多變量統計：Final Report

Hostel in Japan

410578068 統計四 陳威傑

一、 前言

研究動機&目的:

即將到來的外國遊客。預計到 2020 年東京奧林匹克運動會之前，來日本的外國遊客將會增加。因此，考慮除了酒店以外在日本居住的另一種方法將使計劃在不久的將來訪問日本的所有遊客受益。

二、 研究方法

此次研究，我們將進行三項分析。分別是多變量線性迴歸分析、主成分分析、多變量分析進行研究。

多變量線性迴歸分析: 最低房價與地點對細項評分的影響程度。

主成分分析: 所有細項評分之主導關係。

多變量分析: 將房價分群，觀察細項評分對於選擇旅館的影響。

三、資料與變數說明

- 此資料來源為 Kaggle 網站— How to Find a Hostel in Japan，一共有 327 筆資料(移除 2 筆異常資料後剩餘 325 筆資料)，列出前 10 筆資料之內容如下：

Obs	num	name	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	1	"Bike & Bed" CharinCo Hostel	Osaka	3300	2.9	135.5137671	34.682678	9.2	Superb	8.9	9.4	9.3	8.9	9	9.4	9.4
2	2	& And Hostel	Fukuoka-City	2600	0.7	NA	NA	9.5	Superb	9.4	9.7	9.5	9.7	9.2	9.7	9.5
3	3	&And Hostel Akihabara	Tokyo	3600	7.8	139.7774724	35.6974473	8.7	Fabulous	8	7	9	8	10	10	9
4	4	&And Hostel Ueno	Tokyo	2600	8.7	139.783667	35.712716	7.4	Very Good	8	7.5	7.5	7.5	7	8	6.5
5	5	&And Hostel-Asakusa North-	Tokyo	1500	10.5	139.7983712	35.7278979	9.4	Superb	9.5	9.5	9	9	9.5	10	9.5
6	6	1night1980hostel Tokyo	Tokyo	2100	9.4	139.7869499	35.724384	7	Very Good	5.5	8	6	6	8.5	8.5	6.5
7	7	328 Hostel & Lounge	Tokyo	3300	16.5	139.7454672	35.5480439	9.3	Superb	8.7	9.7	9.3	9.1	9.3	9.7	8.9
8	8	36Hostel	Hiroshima	2000	1.6	NA	NA	9.5	Superb	8.8	9.9	9.2	9.6	9.8	9.8	9.5
9	9	Ace Inn Shinjuku	Tokyo	2200	3	139.7243036	35.6925119	7.7	Very Good	6.7	7.2	6.8	8.5	7.8	8.5	8.1
10	10	Air Osaka Hostel	Osaka	1600	9.7	135.4769556	34.6222596	9.2	Superb	9.5	9.1	8.7	8.8	8.9	9.8	9.5

- 上表中各變數之內容則如下表所示：

變數	描述	變數型式
	旅館價格與地點	
X_1	所在城市	非數值
X_2	每晚最低房價	數值
X_3	與市中心之距離	數值
X_4	經度	數值
X_5	緯度	數值
	旅館評分	
X_6	評分總結	數值
X_7	評分域（Rating band）	非數值
X_8	評分項目—氛圍	數值
X_9	評分項目—整潔	數值
X_{10}	評分項目—設施	數值
X_{11}	評分項目—地點	數值
X_{12}	評分項目—安全	數值
X_{13}	評分項目—人員	數值
X_{14}	評分項目—CP 值	數值

- 資料來源網址：
  https://www.kaggle.com/gravitymhxy2/how-to-find-a-hostel-in-japan/data

四、多變量線性迴歸分析 (Multivariate Linear Regression Analysis)

- 目的：了解每晚最低房價(X2)、市中心距離(X3)，對地點評分項目結果(Y1)、氛圍評分項目結果(Y2)是否有影響。

- 分析結果：

應變數 Y1 單變量迴歸分析：

變異數分析

來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	2	33.74	16.87	15	<.0001
誤差	322	362.09	1.125		
已校正的總計	324	395.825			

（表 1）變異數分析

H_0：$\beta_2 = \beta_3 = 0$ vs. H_a：At least one of $\beta_i \neq 0$
$F = 15.00$，p-value $< 0.0001 < \alpha = 0.05$
Reject H_0, the model is significant. $R^2 = 0.0852$

參數估計值					
變數	**自由度**	**參數估計值**	**標準誤差**	**t 值**	**Pr > |t|**
Intercept	1	8.78	0.20421	43	<.0001
X2	1	0.00012	7.19E-05	1.65	0.0995
X3	1	-0.066	0.01275	-5.17	<.0001

（表 2）參數估計值

$H_0：\beta_2 = 0$ vs. $H_a：\beta_2 \neq 0$
$t = 1.65$，p-value $= 0.0995 \nless \alpha = 0.05$
Not reject H_0, x_2 is not significant.

