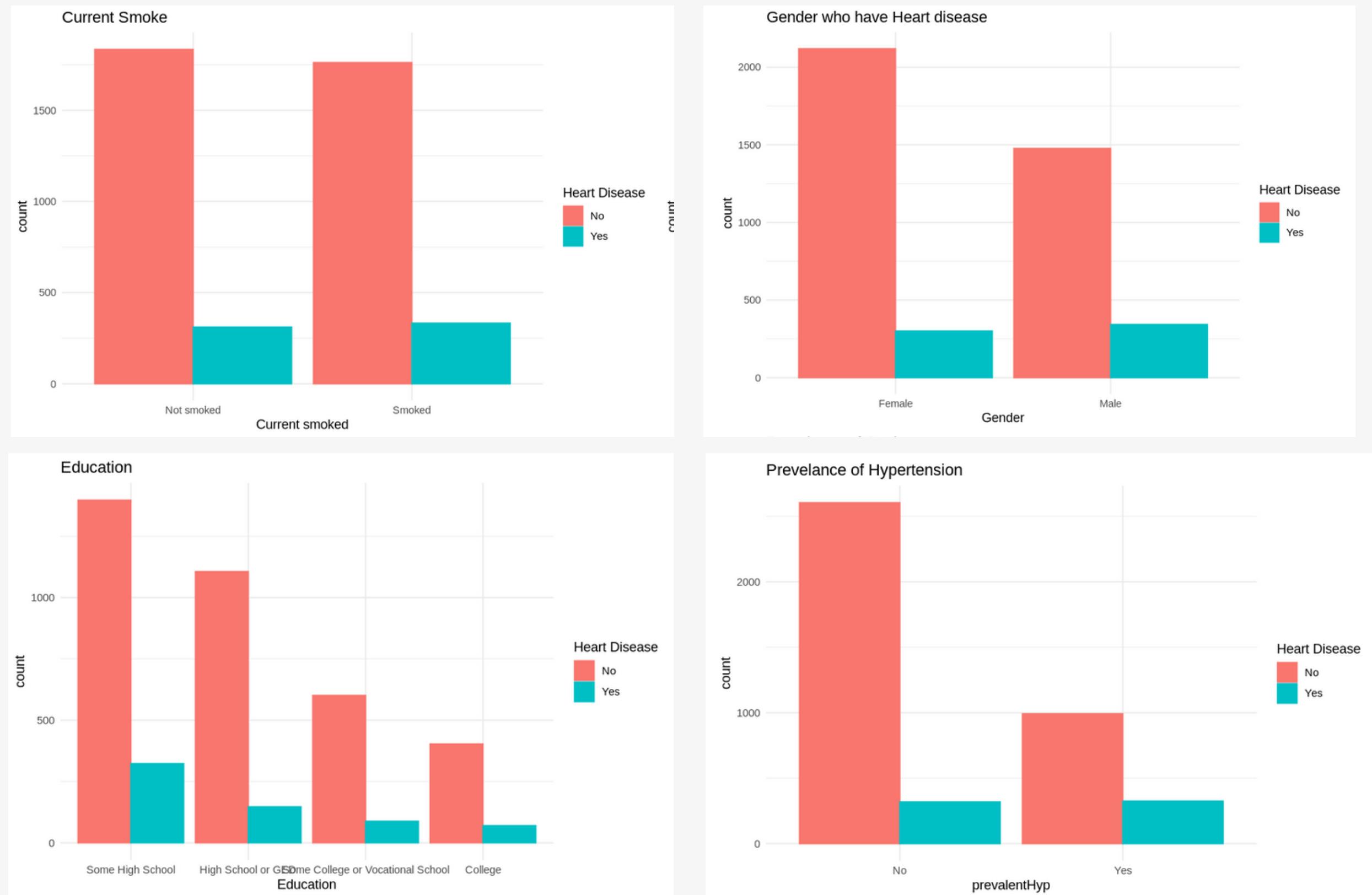
# ข้อมูลใบตารางเบื้องต้น

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int>	<int>	<int>
1	1	39	4	0	0	0	0	0	0	195	106.0	70	26.97	80	77	0
2	0	46	2	0	0	0	0	0	0	250	121.0	81	28.73	95	76	0
3	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
4	0	61	3	1	30	0	0	1	0	225	150.0	95	28.58	65	103	1
5	0	46	3	1	23	0	0	0	0	285	130.0	84	23.10	85	85	0
6	0	43	2	0	0	0	0	1	0	228	180.0	110	30.30	77	99	0

