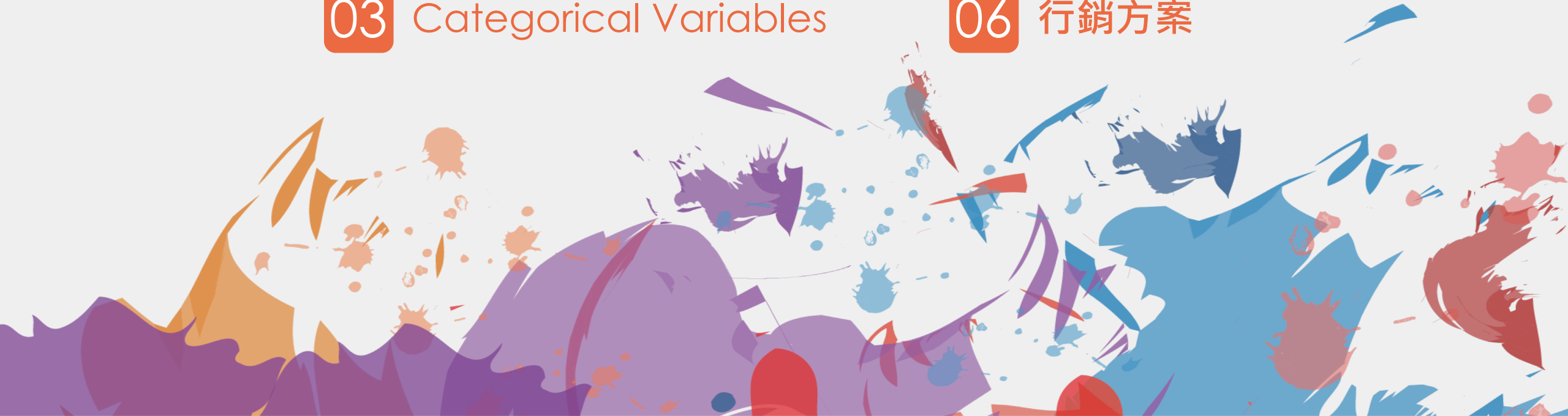
108-1 Statistical Data Analysis for Business and Management Final Project

AsiaYo之新客存活率分析

Group#1
林祐萱 張逸安
陳映蓁 蔡岳勳
鄭子萱 蕭維嘉

CONTENT



PART 01

研究目的&脈絡

研究目的

Customers Retention 是品牌的一個重要指標，有效的瞭解公司用戶並且利用 RFM 做分類並對其進行數據分析，並制定營銷策略。

Recency

Frequency

Monetary

切分出不同客群，並找尋其對應的回購週期（如：3個月）

判斷客群的回購率

判斷客群 Retention Rate

並依此制定相關Marketing Plan

研究目的

Customers Retention 是品牌的一個重要指標，有效的瞭解公司用戶
並且利用 RFM 做分類並對其進行數據分析，並制定營銷策略。

Recency

Frequency

Monetary

切分出不同客群，並找尋其對應的回購週期（如：3個月）

1. 建立**預測回購週期(T)MLR模型**，並代回預測新客的T，以利計算**RT ratio**
2. 建立**判斷留存於平台機率的GLM模型**，並代回預測新客的**存活率**

