Multivariate Analysis Final Report

公車路線資料

統博一　鍾興潔
統碩一　黃三騰
統碩一　林子祥
應數四　蔡伊婷

Outline

- Introduction
- Data
- Objective
- Method
- Conclusion

Introduction

Introduction

- ✓偏鄉地區
- ✓交通方式
- ✓有些停靠站少有人使用

Data

Data

上車資訊　　下車資訊

每筆搭車的紀錄

	A	B	C	D	E	F	G	H	I	J	K
1	路線	卡號	票種	上車日期	上車時間	上車分鐘	上車站名	下車日期	下車時間	下車分鐘	下車站名
2	11	F23A1C03	敬老票	2016/3/7	15:12	912	ZUFC	2016/3/7	15:27:47	927	CSPK
3	11	F218BA02	敬老票	2016/4/6	10:20	620	ZUFC	2016/4/6	10:33:23	633	CSPK
4	11	12271A03	敬老票	2016/9/6	10:20	620	ZUFC	2016/9/6	10:34:15	634	CSPK
5	11	223FBC02	敬老票	2016/10/7	10:37	637	ZUFC	2016/10/7	10:52:27	652	CSPK
6	11	F23A1C03	敬老票	2016/1/20	08:18	498	ZUFC	2016/1/20	08:32:43	512	CSPK
7	11	C2361C03	敬老票	2016/1/19	08:29	509	ZUFC	2016/1/19	08:45:51	525	CSPK
8	11	C2361C03	敬老票	2016/3/21	08:25	505	ZUFC	2016/3/21	08:38:40	518	CSPK
9	11	F23A1C03	敬老票	2016/3/3	08:21	501	ZUFC	2016/3/3	08:34:56	514	CSPK
10	11	C2361C03	敬老票	2016/3/8	08:22	502	ZUFC	2016/3/8	08:38:26	518	CSPK
11	11	C2EB1B03	敬老票	2016/3/15	08:32	512	ZUFC	2016/3/15	08:49:22	529	CSPK
12	11	223FBC02	敬老票	2016/3/14	08:22	502	ZUFC	2016/3/14	08:37:51	517	CSPK
13	11	C2EB1B03	敬老票	2016/3/23	08:28	508	ZUFC	2016/3/23	08:41:38	521	CSPK
14	11	C2EB1B03	敬老票	2016/12/26	08:21	501	ZUFC	2016/12/26	08:37:18	517	CSPK
15	11	F23A1C03	敬老票	2016/12/13	08:23	503	ZUFC	2016/12/13	08:37:36	517	CSPK

- 路線
- 票種
- 上下車時間與日期
- 上下車分鐘數
- 上下車站名
- 資料來源：
  東部運輸中心
  （105年搭乘公車刷卡紀錄）

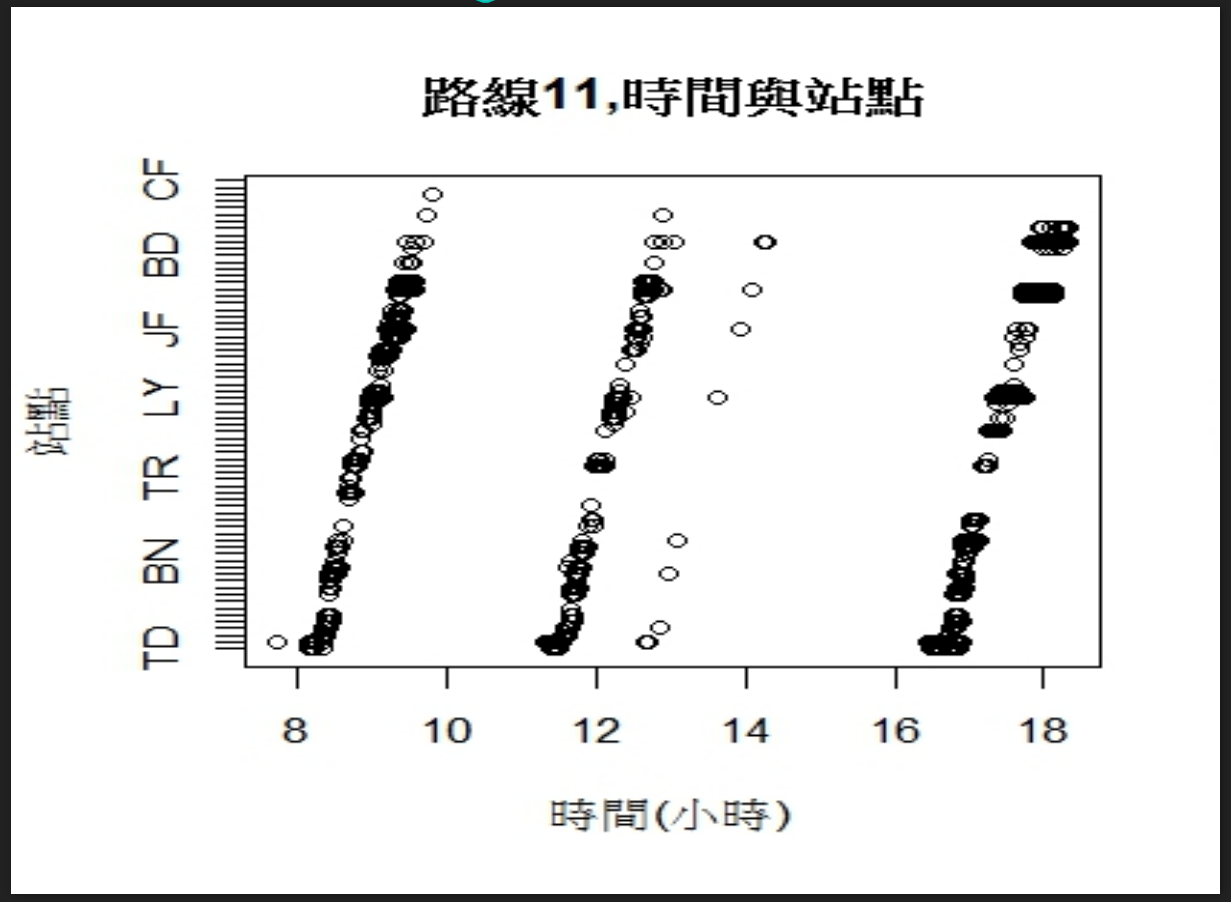
Data

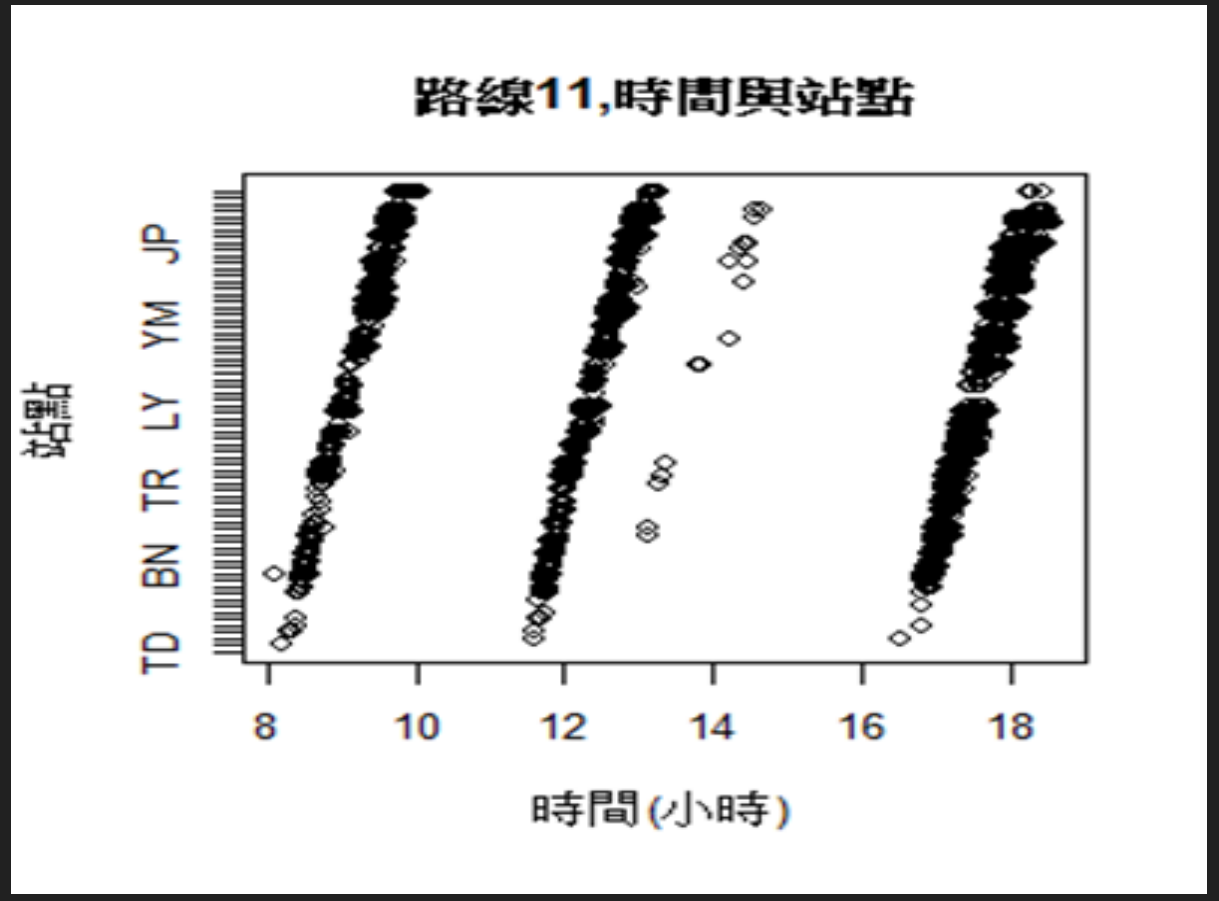
- 路線：11(上車54站, 下車71站)、22(上車51站, 下車68站)、
  33(上車50站, 下車60站)、44(上車51站, 下車62站)
- 票種：敬老票、學生票、一般票(類別)
- 上下車時間：時間轉換為分鐘數(e.g. 09:08→60*9+8=548)
- 上下車站名：各個站點的名稱(類別)
- 去程與回程資料分開
- 資料來源:東部運輸中心(105年搭乘公車刷卡紀錄)

