

高維度資料分析期末報告

俄羅斯房價預測

410578046 張馨云

410578012 羅季安

410578021 李承志

410578053 王勃淵

目錄

目錄	2
圖表目錄	3
1. 導論	5
1.1. 分析背景	5
1.2. 分析動機與目的	5
1.3. 文獻回顧	5
1.4. 所需資料與資料分析之重要性	5
1.5. 資料分析流程	6
1.6. 研究分析困難之處	6
2. 資料處理	7
2.1. 資料結構	7
2.2. 描述性統計與變數摘要	7
2.3. 探索資料 (EDA)	11
2.4. 資料分布與遺失處理	15
2.5. 資料降維	16
2.5.1. $n \gg p$	17
2.5.2. $n = p$	21
2.5.3. $n \ll p$	27
2.5.4. 小結	32
2.6. 模型預測	35
2.6.5. $n \gg p$	35
2.6.6. $n \ll p$	37
3. 結論	41
4. 參考書目與文章	42
附錄 I、資料結構	43
附錄 II、連續變數摘要表	46
附錄 III、使用的 R 套件	51
附錄 IV、工作分配表	52

圖表目錄

圖 1	資料分析流程圖	6
表 1	各變數類型	7
表 2	Summary of Continuous Variables.....	7
表 3	Tables of Categorical Variables.....	8
表 4	計數變數之遺失數與比率	9
圖 2	Level Ratio of ecology.....	10
圖 3	Level Count of sub_area.....	10
圖 4	建造年分 (Count of build_year)	11
圖 5	年月成交量	11
圖 6	交易時間與購屋用途.....	12
圖 7	不同交易目的之價格分布	12
圖 8	交易日期與平均價格.....	12
圖 9	不同地區價格.....	13
圖 10	平均價格前 20 之地區	13
圖 11	價錢與總面積之散佈圖	14
圖 12	以公寓樓層數分類之 log (價格) 盒形圖	14
圖 13	價格與公共運輸站之間的關係	15
圖 14	Missing Pattern of Categorical Variables.....	15
圖 15	Missing Pattern of Continuous Variables.....	16
圖 16	QQ-Plot of price_doc.....	16
圖 17	Scree Plot of PCA.....	17
圖 18	變數在各個 PC 貢獻的比例	18
圖 19	Correlation Circle of PCA.....	18
圖 20	使用 K-Means 分群法對前兩個主成分作圖	19
圖 21	使用 K-Means 分群法對 MDS 作圖.....	19
圖 22	ISOMAP of k=5000.....	19
圖 23	ISOMAP of k=1500.....	19
圖 24	ISOMAP of k=2500.....	20
圖 25	使用 K-Means 分群法對 ISOMAP 作圖.....	20
圖 26	CoRanking Matrix of PCA.....	20
圖 27	Co-Ranking Matrix of MDS.....	21
圖 28	Co-Ranking Matrix of ISOMAP.....	21
圖 29	Comparison of Covariance Matrix	22
圖 30	Scree Plot of PCA.....	22
圖 31	Correlation Circle of PCA.....	23
圖 32	Comparison of Different K of ISOMAP.....	24

圖 33	使用 K-Means 分群法對前兩個主成分作圖	24
圖 34	使用 K-Means 分群法對 MDS 作圖.....	25
圖 35	使用 K-Means 分群法對 ISOMAP 作圖.....	25
圖 36	Co-Ranking Matrix of PCA.....	26
圖 37	Co-Ranking Matrix of MDS.....	26
圖 38	Co-Ranking Matrix of ISOMAP.....	27
圖 39	Comparison of Covariance Matrix.....	28
圖 40	Eigenvalue between result treatment and untreatment result	28
圖 41	Cummulative Proportions of eigenvalues between True & Shrinkage	29
圖 42	PCA.....	30
圖 43	MDS	30
圖 44	ISOMAP.....	30
圖 45	Co-Ranking Matrix of PCA.....	31
圖 46	Co-Rnaking Matrix of MDS.....	31
圖 47	Co-Ranking Matrix of ISOMAP.....	32
圖 48	各維度縮減方法在 $n \gg p$ 之 LCMC	32
圖 49	各維度縮減方法在 $n = p$ 之 LCMC	33
圖 50	各維度縮減方法在 $n \ll p$ 之 LCMC	33
圖 51	不同資料大小進行 PCA 之 LCMC	34
圖 52	不同資料大小進行 MDS 之 LCMC	34
圖 53	不同資料大小進行 ISOMAP 之 LCMC	35
圖 54	房屋價格與以 MDS 降維後前兩維的關係圖	36
圖 55	Pairwise of first 2 MDS & price.....	36
表 5	以 MDS 降維後線性回歸之參數	36
表 6	regression of backward elimination.....	37
圖 56	Mean Square Error of $\text{Log}(\lambda)$	37
表 7	lasso regression.....	37
圖 57	38
圖 58	39
圖 59	heatmap of Correlation Matrix.....	39
圖 60	$\text{Log}(\lambda)$ v.s Lasso Coefficient.....	40
附錄表 1	資料結構表	43
附錄表 2	連續變數摘要表.....	46
附錄表 3	使用 R 套件一覽表.....	51
附錄表 4	工作分配表	52