$H_0：\beta_3 = 0$ vs. $H_a：\beta_3 \neq 0$
$t = -5.17$，p-value $< 0.0001 < \alpha = 0.05$
Reject H_0, x_3 is significant.

應變數 Y2 單變量迴歸分析：

變異數分析					
來源	**自由度**	**平方和**	**平均值平方**	**F 值**	**Pr > F**
模型	2	6.885	3.44	2.94	0.0541
誤差	322	376.655	1.17		
已校正的總計	324	383.54			

（表 3）變異數分析

H_0：$\beta_2 = \beta_3 = 0$ vs. H_a：At least one of $\beta_i \neq 0$
F $= 2.94$，p-value $< 0.0541 \prec \alpha = 0.05$
Not reject H_0, the model is not significant. $R^2 = 0.0180$

參數估計值					
變數	**自由度**	**參數估計值**	**標準誤差**	**t 值**	**Pr > |t|**
Intercept	1	8.655	0.2083	41.56	<.0001
X2	1	0.00016	7.33E-05	2.16	0.0314
X3	1	0.015	0.013	1.16	0.2457

（表 4）參數估計值

H_0：$\beta_2 = 0$ vs. H_a：$\beta_2 \neq 0$
t $= 2.16$，p-value $= 0.0314 < \alpha = 0.05$
Reject H_0, x_2 is significant.

H_0：$\beta_3 = 0$ vs. H_a：$\beta_3 \neq 0$
t $= 1.16$，p-value $= 0.2457 \prec \alpha = 0.05$
Not reject H_0, x_3 is not significant.

多變量檢定迴歸分析：

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=159.5					
統計值		值	F 值	分子自由度	分母自由度	Pr>F
Wilks' Lambda		0.98397342	2.61	2	321	0.0748
Pillai's Trace		0.01602658	2.61	2	321	0.0748
Hotelling-Lawley Trace		0.01628761	2.61	2	321	0.0748
Roy's Greatest Root		0.01628761	2.61	2	321	0.0748

（表 5）多變量分析

H_0：$\boldsymbol{\beta_2 = 0}$ vs. H_a：$\boldsymbol{\beta_2 \neq 0}$
Wilk's $\Lambda = 0.9840$，F $= 2.61$，p-value $= 0.0748 \nless \alpha = 0.05$
Not reject H_0, $\mathbf{x_2}$ is not significant.

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=159.5					
統計值	值	F 值	分子自由度	分母自由度	Pr>F
Wilks' Lambda	0.88430153	21.00	2	321	<.0001
Pillai's Trace	0.11569847	21.00	2	321	<.0001
Hotelling-Lawley Trace	0.13083599	21.00	2	321	<.0001
Roy's Greatest Root	0.13083599	21.00	2	321	<.0001

（表 6）多變量分析

H_0：$\beta_3 = 0$ vs. H_a：$\beta_3 \neq 0$
Wilk's $\Lambda = 0.8843$，F $= 21.00$，p-value $< 0.0001 < \alpha = 0.05$
Reject H_0, x_3 is significant.

Multivariate Statistics and F Approximations					
S=2 M=-0.5 N=159.5					
統計值	值	F 值	分子自由度	分母自由度	Pr>F
Wilks' Lambda	0.86977390	11.60	4	642	<.0001
Pillai's Trace	0.13203542	11.38	4	644	<.0001
Hotelling-Lawley Trace	0.14764388	11.84	4	384.16	<.0001
Roy's Greatest Root	0.13186904	21.23	2	322	<.0001

（表 7）多變量分析

H_0：$\beta_2 = \beta_3 = 0$ vs. H_a：At least one of $\beta_i \neq 0$
Wilk's $\Lambda = 0.8698$，F $= 11.60$，p-value $< 0.0001 < \alpha = 0.05$
Reject H_0, the model is significant.

移除不顯著變數，重新配模

應變數 Y1 單變量迴歸分析：

變異數分析

來源	自由度	平方和	平均值平方	F 值	Pr>F
模型	1	30.67	30.67	27.13	<.0001
誤差	323	365.16	1.131		
已校正的總計	324	395.83			

（表 8）變異數分析

$H_0：\beta_3 = 0$ vs. $H_a：\beta_3 \neq 0$

$F = 27.13$，p-value $< 0.0001 < \alpha = 0.05$

Reject H_0, the model is significant. $R^2 = 0.0775$

參數估計值

變數	自由度	參數估計值	標準誤差	t 值	Pr > |t|
Intercept	1	9.08042	0.09448	96.11	<.0001
X3	1	-0.06655	0.01278	-5.21	<.0001

（表 9）參數估計值

$H_0：\beta_3 = 0$ vs. $H_a：\beta_3 \neq 0$

$t = -5.21$，p-value $< 0.0001 < \alpha = 0.05$

Reject H_0, x_3 is significant.

