資料科學期末報告

組別：第三組

日期：2020.06.1608753204

資科碩一 王柏仁
資科碩一 李鈺祥
資科碩一 唐英哲
資科碩一 林祐丞
資科碩一 蕭郁君

大綱

- 專案簡介
- 資料輸入輸出與前處理
- 模型設定及處理
- 實驗結果
- 參考出處

專案簡介

- 若鐵達尼存活預測是
  - 初入資料科學、Kaggle競賽的熱門入門題
  - 初入特徵工程，了解其過程及特徵選取
  - 瞭解資料科學的運作過程
- 那此次的心臟病分類預測競賽就是
  - 正式實作資料科學的流程
  - 較深入進行特徵工程，用各種假設檢定、熵來了解特徵之間對答案及模型的重要性
  - 遇到實務上的問題並嘗試著解決

資料輸入輸出

- 共有欄位如下：

Id	Max rate
Age	Exercise angina
Sex	St depression
Chest pain	Slope
Resting bp	Vessels
Cholestoral	Thalium scan
High sugar	Heart disease
Ecg	

資料前處理

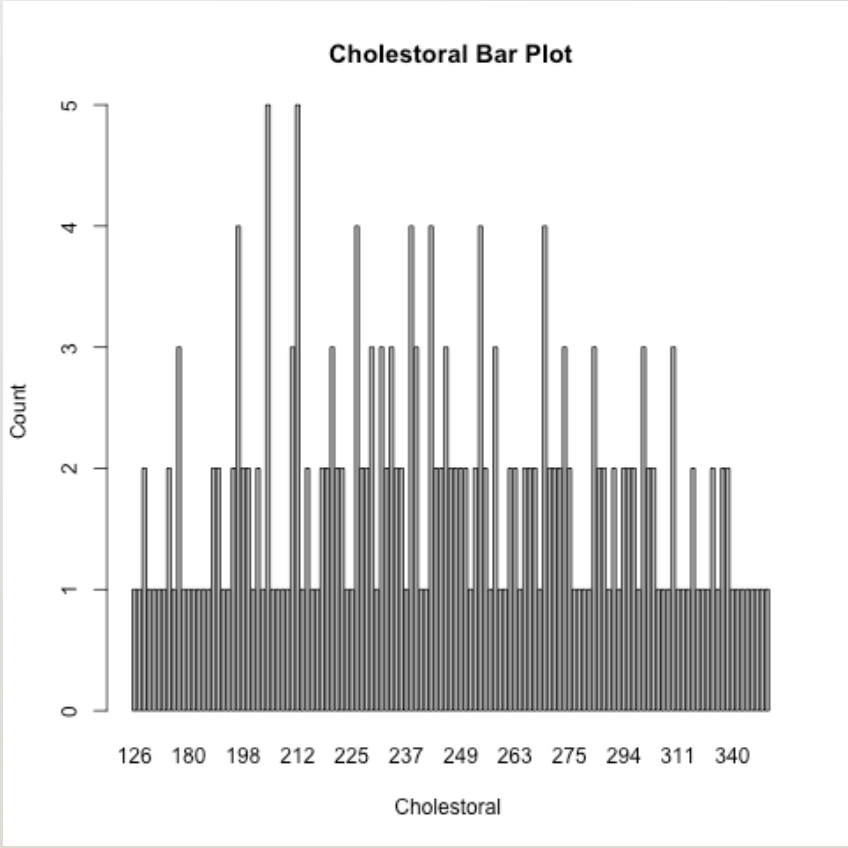
- 轉換之前
- 為什麼要轉換
- 如何轉換與處理資料
- 各欄位原始格式
- 預期各欄位轉換後格式
- 資料觀察
- 轉換後的資料
  - 訓練
  - 測試

資料前處理 – 轉換之前

- 理解、觀察各欄位資料
- 進行適當的轉換於特定欄位
- 觀察轉換後的分佈，例如
  - 受試者的年齡區間分佈
  - 受試者的性別分佈
  - 受試者的總膽固醇區間分佈
  - 各個欄位與心臟病的關係程度為何
  - ...等

資料前處理 – 為什麼要轉換

- 多數欄位原始型態是數值型
- 資料範圍相當大時，觀察其分佈會有些困難
- 因此需要做適當的資料轉換



資料前處理 – 如何轉換與處理資料

- 前處理：
  - Outlier：（ > 第三四分位 + 0.5*IQR； < 第一四分位 - 0.5*IQR ）
    - 將個欄位 outlier 的 index 記錄起來，找將共同的 id 的資料並移除
    - 其他 outlier 則用最大值或最小值補齊，避免訓練時讓模型訓練錯誤
  - NA Numbers：
    - 找到 NA 的欄位資料，並用 median 來填補（因為沒有連續型數值變數有 NA，所以對類別變數則用 median）
  - Normalize：
    - 這部分未處理，已包含在建模裡面

資料前處理 – 如何轉換與處理資料

- 數值變數：
  - 區間化：利用 kmeans 方法，找到最大組間距離、最小組內距離的 K 點。並利用 cut 針對 K 個中心找到相對應的區間，Label 化
- 類別變數：
  - 連結變數與答案之間的關係，算出 contingency table
  - 變數間的重要性與相關性
    - Chi Square 假設檢定：
      - 當 p value 小於 alpha（0.05），則拒絕 H0，也就是兩變數之間相關；反之則接受 H0，即兩變數獨立
    - Mutual Information：
      - 利用相互資訊熵，來觀察變數對於答案 Label 的重要程度

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

資料前處理 – 各欄位原始格式

Input	Format
id	integer
age	numeric
sex	binary
chest_pain	multinomial
resting_bp	numeric
cholestoral	numeric
high_sugar	binary
ecg	multinomial

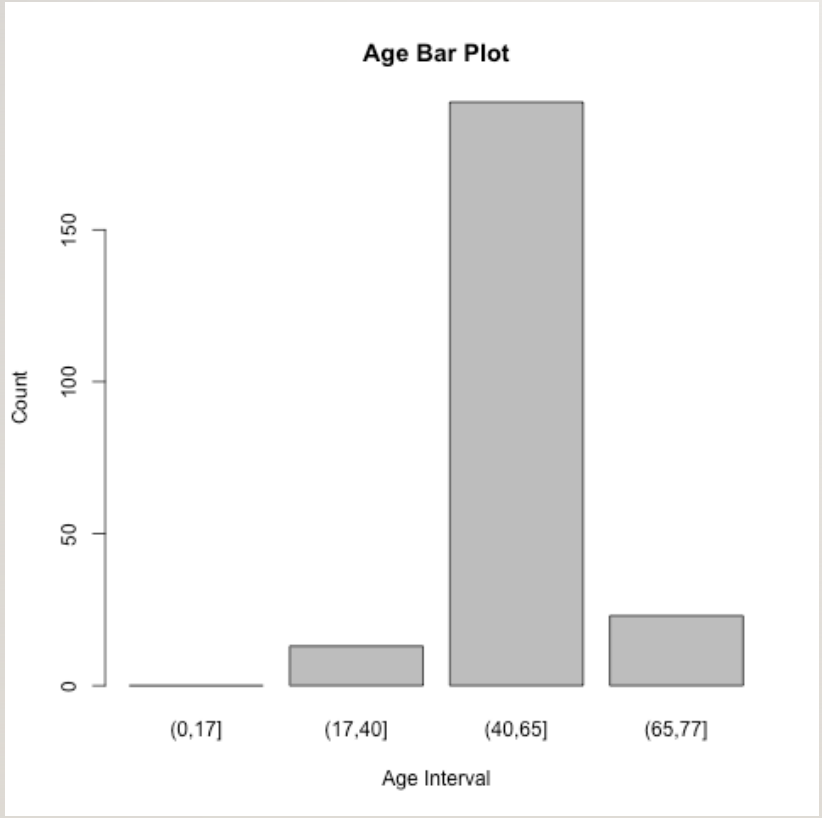
資料前處理 – 各欄位原始格式

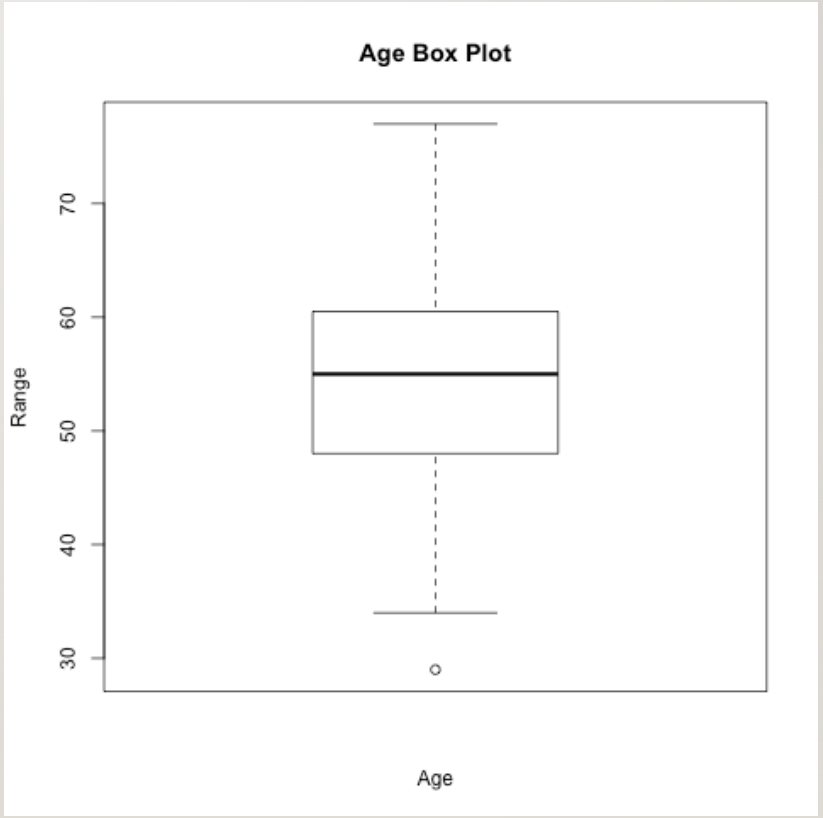
Input	Format
max_rate	numeric
exercise_angina	binary
st_depression	binary
slope	multinomial
vessels	multinomial
thalium_scan	multinomial
Heart_disease	binary