1. 導論

1.1. 分析背景

房屋價格 (Housing Price) 是消費者和開發商極為關注的，在財務預算規劃時 (無論是個人預算還是公司預算)，房屋價格是每個人最重視的也是最不可確定的龐大支出。然而，對於大筆的消費支出，我們總是希望銀行房仲業者能夠先行調查房屋周遭的便利性以提供購買者完整的環境資訊。所以，俄羅斯歷史最為悠久也是規模最大的銀行，Sber Bank 收集了房地產價格以及其他生活指標來幫助客戶的財務規劃，目的是使客戶、開發商和貸方在簽訂租約或購買建築物時更加安心。

1.2. 分析動機與目的

成家立業是每個人成長過程中必須經歷的，而以臺灣來說，近年來都市蓬勃發展，六都以及其他縣市的房屋價格受交通便利性、住宅附近環境等等影響。我們好奇的是，在同為部分亞洲地區的俄羅斯是否如同臺灣一樣，在首都地區莫斯科或是其他工業重地例如貝加爾湖地區的房價有較高的情形。在這筆資料中，Sber Bank 收集了關於房屋至學校、地鐵、公園等公眾場所的距離以及行走所需時間，以及該地區之男女人數與不同工作年齡等指標。我們希望藉由總計超過 200 個變數中，經過篩選並以不同降維的方式探討在變數濃縮後可以確切反應出這些指標變數在房屋價格上的關聯，藉由線性迴歸模型選出最適合代表此筆房屋資料的模式。

1.3. 文獻回顧

根據盛世良在《2001 年前房價趨勢》一文中提到，在 1990 年代全世界受到經濟危機的影響導致房價下跌，俄羅斯也不例外，在大都市聖彼得堡房價下跌 10%~20% 不等，葉卡捷琳堡更跌了 40%，剩下每坪 334 美元。然而，根據聶書嶺在《中亞信息》提到，隨著經濟蕭條後的逐漸復甦，在 2003 年開始俄羅斯房價也逐漸回升，在當年的初級市場中房屋價格上漲了 20%~25%。岳萍更在《俄羅斯房價飛漲》中提及，到了 2006 年俄羅斯房價平均上漲 73%，且有逐年攀升的情形。因此，房價炒作、炒地皮並非只發生在臺灣，而這些也不禁讓我們聯想到，在本次的資料分析中大多的房屋價格登陸時間落在 2010 年至 2016 年，所以，反應房屋價格時是否也受到當時外在因素影響也是我們值得注意的事情。

1.4. 所需資料與資料分析之重要性

在分析此筆資料中，除了統計上對於資料分析的相關知識外，還需要對於俄羅斯莫斯科地理以及一般房地產有足夠的了解，例如：建材、土地

與莫斯科各區等地理與經濟之背景知識。

也如同先前所說明，買房是每個人必經的階段，所以我們在購屋的同時需要考量到房價是否與其坪數、便利性等等成正比，人們不希望自己冤大頭買到不適合的房子，因此分析此方屋價格資料的重要性在於能夠確切反應影響房屋價格之因素，使民眾可以選擇適合自己避風港最適合的地點。

1.5. 資料分析流程

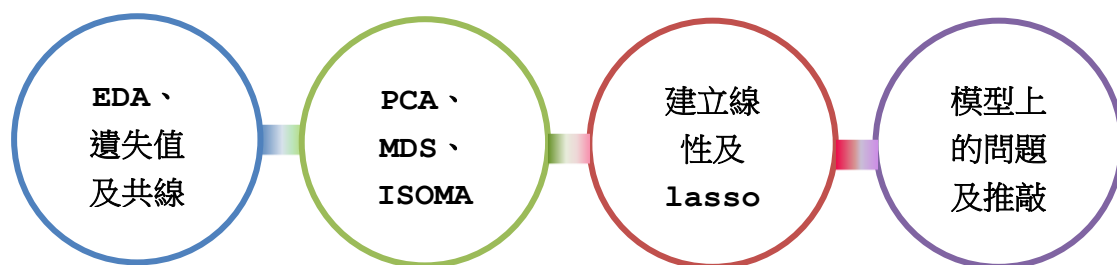


圖 1 資料分析流程圖

1.6. 研究分析困難之處

首先針對資料中的共線性的問題，我們以刪除某些變數以及將變數轉換成另一個變數（例如將男女生的人口數轉換成性別比）進行修正，此外在進行降維的部分因為資料觀測值多，所以在進行 ISOMAP 時發生資料容易分裂的情形，因此最後取一個 $k=2500$ 的結果作為選擇，耗時兩天運行此降維過程。最後在模型配適的部分因為資料的特質導致配適模型上的困難，文內將說明此情況。