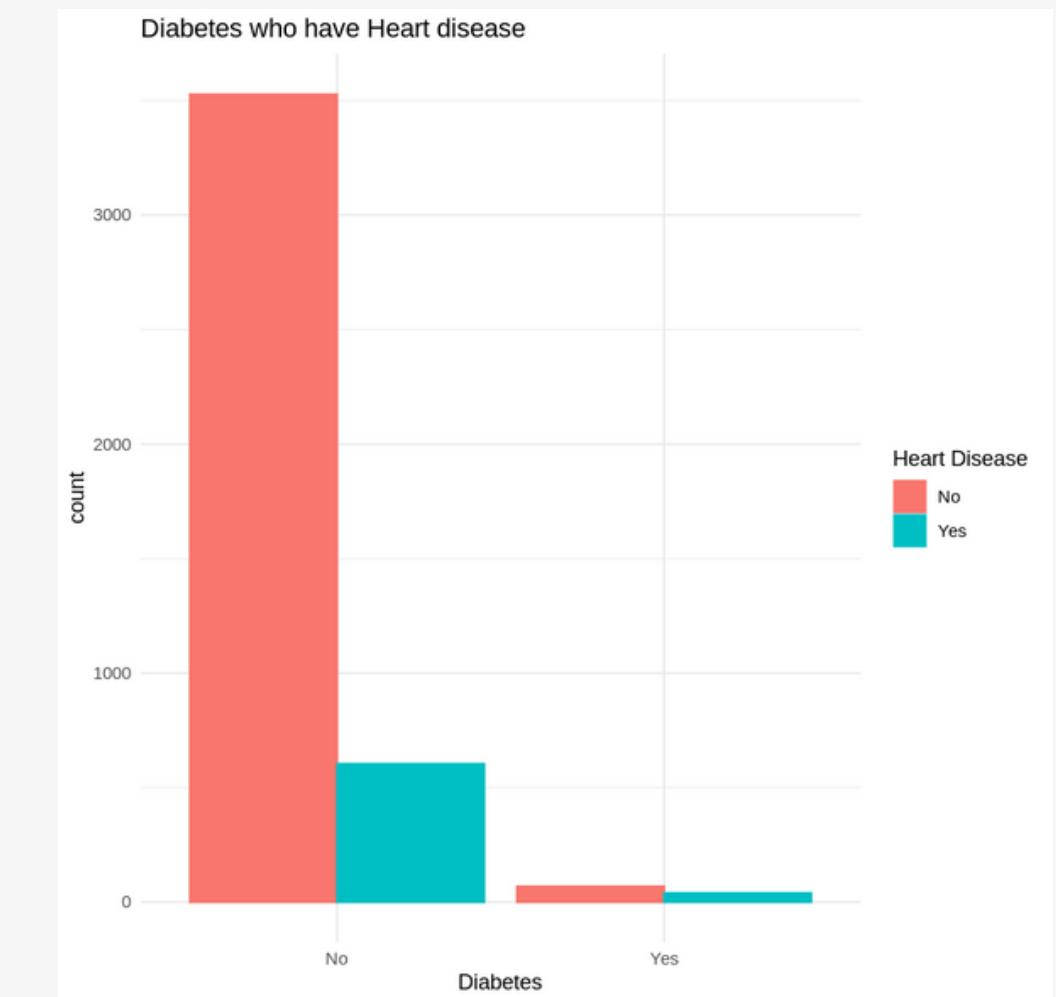
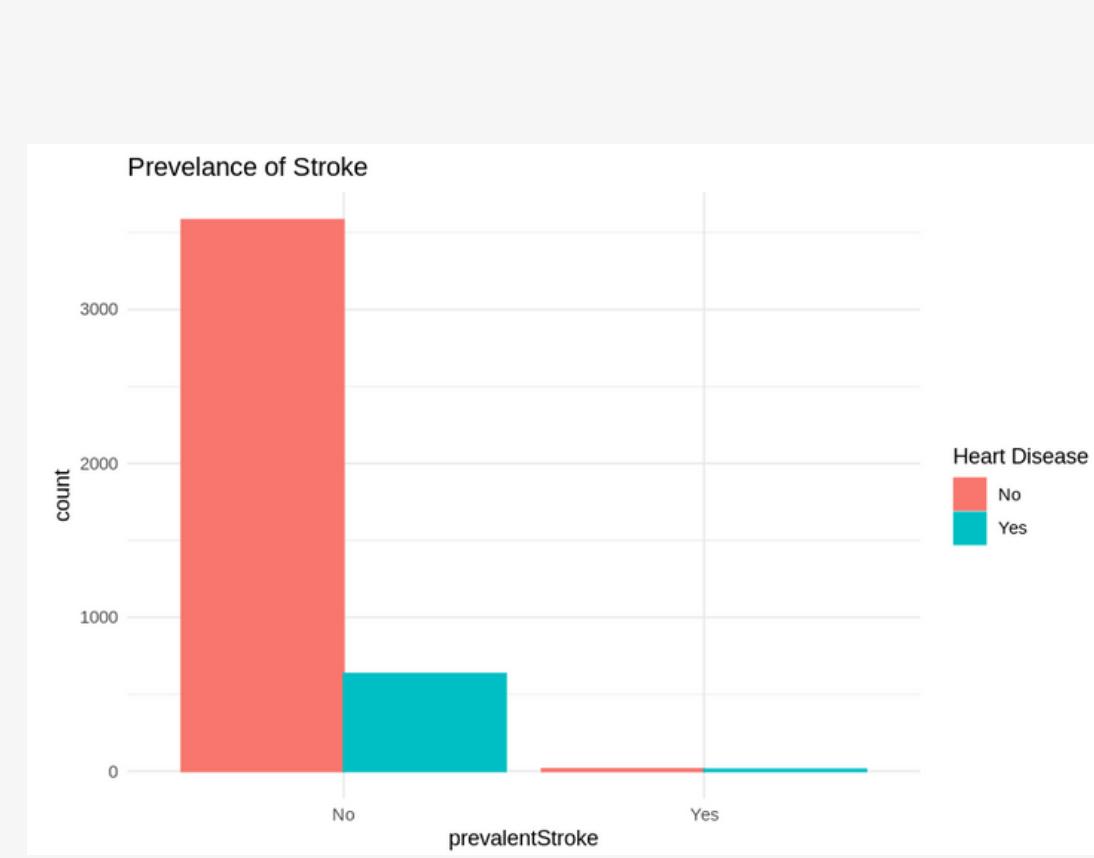
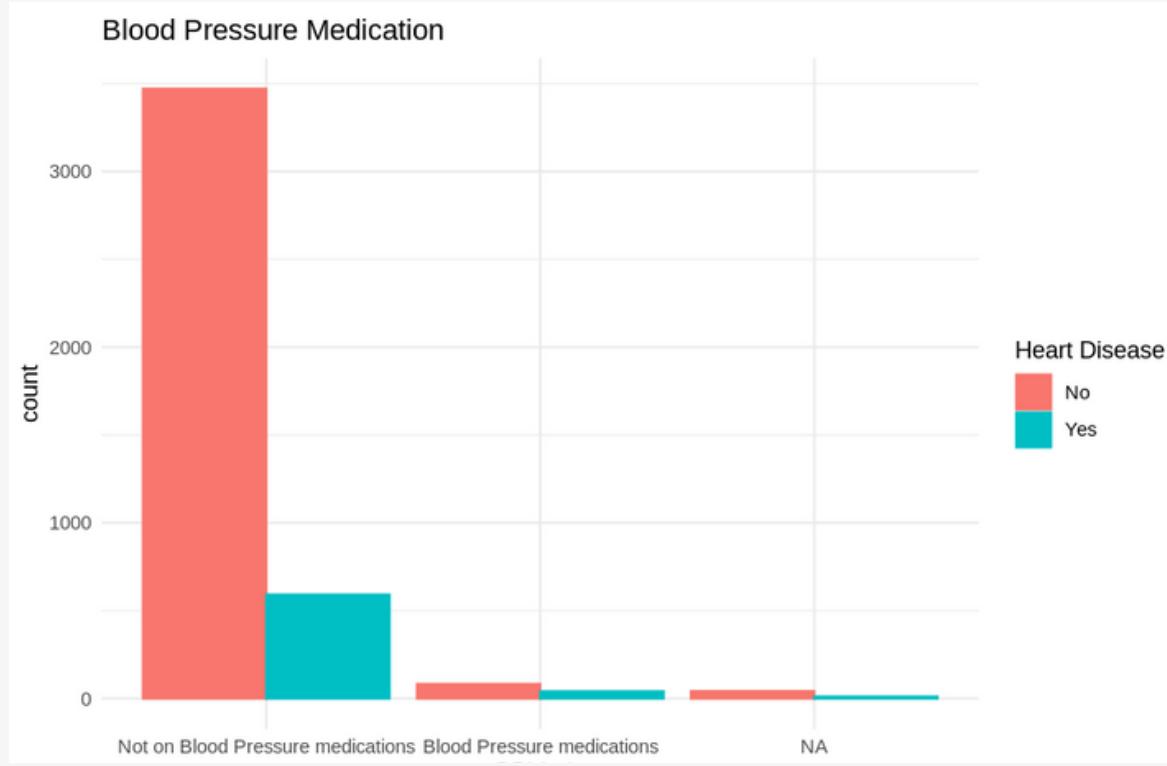
# โครงสร้างของ ข้อมูลเบื้องต้น

```
'data.frame': 4240 obs. of 16 variables:  
 $ male          : int 1 0 1 0 0 0 0 0 1 1 ...  
 $ age           : int 39 46 48 61 46 43 63 45 52 43 ...  
 $ education     : int 4 2 1 3 3 2 1 2 1 1 ...  
 $ currentSmoker : int 0 0 1 1 1 0 0 1 0 1 ...  
 $ cigsPerDay    : int 0 0 20 30 23 0 0 20 0 30 ...  
 $ BPMeds        : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ prevalentStroke: int 0 0 0 0 0 0 0 0 0 0 ...  
 $ prevalentHyp  : int 0 0 0 1 0 1 0 0 1 1 ...  
 $ diabetes       : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ totChol        : int 195 250 245 225 285 228 205 313 260 225 ...  
 $ sysBP          : num 106 121 128 150 130 ...  
 $ diaBP          : num 70 81 80 95 84 110 71 71 89 107 ...  
 $ BMI            : num 27 28.7 25.3 28.6 23.1 ...  
 $ heartRate      : int 80 95 75 65 85 77 60 79 76 93 ...  
 $ glucose         : int 77 76 70 103 85 99 85 78 79 88 ...  
 $ TenYearCHD     : int 0 0 0 1 0 0 1 0 0 0 ...
```