資料前處理 – 預期各欄位轉換後格式

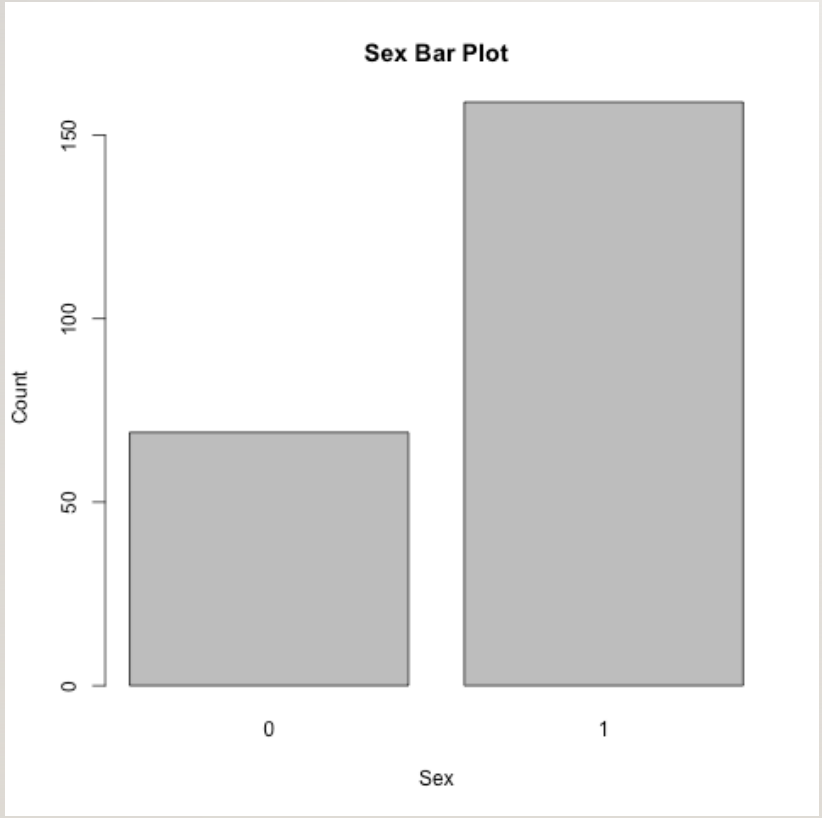
Input	Format	Remark
age_interval	multinomial	利用設定區間進行標籤化 P.S: 0, 17, 40, 65, max_value
resting_bp_interval	multinomial	利用設定區間進行標籤化 P.S: 0, 120, 139, max_value
cholestoral_interval	multinomial	利用設定區間進行標籤化 P.S: 0, 129, 200, 239, max_value
max_rate	multinomial	利用 Kmeans 進行標籤化 P.S: 113, 121, 152, 177, 202
st_depression	multinomial	利用 Kmeans 進行標籤化 P.S: 0, 0.03, 0.51, 0.81, 1.11, 1.48, 1.97, max_value
vessels	multinomial	Filling up NA
thalium_scan	multinomial	Filling up NA, (3, 6, 7) ->(1,2,3)