最終模型：$\hat{y}_1 = 9.08 - 0.07x_3$

應變數 Y2 單變量迴歸分析：

變異數分析					
來源	自由度	平方和	平均值平方	F 值	Pr>F
模型	1	5.303	5.303	4.53	0.0341
誤差	323	378.24	1.171		
已校正的總計	324	383.54			

（表 10）變異數分析

H_0：$\beta_2 = 0$ vs. H_a：$\beta_2 \neq 0$

F $= 4.53$，p-value $= 0.0341 < \alpha = 0.05$

Reject H_0, the model is significant. $R^2 = 0.0138$

參數估計值					
變數	自由度	參數估計值	標準誤差	t 值	Pr > |t|
Intercept	1	8.75	0.19	45.49	<.0001
X2	1	0.00016	7.33E-05	2.13	0.0341

（表 11）參數估計值

H_0：$\beta_2 = 0$ vs. H_a：$\beta_2 \neq 0$

t $= 2.13$，p-value $= 0.0341 < \alpha = 0.05$

Reject H_0, x_3 is significant.

最終模型：$\hat{y}_2 = 8.75 - 0.0002x_3$

- 結論：
  1. 市中心距離每增加 1 單位，則地點評分項目結果會下降 0.07 分。因此，若距離市中心愈遠，地點分數則會愈低。
  2. 每晚最低房價每增加 1 單位，則氛圍評分項目結果會下降 0.0002 分。因此，若每晚最低房價愈高，氛圍分數則會愈低。

五、主成分分析 (Principal Components)

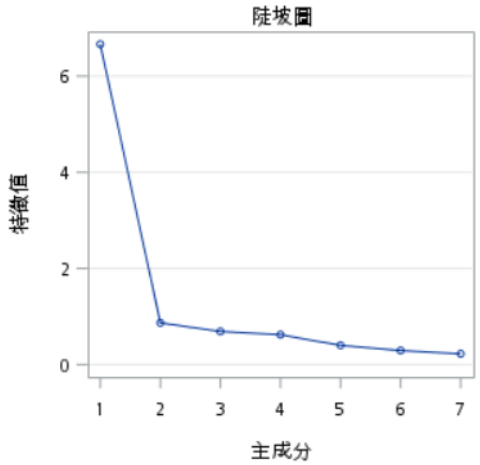
- 目的：透過主成分分析，了解氛圍(X8)、整潔(X9)、設施(X10)、地點(X11)、安全(X12)、人員(X13)、CP 值(X14)評分項目之主導關係。

- 分析結果：

共變異數(Covariance)

共變異數矩陣的特徵值				
	特徵值	差異	比例	累計
1	6.66585302	5.79912235	0.6832	0.6832
2	0.86673067	0.17899023	0.0888	0.7720
3	0.68774044	0.06727534	0.0705	0.8425

（表 12）共變異數矩陣的特徵值



（圖 1）陡坡圖

透過特徵值與陡坡圖，可以知道 Prin1 佔比最高，但解釋能力僅 68.32%，留下 3 個主成分，解釋能力共 84.25%

特徵向量				
		Prin1	Prin2	Prin3
氛圍	X8	0.459123	0.160189	-0.70666
整潔	X9	0.402108	-0.21124	-0.00304
設施	X10	0.454515	-0.11528	-0.01287
地點	X11	0.255649	0.900469	0.336872
安全	X12	0.335156	-0.2876	0.60576
人員	X13	0.34782	-0.13703	0.140319
CP 值	X14	0.349667	-0.06394	-0.01841

（表 13）主成分分析

$$\hat{y}_1 = 0.46x_8 + 0.40x_9 + 0.45x_{10} + 0.26x_{11} + 0.34x_{12} + 0.35x_{13} + 0.35x_{14}$$
$$\hat{y}_2 = 0.16x_8 - 0.21x_9 - 0.12x_{10} + 0.90x_{11} - 0.29x_{12} - 0.14x_{13} - 0.06x_{14}$$
$$\hat{y}_3 = -0.71x_8 - 0.003x_9 - 0.01x_{10} + 0.34x_{11} + 0.61x_{12} + 0.14x_{13} - 0.02x_{14}$$