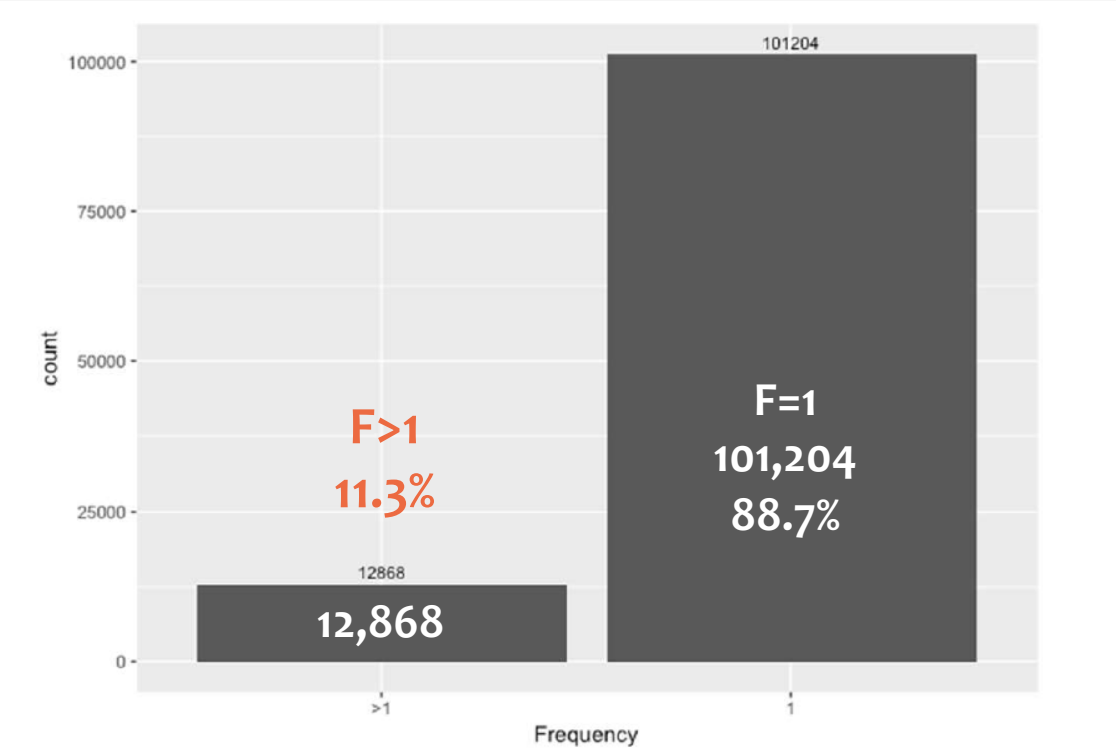
每個人的購物週期不同
無法預測新客的週期，所以無法判定客戶是否會回購

判斷客群的回購率

判斷客群 Retention Rate

並依此制定相關Marketing Plan

研究目的



圖：客戶在平台消費頻率分布

- 定義
  F>1為舊客
  F=1為新客*
- AsiaYo平台的客戶，**新客為舊客的8倍**

→在意新客是否能留於平台繼續消費

*由於AsiaYo於2014年成立，但資料僅起於2018.01.01，因此在此先暫時定義新客為這兩年間消費無超過一次的客戶

研究目的

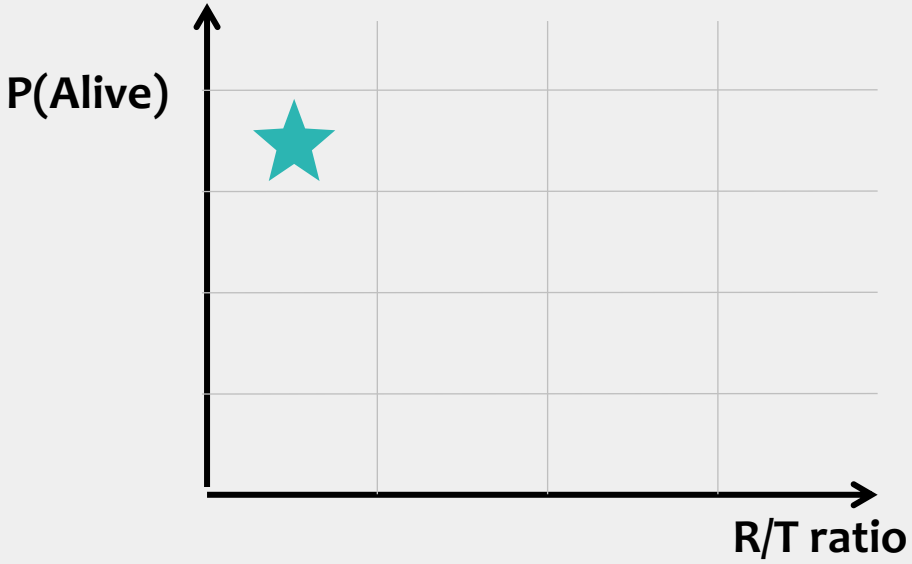
利用舊客的資料來回推新客可能的習性

- 建立**預測回購週期(T)MLR模型**，並代回預測新客的T，以利計算RT ratio

$$\frac{R}{T}\ ratio$$

- 若 ≤1，表示該客戶回購機率高
- 若比值越大，表示客戶想離開或已經離開AsiaYo平台

- 建立**判斷留存於平台機率的GLM模型**，並代回預測新客的存活率

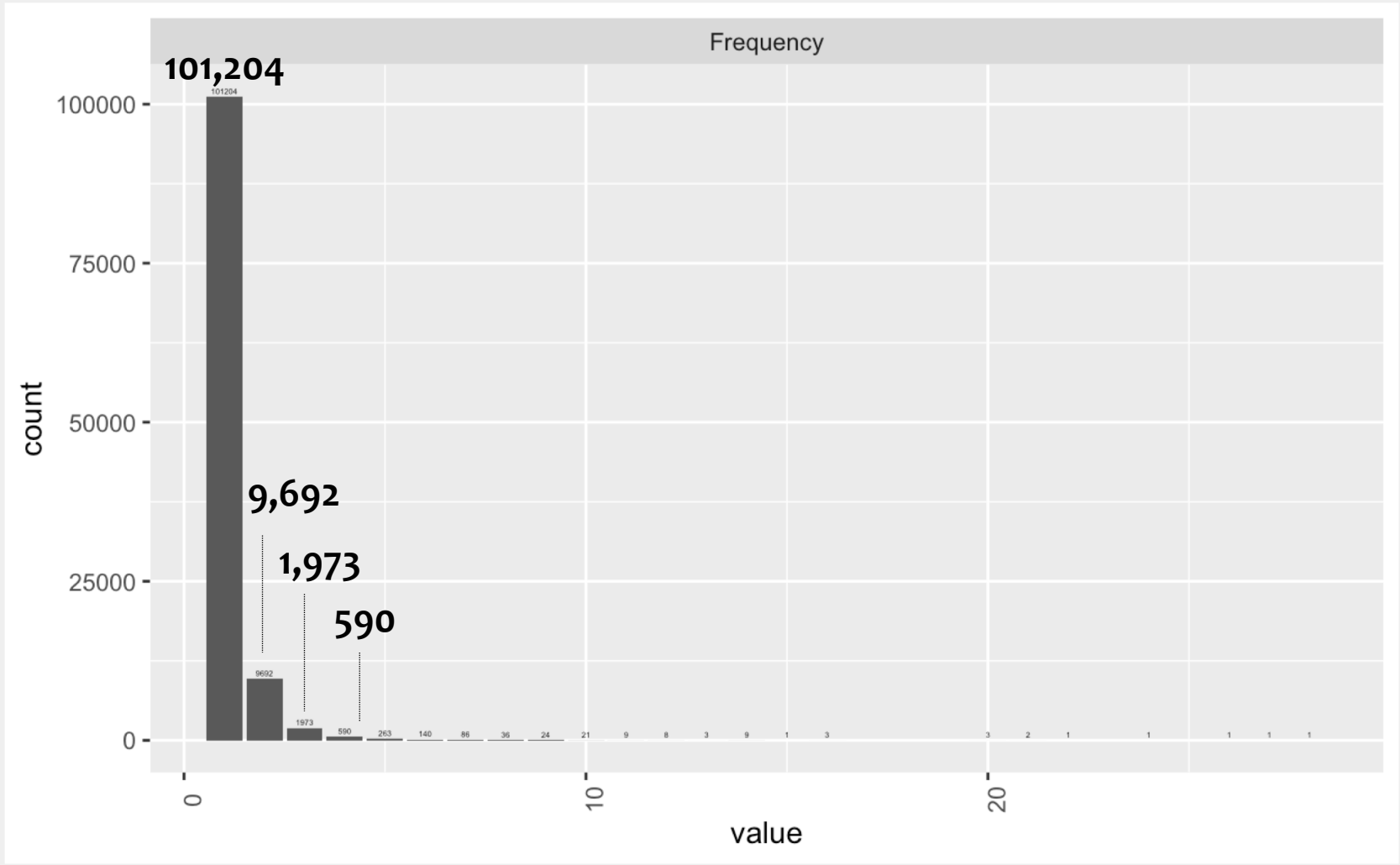


→本組改以RT ratio 與存活率來代表retention rate
並藉此區分新客給予行銷建議

研究脈絡

1. **資料前處理**：用百分比的方式，將「訂單」資料轉換回「用戶」特徵資料集
2. **為用戶貼上標籤**：將用戶特徵資料集的百分比轉換成新的類別變數，為用戶貼上標籤
3. **預測回購週期**：使用MLR模型來預測新客的回購週期
4. **預測存活率**：使用GLM模型來預測新客的存活率
5. **行銷建議**：依據舊客回購週期與存活率，制定優先溝通的客群

Appendix



圖：客戶在平台消費頻率分布

PART 02

資料前處理

資料前處理

- 原始資料
  - 訂單成立日期(order_date)區間：**2018-01-01 ~ 2019-11-09**
  - 共201,694筆訂單
  - 共146,261位不重複會員(user_id)
  - **One row per order id**

- 排除從b2b_partner來的訂單 (∵無法辨認個別消費者)
  - 剩餘**198,928筆訂單**
  - 剩餘**146,257位不重複會員(user_id)**

Numerical Variables (1)

One row per order id → One row per user id

	變數名稱	Description
R	*Recency*	最後一筆訂單日期距離tag date (2019-11-10)的天數。
F	*Frequency*	所有不重複日期數量。(本組假設用戶會在同一天將一次旅行所需的所有訂房完成)
	T	每筆訂單之間間隔的天數(訂購週期)之中位數。
M	*TtlContri*	所有臺幣金額加總。 *Σ (twd_amount)*
	Monetary	平均**每次旅行**的花費。 *(TtlContri / Frequency)*
	AvgPmt	平均**每人每晚**的花費。 *Avg[twd_amount / (Guest * Night)]*
	TtlQpon	總共在訂房時使用Coupon次數。
	DailyQpon	平均每次旅行在訂房時使用Coupon次數。 *(TtlQpon / Frequency)*
	TtlNote	總共在訂房時撰寫文字附註次數。
	DailyNote	平均每次旅行在訂房時撰寫文字附註次數。 *(TtlNote / Frequency)*

Numerical Variables (2)

One row per order id → 機率 → One row per user id

- Book status (訂單狀態)

Accepted	Rejected	Cancelled	Unconfirmed
5	3	2	1

共 5+3+2=10 筆有效訂單

$\times \frac{1}{10}$

PctAcc	PctRej	PctCan
0.5	0.3	0.2

- Platform (操作平台)

Android	iOS	Mobile	PC
2	0	1	2

共 2+0+1+2= 5 筆Accepted訂單

$\times \frac{1}{5}$

Android (Pct)	iOS (Pct)	Mobile (Pct)	PC (Pct)
0.4	0	0.2	0.4

- Guest (人數)
- Child (小孩人數)
- Rooms (房間數)
- Nights (過夜數)
- Bnb Country (旅宿國家)
- Bnb Type (旅宿類型)
- Order week day (星期幾下單)
- Order time (下單時段)
- Check-in Month (入住月份)
- Reservation (提前多久訂房) (Order date ~ Check-in date)

PART 03

Categorical variable

類別變數與各類別命名邏輯(1/3)

■ 新增共10個類別變數：

	類別變數	意義
1.	*NightType*	用戶訂房偏好停留幾晚
2.	*ChildType*	用戶是否有帶小孩出遊
3.	*GuestNum*	用戶偏好出遊的團體大小
4.	*Device*	用戶偏好使用何種管道在該平台訂房
5.	*BnbType*	用戶偏好訂購何種房型
6.	*DoW*	用戶偏好的訂房時間
7.	*OrderPre*	用戶偏好在入住時間多少天前訂房
8.	*Season*	用戶出遊季節偏好
9.	*Area*	用戶「出發地」所處的區域
10.	*Travel_dist*	用戶偏好的旅途距離長短

類別變數與各類別命名邏輯(2/3)

Levels命名邏輯

大於一定比例(50%)為一個Level，目的是希望看出顧客有沒有特別的偏好或習慣。

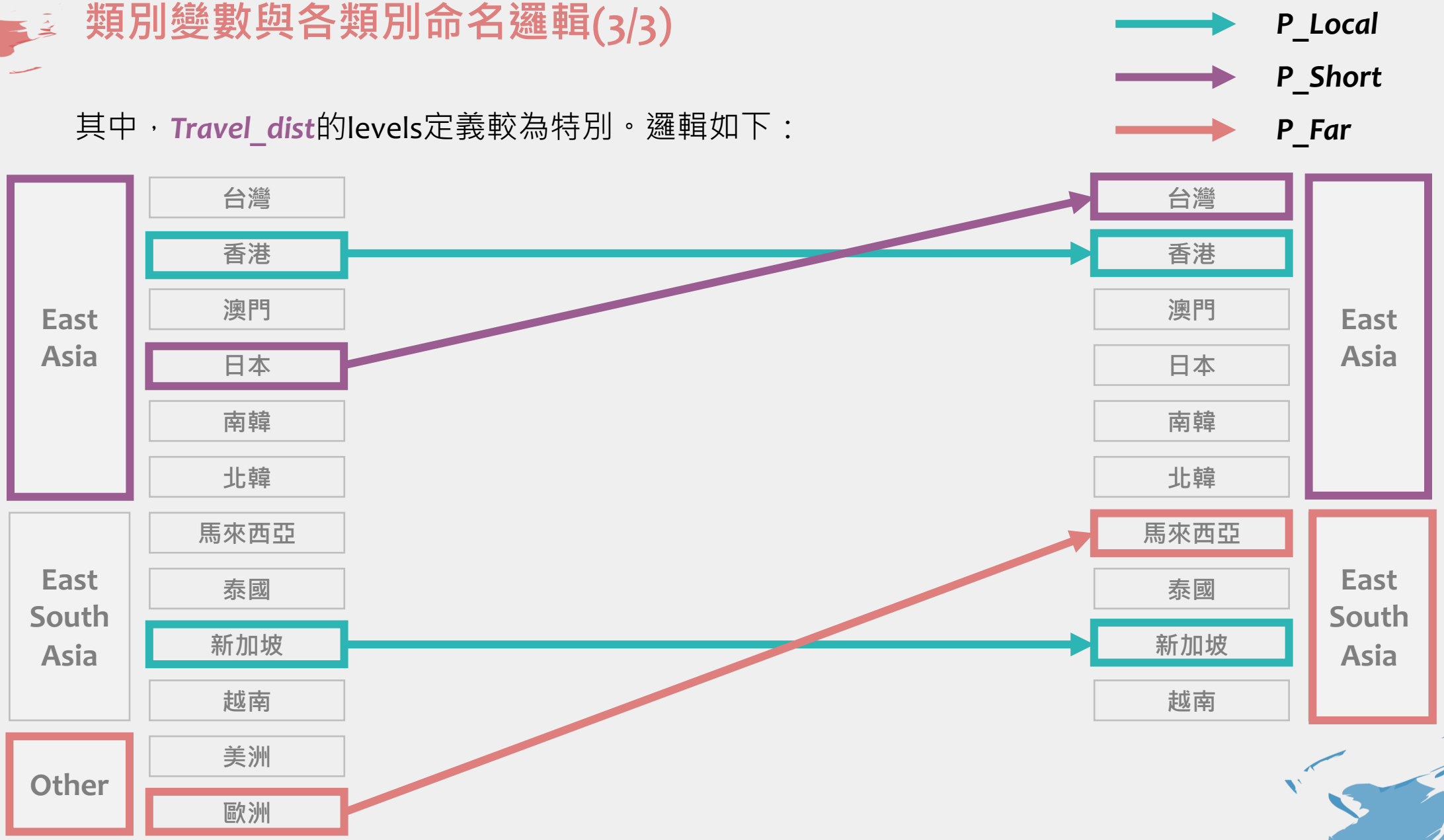
以 *GuestNum* 為例：

Levels of GuestNum	*Rules*
Single	PctGuest1大於等於50%
Double	PctGuest2大於等於50%
SmallGroup	PctGuest3+ PctGuest4大於等於50%
MediumGroup	PctGuest5+ PctGuest6+ PctGuest7大於等於50%
LargeGroup	PctGuest8+ PctGuest9+ PctGuest10大於等於50%
SuperLargeGroup	PctGuest11（含）以上若大於等於50%
Neutral	其他不包含以上分類者