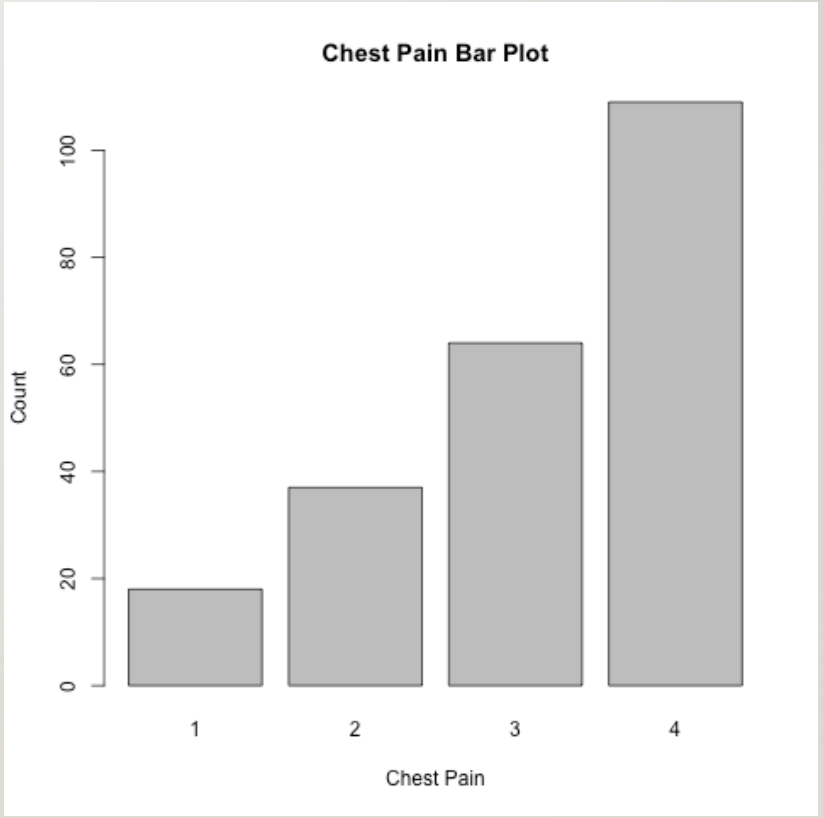
資料前處理 – 資料觀察


Age Bar Plot
Count
0
50
100
150
(0,17]
(17,40]
(40,65]
(65,77]
Age Interval

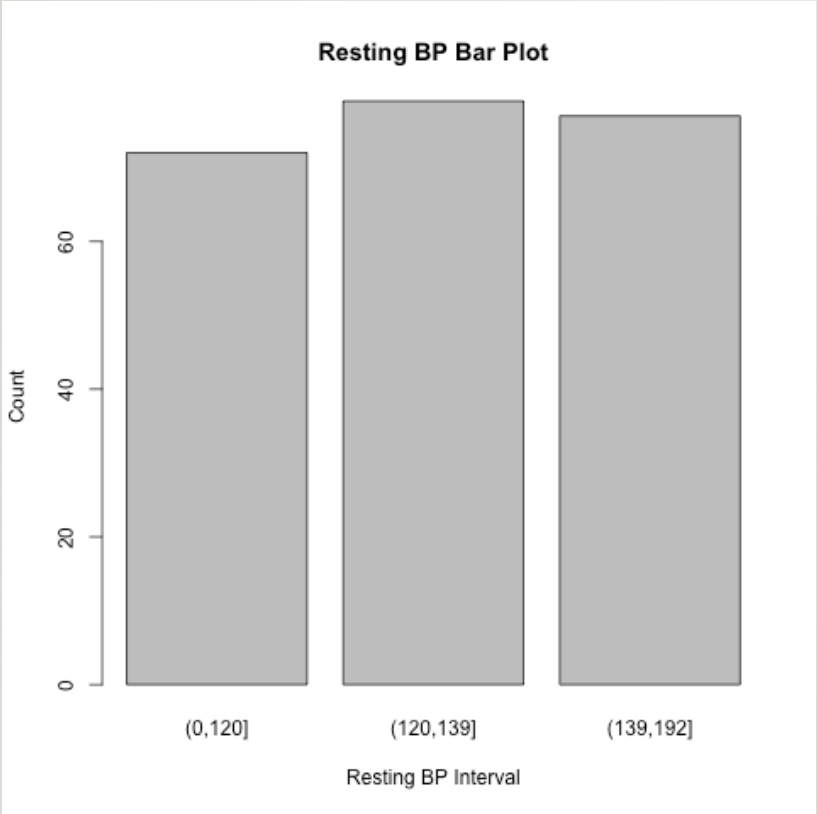

Age Box Plot
Range
30
40
50
60
70
Age

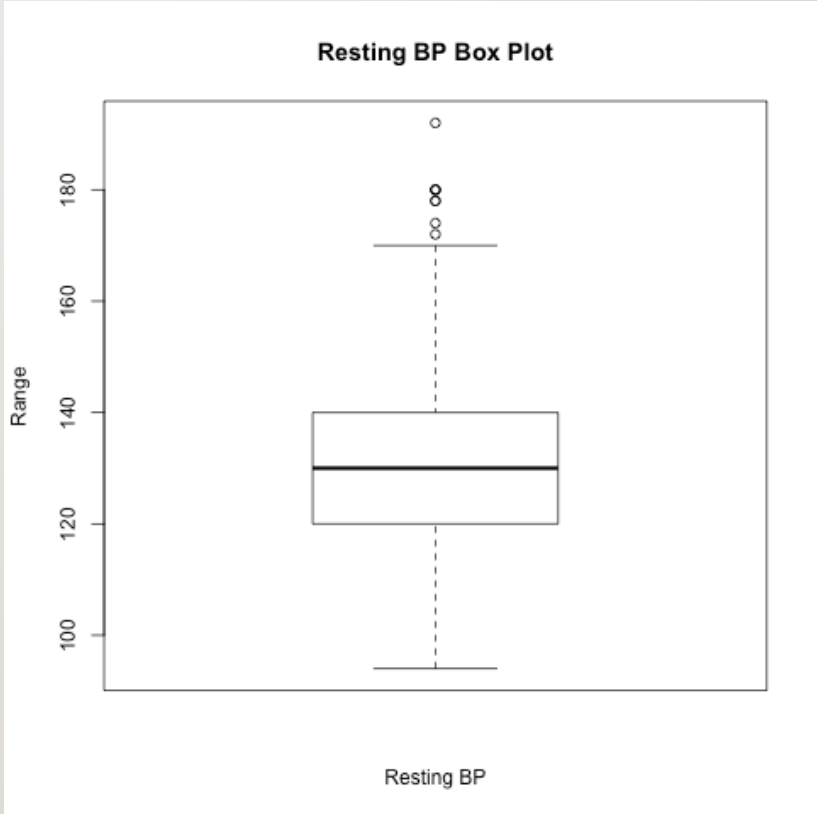
資料前處理 – 資料觀察


Sex Bar Plot
Count
0
50
100
150
0
1
Sex

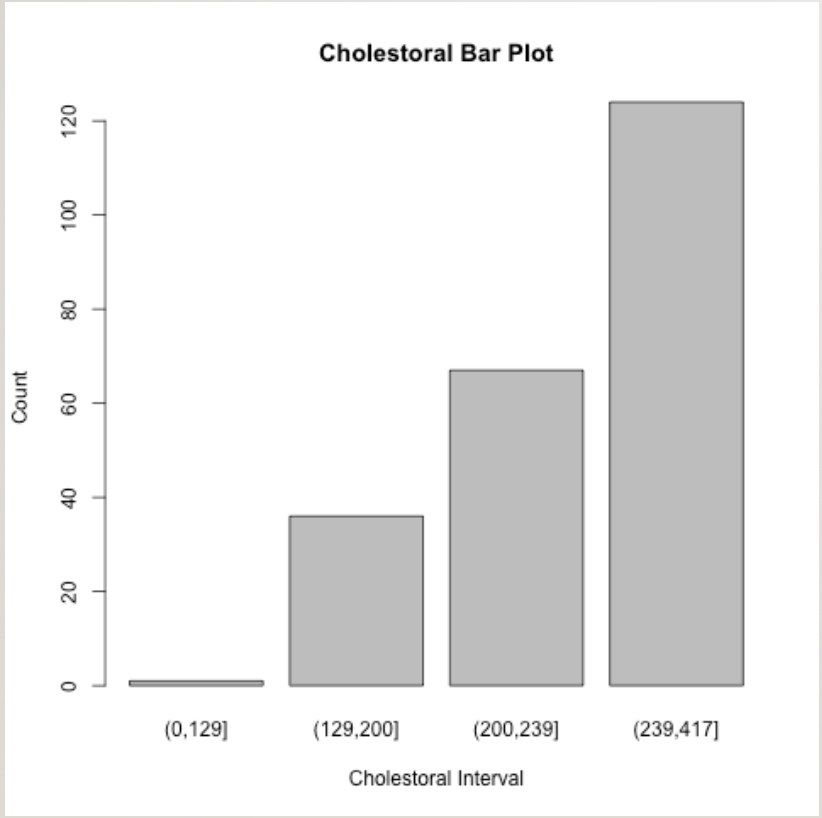

Chest Pain Bar Plot
Count
0
20
40
60
80
100
1
2
3
4
Chest Pain

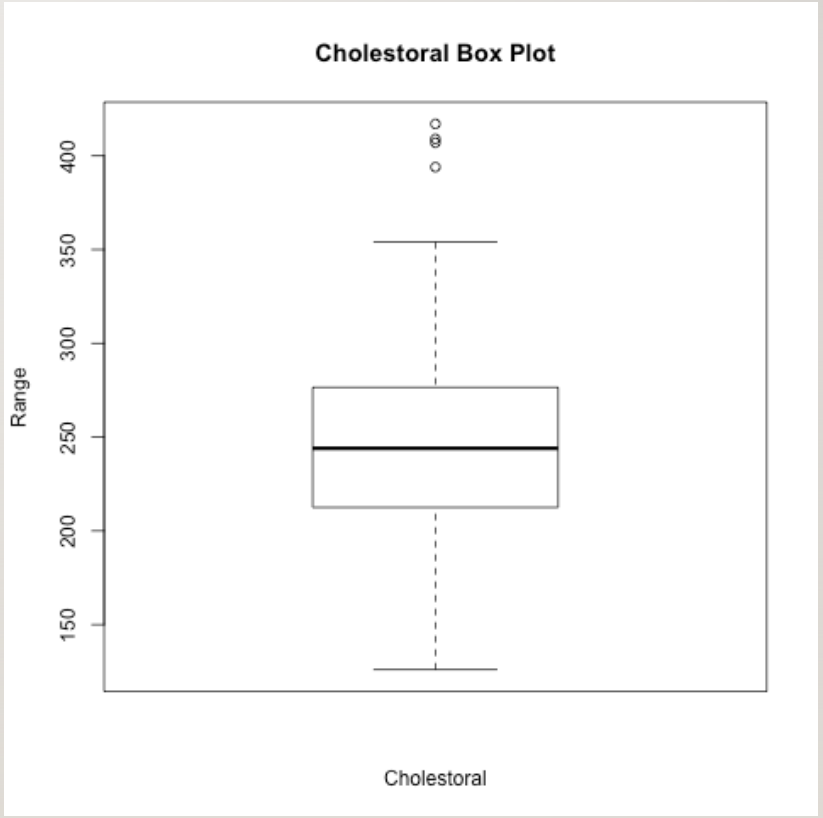
資料前處理 – 資料觀察


Resting BP Bar Plot
Count
0
20
40
60
(0,120]
(120,139]
(139,192]
Resting BP Interval

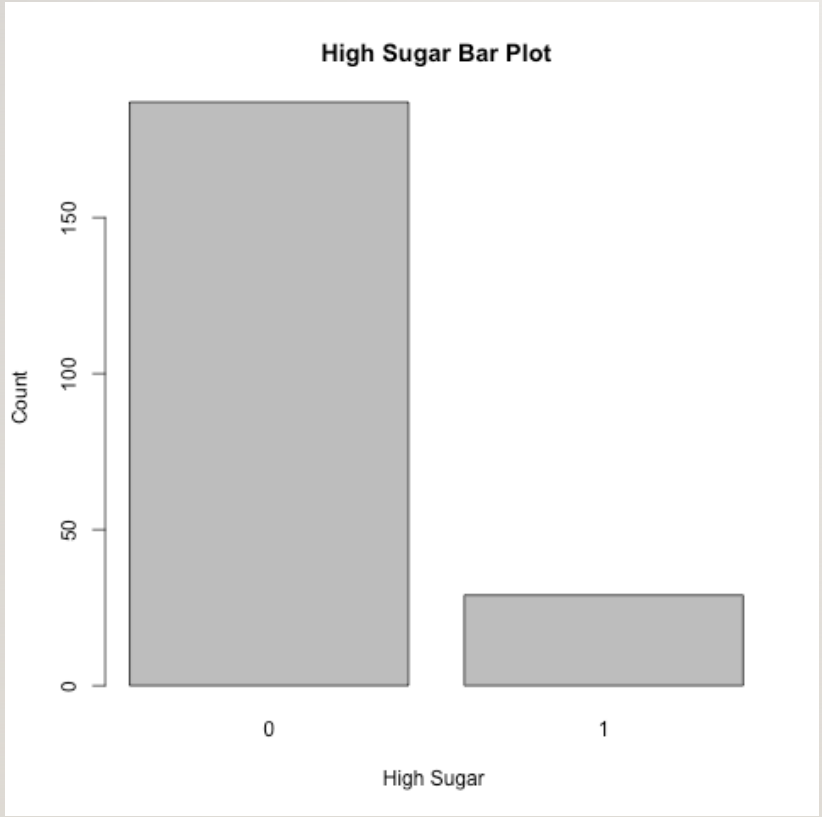

Resting BP Box Plot
Range
100
120
140
160
180
Resting BP

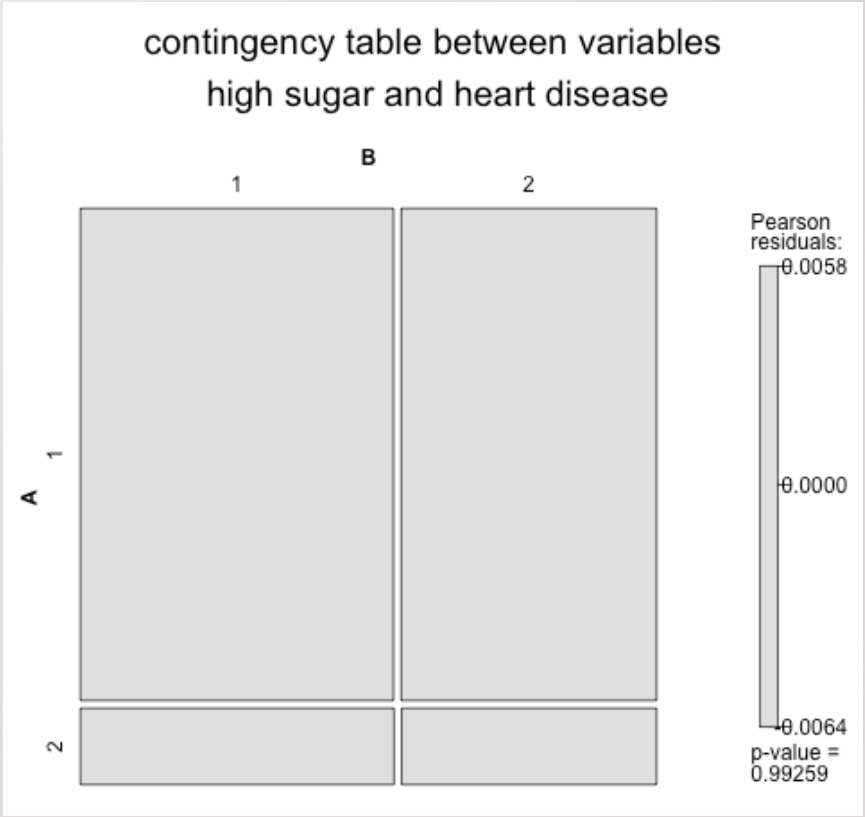
資料前處理 – 資料觀察


Cholestoral Bar Plot
Count
0
20
40
60
80
100
120
(0,129]
(129,200]
(200,239]
(239,417]
Cholestoral Interval