Data

路線(去程)	樣本數	上車站數	下車站數
11	8380	54	71
22	13982	51	68
33	4771	50	60
44	5337	51	62
總共	32470		
路線(回程)	樣本數	上車站數	下車站數
11	2392	62	60
22	14589	66	61
33	2160	55	53
44	4456	63	52
總共	23597		

路線刷卡站點與時間(整年-去回程)





Objective

Objective

- 路線整併、區間直達車
  - 各路線重要站點–使用率高
  - 不同客群–老人 → 不能廢除
  - 不同客群–學生 → 學生專車

Method

路線11站點與使用次數

	u_1	u_2	u_3	u_4	…	u_{8380}	總使用次數
TD	1	1	0	0	…	0	128
SCS	1	0	1	0	…	0	1693
CM	0	1	0	0	…	1	965
SDH	0	0	1	1	…	0	578
⋮	⋮	⋮	⋮	⋮	⋱	⋮	
	0	0	0	1	…	1	

72 站X 8380筆

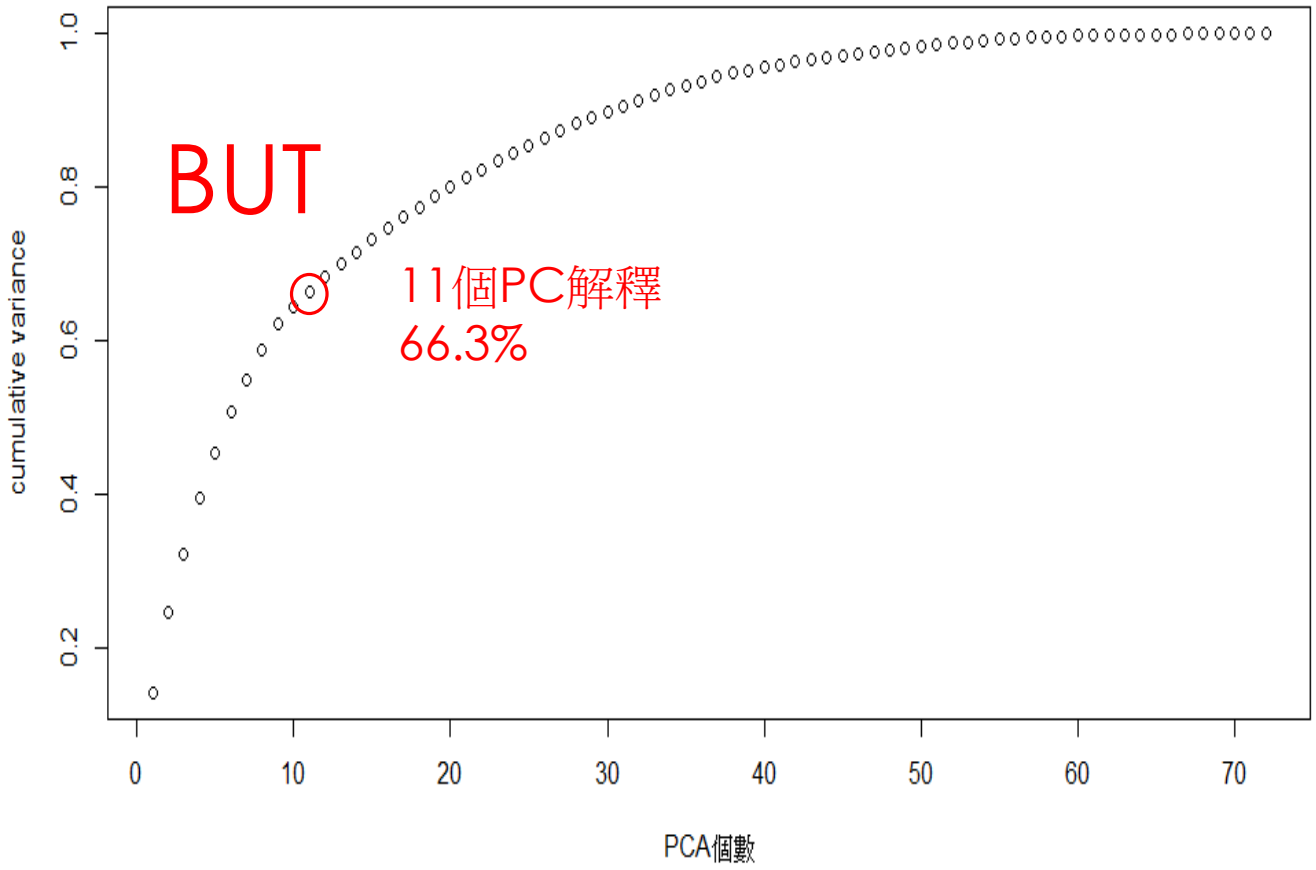
路線11站點與使用次數

- 每一站的使用次數 = 8380筆刷卡的貢獻
- Want: 把72個點畫出來

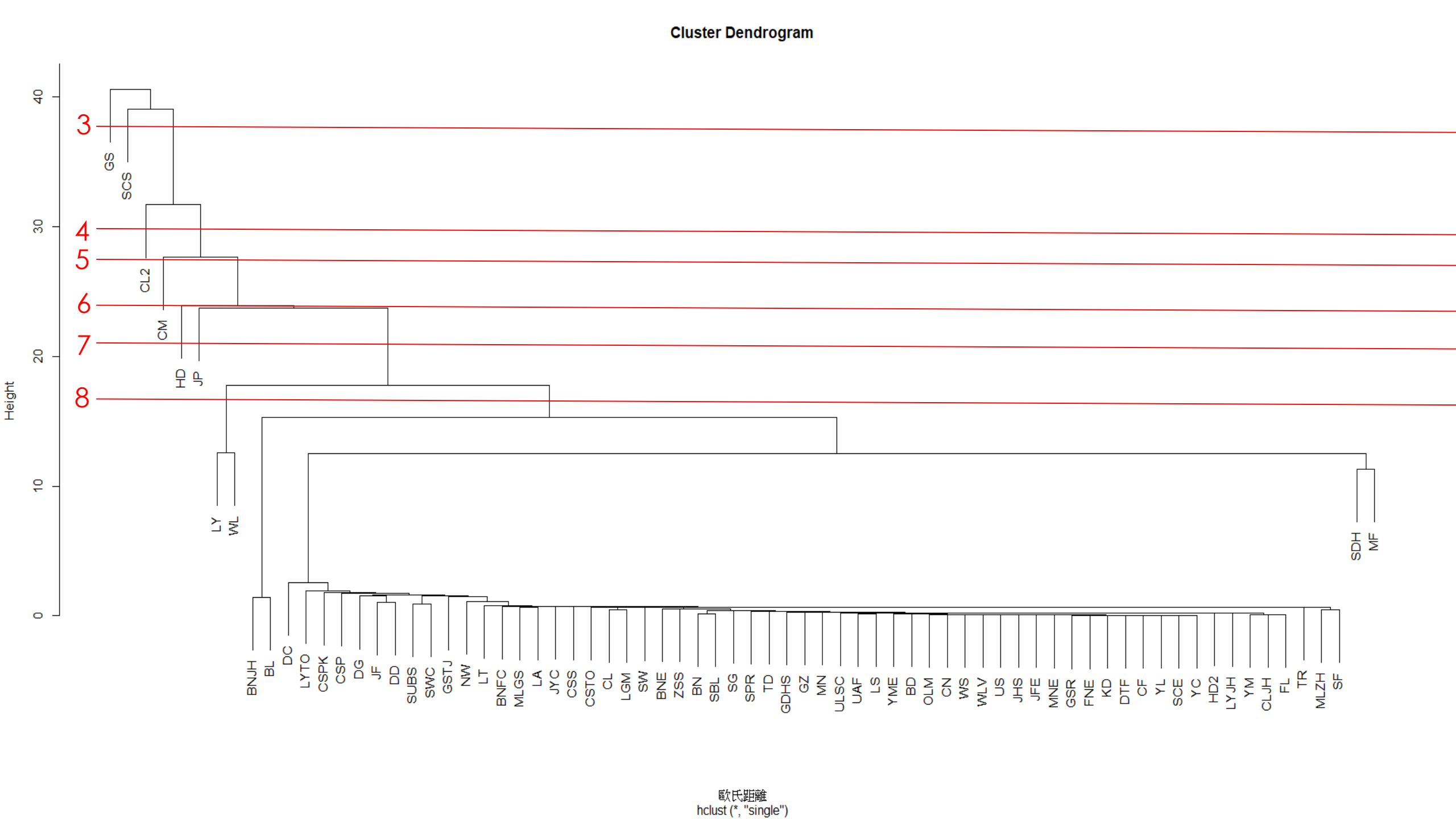
  → PCA降維 (n=72 ,p=8380)

  → 保持原始資料變異的資訊

路線11站點與使用次數



- Hierarchical Clustering
- 每一站視為一個cluster
- 將距離(歐氏)最近站點合併


Cluster Dendrogram
Height
0
10
20
30
40
3
4
5
6
7
8
GS
SCS
CL2
CM
HD
JP
LY
WL
SDH
MF
BNJH
BL
DC
LYTO
CSPK
CSP
DG
JF
DD
SUBS
SWC
GSTJ
NW
LT
BNFC
MLGS
LA
JYC
CSS
CSTO
CL
LGM
SW
BNE
ZSS
BN
SBL
SG
SPR
TD
GDHS
GZ
MN
ULSC
UAF
LS
YME
BD
OLM
CN
WS
WLV
US
JHS
JFE
MNE
GSR
FNE
KD
DTF
CF
YL
SCE
YC
HD2
LYJH
YM
CLJH
FL
TR
MLZH
SF
歐氏距離
hclust (*, "single")

路線11站點與搭車模式

搭車模式

	m_1	m_2	m_3	m_4	…	m_{591}
A	15	0	0	5	…	0
B	15	3	0	0	…	0
C	0	0	9	0	…	0
D	0	3	0	0	…	1
⋮	⋮	⋮	⋮	⋮	⋱	⋮
	0	0	9	0	…	1