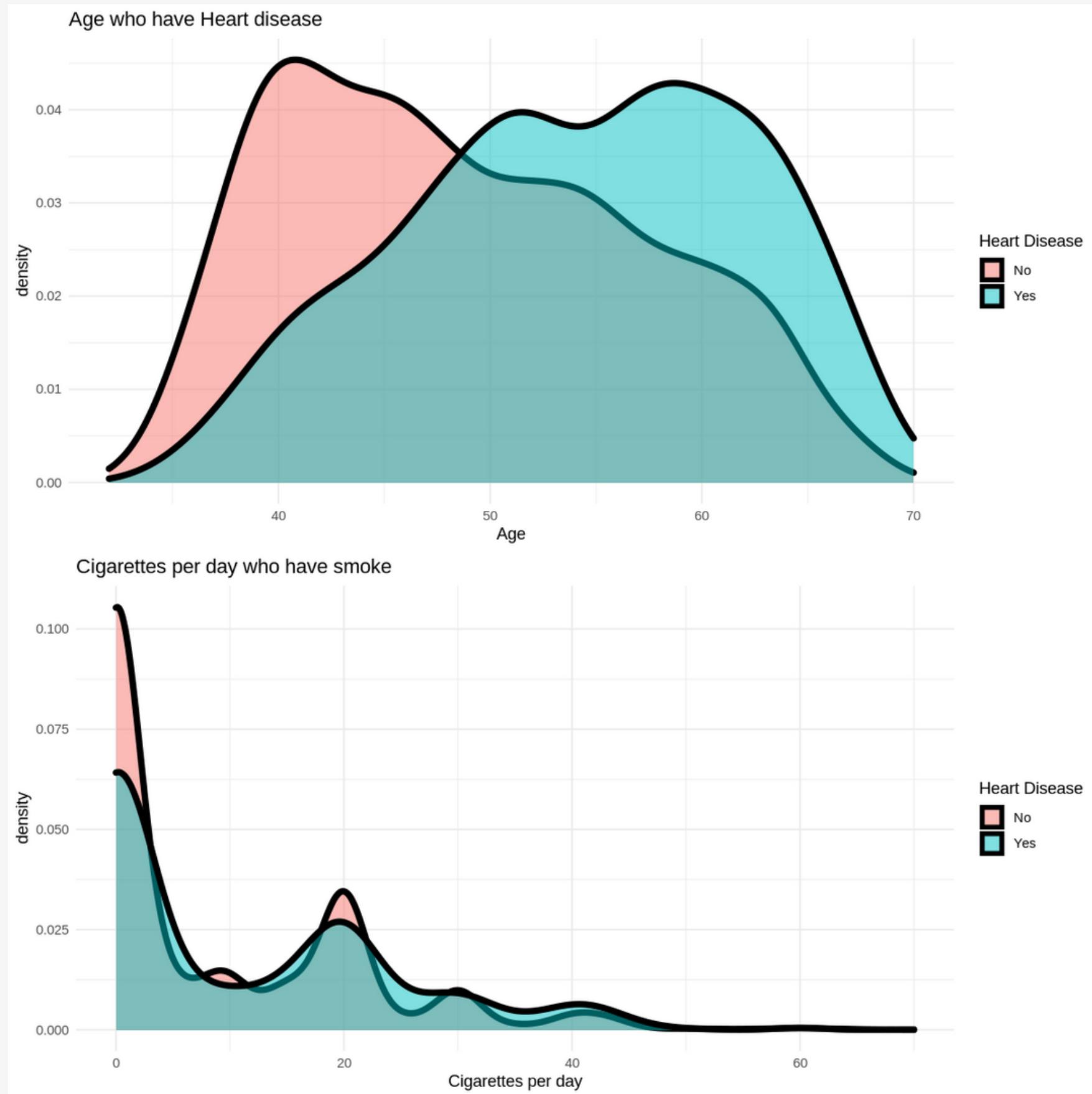
# Bar graph



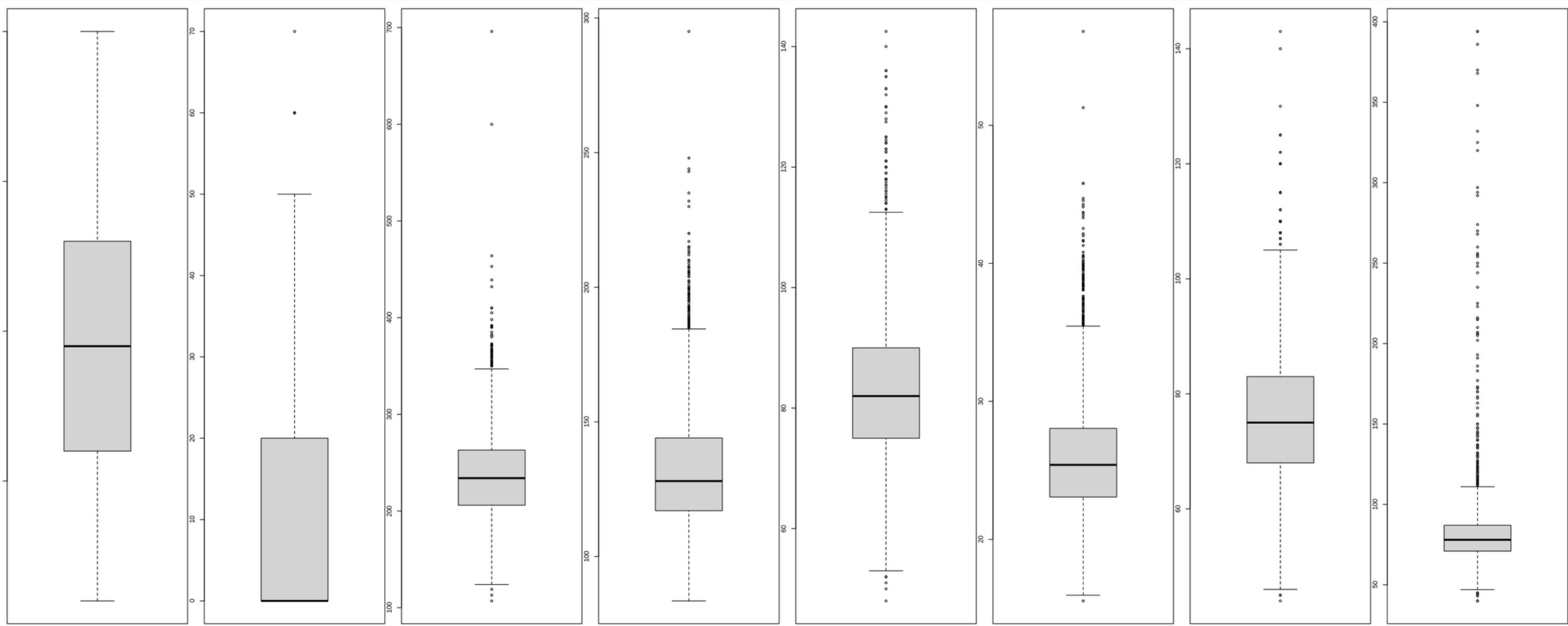
# Bar graph



# Line graph



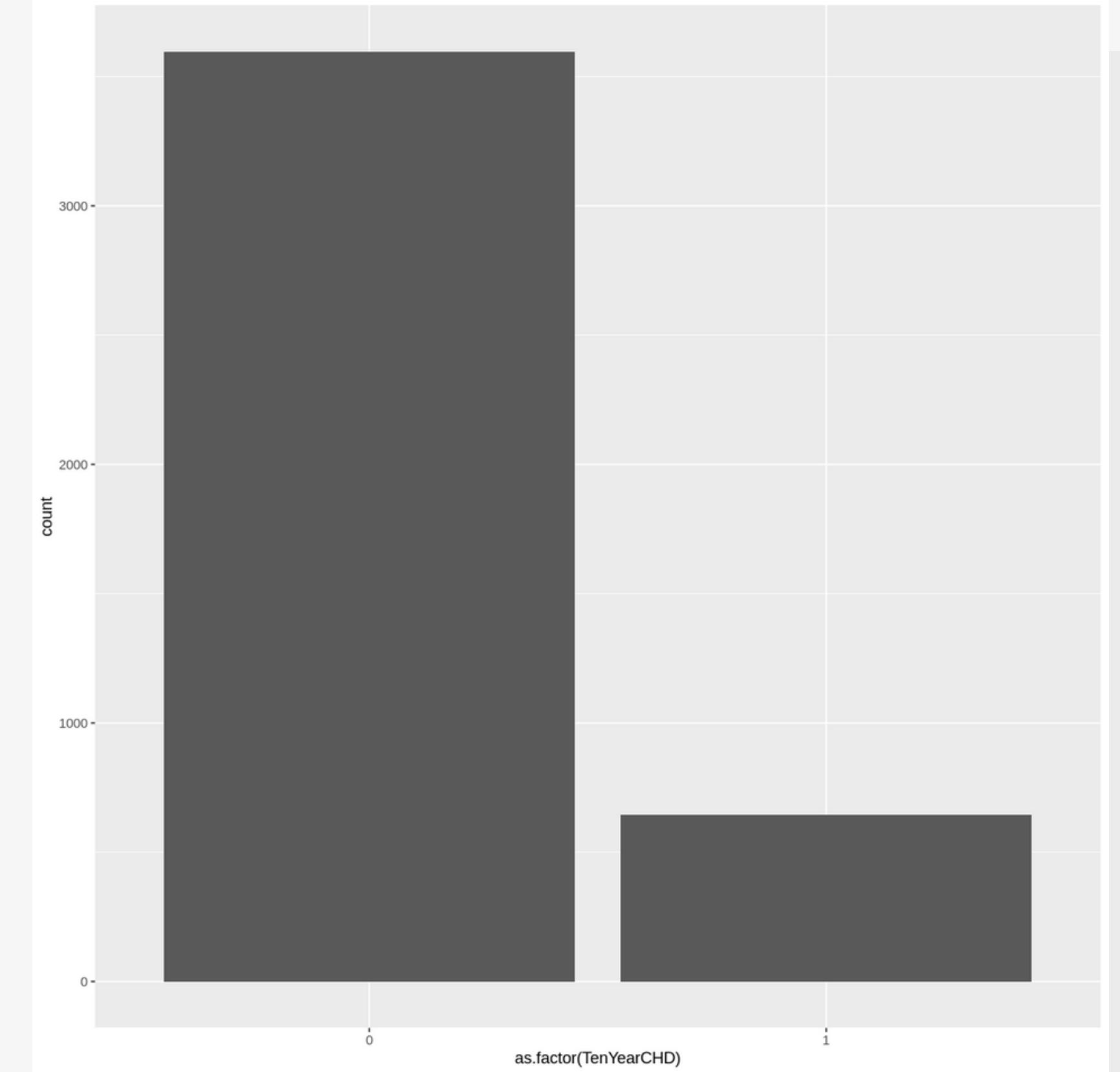
# Boxplot graph



# ຈຳນວນຂອງຂໍ້ມູນລວກ (NA)

male	age	education	currentSmoker	cigsPerDay
0	0	105	0	29
BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol
53	0	0	0	50
sysBP	diaBP	BMI	heartRate	glucose
0	0	19	1	388
TenYearCHD				
0				