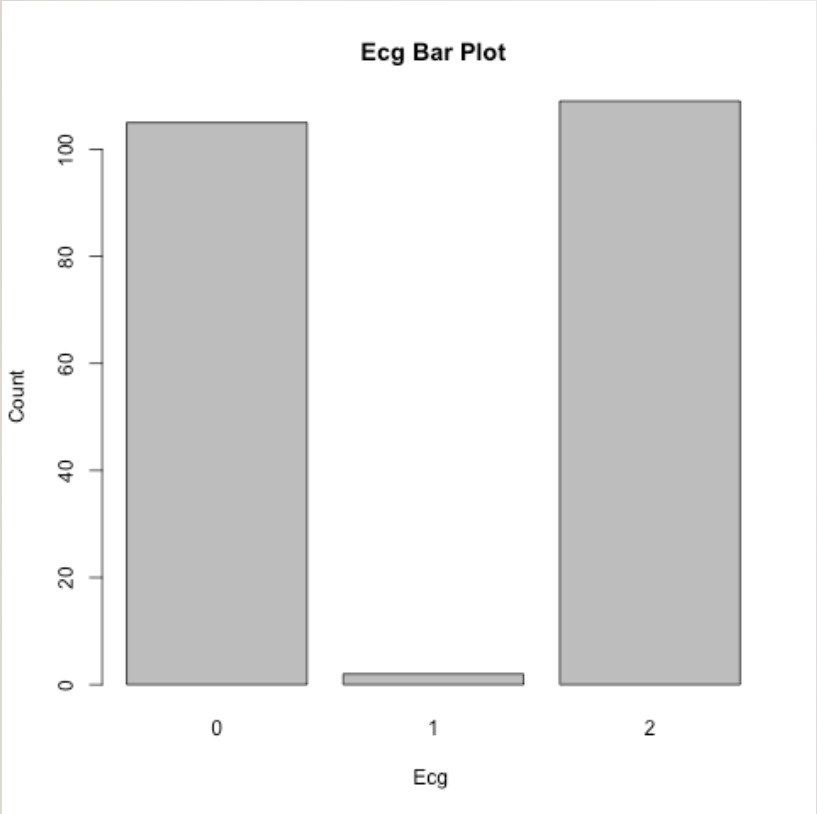

Cholestoral Box Plot
Range
150
200
250
300
350
400
Cholestoral

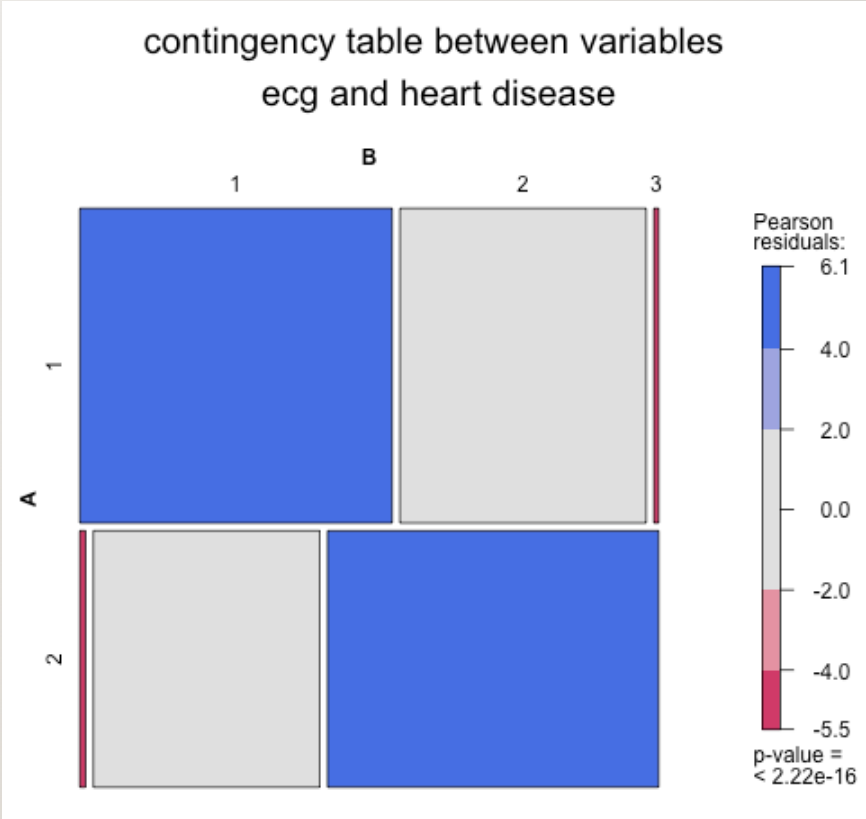
資料前處理 – 資料觀察


High Sugar Bar Plot
Count
0
50
100
150
0
1
High Sugar

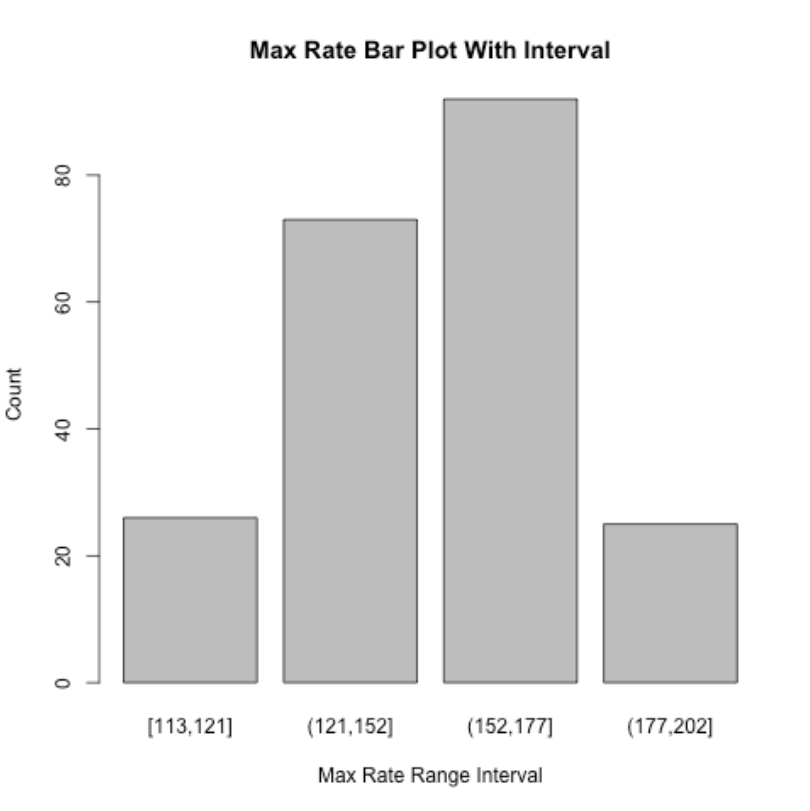

contingency table between variables
high sugar and heart disease
B
1
2
A
1
2
Pearson residuals:
0.0058
0.0000
-0.0064
p-value = 0.99259