2. 資料處理

2.1. 資料結構

表 1 各變數類型（擷取部分，其餘參見附錄 I）

類型	變數名稱		
連續資料	numeric		
	additional_education_km	area_m	basketball_km
	incineration_km	indust_part	industrial_km
	metro_min_walk	mkad_km	mosque_km
	museum_km	nuclear_reactor_km	office_km
	water_treatment_km	workplaces_km	zd_vokzaly_avto_km
計數資料	integer		
	additional_education_raion	afe_count_3000_price_2500	afe_count_5000_na_price
	children_preschool	children_school	church_count_1000
	full_sq	healthcare_centers_raion	hospital_beds_raion
	id	ID_big_road1	ID_big_road2
	X7_14_all	X7_14_female	X7_14_male
Date			
	build_year	timestamp	
Factor			
	big_market_raion	big_road_line	culture_objects_top_
	detention_facility_raion	incineration_raion	nuclear_reactor_raion
	oil_chemistry_raion	radiation_raion	railroad_line
	railroad_terminal_raion	sub_area	ecology
	material	product_type	state
	thermal_power_plant_raion	water_line	

2.2. 描述性統計與變數摘要

表 2 Summary of Continuous Variables（僅含有遺失值之變數，其餘見附錄 II）

Continuous Variables	min	1 st Qu.	Median	Mean	3 rd Qu.	Max	NMiss	MissRate
life_sq	0	20	30	34.4	43	7478	6383	0.2095
floor	0	3	6.5	7.671	11	77	167	0.0055
max_floor	0	9	12	12.56	17	117	9572	0.3141
num_room	0	1	2	1.91	2	19	9572	0.3141
kitch_sq	0	1	6	6.399	9	2014	9572	0.3141

preschool_quota	0	1874	2854	3271	4050	11926	6688	0.2195
cafe_sum_1000_min_price_avg	300	543.2	669.2	710.9	839.3	2500	6524	0.2141
school_quota	1012	5782	7377	8325	9891	24750	6685	0.2194
hospital_beds_raion	0	520	990	1191	1786	4849	14441	0.4739
metro_min_walk	0	11.48	20.45	42.74	45.32	711.22	25	0.0008
metro_km_walk	0	0.957	1.704	3.561	3.777	59.268	25	0.0008
railroad_station_walk_km	0.0282	1.9314	3.23554	4.38694	5.14764	24.65304	25	0.0008
railroad_station_walk_min	0.3378	23.1765	38.82650	52.6433	61.7717	295.8365	25	0.0008
cafe_sum_500_min_price_avg	300	500	666.7	741.3	954.8	4000	13281	0.4359
cafe_sum_500_max_price_avg	500	1000	1167	1247	1500	6000	13281	0.4359
cafe_avg_price_500	400	750	916.7	994.2	1250	5000	13281	0.4359
cafe_sum_1000_max_price_avg	500	1000	1143	1207	1400	4000	6524	0.2141
cafe_avg_price_1000	400	750	912.5	958.8	1120	3250	6524	0.2141
cafe_sum_1500_min_price_avg	300	585.7	692.3	714.1	821.4	2500	4199	0.1378
cafe_sum_1500_max_price_avg	500	1000	1167	1206	1367	4000	4199	0.1378
cafe_avg_price_1500	400	795	926.3	960	1093.8	3250	4199	0.1378
cafe_sum_2000_min_price_avg	300	607.7	683.3	720	791.7	2166.7	1725	0.0566
cafe_sum_2000_max_price_avg	500	1000	1156	1211	1322	3500	1725	0.0566
cafe_avg_price_2000	400	823.5	919.2	965.4	1057.2	2833.3	1725	0.0566
cafe_sum_3000_min_price_avg	300	650	711.1	765.9	815.6	1833.3	991	0.0325
cafe_sum_3000_max_price_avg	500	1102	1212	1283	1333	3000	991	0.0325
cafe_avg_price_3000	400	875.8	961.1	1024.6	1083.3	2416.7	991	0.0325
prom_part_5000	0.21	6.05	8.98	10.35	14	28.56	178	0.0058
cafe_sum_5000_min_price_avg	300	670.9	721.7	765.1	816.7	1875	297	0.0097
cafe_sum_5000_max_price_avg	500	1144	1212	1278	1346	3000	297	0.0097
cafe_avg_price_5000	400	909.4	966.7	1021.7	1091.7	2437.5	297	0.0097
price_doc	100000	4740002	6274411	7123035	8300000	11111112		