# Data Imbalance

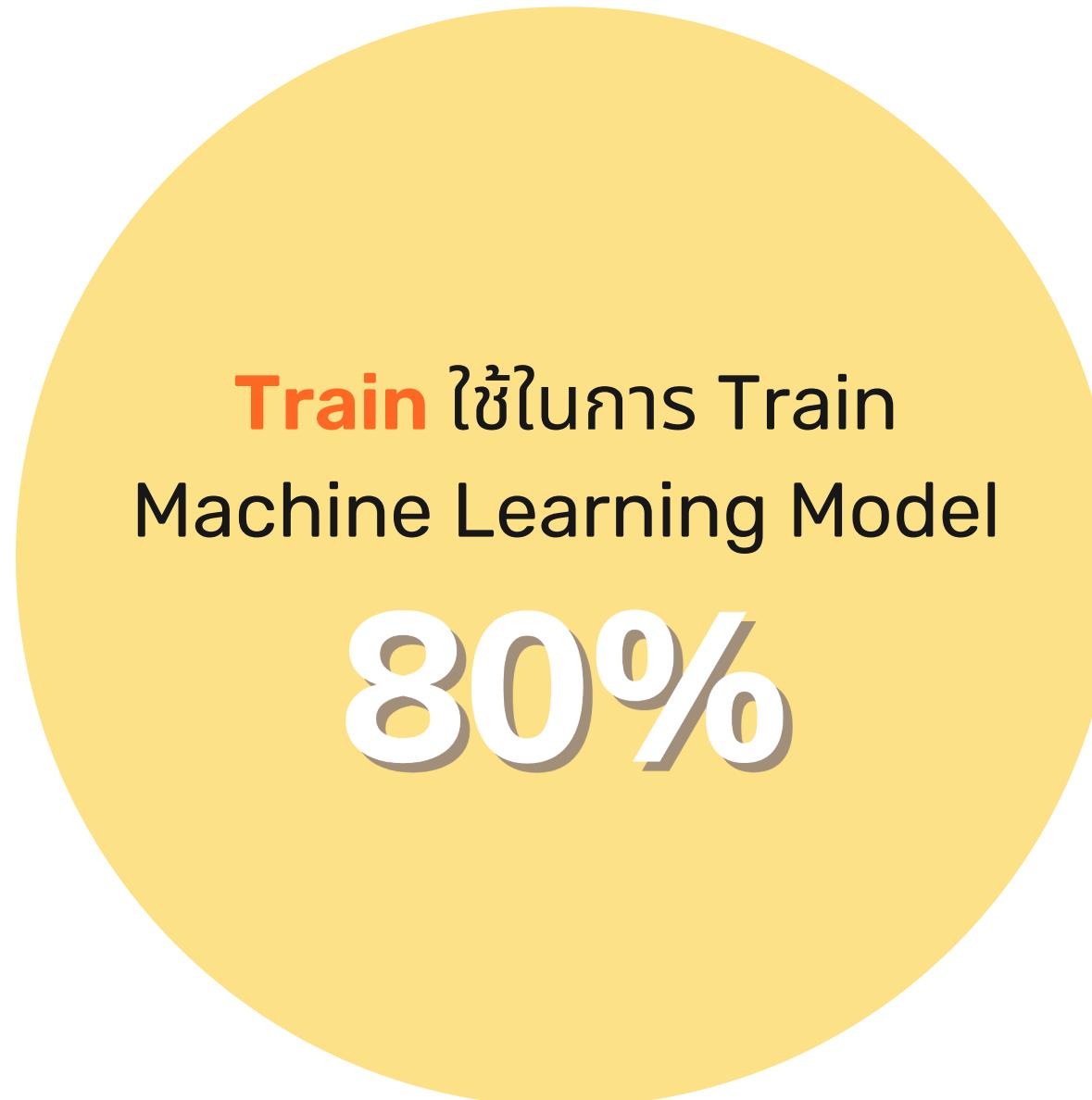




Part 5

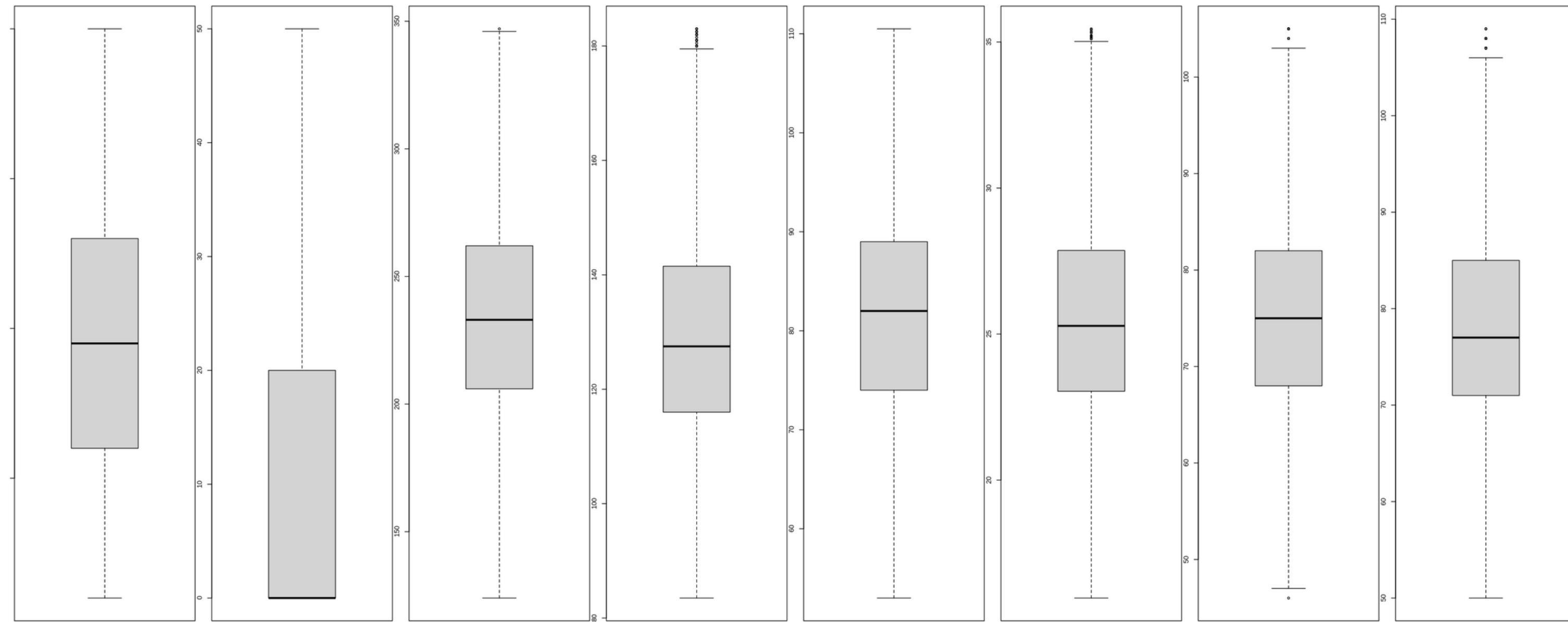
# Data Preprocessing

# Train Test Split



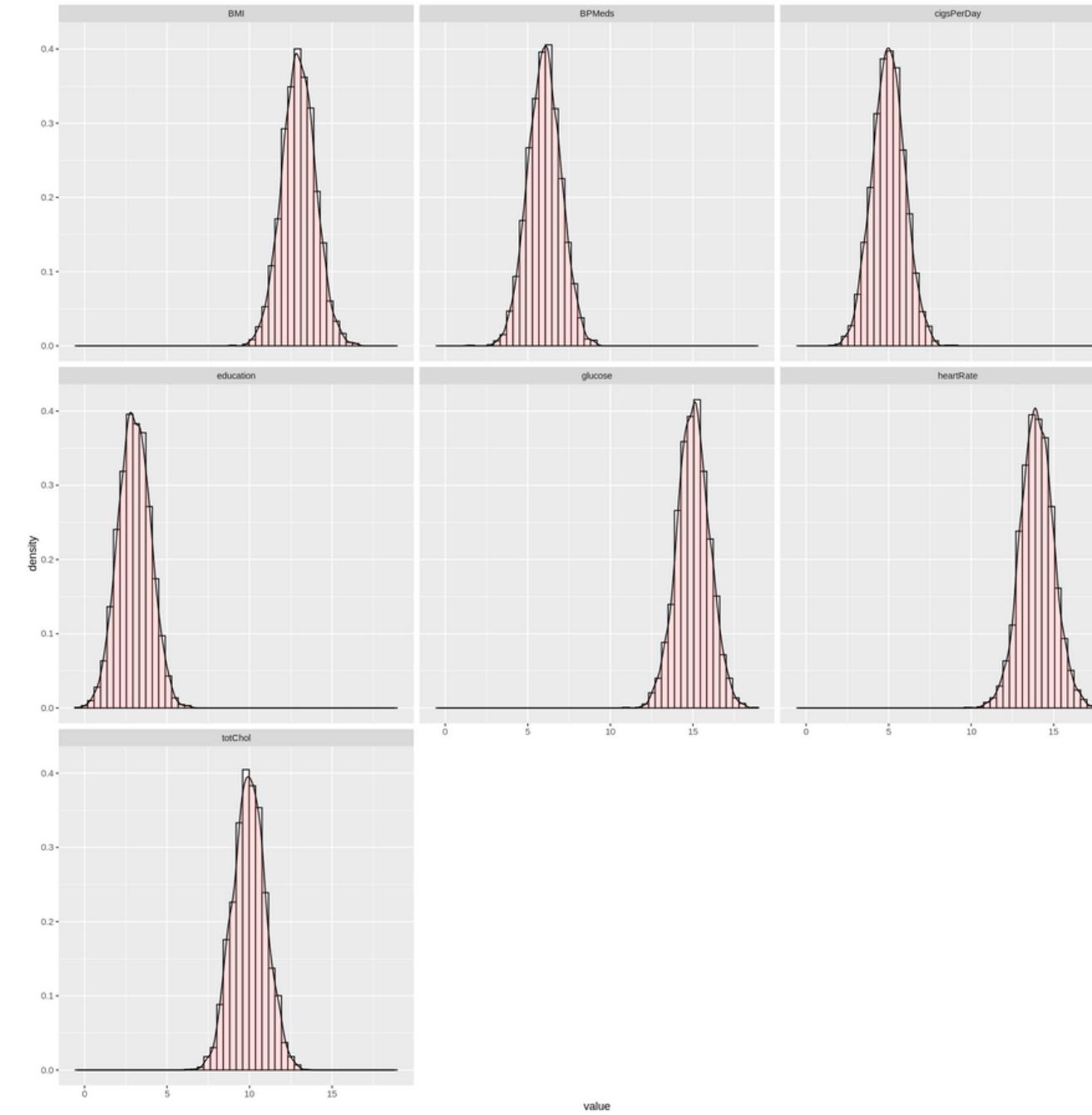
# Outlier Removal

- กำจัด Outlier ใน Train set

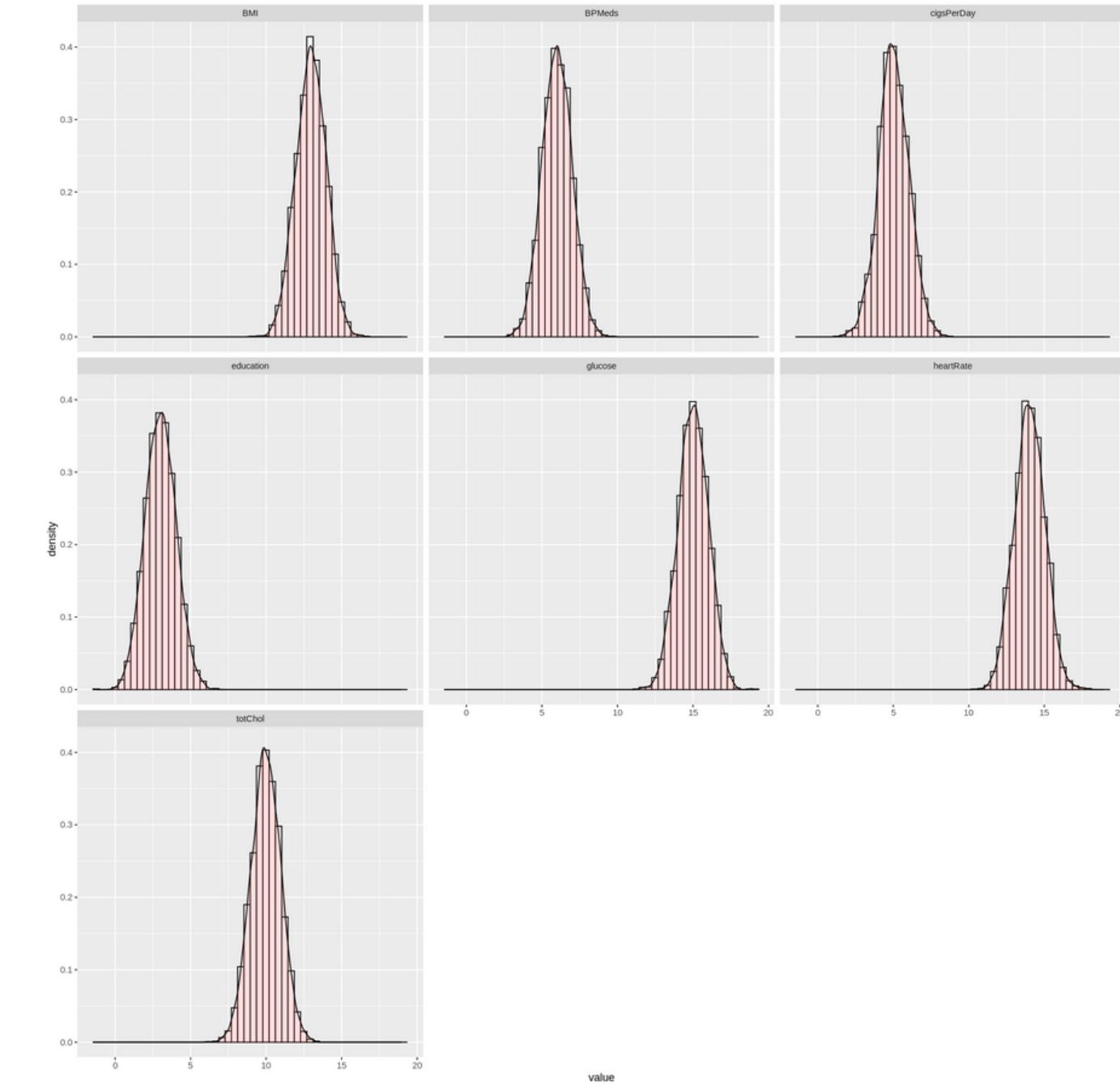


# Dealing with NA

- Plot Distribution ระหว่าง แต่ละ Column กับ Target ก่อน และ หลัง Fill NA



ก่อน Fill NA



หลัง Fill NA

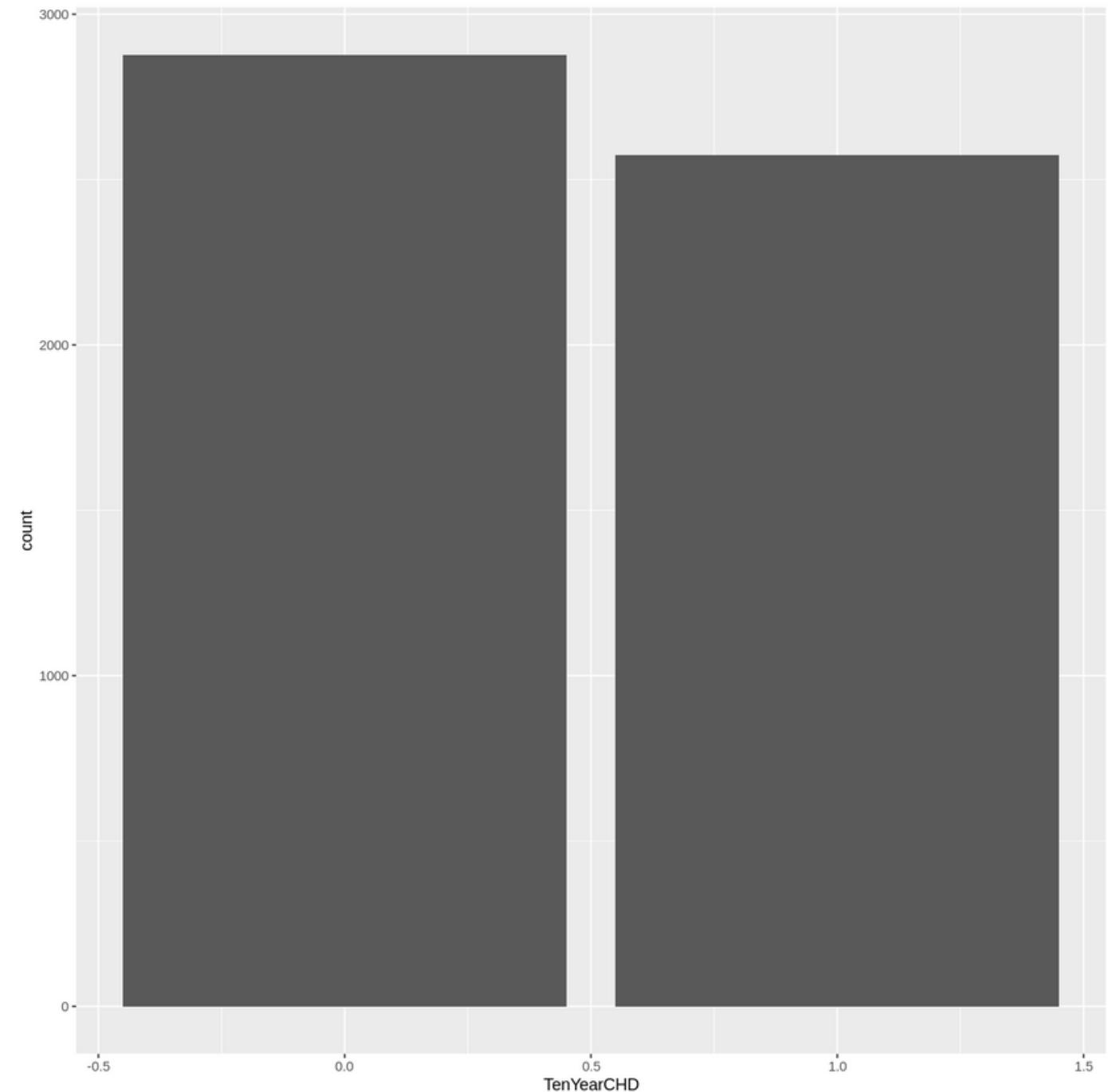
# Over Sampling

with Synthetic Minority  
Oversampling Technique(SMOTE)

- จัดการกับ Imbalance Data ด้วยการทำ Over Sampling with SMOTE โดยมีค่า Parameters K = 4, dup\_size = 4

```
[23] ### Over Sampling with SMOTE
train_balanced <- SMOTE(X = train, target = train$TenYearCHD, K = 4, dup_size = 4)$data
train_balanced <- train_balanced[,-length(train_balanced)]