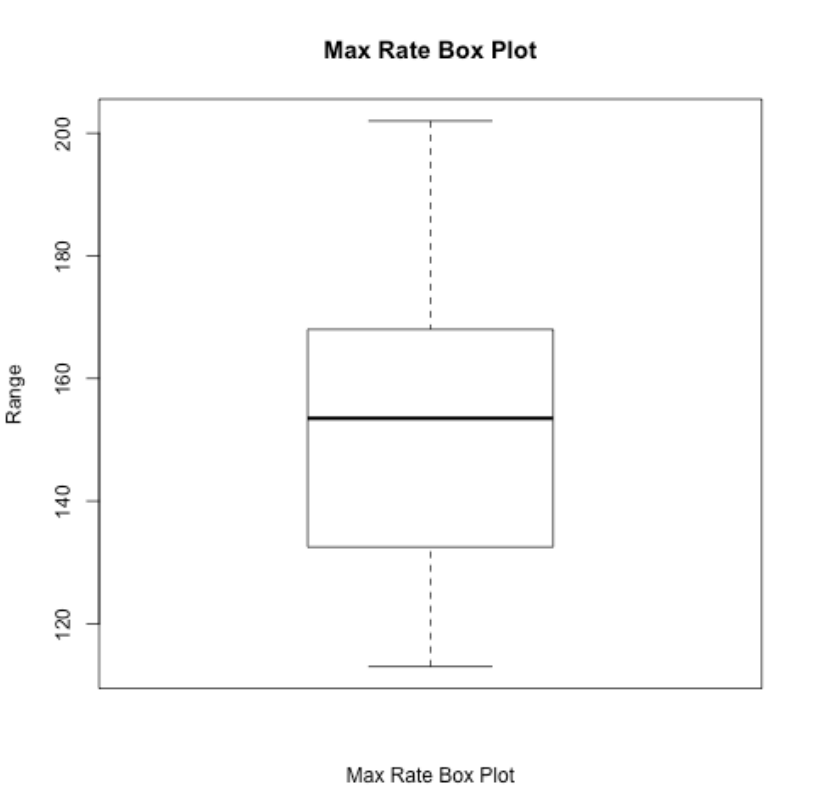
資料前處理 – 資料觀察


Ecg Bar Plot
Count
0
20
40
60
80
100
0
1
2
Ecg

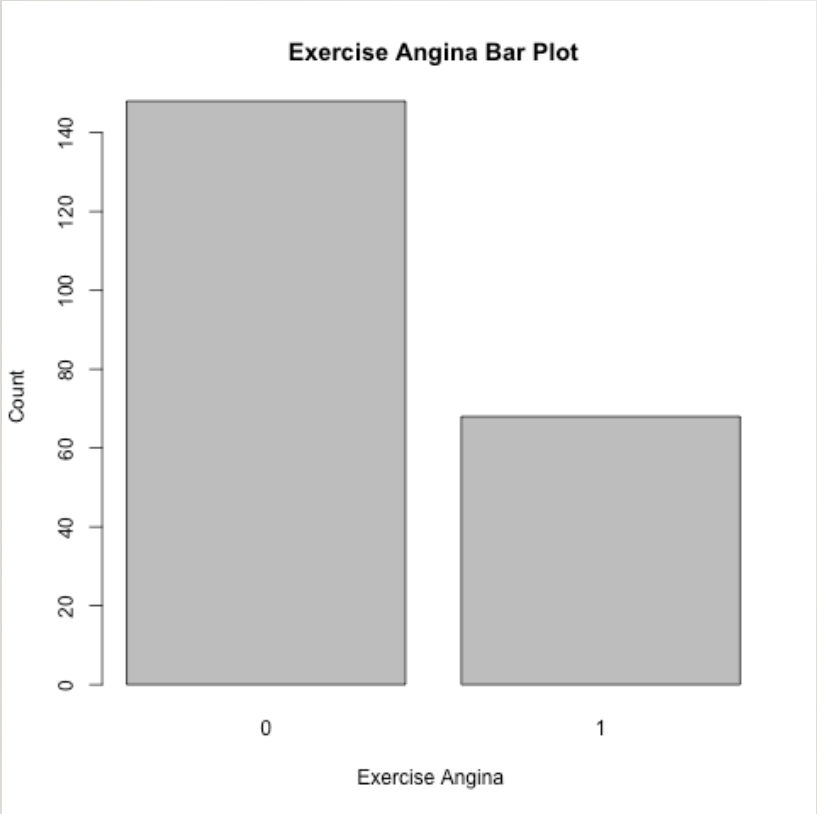

contingency table between variables
ecg and heart disease
B
1
2
3
A
1
2
Pearson
residuals:
6.1
4.0
2.0
0.0
-2.0
-4.0
-5.5
p-value =
< 2.22e-16

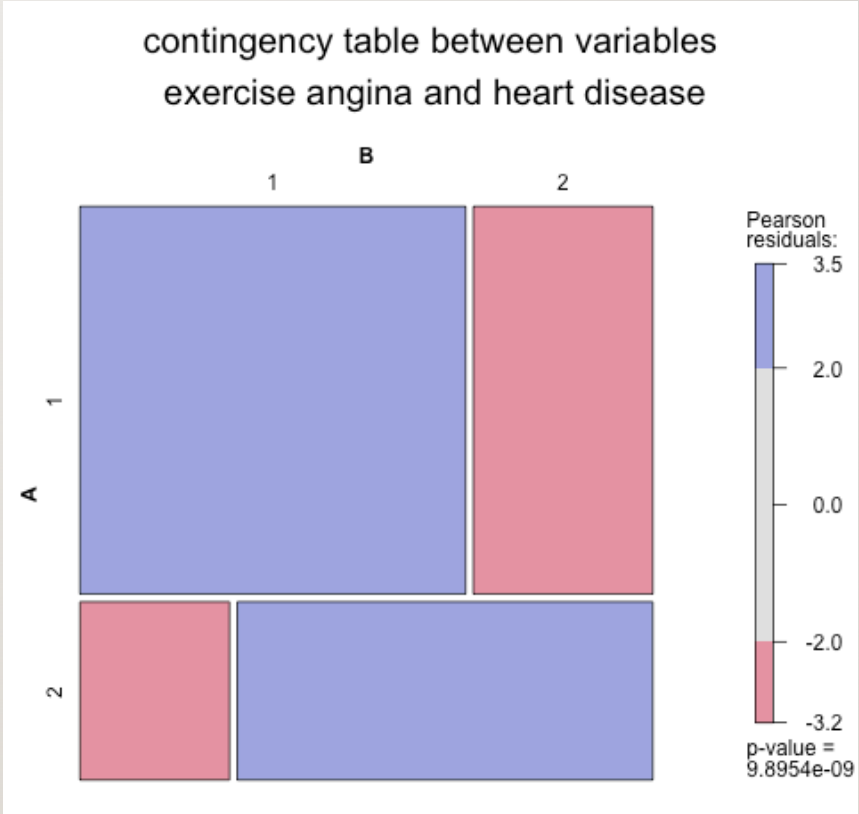
資料前處理 – 資料觀察


Max Rate Bar Plot With Interval
Count
0
20
40
60
80
[113,121]
(121,152]
(152,177]
(177,202]
Max Rate Range Interval

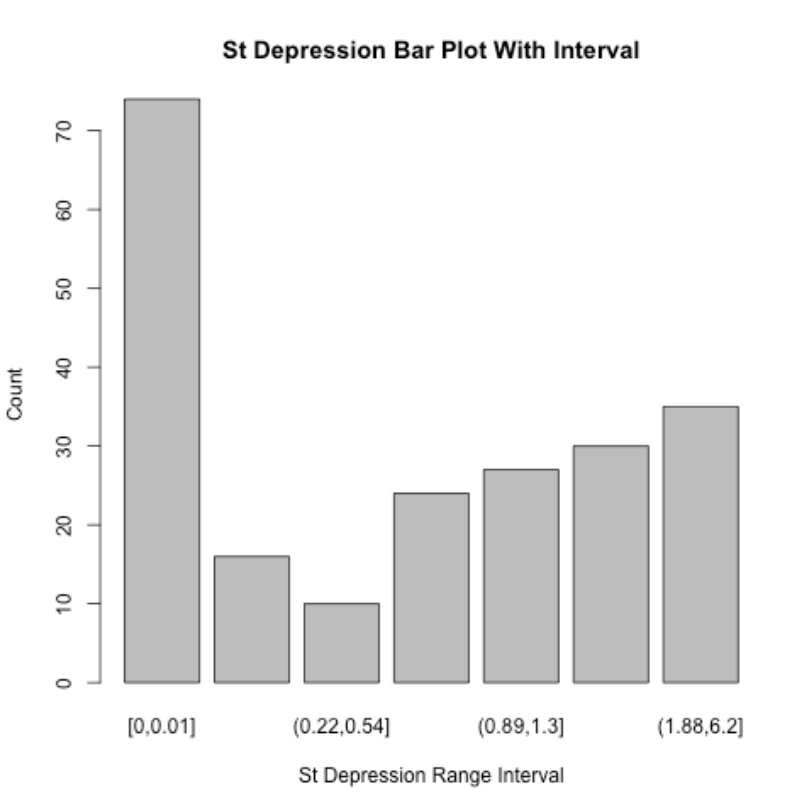

Max Rate Box Plot
Range
120
140
160
180
200
Max Rate Box Plot

資料前處理 – 資料觀察


Exercise Angina Bar Plot
Count
0
20
40
60
80
100
120
140
0
1
Exercise Angina

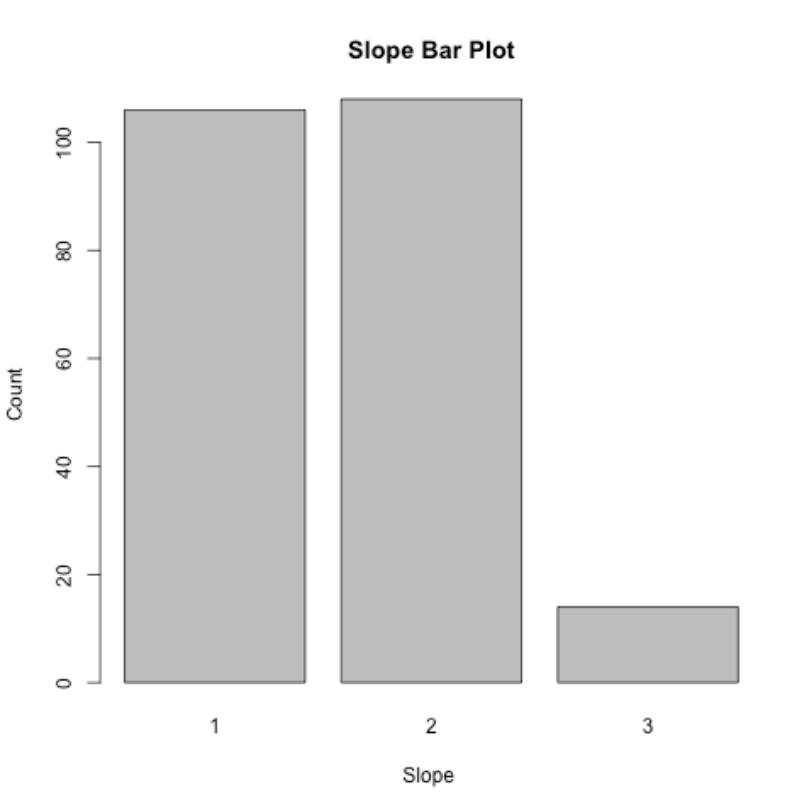

contingency table between variables
exercise angina and heart disease
B
1
2
A
1
2
Pearson residuals:
3.5
2.0
0.0
-2.0
-3.2
p-value = 9.8954e-09