Device, *ChildType*, *GuestNum*, *Device*, *BnbType*, *DoW*, *Area*, *Season*, 也是遵照此邏輯。

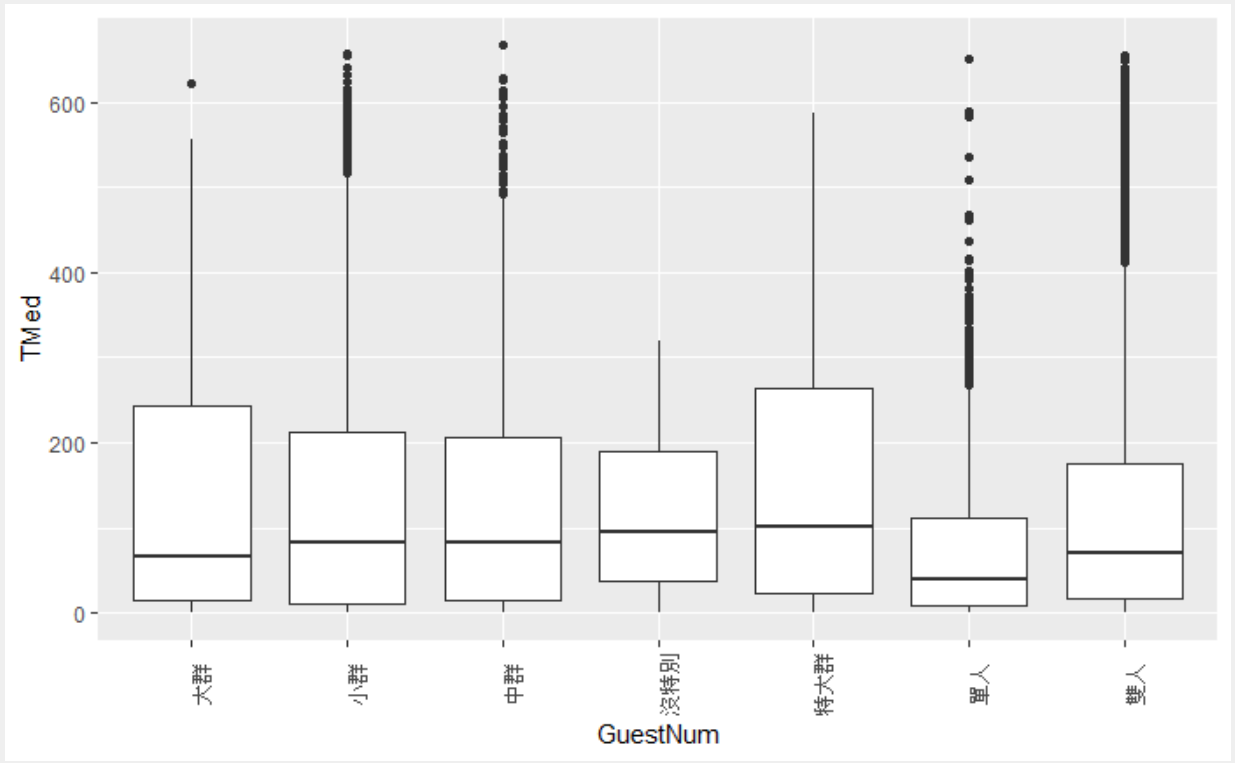
類別變數與各類別命名邏輯(3/3)

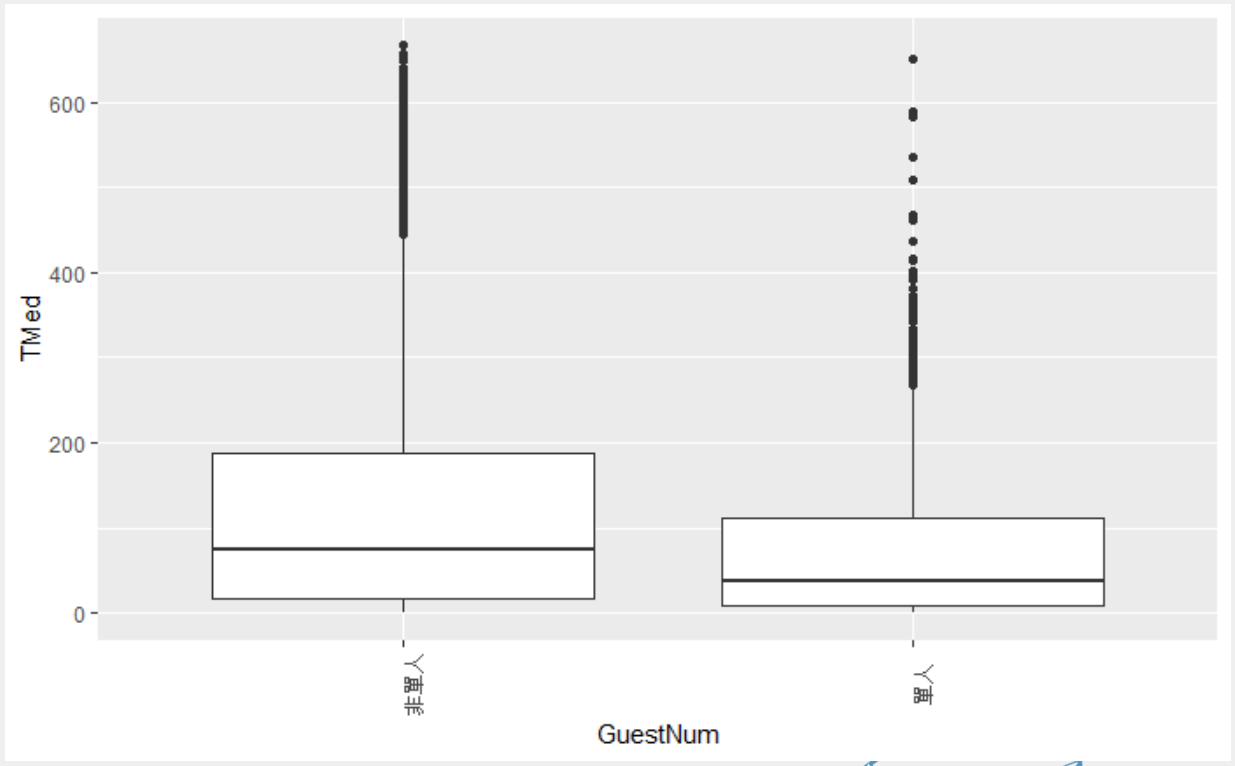
其中，*Travel_dist*的levels定義較為特別。邏輯如下：



調整變數：合併與重新命名

- 最初變數levels的劃分是以本組的想法與直覺，有**切分過細**的疑慮。
- 將各個變數與 ***T (訂購週期)*** 繪製**Boxplot**與**Multiple Comparison**了解levels切分狀況，並思考若合併是否會不合邏輯，最後得到目前的類別。
- 合併方式，舉 ***GuestNum*** 之**Boxplot**為例：





新變數對Tmed影響：預期與實際比較(1/4)

■ 下表為本組最初設計變數時預期變數對 ***T (訂購週期)*** 的影響以及實際模型回歸後比較：

- Reference：每個變數中沒有特別傾向Level
- 正負號數量代表對 ***T (訂購週期)*** 的影響大小
- ▵表示不確定影響程度，為與Reference比較的相對程度

變數	levels	預期對 T 影響	實際對 T 影響
NightType	most_1~3	+	-
	most_4up	+++	+
	other	Reference	
ChildType	a_child	+++	---
	xa_child	Reference	
GuestNum	Single	+	-
	nonsingle	Reference	

新變數對Tmed影響：預期與實際比較(2/4)

變數	levels	預期對 T 影響	實際對 T 影響
Device	app	---	---
	mobile	---	--
	pc	-	-
	neutral	Reference	
BnbType	apt	-	+++
	else.apt	△	++
	neutral	Reference	
Dow	weekday	-	+
	weekend	-	+++
	xspecial	Reference	

新變數對Tmed影響：預期與實際比較(3/4)

變數	levels	預期對 T 影響	實際對 T 影響
Season	spring	+++	---
	summer	+++	-
	autumn	+++	-
	winter	+++	+++
	xspecial	Reference	
OrderPre	near	Reference	
	normal	++	---
	normal+	++	+++
	early	+++	---
	early+	+++	-