表 3 Tables of Categorical Variables

Categorical Variables	Levels	MissRate
ecology	Excellent, good, no data, poor, satisfactory	0
sub_area	Ajeroport,Akademicheskoe,etc	0
material	1~6	0.3141
state	1~4	0.4449
build_year	1961~2018	0.4465
	Levels(Ratio)	
	Investment	OwnerOccupier
product_type	0.6382	0.3618
		0

	No	Yes	
culture_objects_top_25	0.9367	0.0633	0
thermal_power_plant_raion	0.9457	0.0542	0
incineration_raion	0.9240	0.0760	0
oil_chemistry_raion	0.9903	0.0097	0
radiation_raion	0.6432	0.3568	0
railroad_terminal_raion	0.9627	0.0373	0
big_market_raion	0.9074	0.0926	0
nuclear_reactor_raion	0.9717	0.0283	0
detention_facility_raion	0.9001	0.0999	0
water_1line	0.9233	0.0767	0
big_road1_1line	0.9744	0.0256	0
railroad_1line	0.9706	0.0293	0

表 4 計數變數之遺失數與比率

Count Variables	NMiss	MissRate
build_count_block	4991	0.1638
build_count_wood	4991	0.1638
build_count_frame	4991	0.1638
build_count_brick	4991	0.1638
build_count_monolith	4991	0.1638
build_count_panel	4991	0.1638
build_count_foam	4991	0.1638
build_count_slag	4991	0.1638
build_count_mix	4991	0.1638
build_count_before_1920	4991	0.1638
build_count_1921.1945	4991	0.1638
build_count_1946.1970	4991	0.1638
build_count_1971.1995	4991	0.1638
build_count_after_1995	4991	0.1638

由圖 3 可知，房屋的地區分布非常零散，房屋數量超過 1000 的地區以 Poselenie Sosenskoe 為最，Nekrasovka 次之，緊接著是 Poselenie Vnukovskoe；房屋數量超過 500 的以 Poselenie Moskovskij 為最，依序為 Poselenie Voskresenskoe、Tverskoe、Krjukovo 以及 Mar'ino；其餘地區房屋數量皆低於 500：如 Poselenie Mosrentgen 及 Poselenie Rjazanovskoe 等諸多地區房屋數量低於 50，Poselenie Klenovskoe 的房屋數量甚至僅有 1，各地區房屋數量相距懸殊。在圖 4 中可