72站 X 591種

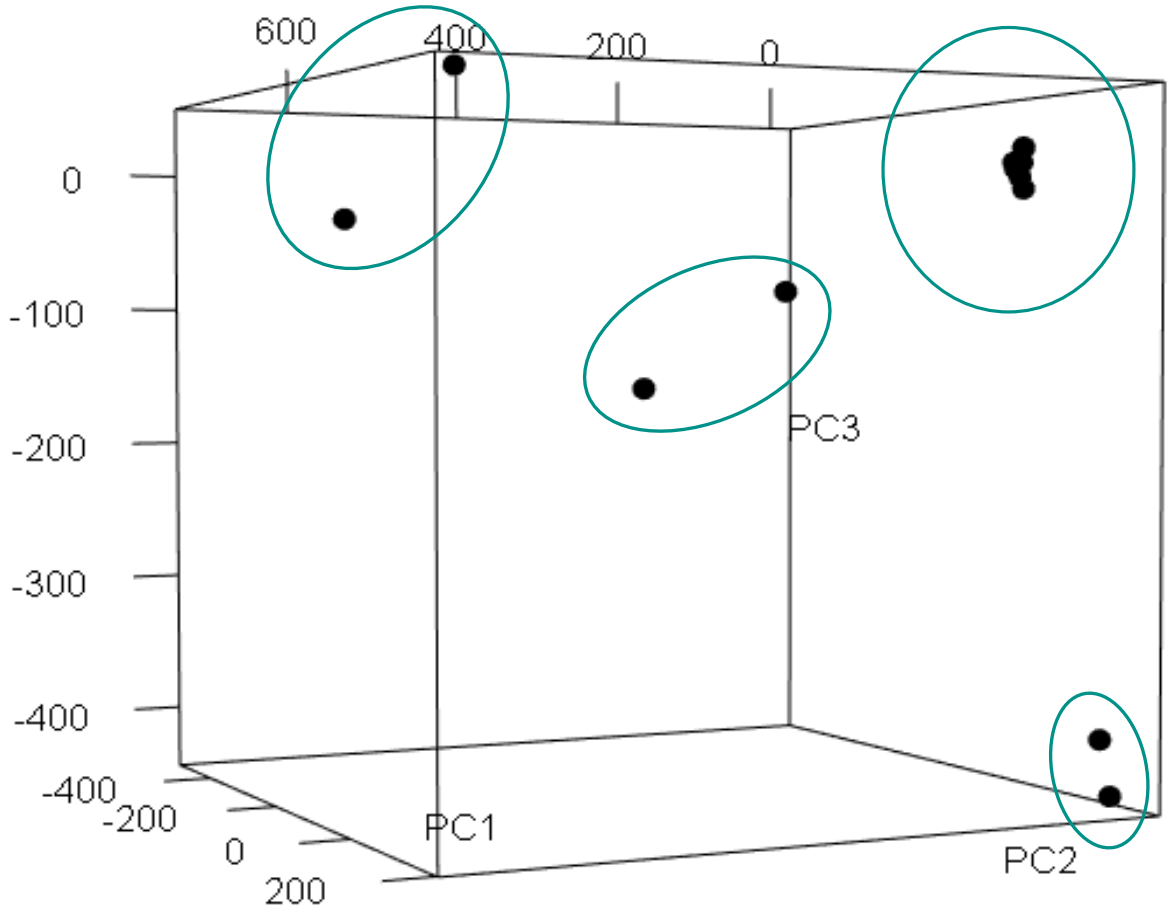
路線11站點與搭車模式

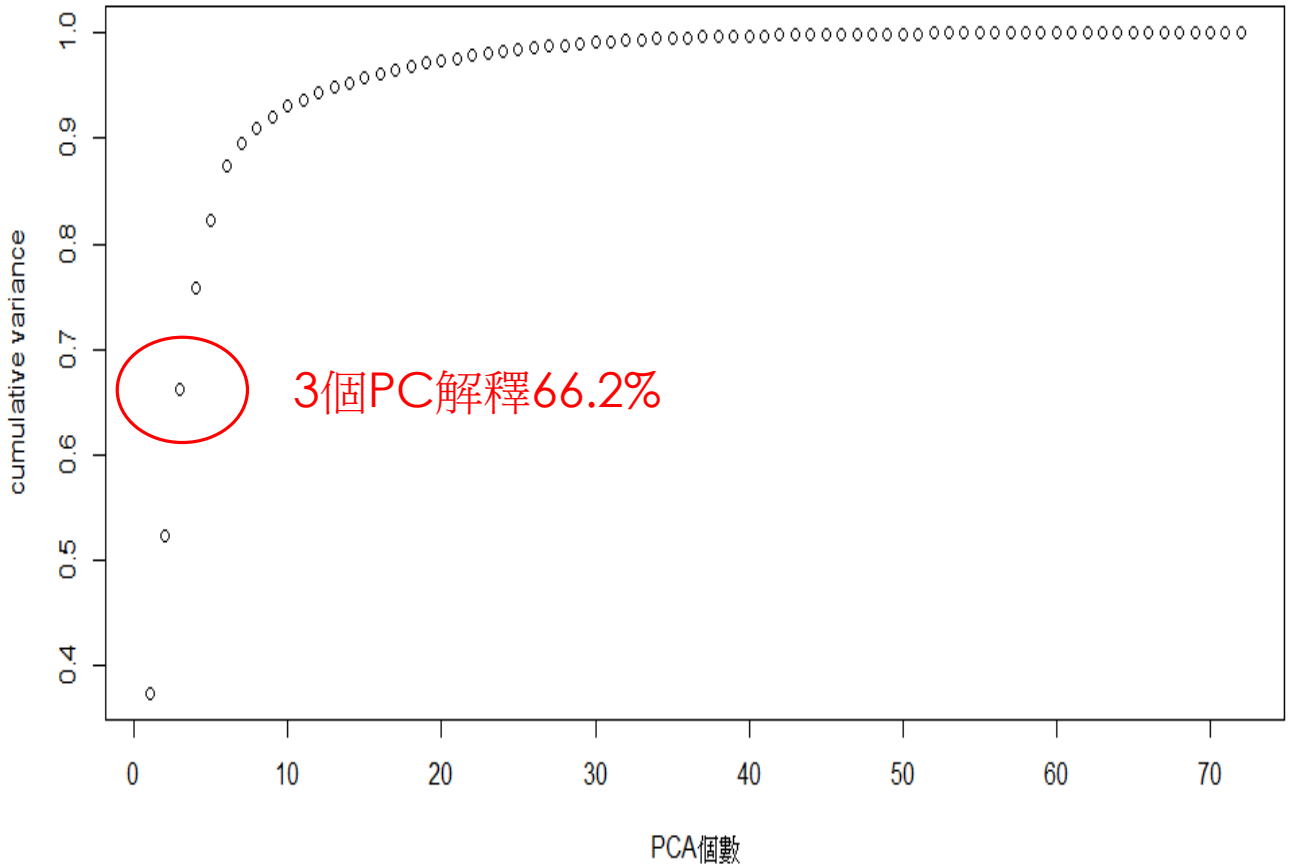
- 每一站的使用次數 = 591種搭車模式的貢獻
- Want: 把72個點畫出來

  → PCA降維 (n=72 ,p=591)