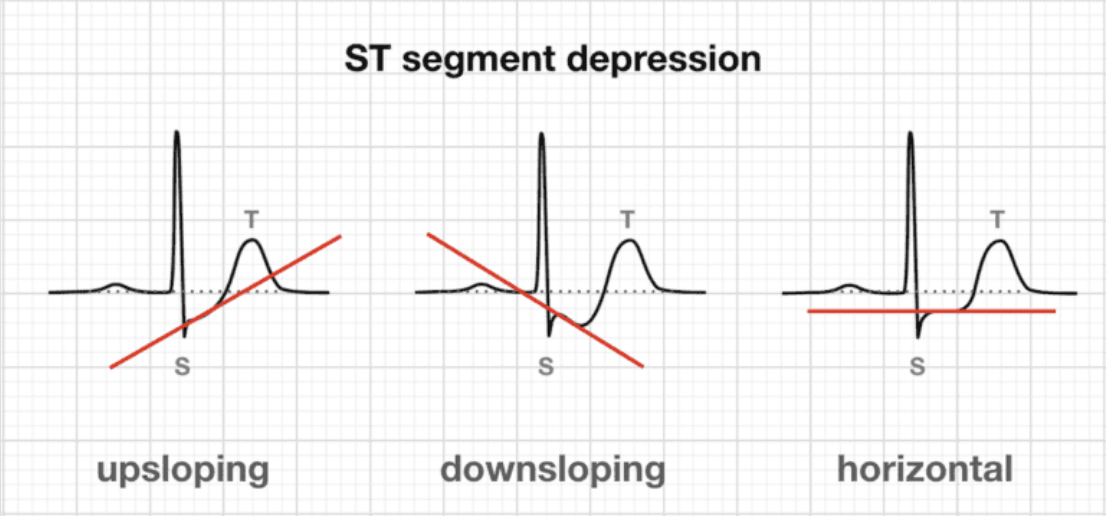
資料前處理 – 資料觀察


St Depression Bar Plot With Interval
Count
0
10
20
30
40
50
60
70
[0,0.01]
(0.22,0.54]
(0.89,1.3]
(1.88,6.2]
St Depression Range Interval