print(paste0("Add data: ", nrow(train_balanced) - nrow(train), " rows"))

[1] "Add data: 2060 rows"
```



# Normalization

- ทำการ Normalization ด้วย Standardization

```
male          age      education currentSmoker
Min. :0.0000  Min. :32.00  Min. :1.000  Min. :0.0000
1st Qu.:0.0000 1st Qu.:44.66 1st Qu.:1.000 1st Qu.:0.0000
Median :0.2637 Median :51.63 Median :2.000 Median :0.4118
Mean   :0.4591 Mean   :51.40 Mean   :1.951 Mean   :0.4924
3rd Qu.:1.0000 3rd Qu.:58.50 3rd Qu.:2.695 3rd Qu.:1.0000
Max.  :1.0000  Max.  :70.00  Max.  :4.000  Max.  :1.0000
cigsPerDay      BPMeds    prevalentStroke prevalentHyp
Min. : 0.000  Min. :0.00000  Min. :0.00000  Min. :0.00000
1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
Median : 1.000  Median :0.00000  Median :0.00000  Median :0.00000
Mean   : 9.064  Mean   :0.03963  Mean   :0.01195  Mean   :0.3793
3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000
Max.  :50.000  Max.  :1.00000  Max.  :1.00000  Max.  :1.0000
diabetes        totChol    sysBP      diaBP
Min. :0.00000  Min. :124.0  Min. : 83.5  Min. : 53.00
1st Qu.:0.00000 1st Qu.:210.0 1st Qu.:120.0 1st Qu.: 76.00
Median :0.00000  Median :235.0  Median :129.0  Median : 82.50
Mean   :0.04198  Mean   :238.1  Mean   :132.7  Mean   : 83.19
3rd Qu.:0.00000 3rd Qu.:264.4 3rd Qu.:144.6 3rd Qu.: 90.00
Max.  :1.00000  Max.  :347.0  Max.  :183.0  Max.  :110.50
BMI            heartRate  glucose    TenYearCHD
Min. :15.96   Min. : 46.00  Min. : 50.00  no :2877
1st Qu.:23.48 1st Qu.: 68.00 1st Qu.: 73.00  yes:2575
Median :25.49  Median : 75.00  Median : 77.00
Mean   :25.70  Mean   : 75.26  Mean   : 78.46
3rd Qu.:27.85 3rd Qu.: 81.90 3rd Qu.: 83.24
Max.  :35.45  Max.  :105.00 Max.  :109.00
```