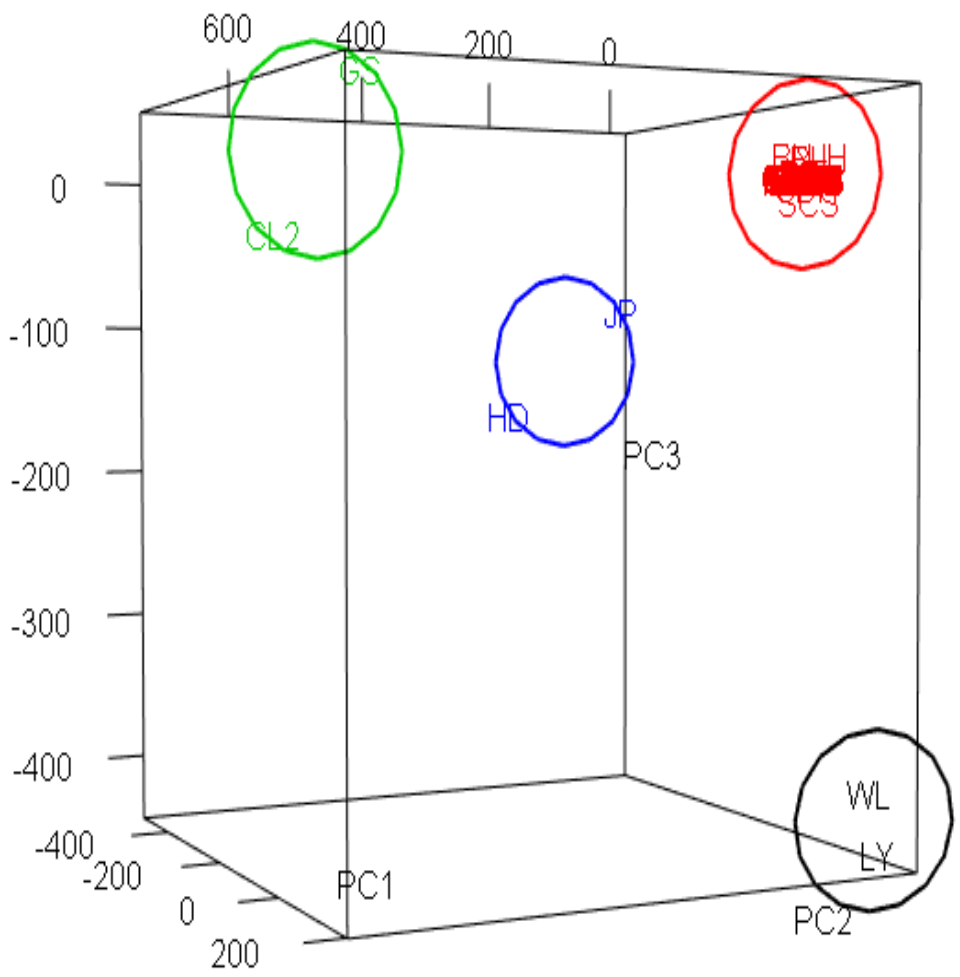
  → 保持原始資料變異的資訊

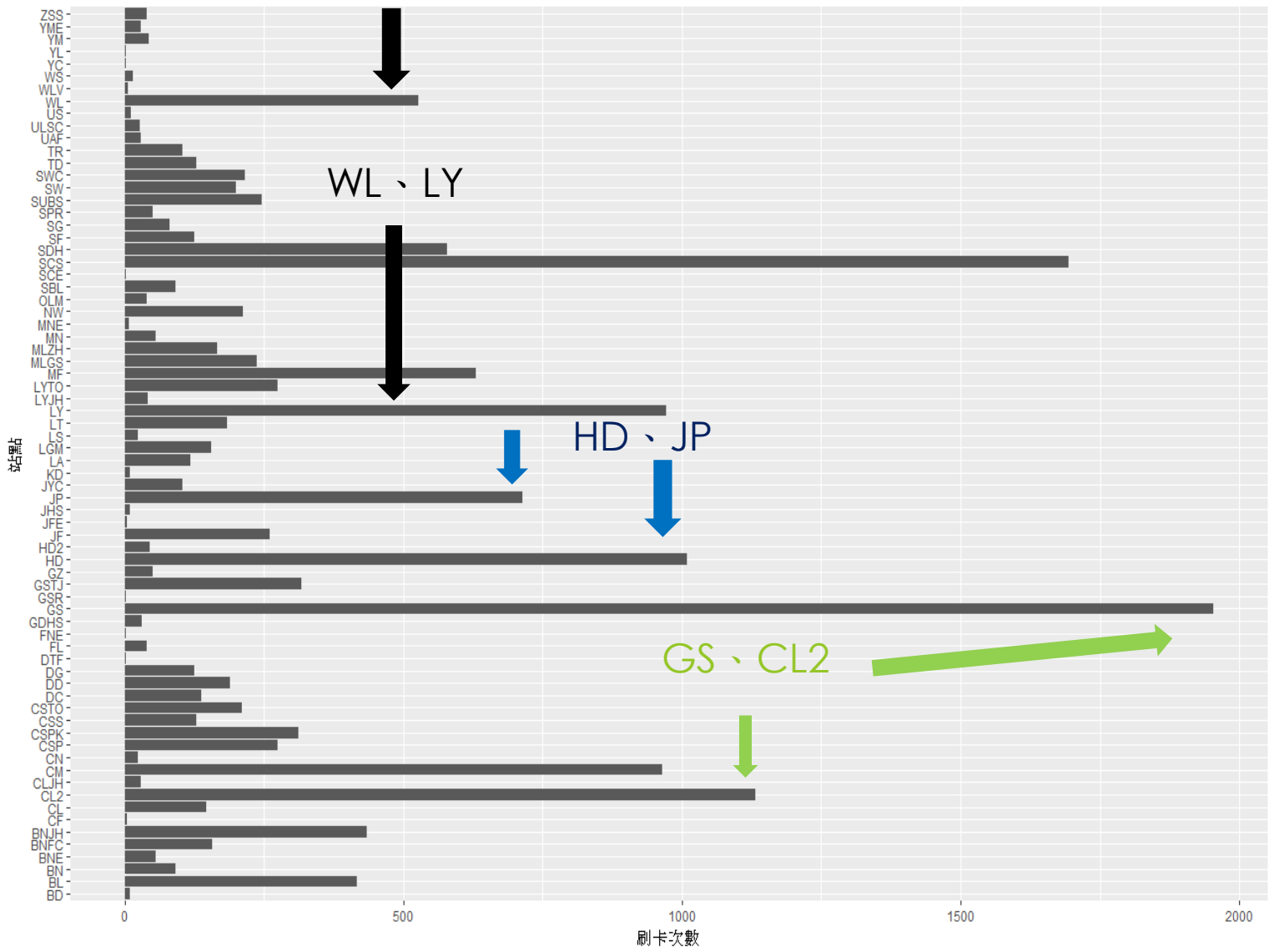
PCA Plot


600
400
200
0
0
-100
-200
-300
-400
-400
-200
0
200
PC1
PC2
PC3

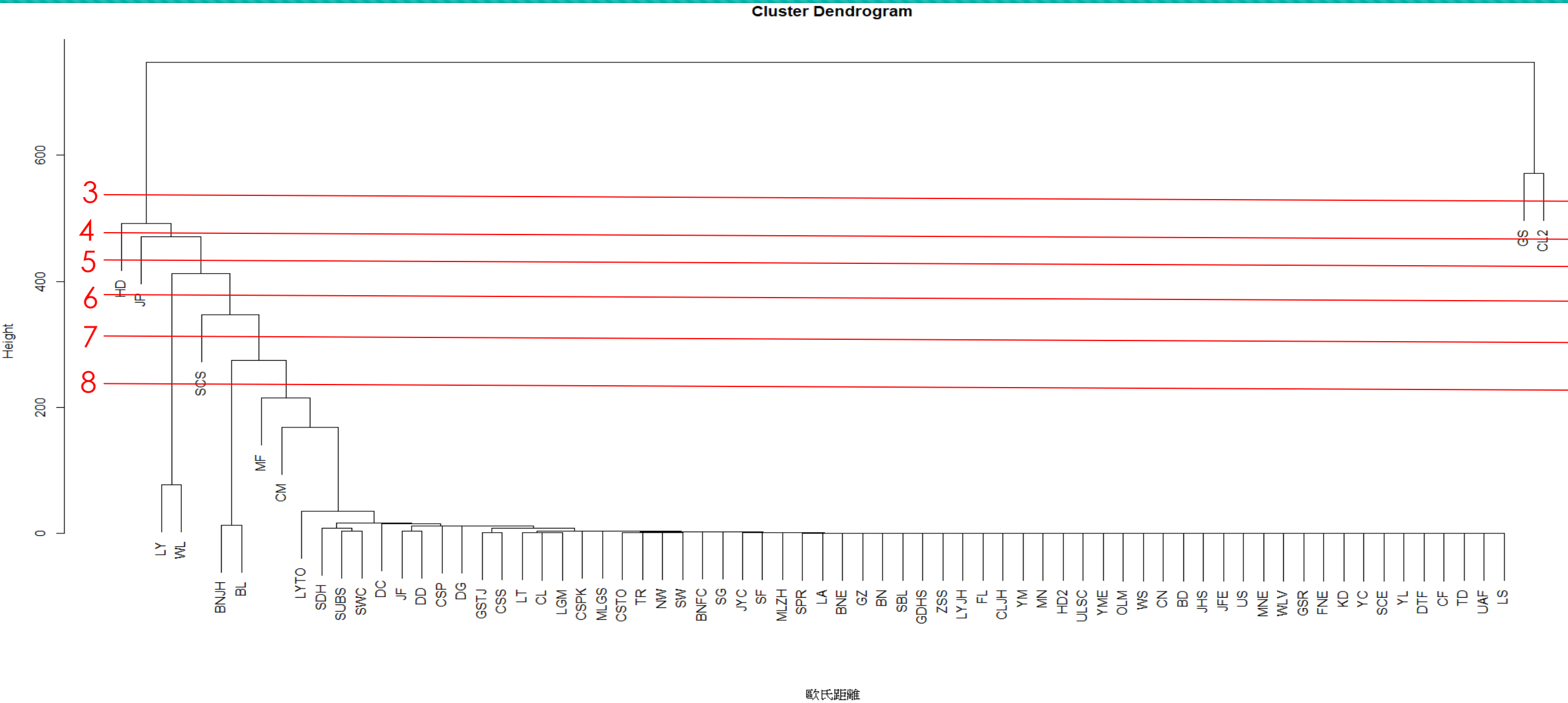

cumulative variance
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50
60
70
PCA個數
3個PC解釋66.2%