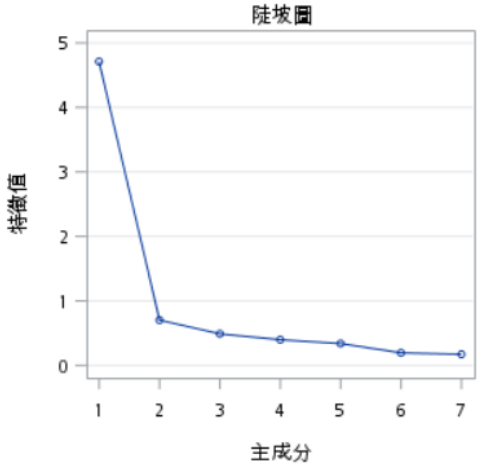
結論：

1. Prin1 為綜合指標，沒有哪個變數特別主導。
2. Prin2 由地點(X11)主導，表示地點評分影響較大。
3. Prin3 由安全(X12)主導，另外氛圍(X8)雖為負值但影響力也大，表此二者對於評分的影響力較大。

標準化(Standardized)

相關矩陣的特徵值				
	特徵值	差異	比例	累計
1	4.7108754	4.010191	0.673	0.673
2	0.7006839	0.211905	0.1001	0.7731
3	0.4887794	0.090962	0.0698	0.8429

（表 14）共變異數矩陣的特徵值



（圖 2）陡坡圖

透過特徵值與陡坡圖，可以知道 Prin1 佔比最高，但解釋能力僅 67.30%，留下 3 個主成分，解釋能力共 84.29%

特徵向量				
		Prin1	Prin2	Prin3
氛圍	X8	0.387677	0.053758	-0.32859
整潔	X9	0.390174	-0.14185	-0.3941
設施	X10	0.416389	-0.07145	-0.26458
地點	X11	0.279858	0.929212	0.194706
安全	X12	0.366177	-0.27391	0.690107

人員	X13	0.38448	-0.16461	0.365054
CP 值	X14	0.404797	-0.07956	-0.13891

（表 15）主成分分析

$$\hat{y}_1 = 0.39\,z_8 + 0.39\,z_9 + 0.42\,z_{10} + 0.28\,z_{11} + 0.37\,z_{12} + 0.38\,z_{13} + 0.40\,z_{14}$$
$$\hat{y}_2 = 0.05\,z_8 - 0.14\,z_9 - 0.07\,z_{10} + 0.93\,z_{11} - 0.27\,z_{12} - 0.16\,z_{13} + 0.08\,z_{14}$$
$$\hat{y}_3 = -0.33\,z_8 - 0.39\,z_9 - 0.26\,z_{10} + 0.19z_{11} + 0.69z_{12} + 0.37\,z_{13} - 0.14\,z_{14}$$

結論：

1. Prin1 為綜合指標，沒有哪個變數特別主導。
2. Prin2 由地點(X11)主導，表示地點評分影響較大。
3. Prin3 由安全(X12)主導，表示安全評分影響較大。

六、多變量分析（MANOVA）

- 目的：將資料以每晚最低房價分為兩群，第 1 群為房價小於 2000 元的旅店，第 2 群為房價大於(含等於)2000 元的旅店，並將氛圍(X8)、整潔(X9)、設施(X10)、地點(X11)、安全(X12)、人員(X13)、CP 值(X14)評分項目透過多變量分析找出與該房價較具差異的旅店。

- 分析結果：

來源	自由度	平方和	均方	F 值	Pr>F
模型	7	3.029	0.4327	2.64	0.0114
誤差	317	51.89	0.164		
已校正的總計	324	54.92			

（表 16）變異數分析

來源	自由度	類型 Ⅲ SS	均方	F 值	Pr>F
氛圍	1	0.5251	0.5251	3.21	0.0743
整潔	1	0.4737	0.4737	2.89	0.0899
設施	1	0.0177	0.0177	0.11	0.7424
地點	1	0.2271	0.2271	1.39	0.2398
安全	1	0.4545	0.4545	2.78	0.0967
人員	1	0.0039	0.0039	0.02	0.8778
CP 值	1	1.3114	1.3114	8.01	0.0049

（表 17）變異數分析

- 結論：

  透過多變量分析，我們可以發現整體模型的表現是顯著的，但是將來源細分，我們可以發現僅有 CP 值（X14）是顯著的，不適合進行判別分析（Discrimination & Classification），也意味者可以在日本以較低房價卻住到品質較好之飯店。

七、結論

透過多變量線性迴歸、主成分、多變量分析，整體綜合評分對於選擇飯店是一個重要指標。地點、安全是細項評分中具有影響力的。另外，透過此次分析，我們發現在日本以較低房價卻住到品質較好之飯店。