新變數對Tmed影響：預期與實際比較(4/4)

變數	levels	預期對 T 影響	實際對 T 影響
Area	EastAsia	△	++
	EastSouthAsia	△	+
	Others	Reference	
Travel_dist	P_Local	---	---
	P_Short	-	+++
	P_Far	+++	+++
	No Tendency	Reference	

PART 04

Interaction Items

交互項

■ 不同變數之間存在交互關係，單一變數不足描述消費者樣態，
因此加入**交互項(Interaction Items)**

Interaction Items	*Description*
OrderPre : TravelDist	消費者傾向提前多久訂房，可能會因旅行距離遠近而異。 例如：越是長距離旅行，會越早提前訂房。
Season : TravelDist	消費者偏好旅行距離之遠近，可能會因季節不同而異。 例如：偏好在夏季進行遠程旅行，在冬季進行國內旅行。
Monetary : Area	消費者每次旅行之平均花費，可能會因旅行地點而異。 例如：東北亞旅行之花費可能會高於東南亞旅行之花費。

■ 由於變數之間可能有許多直覺上難以想到之交互關係，因此本組列出
各種變數組合後，以*step()*函數篩選，最終得出29個交互項

PART 05

MLR & GLM

MLR

Multiple Linear Regression

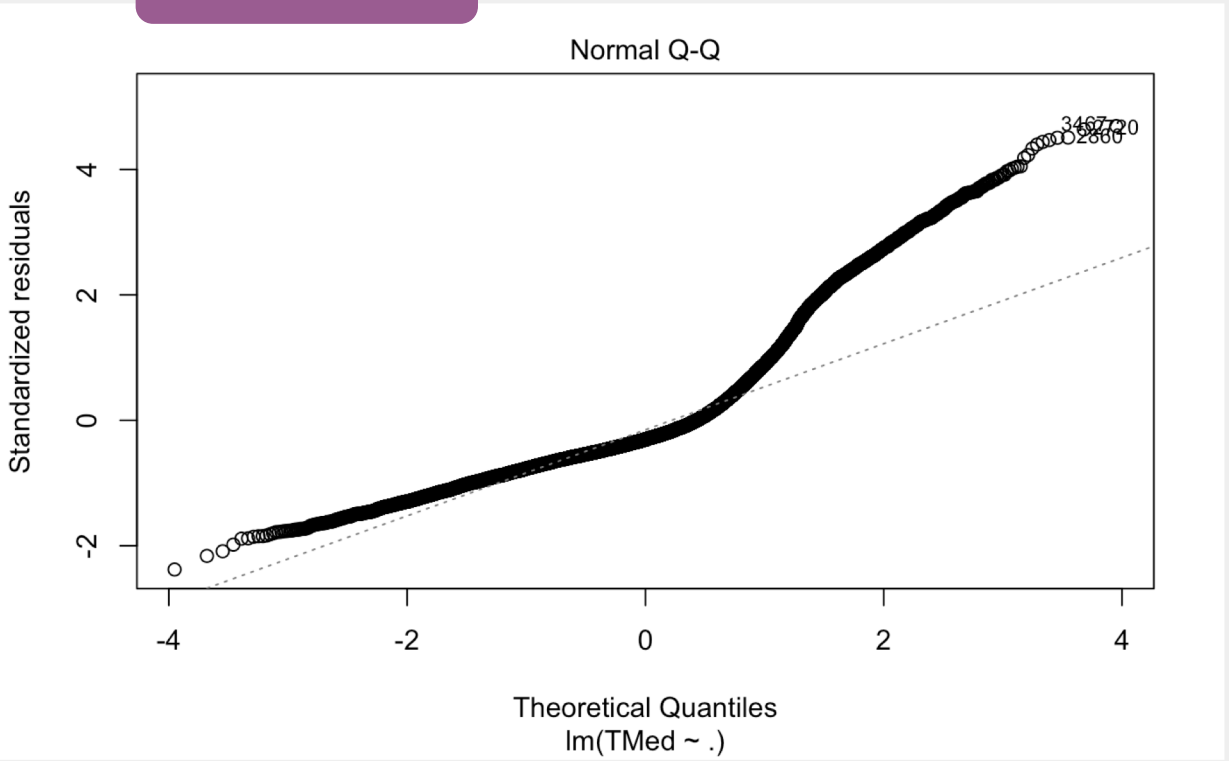
篩選舊客 (Frequency > 1) 的客戶資料

預測**顧客消費週期(T)**

MLR – Model 1

Y	X	
T	• Recency • ~~Frequency~~ • Monetary • AvgPmt • DailyQpon • DailyNote	• NightType • ChildType • GuestNum • Device • BnbType • DoW • OrderPre • Season • Area • TravelDist

違背模型假設



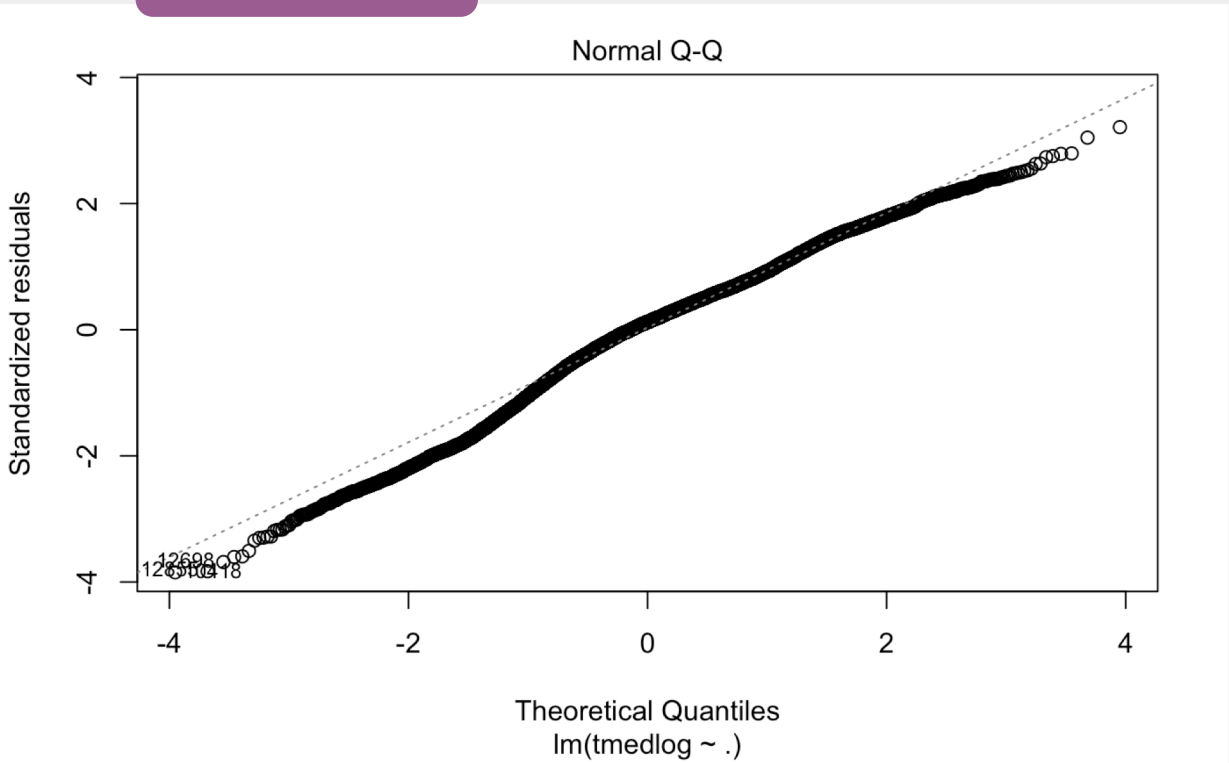
Residual standard error	112.8
R-squared	0.2862
Adjusted R-squared	0.2846

MLR – Model 2

Box-Cox 轉換

Y	X	
Log(T)	• *Recency* • ~~*Frequency*~~ • *Monetary* • *AvgPmt* • *DailyQpon* • *DailyNote*	• *NightType* • *ChildType* • *GuestNum* • *Device* • *BnbType* • *DoW* • *OrderPre* • *Season* • *Area* • *TravelDist*