Kmeans Plot


600
400
200
0
GS
CL2
JP
HD
PC3
0
-100
-200
-300
-400
-400
-200
0
200
PC1
WL
LY
PC2


WL、LY
HD、JP
GS、CL2
站點
刷卡次數
0
500
1000
1500
2000

Hierarchical Plot



路線11站點與搭車模式

- 使用次數高卻沒被分出來 e.g. SCS
  - ➡ 零散的搭車模式所貢獻
- 客群之間搭車模式不同

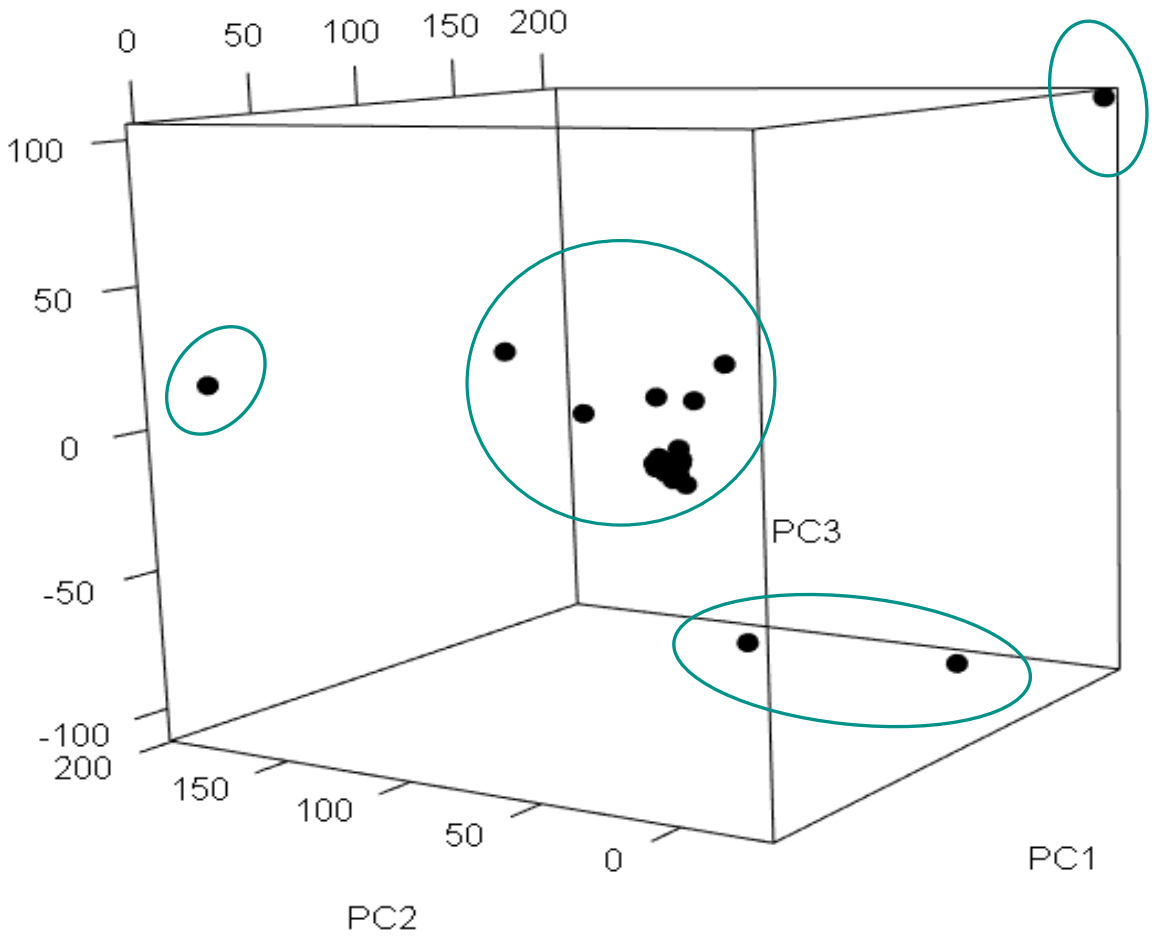
路線11(敬老票)站點與搭車模式

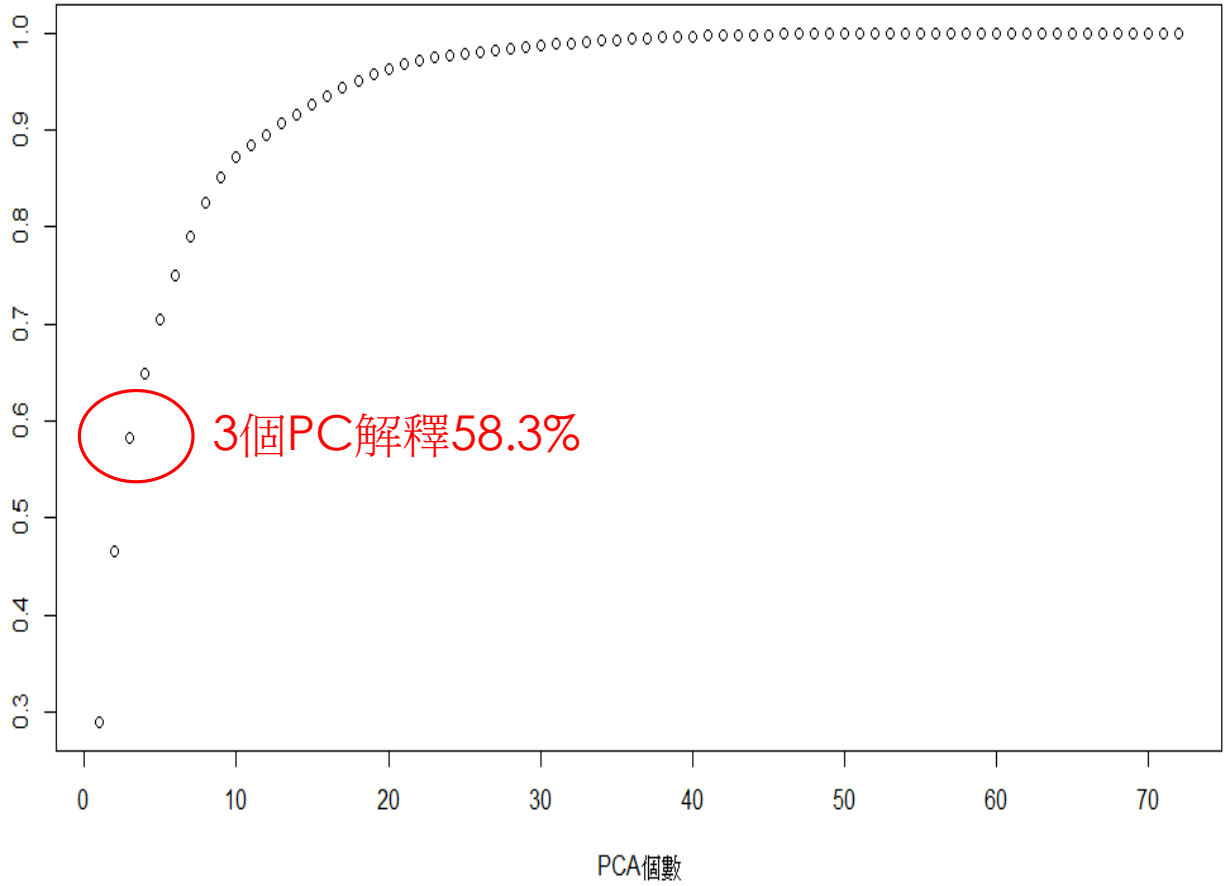
- 每一站的使用次數 = 444種搭車模式的貢獻
- Want: 把72個點畫出來

  → PCA降維 (n=72 ,p=444)

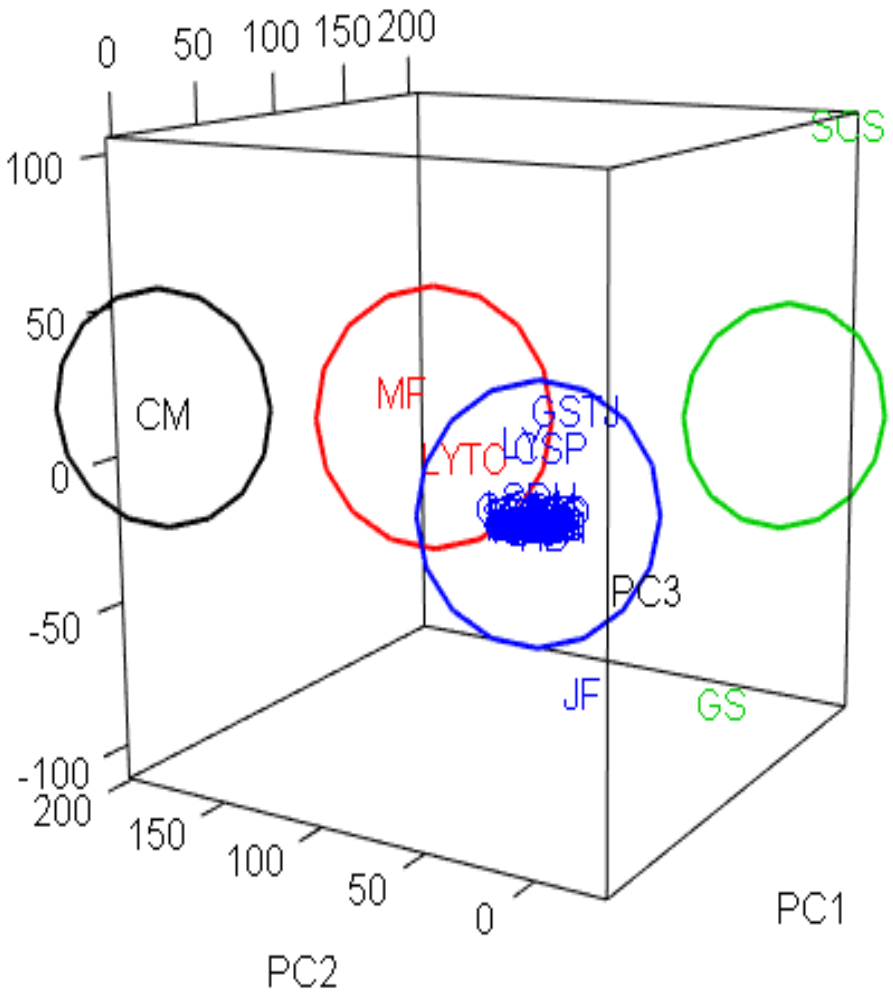
  → 保持原始資料變異的資訊

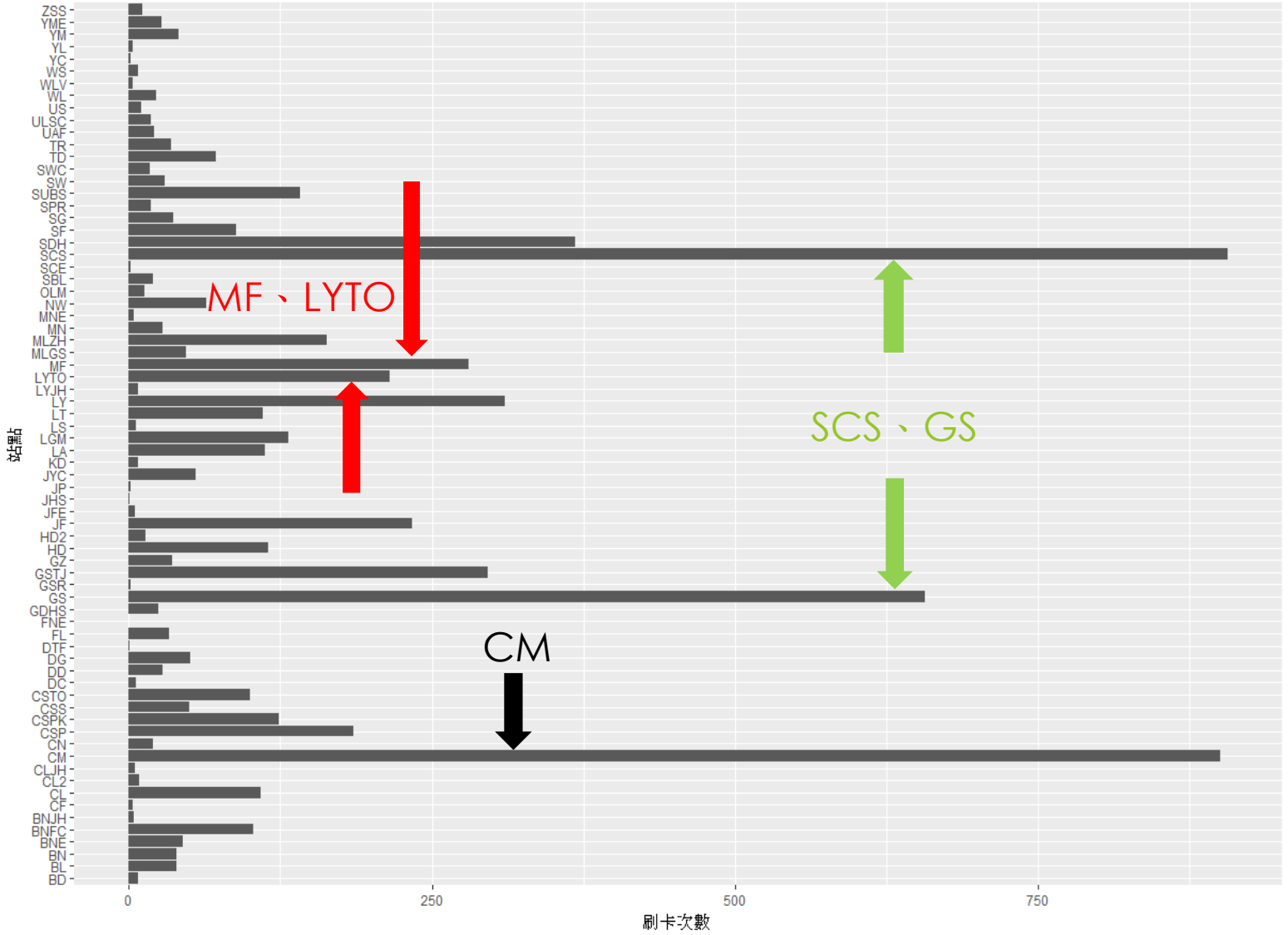
PCA Plot


0
50
100
150
200
100
50
0
-50
-100
200
150
100
50
0
PC2
PC3
PC1


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
10
20
30
40
50
60
70
PCA個數
3個PC解釋58.3%