ก่อน Standardization

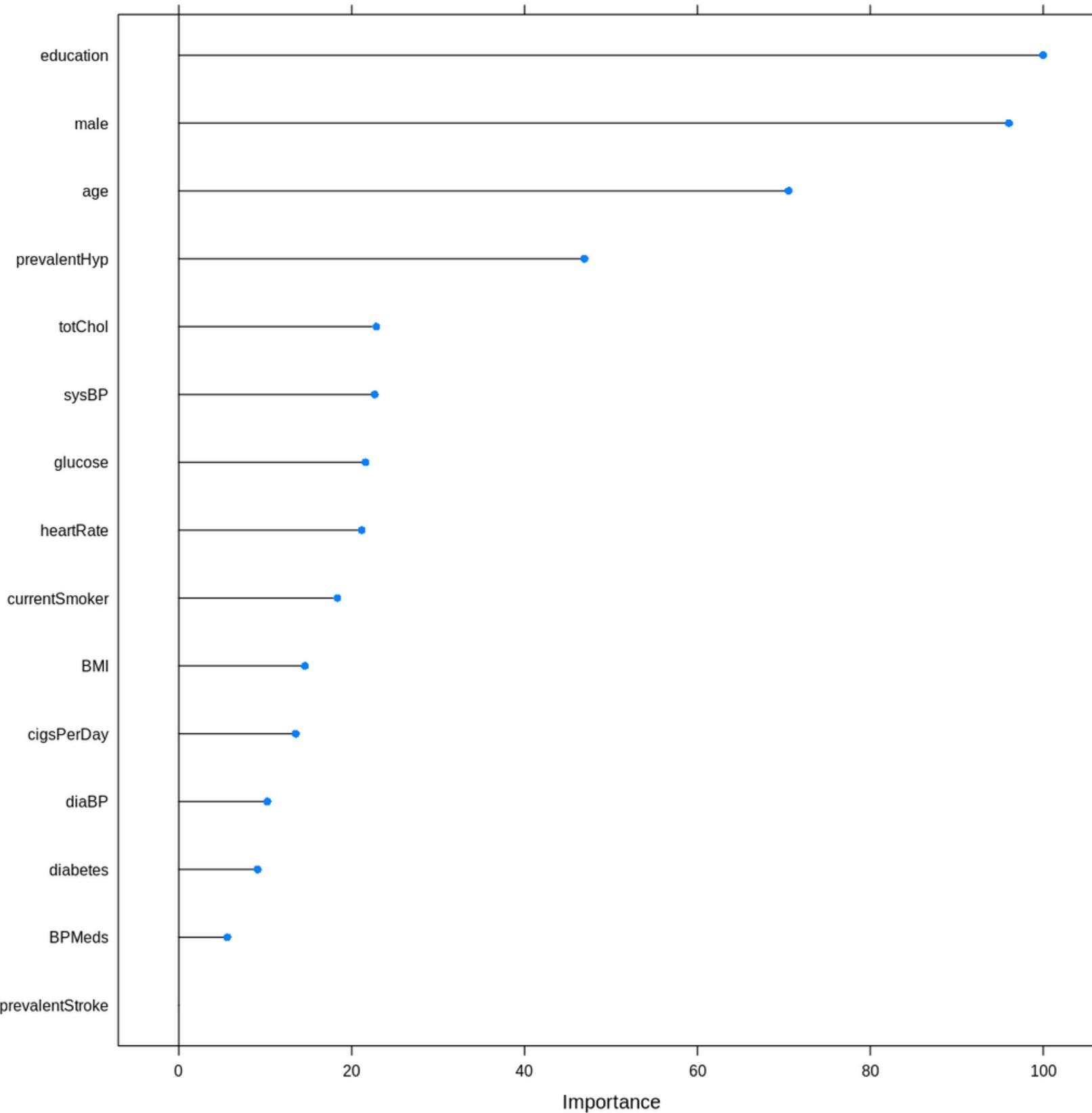
```
male.male      age.age      education.education
Min. :-0.9789220  Min. :-2.3061375  Min. :-0.9988025
1st Qu.:-0.9789220 1st Qu.:-0.8012673 1st Qu.:-0.9988025
Median :-0.4166076 Median : 0.0271679 Median : 0.0513032
Mean   : 0.0000000 Mean   : 0.0000000 Mean   : 0.0000000
3rd Qu.: 1.1532960 3rd Qu.: 0.8439637 3rd Qu.: 0.7814850
Max.  : 1.1532960  Max.  : 2.2108615  Max.  : 2.1515147
currentSmoker.currentSmoker cigsPerDay.cigsPerDay BPMeds.BPMeds
Min. :-1.0120737  Min. :-0.770576  Min. :-0.224406
1st Qu.:-1.0120737 1st Qu.:-0.770576 1st Qu.:-0.224406
Median :-0.1656752  Median :-0.685564  Median :-0.224406
Mean   : 0.0000000 Mean   : 0.0000000 Mean   : 0.0000000
3rd Qu.: 1.0434489 3rd Qu.: 0.929676 3rd Qu.: 0.224406
Max.  : 1.0434489  Max.  : 3.480054  Max.  : 5.437728
prevalentStroke.prevalentStroke prevalentHyp.prevalentHyp diabetes.diabetes
Min. :-0.122385  Min. :-0.813569  Min. :-0.235018
1st Qu.:-0.122385 1st Qu.:-0.813569 1st Qu.:-0.235018
Median :-0.122385  Median :-0.813569  Median :-0.235018
Mean   : 0.000000  Mean   : 0.000000  Mean   : 0.000000
3rd Qu.: 0.122385 3rd Qu.: 1.331143 3rd Qu.: 0.235018
Max.  :10.117408  Max.  : 1.331143  Max.  : 5.363632
totchol.totChol    sysBP.sysBP    diaBP.diaBP
Min. :-2.8724589  Min. :-2.6379743  Min. :-2.9313694
1st Qu.:-0.7065203 1st Qu.:-0.6806074 1st Qu.:-0.6978866
Median :-0.0768870  Median :-0.1979689  Median :-0.0666849
Mean   : 0.0000000 Mean   : 0.0000000 Mean   : 0.0000000
3rd Qu.: 0.6628540 3rd Qu.: 0.6359474 3rd Qu.: 0.6616247
Max.  : 2.7438702  Max.  : 2.6978615  Max.  : 2.6523377
BMI.BMI          heartRate.heartRate  glucose.glucose  TenYearCHD
Min. :-2.9575089  Min. :-2.8027426  Min. :-2.950121  no :2877
1st Qu.:-0.6747032 1st Qu.:-0.6956399 1st Qu.:-0.566102  yes:2575
Median :-0.0636264  Median :-0.0251981  Median :-0.151490
Mean   : 0.0000000 Mean   : 0.0000000 Mean   : 0.0000000
3rd Qu.: 0.6516793 3rd Qu.: 0.6357089 3rd Qu.: 0.495742
Max.  : 2.9589648  Max.  : 2.8481238  Max.  : 3.165407
```

หลัง Standardization

# Feature Selection

- เลือก Feature ที่มีความสำคัญมากกว่า 0.4 มาใช้ในการ Train Machine Learning Model ด้วย Feature Importance ที่ได้จาก XGBoost Model

xgbTree variable importance	Overall
education	100.000
male	96.037
age	70.546
prevalentHyp	46.930
totChol	22.852
sysBP	22.648
glucose	21.581
heartRate	21.158
currentSmoker	18.337
BMI	14.594
cigsPerDay	13.536
diaBP	10.232
diabetes	9.113
BPMeds	5.612
prevalentStroke	0.000



Part 6

# Train model

# Cross Validation

- ກໍາ Crossvalidation ໃນການ Train ແລ້ວ 5-Repeated 10-Fold Cross Validation With F1-score Matric

```
### 5-Repeated 10-Fold Cross Validation
f1 <- function(data, lev = NULL, model = NULL) {
  f1_val <- F1_Score(y_pred = data$pred, y_true = data$obs, positive = "yes")
  c(F1 = f1_val)
}

fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5,
                           summaryFunction = f1, classProbs = TRUE)
```

# Logistic Regression Model

glm\_model

Generalized Linear Model

5452 samples  
4 predictor  
2 classes: 'no', 'yes'

No pre-processing  
Resampling: Bootstrapped (25 reps)  
Summary of sample sizes: 5452, 5452, 5452, 5452, 5452, 5452, ...  
Resampling results:

Accuracy Kappa  
0.6591305 0.3154071

## • မာခြော်ခွဲ့သူ့ Model

```
[1] "Accuracy : 0.642883345561262"  
[1] "Precision: 0.691697191697192"  
[1] "Recall : 0.44"  
[1] "F1-Score : 0.537859007832898"
```

## • Confusion Matrix သူ့ Model

Confusion Matrix and Statistics

		pred	
		no	yes
no	no	2372	505
	yes	1442	1133

Accuracy : 0.6429  
95% CI : (0.63, 0.6556)  
No Information Rate : 0.6996  
P-Value [Acc > NIR] : 1

Kappa : 0.2696

McNemar's Test P-Value : <2e-16

Sensitivity : 0.6917  
Specificity : 0.6219  
Pos Pred Value : 0.4400  
Neg Pred Value : 0.8245  
Prevalence : 0.3004  
Detection Rate : 0.2078  
Detection Prevalence : 0.4723  
Balanced Accuracy : 0.6568  
'Positive' Class : yes

# Neural Network Model

## Neural Network

```
5452 samples
 4 predictor
 2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 4908, 4907, 4906, 4906, 4907, 4906, ...
Resampling results across tuning parameters:

size  decay  F1
1    0e+00  0.6578806
1    1e-04  0.6590477
1    1e-01  0.6581187
3    0e+00  0.6835214
3    1e-04  0.6877576
3    1e-01  0.6822439
5    0e+00  0.7180770
5    1e-04  0.7140470
5    1e-01  0.7203747

F1 was used to select the optimal model using the largest value.
The final values used for the model were size = 5 and decay = 0.1.
```

## • ผลลัพธ์ของ Model

```
[1] "Accuracy : 0.747982391782832"
[1] "Precision: 0.717493661716769"
[1] "Recall   : 0.769320388349515"
[1] "F1-Score : 0.742503748125937"
```

## • Confusion Matrix ของ Model

```
Confusion Matrix and Statistics

pred
      no yes
no  2097 780
yes  594 1981

Accuracy : 0.748
95% CI : (0.7362, 0.7595)
No Information Rate : 0.5064
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4963

McNemar's Test P-Value : 6.01e-07

Sensitivity : 0.7175
Specificity : 0.7793
Pos Pred Value : 0.7693
Neg Pred Value : 0.7289
Prevalence : 0.5064
Detection Rate : 0.3634
Detection Prevalence : 0.4723
Balanced Accuracy : 0.7484

'Positive' Class : yes
```

# Extreme Gradient Boosting Model (XGBoost)

- Confusion Matrix ของ Model

```
Confusion Matrix and Statistics
```

```
pred
  no yes
no 2017 860
yes 178 2397

Accuracy : 0.8096
95% CI : (0.7989, 0.82)
No Information Rate : 0.5974
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6233
```

```
McNemar's Test P-Value : < 2.2e-16
```

```
Sensitivity : 0.7360
Specificity : 0.9189
Pos Pred Value : 0.9309
Neg Pred Value : 0.7011
Prevalence : 0.5974
Detection Rate : 0.4397
Detection Prevalence : 0.4723
Balanced Accuracy : 0.8274
```

```
'Positive' Class : yes
```