看出，房屋建造時間大多於 1960 年代與 2010 年後，呈現雙峰的分佈。

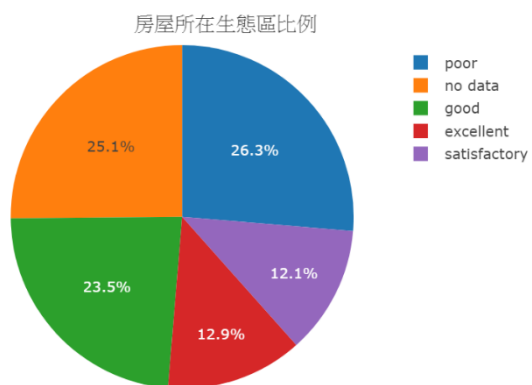


圖 2 Level Ratio of ecology

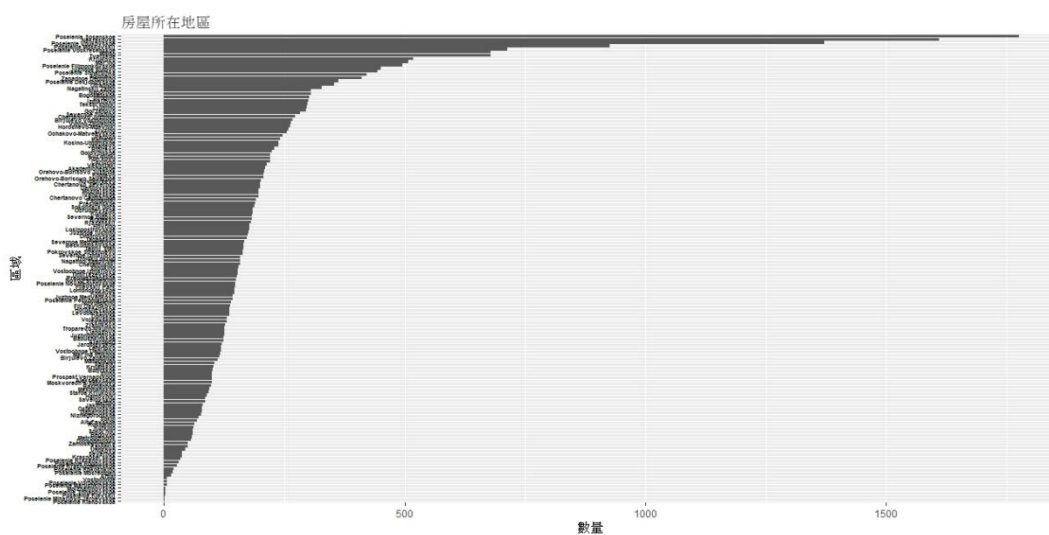


圖 3 Level Count of sub_area

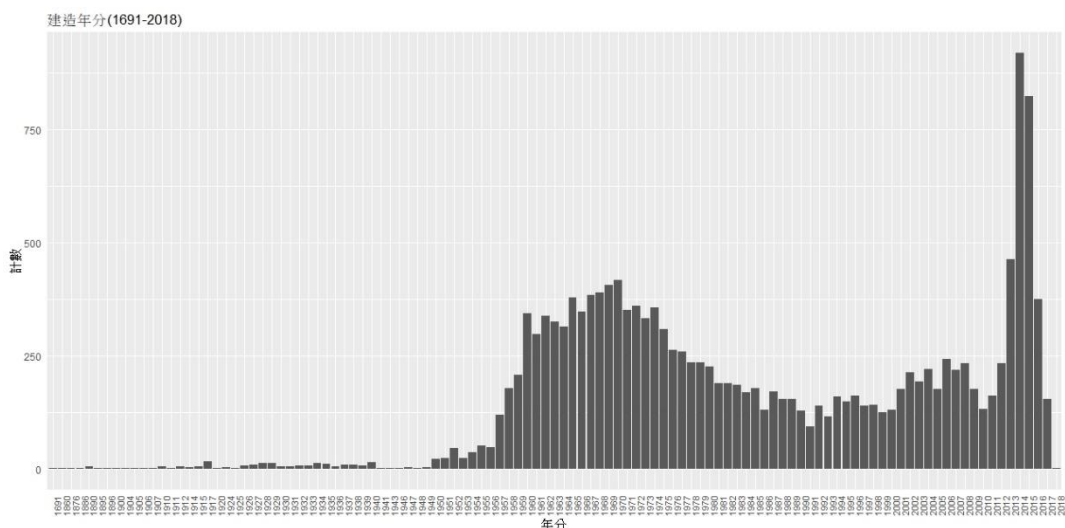


圖 4 建造年分(Count of build_year)

2.3. 探索資料 (EDA)

圖 6 可以發現：平均而言，購屋多半為投資所用，尤其 2012 年中之前以自住為目的購屋者微乎其微，而 2012 年中至 2013 年中，以自住為途之購屋者短暫地高於以投資為途之購屋者。綜合圖 5 以及圖 6，於 2014 年終可見有股購屋潮，以投資為途購屋者購屋數目達到高峰，而以自住為途購屋者也有明顯增長，相比之下，2015 年開始，不論購屋用途，購屋數量皆急劇地下滑，或許與 2014 年國際制裁後 2015 年俄羅斯當局進口替代政策有關：雖然以美金計價的莫斯科房市價格下跌，由於平均工資低且盧布貶值，經濟危機與家庭收入下降之夾擊下，購屋者貸款償付能力價降，進而影響購屋數量。且從圖 8 可知房屋交易價格雖有短暫起色，然 2015 年中平均交易價格到達歷年高峰，致使房屋交易數量動盪。比較不同購屋目的之價錢分布（圖 7），發現兩種購屋目的之價錢分布與皆右偏且近乎重疊，顯示購屋目的與價錢並無關聯。