Kmeans Plot


0
50
100
150
200
100
50
0
-50
-100
200
150
100
50
0
PC2
PC1
PC3
CM
MF
LYTO
GSTJ
LYSP
JF
GS
SCS

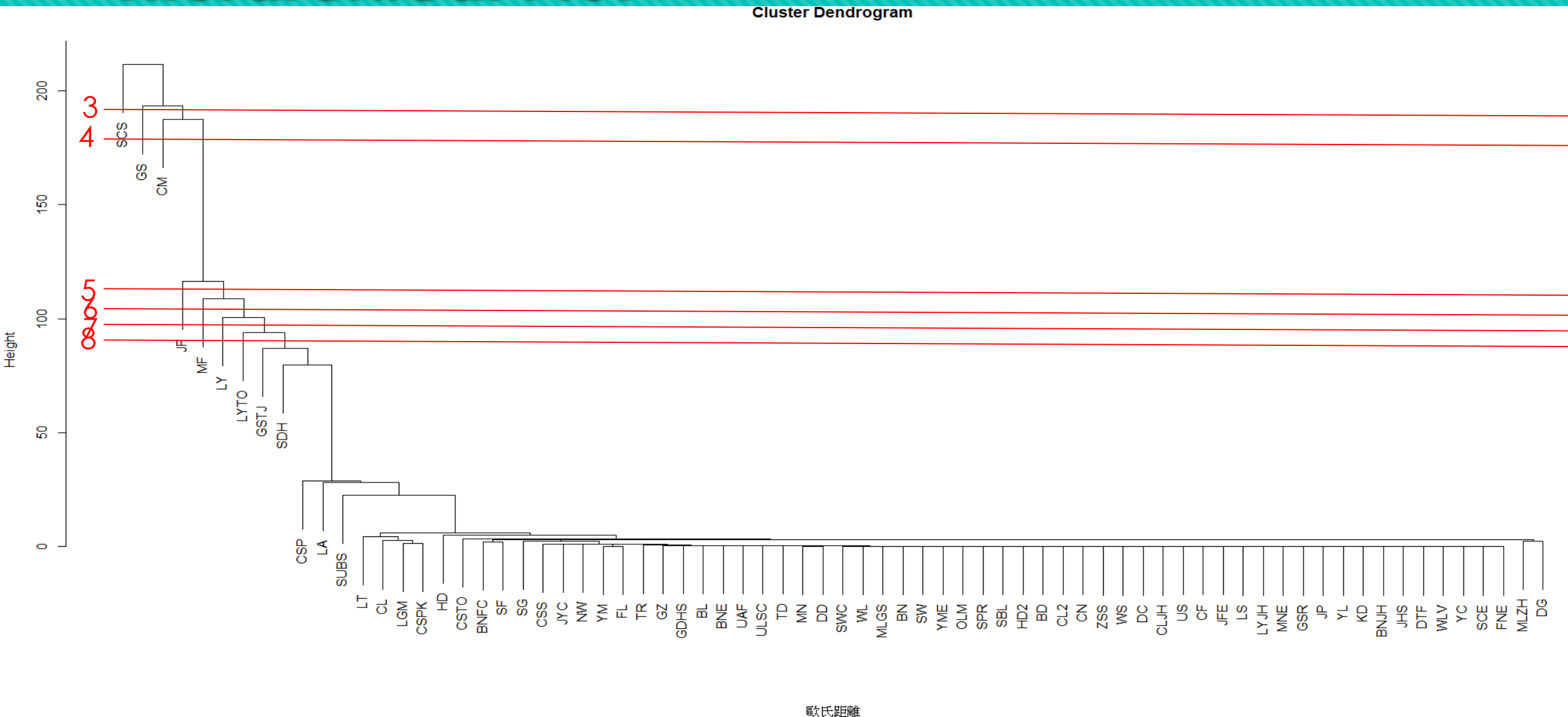

ZSS
YME
YM
YL
YC
WS
WLV
WL
US
ULSC
UAF
TR
TD
SWC
SW
SUBS
SPR
SG
SF
SDH
SCS
SCE
SBL
OLM
NW
MNE
MN
MLZH
MLGS
MF
LYTO
LYJH
LY
LT
LS
LGM
LA
KD
JYC
JP
JHS
JFE
JF
HD2
HD
GZ
GSTJ
GSR
GS
GDHS
FNE
FL
DTF
DG
DD
DC
CSTO
CSS
CSPK
CSP
CN
CM
CLJH
CL2
CL
CF
BNJH
BNFC
BNE
BN
BL
BD
MF、LYTO
SCS、GS
CM
0
250
500
750
刷卡次數