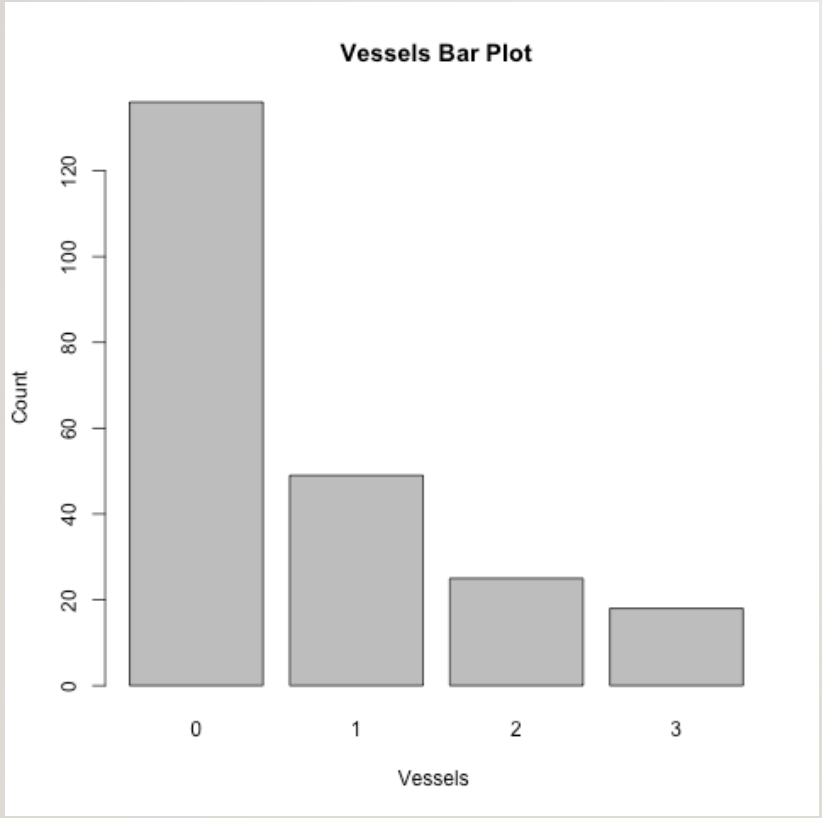

St Depression Box Plot
Range
0
1
2
3
4
5
6
St Depression Box Plot

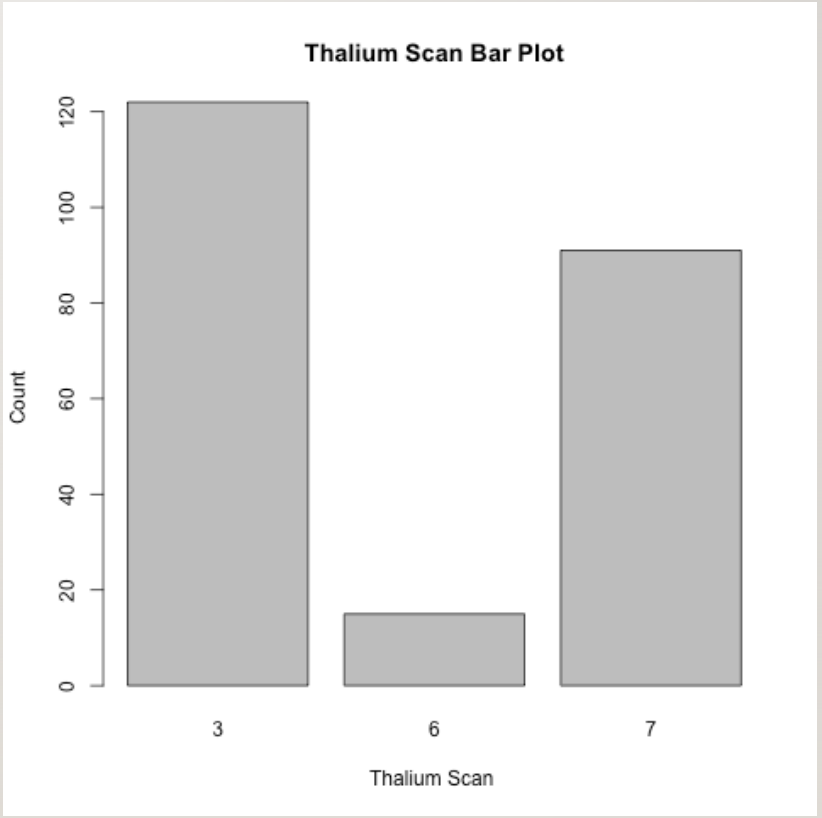
資料前處理 – 資料觀察

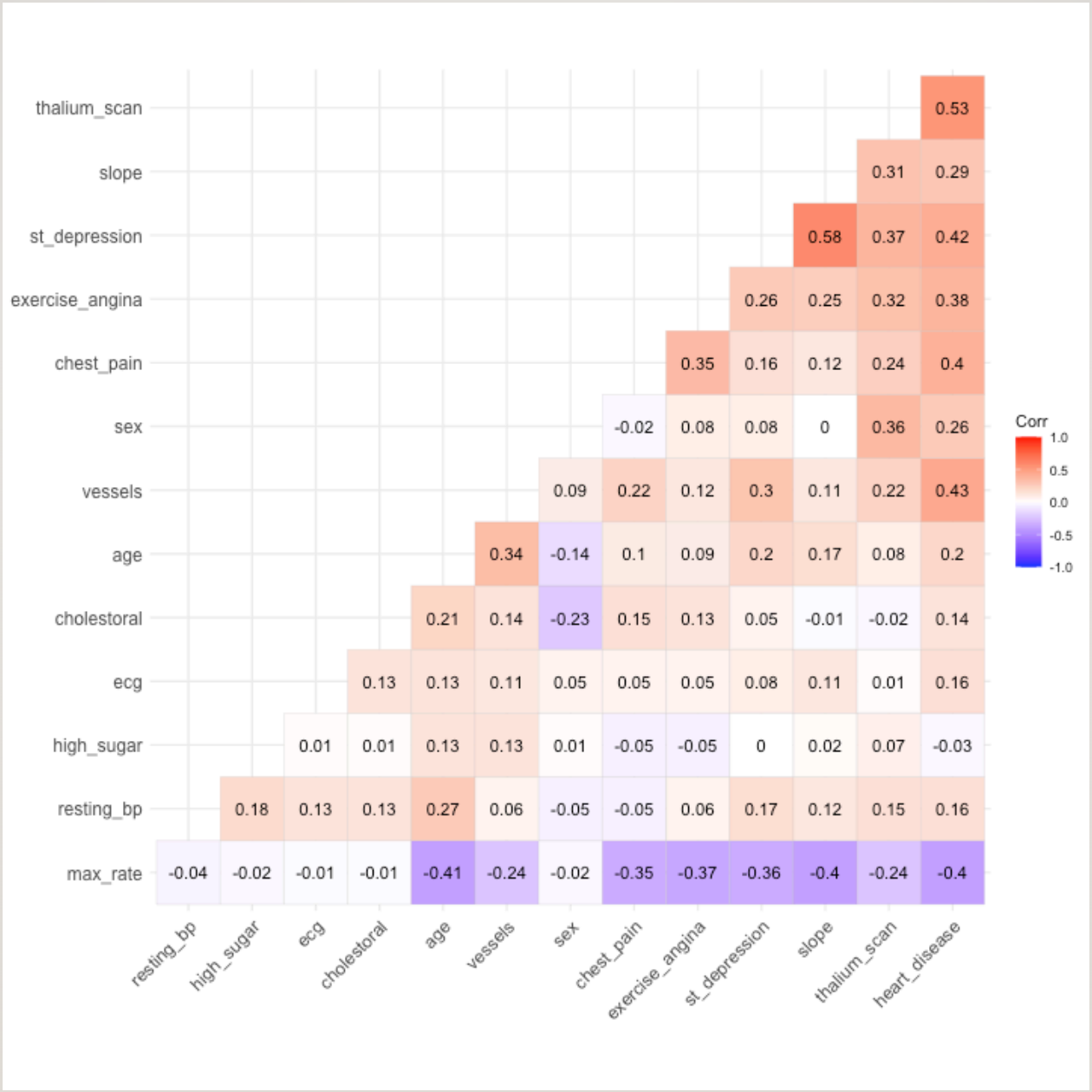

Slope Bar Plot
Count
0
20
40
60
80
100
1
2
3
Slope


ST segment depression
T
S
T
S
T
S
upsloping
downsloping
horizontal

資料前處理 – 資料觀察


Vessels Bar Plot
Count
0
20
40
60
80
100
120
0
1
2
3
Vessels


Thalium Scan Bar Plot
Count
0
20
40
60
80
100
120
3
6
7
Thalium Scan


thalium_scan 0.53
slope 0.31 0.29
st_depression 0.58 0.37 0.42
exercise_angina 0.26 0.25 0.32 0.38
chest_pain 0.35 0.16 0.12 0.24 0.4
sex -0.02 0.08 0.08 0 0.36 0.26
vessels 0.09 0.22 0.12 0.3 0.11 0.22 0.43
age 0.34 -0.14 0.1 0.09 0.2 0.17 0.08 0.2
cholestoral 0.21 0.14 -0.23 0.15 0.13 0.05 -0.01 -0.02 0.14
ecg 0.13 0.13 0.11 0.05 0.05 0.05 0.08 0.11 0.01 0.16
high_sugar 0.01 0.01 0.13 0.13 0.01 -0.05 -0.05 0 0.02 0.07 -0.03
resting_bp 0.18 0.13 0.13 0.27 0.06 -0.05 -0.05 0.06 0.17 0.12 0.15 0.16
max_rate -0.04 -0.02 -0.01 -0.01 -0.41 -0.24 -0.02 -0.35 -0.37 -0.36 -0.4 -0.24 -0.4
resting_bp high_sugar ecg cholestoral age vessels sex chest_pain exercise_angina st_depression slope thalium_scan heart_disease
Corr
1.0
0.5
0.0
-0.5
-1.0

資料前處理 – 轉換後的訓練資料

id	cholestoral	cholestoral_with_label	cholestoral_without_label	high_sugar	ecg	max_rate	max_rate_with_label	max_rate_without_label	exercise_angina	st_depression	st_depression_with_label	st_depression_without_label	heart_disease
1	233	2	(198,239]	1	2	150	2	(121,152]	0	2.3	7	(1.88,6.2]	0
2	286	4	(282,417]	0	2	113	1	[113,121]	1	1.5	6	(1.3,1.88]	1
3	229	2	(198,239]	0	2	129	2	(121,152]	1	6.2	7	(1.88,6.2]	1
5	204	2	(198,239]	0	2	172	3	(152,177]	0	1.4	6	(1.3,1.88]	0
6	236	2	(198,239]	0	0	178	4	(177,202]	0	0.8	4	(0.54,0.89]	0
8	417	4	(282,417]	0	0	163	3	(152,177]	1	0.6	4	(0.54,0.89]	0
9	254	3	(239,282]	0	2	147	2	(121,152]	0	1.4	6	(1.3,1.88]	1
10	203	2	(198,239]	1	2	155	3	(152,177]	1	6.2	7	(1.88,6.2]	1
11	192	1	[182,198]	0	0	148	2	(121,152]	0	0.4	3	(0.22,0.54]	0
12	294	4	(282,417]	0	2	153	3	(152,177]	0	1.3	5	(0.89,1.3]	0
15	199	2	(198,239]	1	0	162	3	(152,177]	0	0.5	3	(0.22,0.54]	0
17	229	2	(198,239]	0	0	168	3	(152,177]	0	1	5	(0.89,1.3]	1
18	239	3	(239,282]	0	0	160	3	(152,177]	0	1.2	5	(0.89,1.3]	0
19	275	3	(239,282]	0	0	139	2	(121,152]	0	0.2	2	(0.01,0.22]	0
20	266	3	(239,282]	0	0	171	3	(152,177]	0	0.6	4	(0.54,0.89]	0
21	211	2	(198,239]	0	2	144	2	(121,152]	1	1.8	6	(1.3,1.88]	0
24	224	2	(198,239]	0	2	173	3	(152,177]	0	6.2	7	(1.88,6.2]	1
27	417	4	(282,417]	0	0	172	3	(152,177]	0	0	1	[0,0.01]	0
28	226	2	(198,239]	0	0	114	1	[113,121]	0	6.2	7	(1.88,6.2]	0
29	247	3	(239,282]	0	0	171	3	(152,177]	0	1.5	6	(1.3,1.88]	0
31	239	3	(239,282]	0	0	151	2	(121,152]	0	1.8	6	(1.3,1.88]	0
33	417	4	(282,417]	0	0	158	3	(152,177]	0	0	1	[0,0.01]	1

資料前處理 – 轉換後的測試資料

id	cholestoral	cholestoral_with_label	cholestoral_without_label	high_sugar	ecg	max_rate	max_rate_with_label	max_rate_without_label	exercise_angina	st_depression	st_depression_with_label	st_depression_without_label
179	564	6	(287,564]	0	0	162	6	(161,166]	0	1.9	3	(1.26,2.15]
14	263	5	(260,287]	0	0	173	8	(172,187]	0	0	1	[0,0.14]
195	211	3	(209,234]	0	2	131	1	[131,132]	0	1.5	3	(1.26,2.15]
118	183	1	[177,186]	0	0	187	8	(172,187]	0	1.4	3	(1.26,2.15]
299	264	5	(260,287]	0	0	132	2	(132,142]	0	1.2	2	(0.14,1.26]
229	206	2	(186,209]	0	2	131	1	[131,132]	1	0	1	[0,0.14]
244	234	3	(209,234]	0	0	145	3	(142,151]	0	2.6	4	(2.15,4]
302	236	4	(234,260]	0	2	174	8	(172,187]	0	0	1	[0,0.14]
153	564	6	(287,564]	0	2	160	5	(156,161]	0	1.6	3	(1.26,2.15]
90	256	4	(234,260]	0	2	149	3	(142,151]	0	0.5	2	(0.14,1.26]
91	564	6	(287,564]	0	2	151	4	(151,156]	0	0.4	2	(0.14,1.26]
256	209	2	(186,209]	0	0	173	8	(172,187]	0	0	1	[0,0.14]
197	234	3	(209,234]	1	2	131	1	[131,132]	0	0.1	1	[0,0.14]
26	219	3	(209,234]	0	0	158	5	(156,161]	0	1.6	3	(1.26,2.15]
7	268	5	(260,287]	0	2	160	5	(156,161]	0	4	4	(2.15,4]
289	221	3	(209,234]	0	2	163	6	(161,166]	0	0	1	[0,0.14]
254	295	6	(287,564]	0	2	157	5	(156,161]	0	0.6	2	(0.14,1.26]
211	215	3	(209,234]	0	0	170	7	(166,172]	0	0	1	[0,0.14]
78	564	6	(287,564]	0	2	142	3	(142,151]	0	1.5	3	(1.26,2.15]
81	208	2	(186,209]	0	2	148	3	(142,151]	1	4	4	(2.15,4]
43	564	6	(287,564]	0	0	162	6	(161,166]	0	0.4	2	(0.14,1.26]

模型設定及處理

- XGBoost
- Logistic Regression

模型設定及處理 – XGBOOST

- 特徵選取與特徵轉換：
  - 將 slope 轉換成「binary」型式
  - 將 thalium_scan 轉換成「one-hot encoding」
  - 利用 xgb.importance 的套件，找出較重要的特徵，並訓練模型
- 實驗設定：
  - Cross Validation：20

Feature	Gain	Frequency
thalium_scan	0.275342017	0.01639344
chest_pain	0.133716699	0.06557377
cholestoral	0.114106357	0.19672131
exercise_angina	0.010420478	0.01639344
High_sugar	0.007573798	0.01639344

模型設定及處理 – LOGISTIC REGRESSION

- 特徵選取與特徵轉換：
  - 原始資料：
    - 根據全部資料，利用 student t test 來找到對應的 t 值（P value），選取重要特徵：
      - chest_pain
      - vessels
      - thalium_scan

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.198600   4.765700  -1.720 0.085372 .
age            -0.003836   0.043680  -0.088 0.930025
sex             1.204706   0.847649   1.421 0.155249
chest_pain      0.631230   0.319222   1.977 0.047997 *
resting_bp      0.035300   0.019485   1.812 0.070037 .
cholestoral     0.004815   0.007044   0.684 0.494254
high_sugar     -0.244334   0.853651  -0.286 0.774708
ecg             0.377170   0.315155   1.197 0.231394
max_rate       -0.029827   0.017083  -1.746 0.080815 .
exercise_angina -0.055663   0.689874  -0.081 0.935691
st_depression   0.302998   0.404717   0.749 0.454058
slope          -0.235356   0.639480  -0.368 0.712842
vessels         1.402988   0.404926   3.465 0.000531 ***
thalium_scan    0.537693   0.179171   3.001 0.002691 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 181.047  on 132  degrees of freedom
Residual deviance:  81.029  on 119  degrees of freedom
AIC: 109.03
```

模型設定及處理 – LOGISTIC REGRESSION

- 特徵選取與特徵轉換：
  - 轉換後資料（非最終模型）：
    - One hot encoding
      - 各別考慮 chest_pain, ecg, , thalium_scan 三變數：0/1
    - Age：
      - cut into bins => nonsignificant
    - linear combinations
      - examine p-value

模型設定及處理 – LOGISTIC REGRESSION

- 特徵選取與特徵轉換：
  - 轉換後資料（最終模型）：
    - 以decision tree, 拿沒有 NA 的 column 項來預測遺失值
      - Vessels ； thalium_scan
    - resting_bp_without_label：
      - 用kmeans 跑區間，最後分出5個區間：(93, 113, 133, 160, 192)