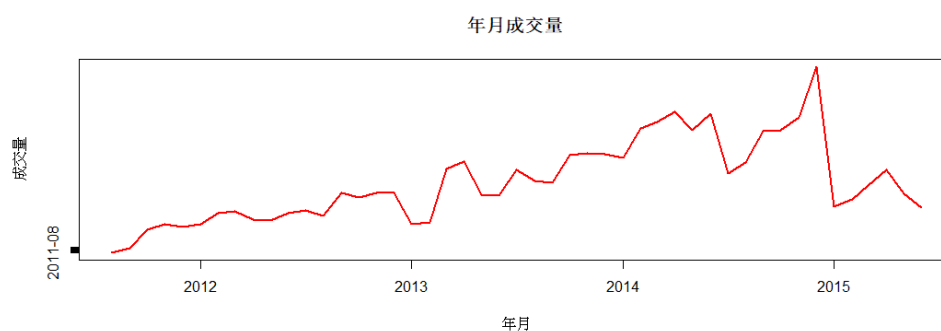


圖 5 年月成交量

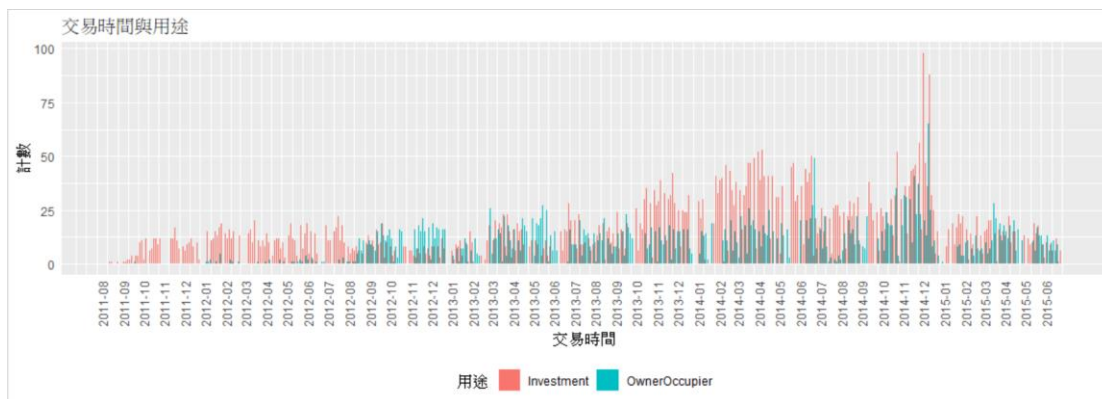


圖 6 交易時間與購屋用途

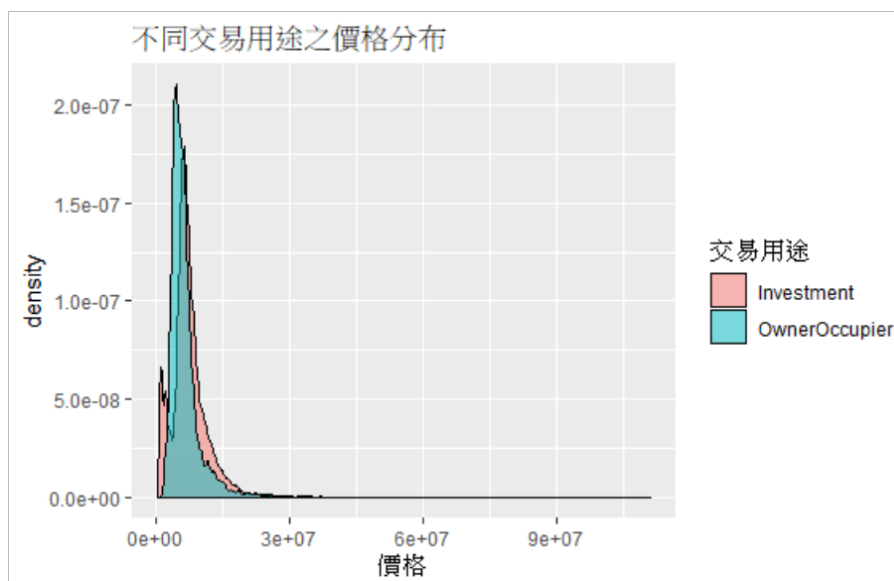


圖 7 不同交易目的之價格分布

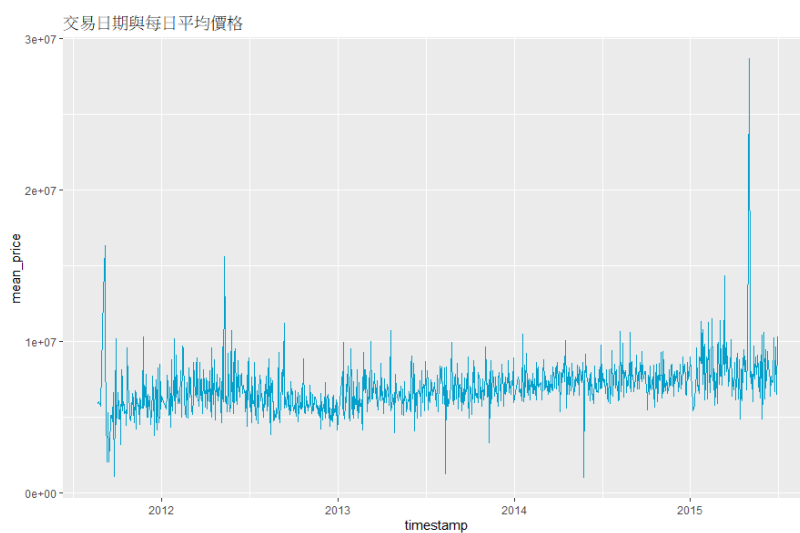


圖 8 交易日期與平均價格

圖 9 可以發現不同地區的平均價格落差其實並不大，於是我們進一步繪製平均價格前 20 地區之長條圖（圖 10），對比圖 3，價錢反應於交易

量上，前幾名的地區交易量的確較低。

圖 11、圖 12 以及圖 13 探討一些可能影響價格的因素：從價格與總面積之散佈圖（圖 11）可以看出，房屋居住面積大多落在 200 平方公尺以下，價格大致隨總面積增加而急速增長。一般而言，公寓會因景觀與安寧等因素隨樓層越高價位越高，而莫斯科的房價並沒有明顯反應出此一現象。從圖 12 可以發現，大約在 26 樓（已經算是高樓層）以下樓層之房價都沒有太大的波動，而超過 27 樓之樓層房價才有明顯地提升。圖 13A 以及圖 13B 分別表示價格與火車站之距離關係以及價格與公車站距離關係：無論是公車站或火車站，同一終點站之房價皆呈現帶狀分布，顯示其並無明顯地影響房屋價格。