路線11(敬老票)站點與搭車模式

- SCS、GS、CM、MF、LYTO → K means
  SCS、GS、CM、JF、MF、LY、LYTO → Hierarchical

- SCS、GS、CM、JF、MF、LY、LYTO ：不能廢除

Hierarchical Plot



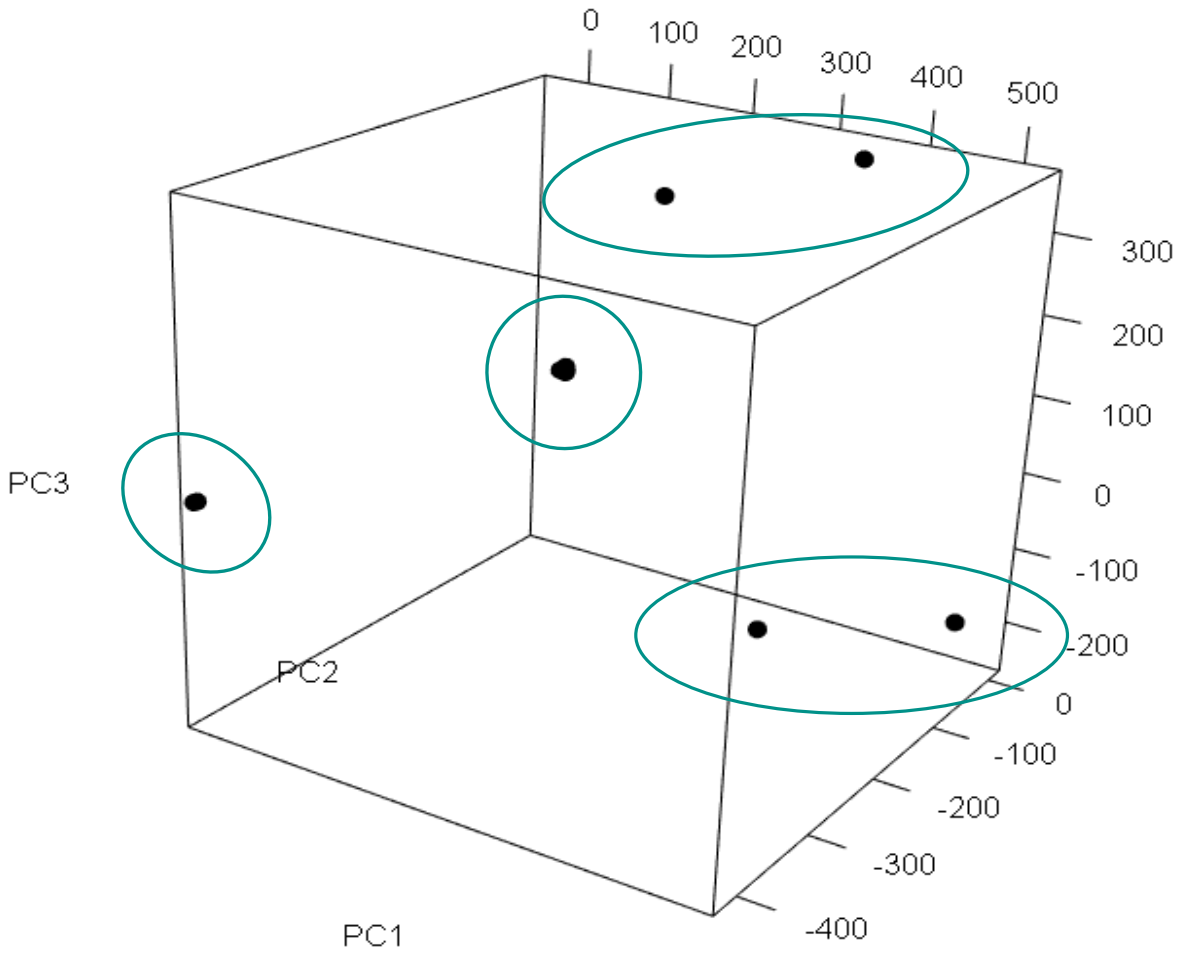
路線11(學生票)站點與搭車模式

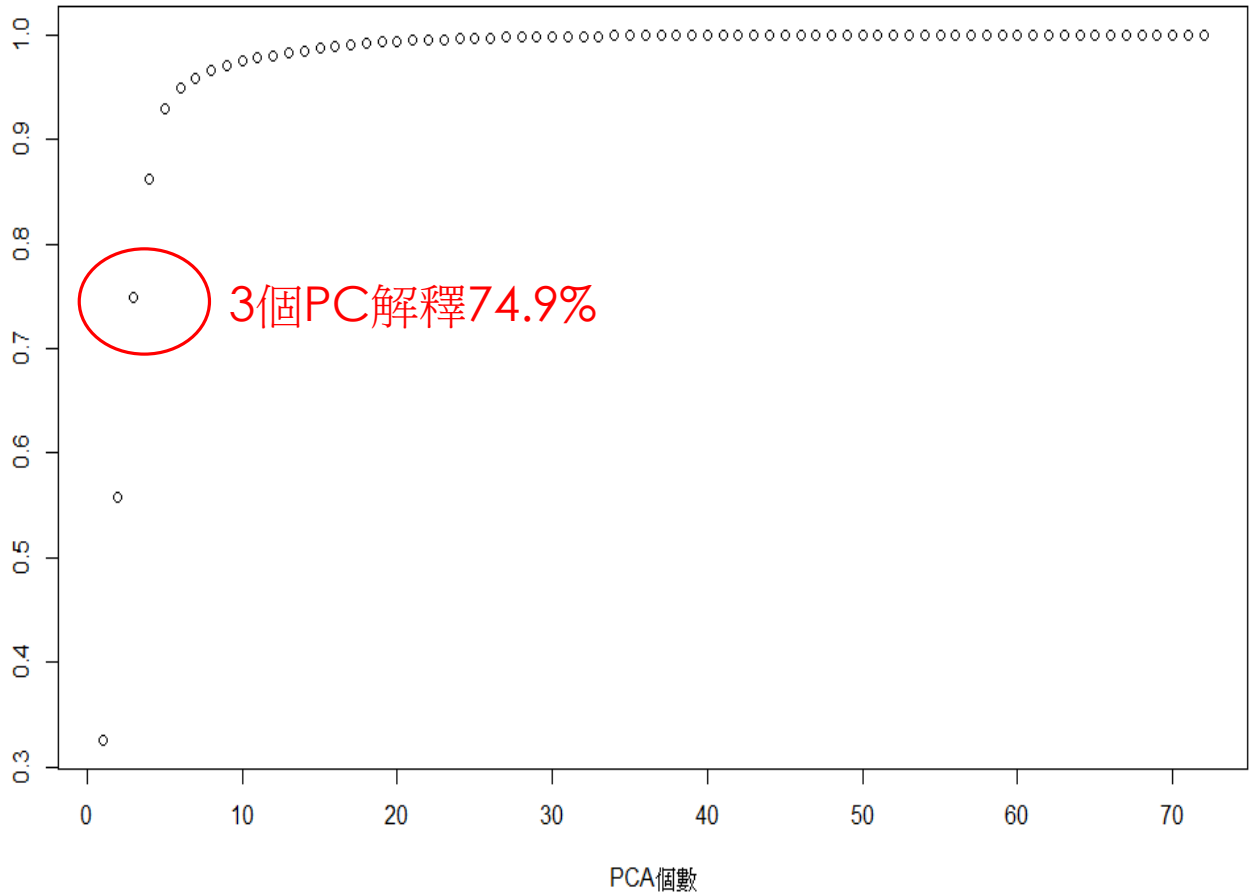
- 每一站的使用次數 = 230種搭車模式的貢獻
- Want: 把72個點畫出來

  → PCA降維 (n=72 ,p=230)