符合模型假設



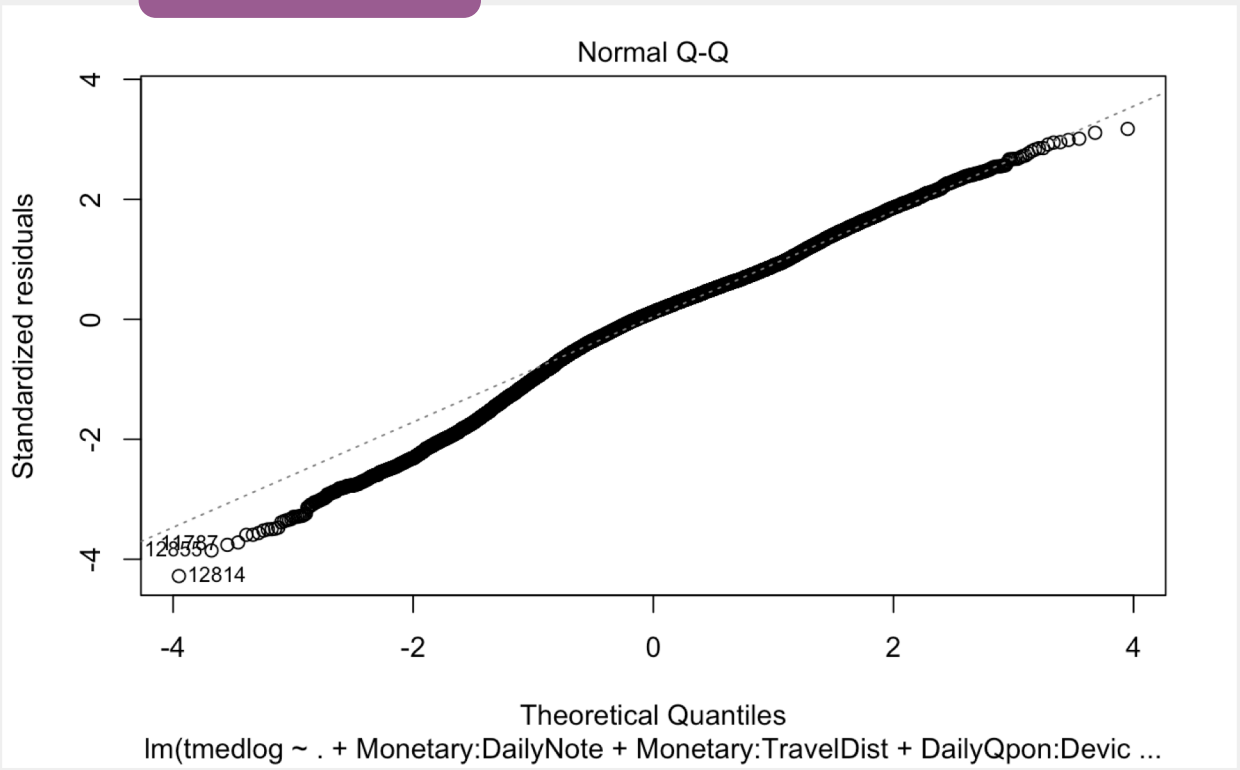
Residual standard error	1.432
R-squared	0.3721
Adjusted R-squared	0.3707

MLR – Model 3

加入交互項、增加模型解釋能力

Y	X	
Log(T)	• *(原變數) +* • *Monetary:DailyNote* • *Monetary:TravelDist* • *DailyQpon:Device* • *DailyQpon:BnbType* • *DailyQpon:Season* • *NightType:TravelDist* • *ChildType:Season* • *GuestNum:OrderPre* • *Device:DoW* • *OrderPre:Season* • *OrderPre:TravelDist* • *Season:TravelDist* • *Monetary:AvgPmt* • *Monetary:GuestNum*	• *Monetary:Area* • *Monetary:Season* • *AvgPmt:DailyQpon* • *AvgPmt:Season* • *AvgPmt:Area* • *DailyQpon:DailyNote* • *DailyQpon:TravelDist* • *DailyNote:GuestNum* • *DailyNote:Device* • *DailyNote:Season* • *DailyNote:Area* • *GuestNum:TravelDist* • *Device:OrderPre* • *DoW:Season* • *Area:TravelDist*

符合模型假設



Model 3 is better !

	Model 3	Model 2
Residual standard error	1.383	1.432
R-squared	0.4203	0.3721
Adjusted R-squared	0.4133	0.3707

GLM

Generalized Linear Model

篩選舊客戶(Frequency > 1)的客戶資料
預測**顧客存活率**

GLM模型架構

定義顧客的
死亡與存活狀態

定義
RTRatio

Binomial Logistic
Regression Model
顧客存活模型

GLM模型架構

顧客特性

- Asiayo顧客屬於**非契約顧客(Non-Contractual Customer)**

RT Ratio

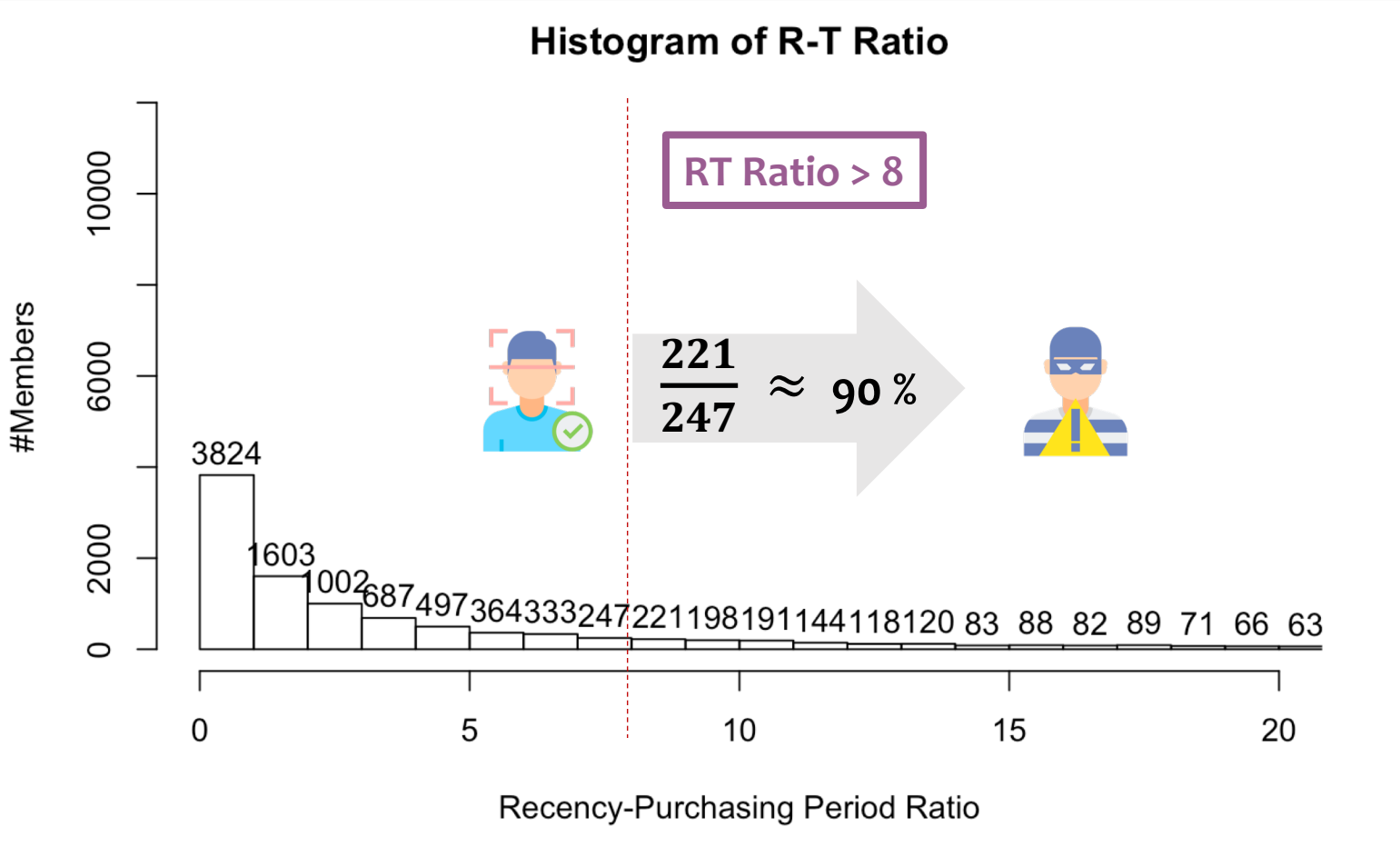
- 假設：每位顧客的回購行為都存在自己的購物週期

$$\text{RT Ratio} = \frac{\text{Recency (最近一次購買期間)}}{\text{T (訂房週期)}}$$

RTRatio 趨近於 1	RT Ratio 愈大
	
顧客回購機率越高	顧客可能已經想離開 或已經離開 AsiaYo

Status

如何設定判斷標準？



RT Ratio	Status	
> 8	0	死亡
< 8	1	存活

Model 1

Binomial Logistic Regression

Y	X	
Status	~~Recency~~ ~~Frequency~~ Monetary AvgPmt DailyQpon DailyNote	NightType ChildType GuestNum Device BnbType DoW OrderPre Season Area TravelDist

Summary

Null deviance	16411 on 12867 degrees of freedom
Residual deviance	12845 on 12839 degrees of freedom
AIC	12903

Good Fit Model Test

P – Value = 0.4827627

The Model is Good !

MLR – Model 3

加入交互項

Y	X	
Status	• (原變數) + • Monetary:DailyNote • Monetary:TravelDist • DailyQpon:Device • DailyQpon:BnbType • DailyQpon:Season • NightType:TravelDist • ChildType:Season • GuestNum:OrderPre • Device:DoW • OrderPre:Season • OrderPre:TravelDist • Season:TravelDist • Monetary:AvgPmt • Monetary:GuestNum	• Monetary:Area • Monetary:Season • AvgPmt:DailyQpon • AvgPmt:Season • AvgPmt:Area • DailyQpon:DailyNote • DailyQpon:TravelDist • DailyNote:GuestNum • DailyNote:Device • DailyNote:Season • DailyNote:Area • GuestNum:TravelDist • Device:OrderPre • DoW:Season • Area:TravelDist

Summary

	Model 1	DF	Model 2	DF
Null deviance	16411	12867	16411	12867
Residual deviance	12845	12839	12435	12714
AIC	12903		12743	

Model 2 is better !