- ผลลัพธ์ของ Model

```
[1] "Accuracy : 0.809611151870873"
[1] "Precision: 0.73595333128646"
[1] "Recall   : 0.930873786407767"
[1] "F1-Score : 0.82201646090535"
```

# Stacking Model

- นำ 3 Models ที่ได้ train มาทำ Stacking และทำ Hard Voting

Confusion Matrix and Statistics

```
pred
  no yes
no 2182 695
yes 598 1977
```

```
Accuracy : 0.7628
95% CI : (0.7513, 0.7741)
No Information Rate : 0.5099
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.5252
```

```
McNemar's Test P-Value : 0.007591
```

```
Sensitivity : 0.7399
Specificity : 0.7849
Pos Pred Value : 0.7678
Neg Pred Value : 0.7584
Prevalence : 0.4901
Detection Rate : 0.3626
Detection Prevalence : 0.4723
Balanced Accuracy : 0.7624
```

```
'Positive' Class : yes
```

Confusion Matrix ของ Model

- ผลลัพธ์ของ Model

```
[1] "Accuracy : 0.762839325018342"
[1] "Precision: 0.739895209580838"
[1] "Recall    : 0.767766990291262"
[1] "F1-Score : 0.753573470554603"
```



Part 7

# Evaluation

# ทดสอบประสิทธิภาพของ Model กับ Test set

Confusion Matrix and Statistics

```
pred
  no yes
no  597 122
yes 70  59
```

```
Accuracy : 0.7736
95% CI  : (0.7439, 0.8014)
No Information Rate : 0.7866
P-Value [Acc > NIR] : 0.8325924
```

```
Kappa : 0.2469
```

```
McNemar's Test P-Value : 0.0002327
```

```
Sensitivity : 0.32597
```

```
Specificity : 0.89505
```

```
Pos Pred Value : 0.45736
```

```
Neg Pred Value : 0.83032
```

```
Prevalence : 0.21344
```

```
Detection Rate : 0.06958
```

```
Detection Prevalence : 0.15212
```

```
Balanced Accuracy : 0.61051
```

```
'Positive' Class : yes
```

Confusion Matrix ของ Model

- ผลลัพธ์ของ Model

```
[1] "Accuracy : 0.773584905660377"
[1] "Precision: 0.325966850828729"
[1] "Recall    : 0.457364341085271"
[1] "F1-Score  : 0.380645161290323"
```

Ready to last topic?

# Discussion and Conclusion

# Summary

ใน Project นี้ กลุ่มของเราระบุไปด้วยกระบวนการ **เพื่อหาความเสี่ยงในการเกิดโรคหัวใจและหลอดเลือดภายในร่างกายในระยะเวลา 10 ปี** ซึ่งประกอบไปด้วยกระบวนการ Data science 5 ขั้นตอน

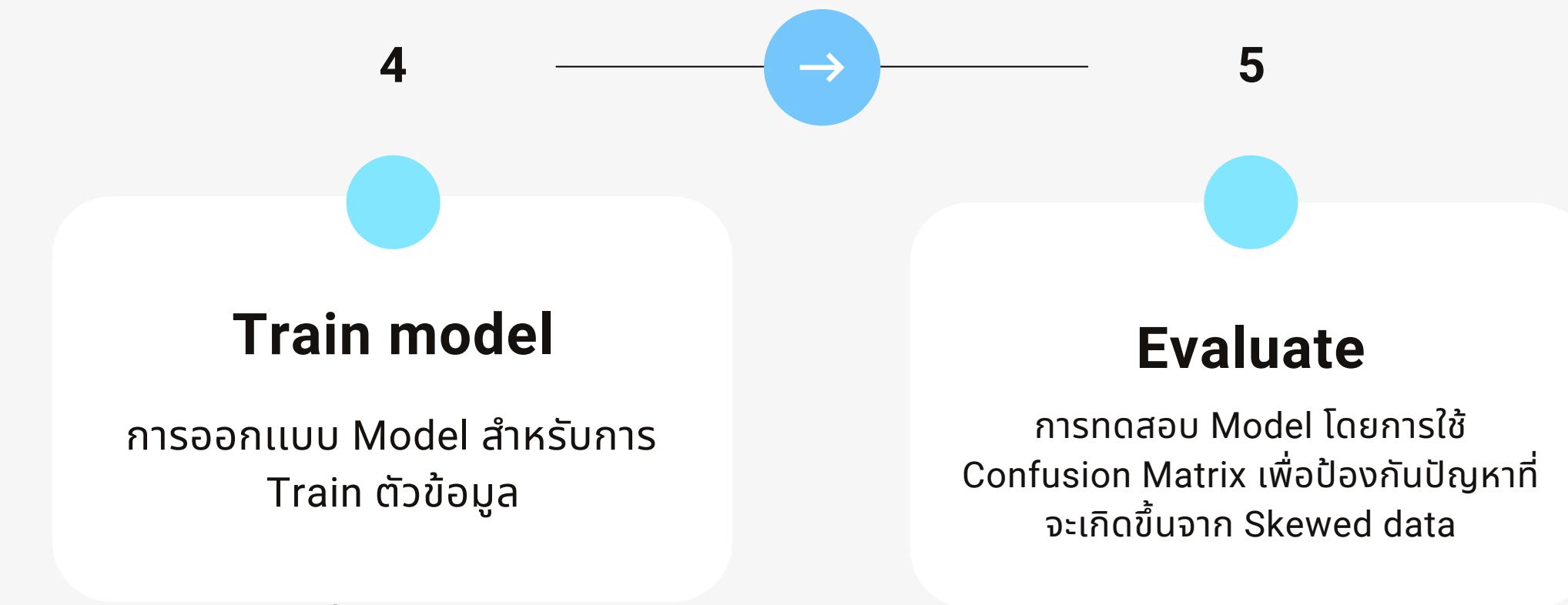


\*ปัญหาที่พบ : Outlier, NaN และ Data imbalance



# Summary

ใน Project นี้ กลุ่มของเรารaid ทำการศึกษาเพื่อหาความเสี่ยงในการเกิดโรคหัวใจและหลอดเลือดภายในระยะเวลา 10 ปี ซึ่งประกอบไปด้วยกระบวนการทาง Data science 5 ขั้นตอน



\*มี 3 Model ดังนี้  
1.Logisical Regression Model  
2.Neural network Model  
3.XGBoost Model\*

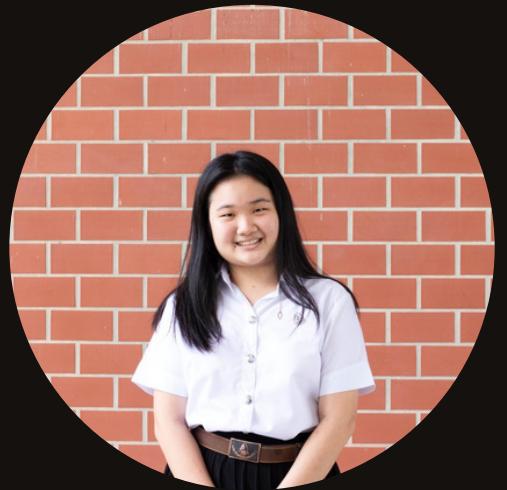
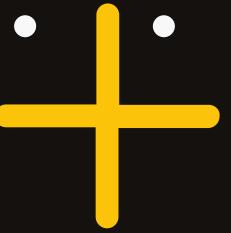


# Discussion

ได้ผลลัพธ์ **Precision 0.32, Recall 0.45**  
และ **F1-score** ที่ **0.38** ซึ่งเป็นผลลัพธ์ที่ยัง  
ไม่สามารถนำไปใช้งานจริงกับโรงพยาบาลได้

- คิดว่าสามารถเพิ่มความถูกต้องของ **Model** ได้ในกรณีที่มีข้อมูลมากกว่านี้จาก  
การสำรวจลุ่มผู้ป่วยโรคหัวใจในเมืองอื่น
- **เปลี่ยน Model** ให้มีความซับซ้อนมากขึ้น

# The Team



**Ms.Natchariya  
Wongamnuayporn**  
61070507204



**Mr.Natchapol  
Patamawisut**  
61070507205



**Mr.Rungwigrai  
Payakkanuwat**  
61070507219



**Mr.Sahassawas  
Srichan**  
61070507223



**Mr.Thanachart  
Wongkum**  
61070507228



# Thank you!