```
# fill na's (vessels,thalium_scan)
vessels.model <- rpart(vessels ~ age+sex+chest_pain+resting_bp+cholestoral+high_sugar
                       +ecg+max_rate+exercise_angina+st_depression+slope+heart_disease
                       ,data=train[!is.na(train$vessels), ],method='class')

train$vessels[is.na(train$vessels)] <-predict(vessels.model, newdata=train[is.na(train$vessels),]
                                              ,type = "class")

thalium_scan.model <- rpart(thalium_scan ~ age+sex+chest_pain+resting_bp+cholestoral+high_sugar+ecg
                            +max_rate+exercise_angina+st_depression+slope+vessels+heart_disease
                            , data=train[!is.na(train$thalium_scan), ], method='class')

train$thalium_scan[is.na(train$thalium_scan)] <- predict(thalium_scan.model
                                                         , newdata=train[is.na(train$thalium_scan),]
                                                         ,type="class")
```

模型設定及處理 – LOGISTIC REGRESSION

- 轉換後資料（最終模型，承上）：
  - slope_label：
    - origin category: 0,1,2
    - Since distribution like the table
      - category = 0 => 0 ； category = 1,2 => 1
  - thalium_scan
    - Origin: 3,6,7 => transform to 1,2,3
    - One – hot encoding
      - 雖然有三類，但實際上只需要用兩類來概括資料：thalium_scan_1 + thalium_scan_3 as feature
- 實驗設定：
  - Cross Validation：5~10
  - 根據 proposed-final 模型，我們設不同的 seed

	slope	Heart_disease
0	1	0.283
1	2	0.657
2	3	0.500

模型設定及處理

小結：

- 對於此次比賽我們使用模型有：XGBoost / Logistic Regression
- 針對模型訓練，我們採取以下方式：
  - 利用 xgb.importance 找出較重要的特徵並訓練模型
  - 相對於 Null Model，我們有針對飽和模型進行 Baseline 模型之前的訓練，並利用 T Test 找到較重要的特徵後，進行該特徵的強化與處理
  - 用 try and error 的精神，根據不同且較重要的特徵選取（結合），進行訓練，挑選較好的模型
  - 根據 proposed-final 模型，我們設不同的 seed 來訓練模型
- 我們使用 K-Fold Cross Validation 來進行模型驗證

實驗結果 – LOGISTIC REGRESSION

model	Item set	Private Score	Public Score
XGBoost	thalium_scan、chest_pain、cholestoral、exercise_angina、High_sugar	0.86792	0.90909
Baseline（特徵未經處理）	chest_pain、vessels、thalium_scan、resting_bp、max_rate	0.86363	0.83018
Proposed-A（Slope_label 改為二類）	resting_bp_without_label、slope_label、vessels、chest_pain、thalium_scan	0.81132	0.90909
Proposed-C（chest_pain 改為 one-hot encoding）	resting_bp_without_label、slope_label、vessels、chest_pain_2、chest_pain_3、chest_pain_4、thalium_scan	0.77358	0.86363
Proposed-Final（thalium_scan 改為 one-hot encoding）	resting_bp_without_label、slope_label、vessels、chest_pain、thalium_scan_1、thalium_scan_3	0.86792	0.90909

Accuracy is not always the best choice to choose the test mode !!!!

實驗結果 – LOGISTIC REGRESSION

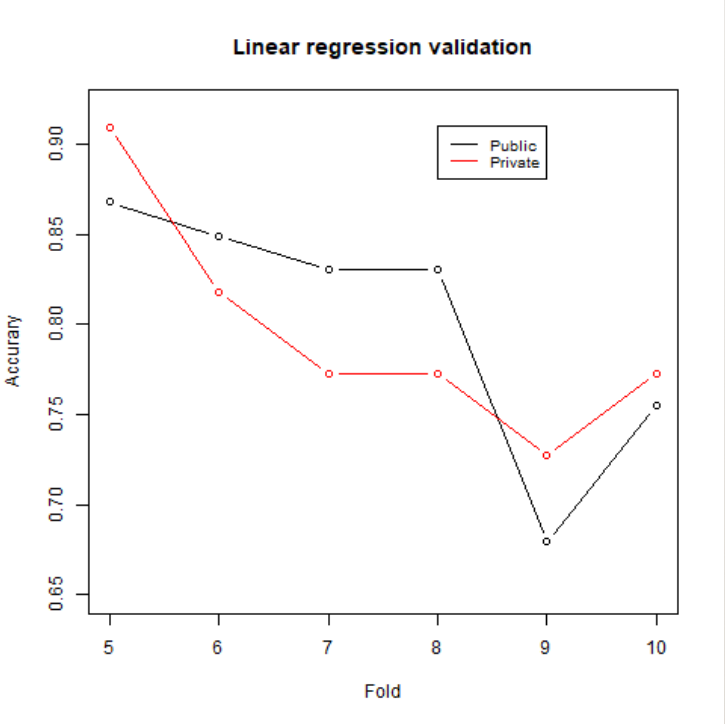
- 不同的 seed：

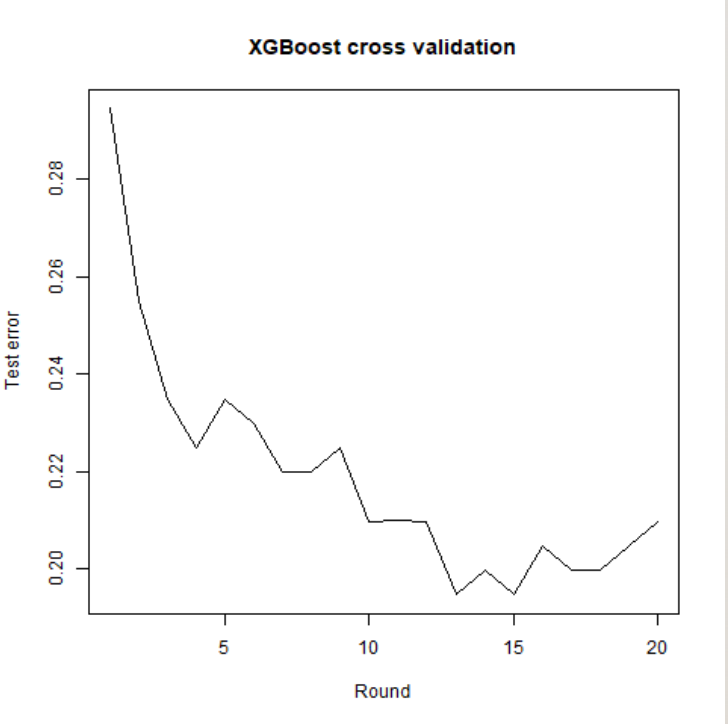
seed	public score	private score
4	0.86792	0.90909
10	0.83018	0.81818

predict_final.csv 34 minutes ago by 流 check - seed: 4 (prev: 10)	0.86792	0.90909

The way we cut into fold is important, it can be improved if we make some tricks before cut

實驗結果


Linear regression validation
Public
Private
Accurary
0.65
0.70
0.75
0.80
0.85
0.90
5
6
7
8
9
10
Fold


XGBoost cross validation
Test error
0.20
0.22
0.24
0.26
0.28
5
10
15
20
Round

實驗結果

小結：

- 我們使用了 Accuracy, Test Error 來進行衡量，以找到最佳之模型
- 針對不同的資料處理及標籤設定，會有不同訓練結果。加上我們採取不同標籤的結合來訓練模型，以及改變 seed 也會影響結果。綜合以上，我們最後且最佳模型有很大進步
- 對於此次比賽的結果， 我們認為資料的特徵工程及選擇較為困難，因為不知道何者是最重要且必要的特徵。此外，訓練時的抽樣方法也需多加要研究

參考出處

- Age interval（年齡區間劃分）
  - 出自《老年性生理學和老年的性生活》一書
- Blood pressure interval（血壓區間劃分）
- Cholestoral interval（總膽固醇區間劃分）
  - 啟新診所
  - 馬偕醫院
- 資料處理：
  - https://www.datacamp.com/community/tutorials/contingency-tables-r
  - https://www.pluralsight.com/guides/cleaning-up-data-from-outliers
  - https://www.gastonsanchez.com/r4strings/formatting.html
  - https://www.guru99.com/r-data-frames.html

參考出處

- 套件引用：
  - https://cran.r-project.org/web/packages/hash/hash.pdf
  - https://stackoverflow.com/questions/23765996/get-all-keys-from-ruby-hash
  - https://www.rdocumentation.org/packages/tibble/versions/1.4.2/topics/add_column
  - https://stackoverflow.com/questions/45741498/add-column-in-tibble-with-variable-column-name
  - https://statmath.wu.ac.at/projects/vcd/
  - https://rdrr.io/cran/infotheo/man/mutinformation.html
  - https://cran.r-project.org/web/packages/infotheo/infotheo.pdf
  - https://www.rdocumentation.org/packages/stringr
  - https://stringr.tidyverse.org/reference/str_detect.html
  - https://www.rdocumentation.org/packages/vcd/

ANY QUESTIONS

END