Good Fit Model Test

P – Value = 0.960502

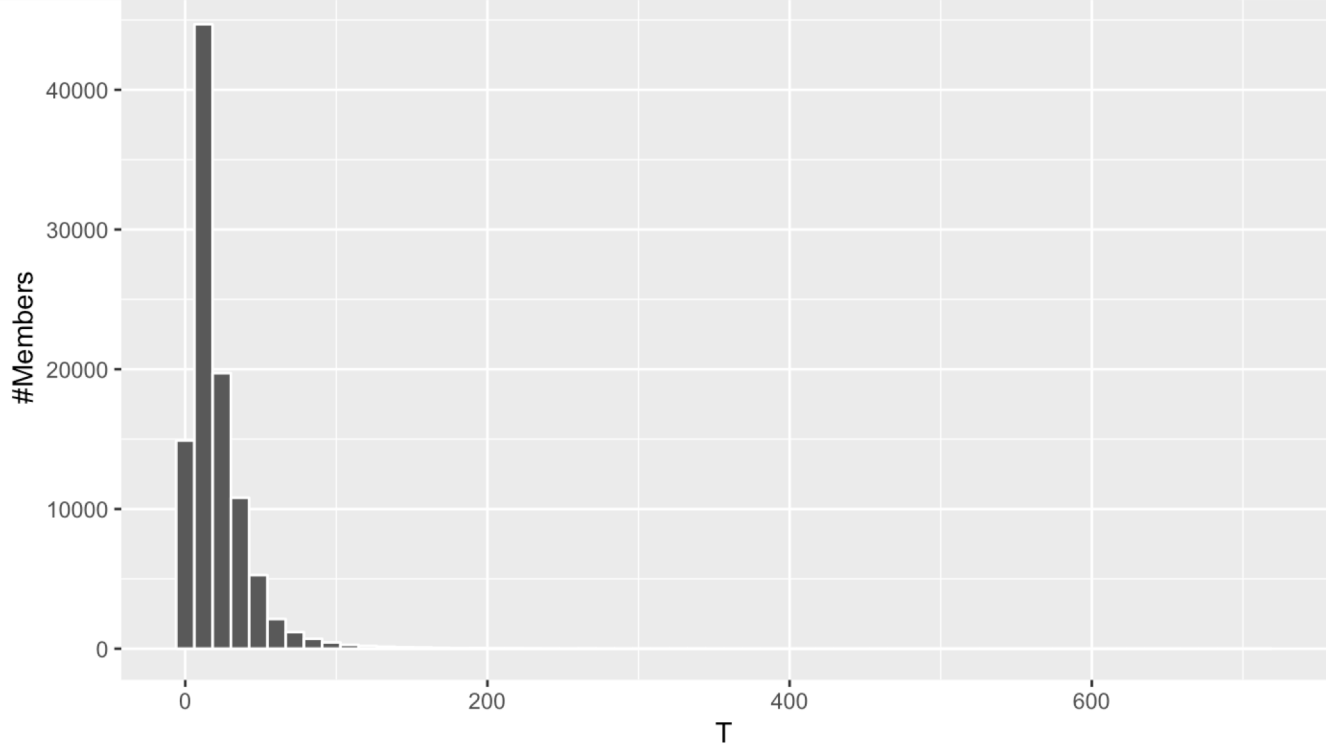
The Model is Good !

Prediction

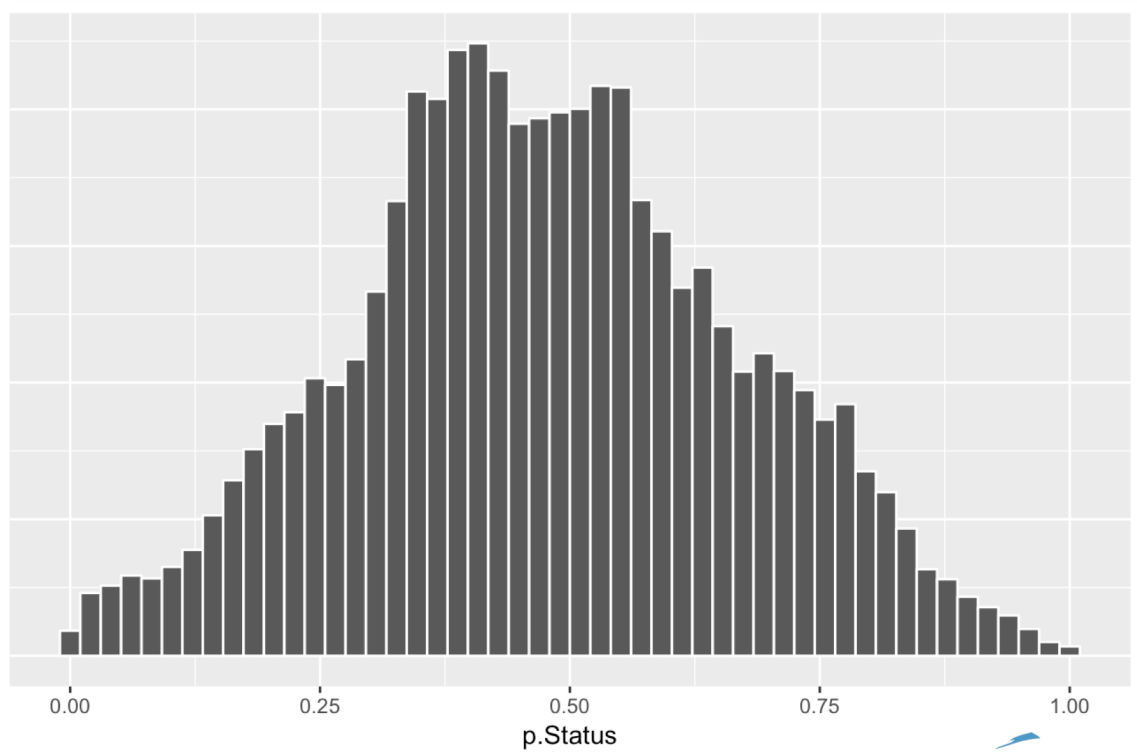
預測 **Frequency = 1** 的客戶

Frequency = 1 的顧客

MLR 預測顧客消費週期(T)



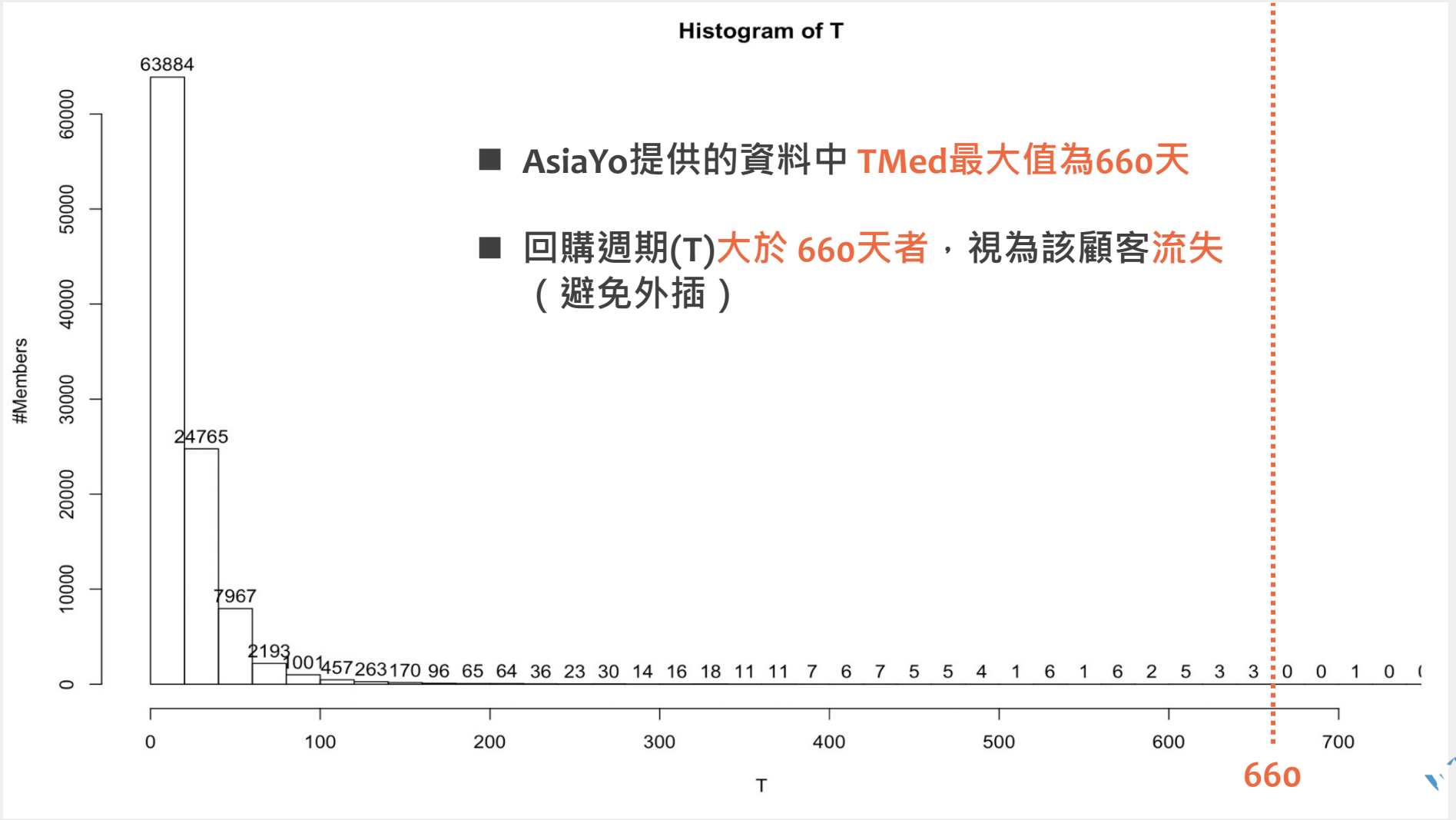
GLM 預測顧客存活率P(Status = 1)