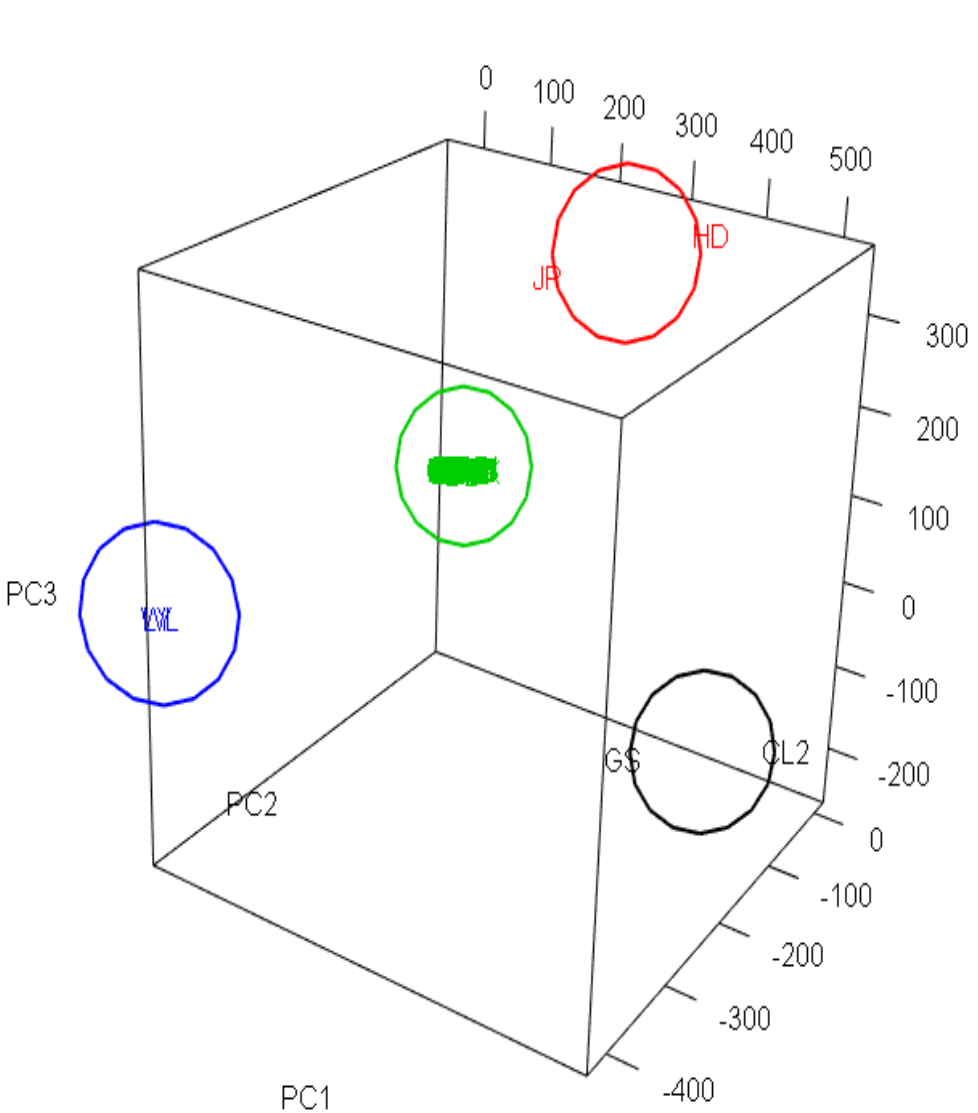
  → 保持原始資料變異的資訊

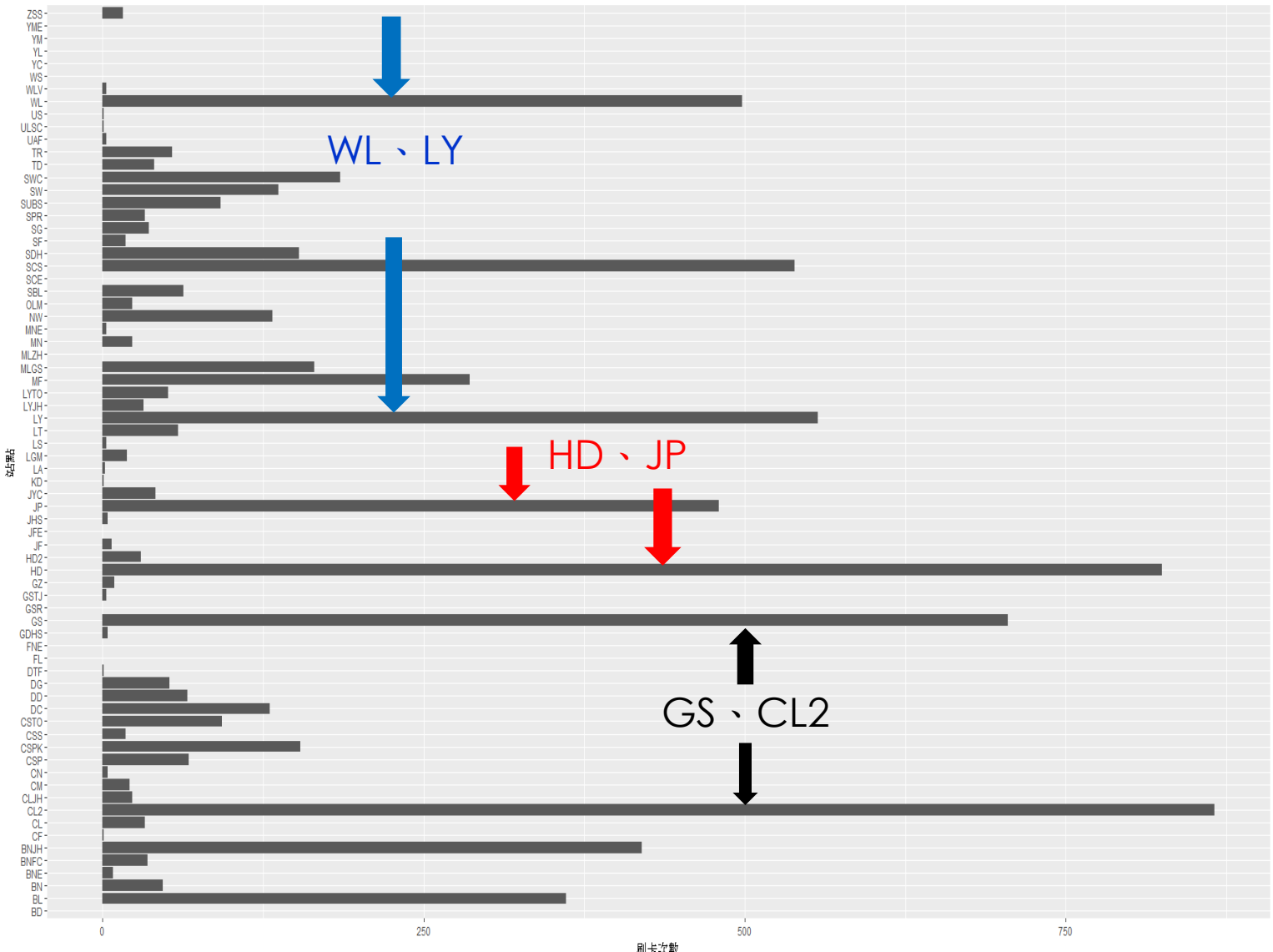
PCA Plot


0
100
200
300
400
500
300
200
100
0
-100
-200
-300
-400
PC3
PC2
PC1

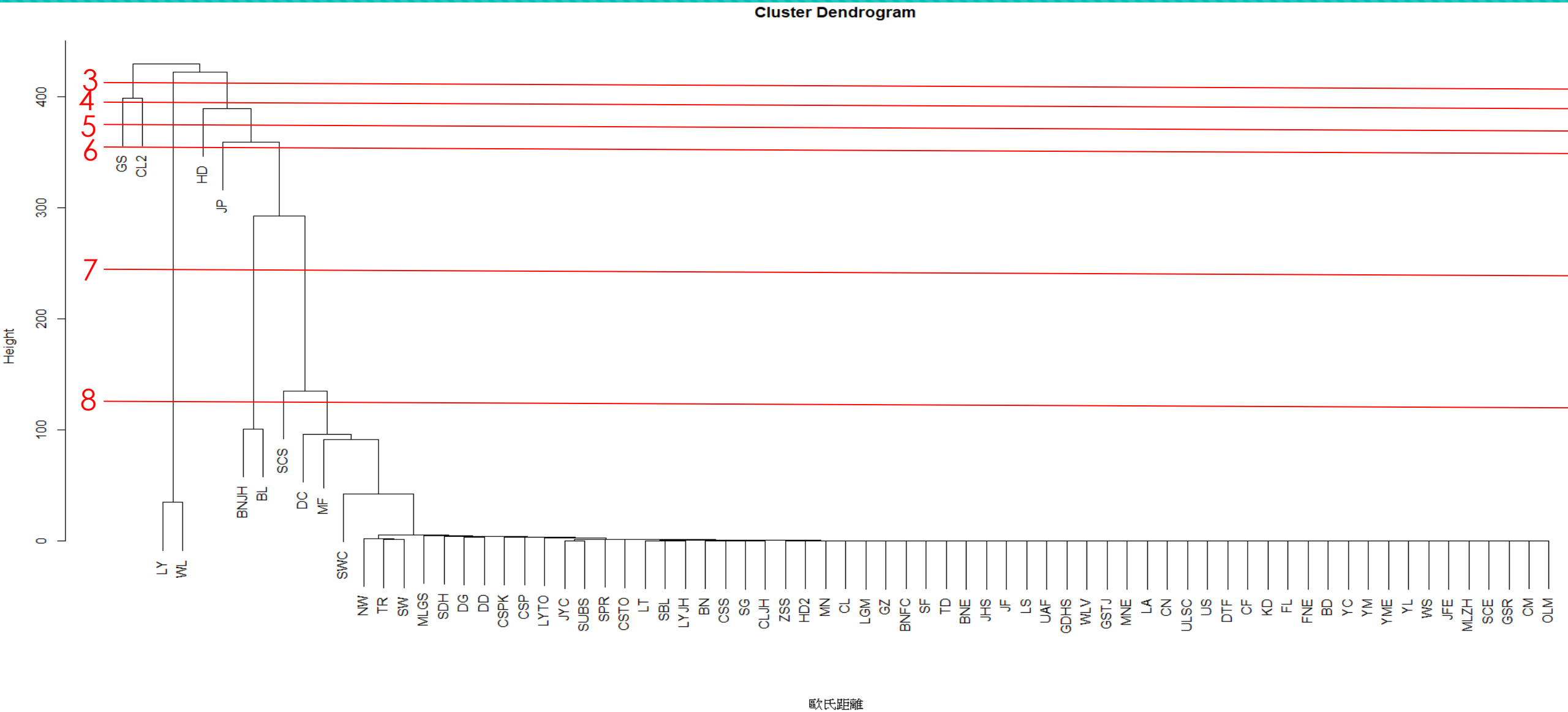

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
10
20
30
40
50
60
70
PCA個數
3個PC解釋74.9%

Kmeans Plot


0
100
200
300
400
500
HD
JP
300
200
100
0
-100
-200
PC3
YWL
GS
CL2
PC2
0
-100
-200
-300
-400
PC1


WL、LY
HD、JP
GS、CL2
0
250
500
750

Hierarchical Plot



路線11(學生票)站點與搭車模式

- 3個群 (WL、LY), (HD、JP), (GS、CL2) 附近有國中小學
- 模式:
  1. WL ↔ LY : 424
  2. HD ↔ JP : 377
     HD ↔ CL2 : 349
  3. GS ↔ CL2 : 418
- 專車: GS ↔ HD ↔ CL2 ↔ JP

K means

- Initial set of k means $m_1^{(1)}, m_2^{(1)}, \ldots, m_k^{(1)}$.
- Step 1: consider $S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall\, 1 \leq j \leq k \right\}, \quad i = 1, \ldots, k$

  each $x_p \in some\ S_i^{(t)}$.
- Step 2: update new means by $m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in s_i^{(t)}} x_j \ , \quad i = 1, \ldots, k$

PCA

PCA(Principal Component Analysis)，主成分分析。去除多餘資訊，將原有複雜的數據降維但保留數據對變異數貢獻最大的特徵。主要原理為降低資料維度，又希望遺失的訊息能降到最低。

PCA

- Let $X_{n\times p}$, find $u_1 \in R^p$, $\left|\left|u_1\right|\right| = 1$
- To $$\min_{u_1} \sum_{i=1}^{n} \left|\left|X_i - \widehat{X_i}\right|\right|^2$$
$$\Leftrightarrow \max_{u_1} u_1^T X^T X u_1$$
$$\Leftrightarrow \quad find\ largest\ eigenvalue\ of\ X^T X$$
- The projection coordinate $Y_1 = Xu_1 \quad \Longrightarrow \ PC1$

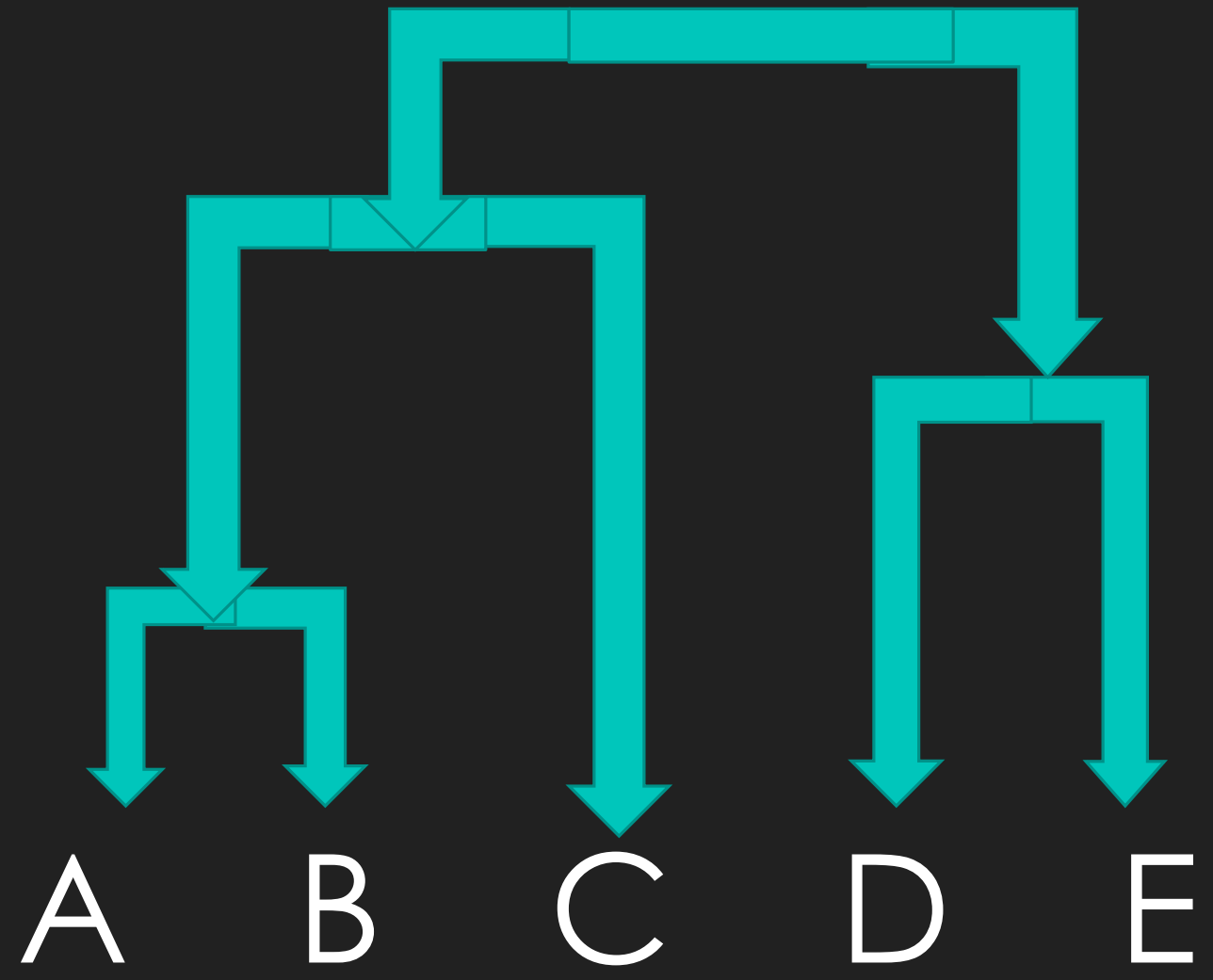
Hierarchical Clustering

Example

	PCA1	PCA2
A站	1	1
B站	1	0
C站	0	2
D站	2	4
E站	3	5

	A站	B站	C站	D站	E站
A站	0	1	1.4	3.2	4.5
B站	1	0	2.2	4.1	5.4
C站	1.4	2.2	0	2.8	4.2
D站	3.2	4.1	2.8	0	1.4
E站	4.5	5.4	4.2	1.4	0

	AB站	C站	D站	E站
AB站	0	1.8	3.6	4.9
C站	1.8	0	2.8	4.2
D站	3.6	2.8	0	1.4
E站	4.9	4.2	1.4	0



Conclusion

Conclusion

- 資訊: 使用次數長條圖無法得知
- 路線11 重要站點分析

  學生族群 ➡ 學生專車

  老人族群 ➡ 不能廢除

Conclusion

- Next:
- 路線22、33、44的重要站點分析
- 使用率高 ≠ 特定模式高

Q & A

Thanks for listening