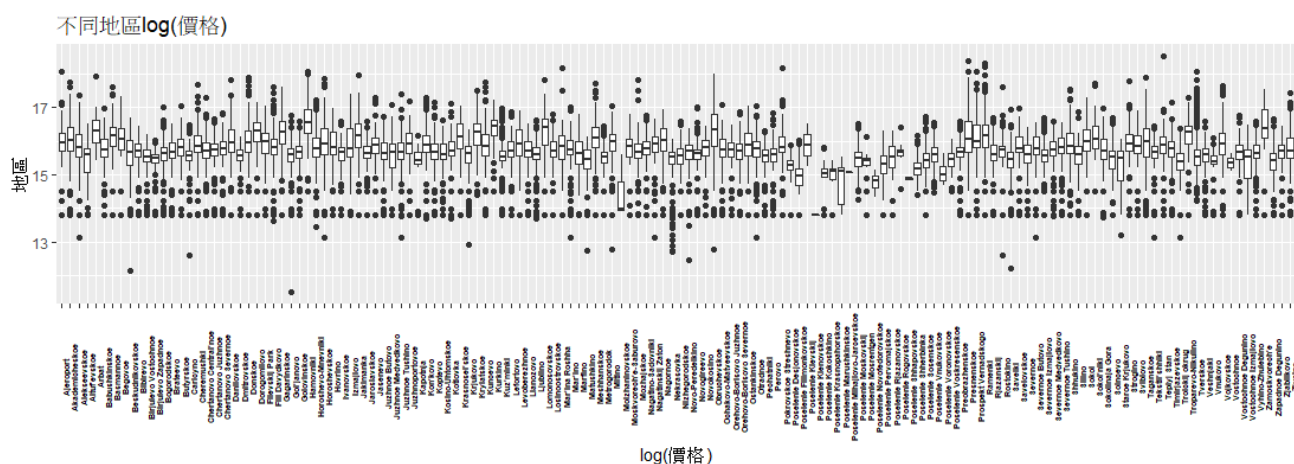


圖 9 不同地區價格

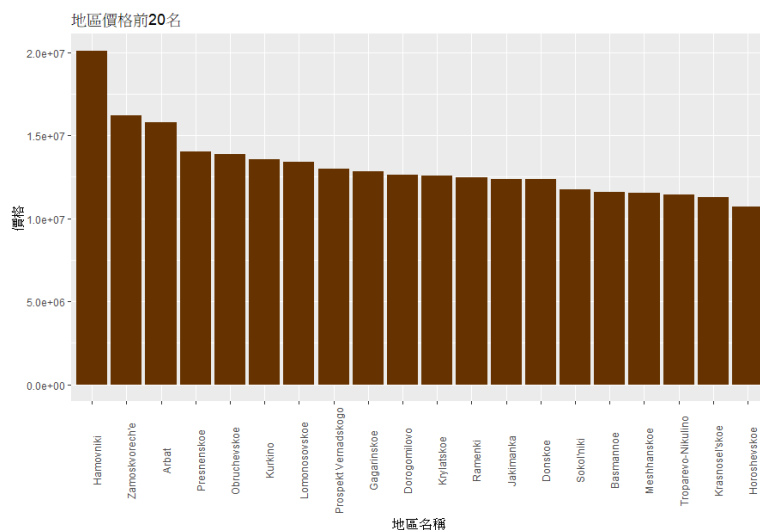


圖 10 平均價格前 20 之地區

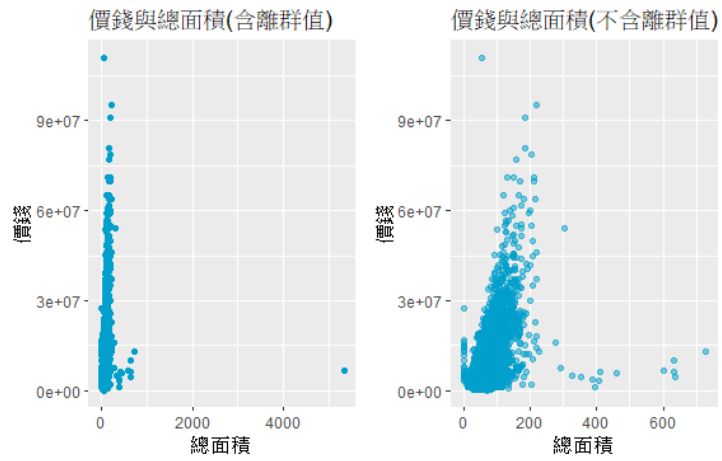


圖 11 價錢與總面積之散佈圖(scatter plot of priice_doc & full_sq)