PART 06

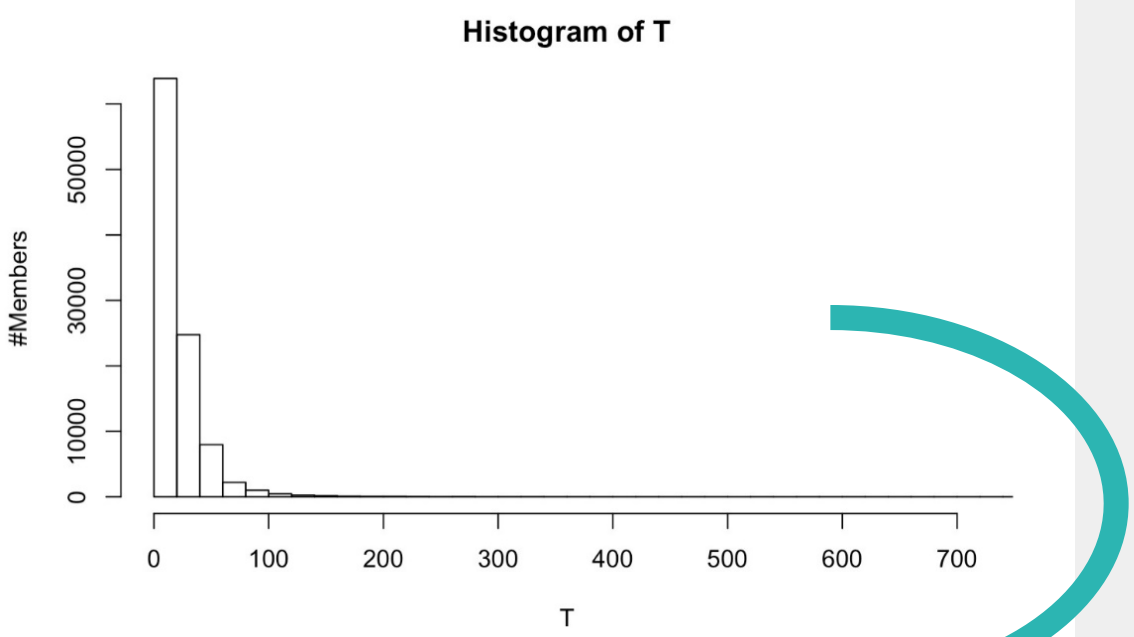
行銷建議

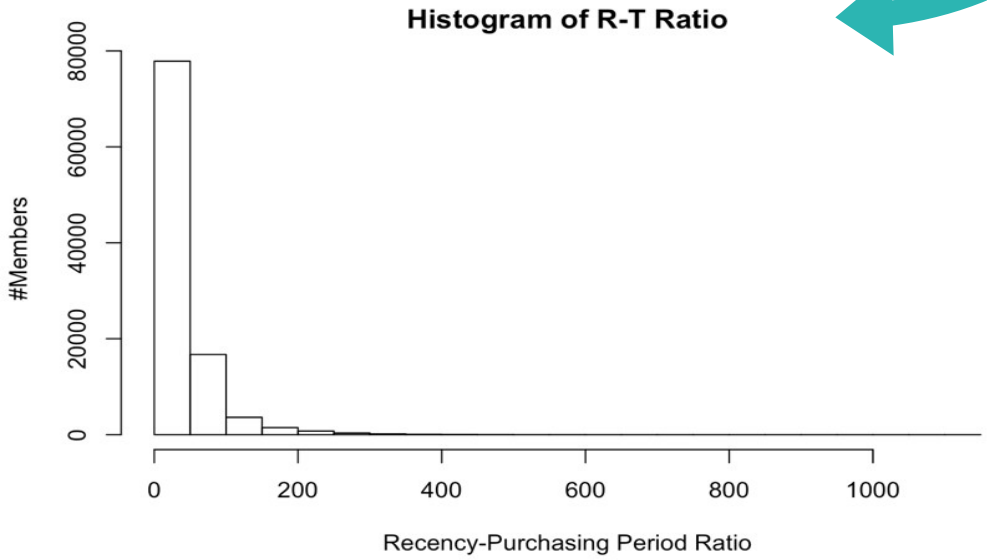
定義目標客群（1/2）



- AsiaYo提供的資料中 TMed最大值為660天
- 回購週期(T)大於 660天者，視為該顧客流失（避免外插）

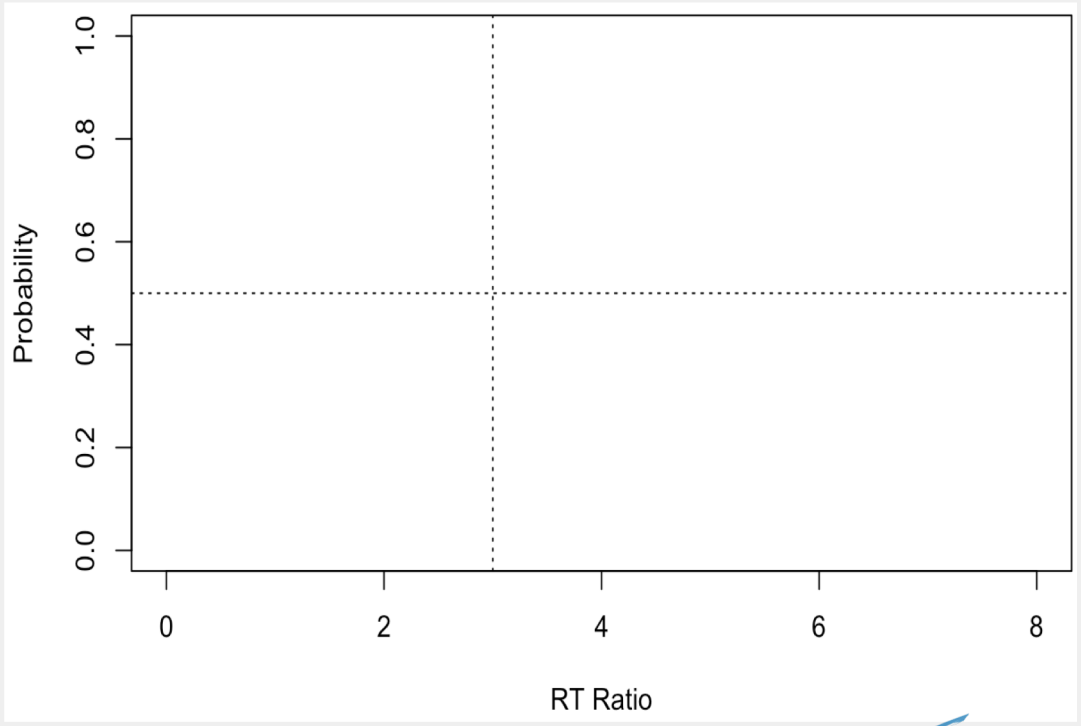
定義目標客群（2/2）





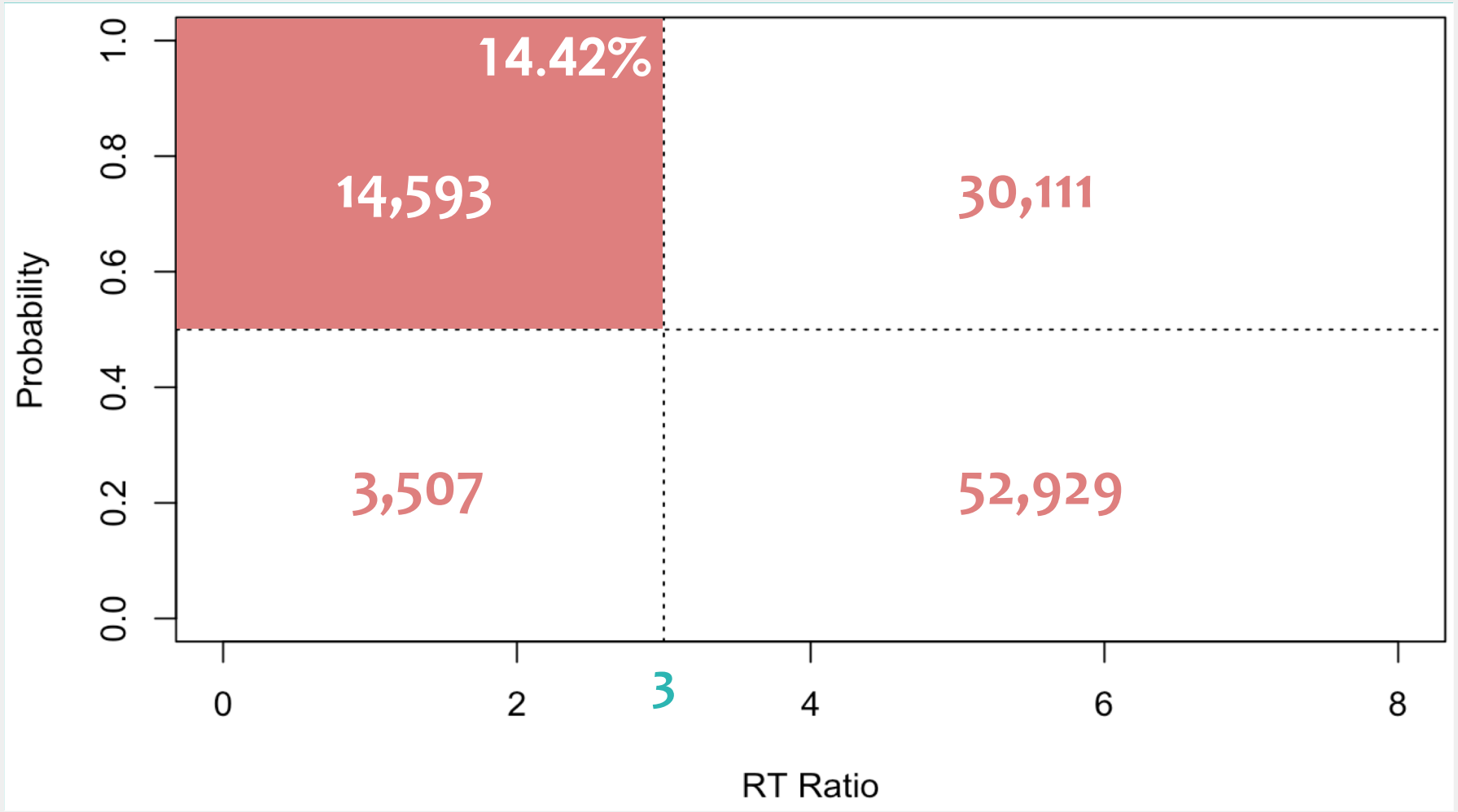
1. 將回購週期的人數轉換成**RT Ratio**

2. 結合 GLM預測的新客存活率
   - 對 **RT Ratio**及**存活率**作圖並定義分界

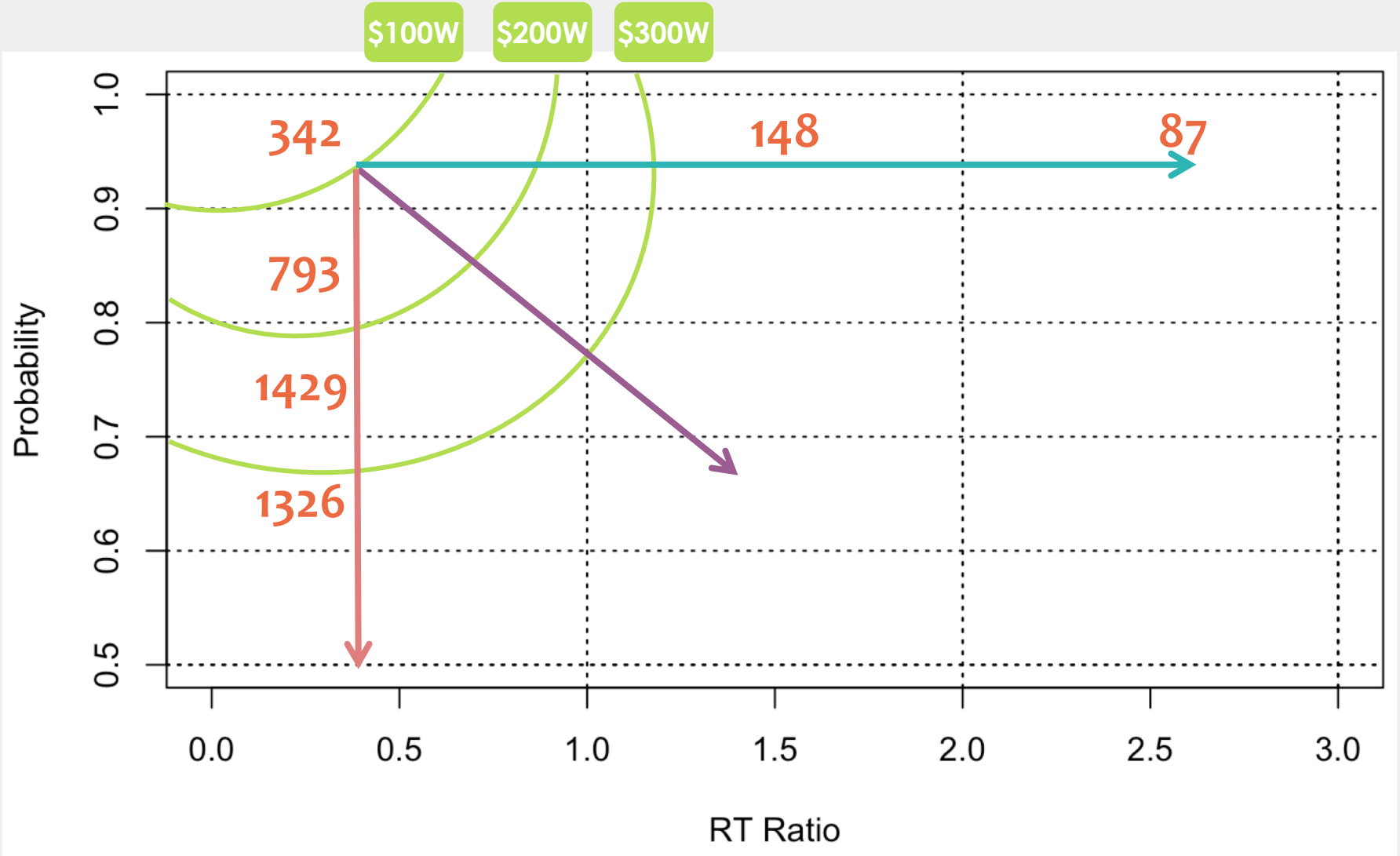


新客之客群分佈（存活機率 與 RT Ratio)

1. RT Ratio > 8 視為死亡（0.06%）
2. 行銷目標鎖定在 P > 0.5 且 RT Ratio < 3 (類似三倍標準差）



行銷目標客群（ P > 0.5 & RT Ratio < 3)



RT Ratio比較重要

針對不同RT Ratio做行銷

存活率比較重要

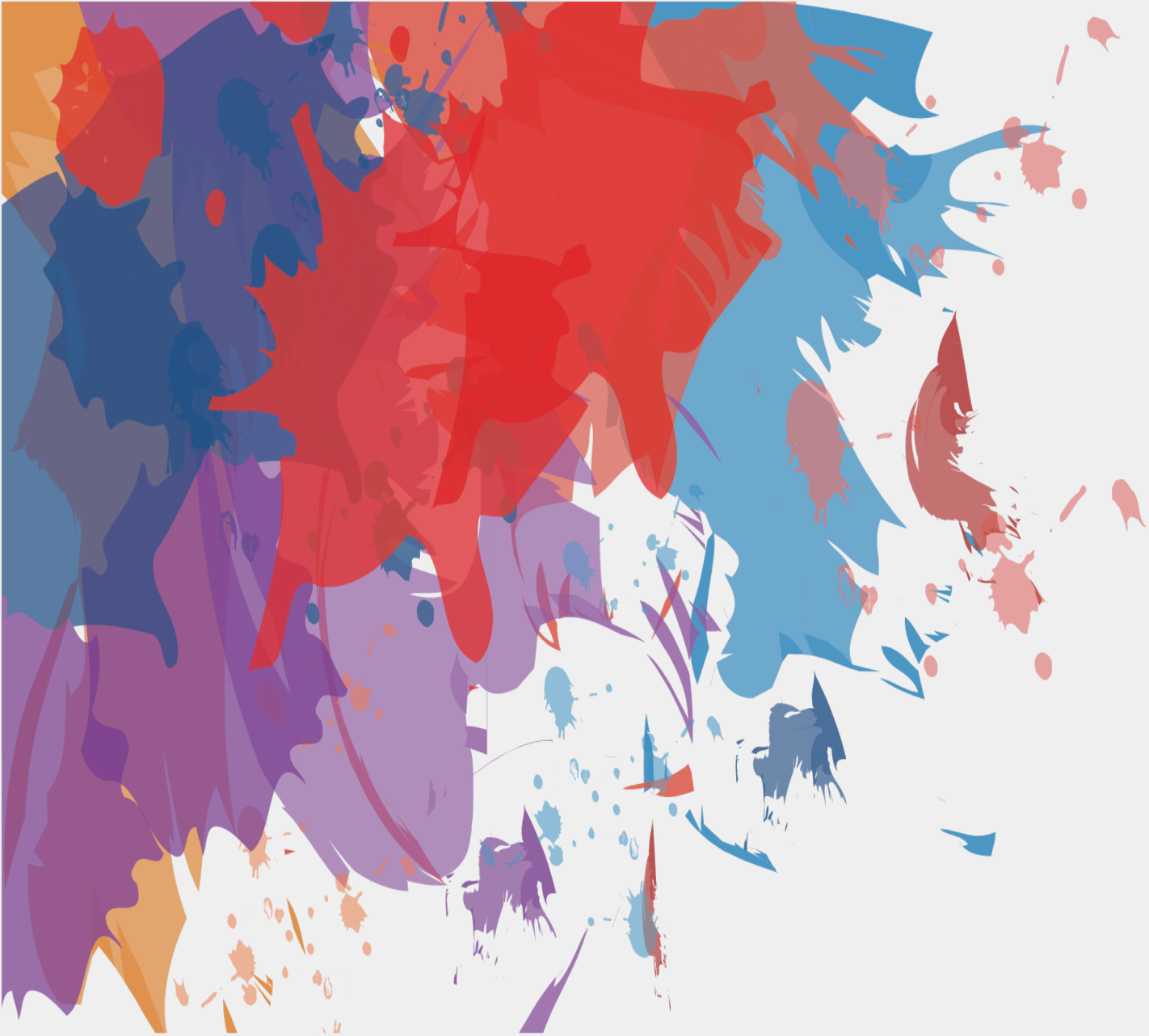
針對不同存活率做行銷

兩者重要性相當

依照行銷預算選擇最適合路徑

總結

1. RT Ratio 與 存活率 取代某固定時間的 Retention rate
2. One Number One Row 的型態（Order ID 轉成 User ID）
3. 以MLR模型預測新客回購週期T
4. 以GLM模型預測新客存活率
5. 依照行銷預算選擇適當的路徑
   - RT Ratio 低到高
   - 存活率 高到低
   - 兩者並重



Thank You