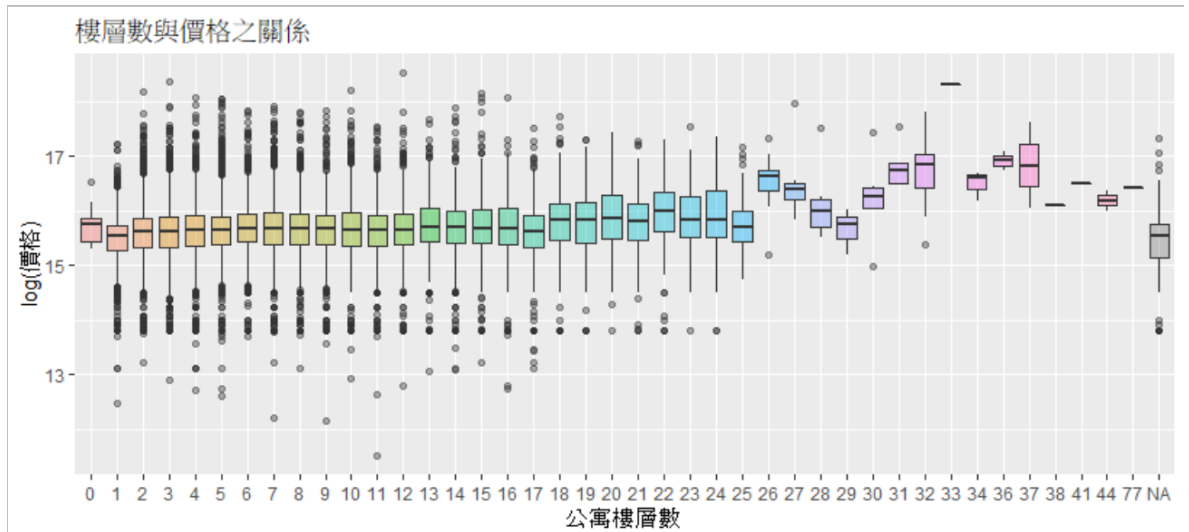


圖 12 以公寓樓層數分類之 $\log(\text{價格})$ 盒形圖

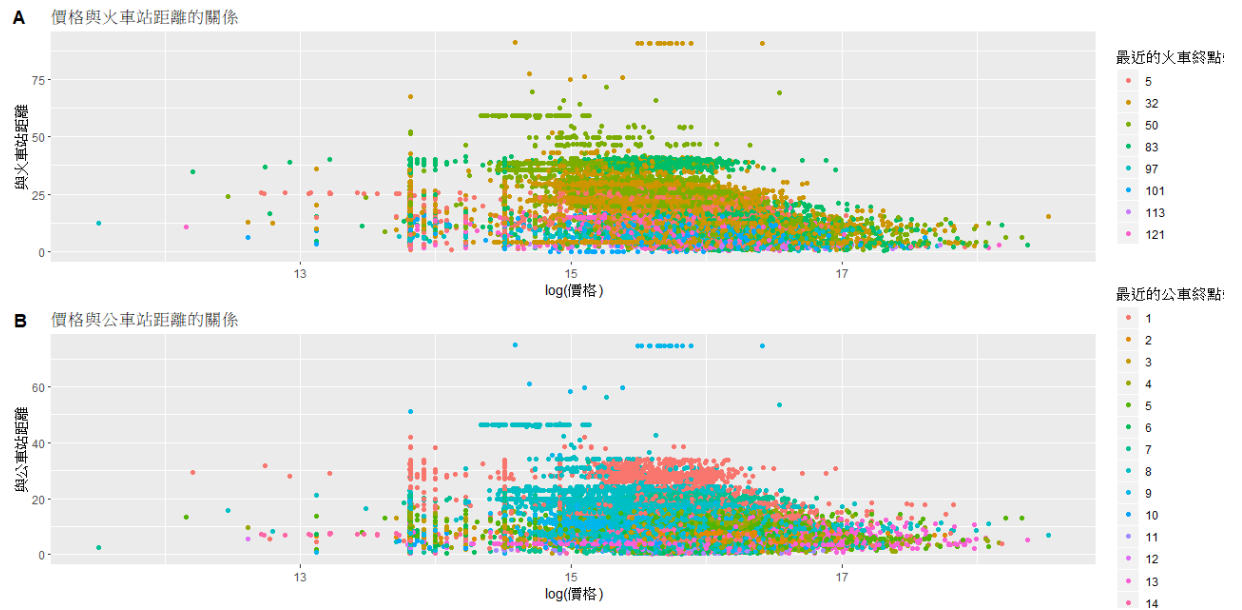


圖 13 價格與公共運輸站之間的關係

2.4. 資料分布與遺失處理

假設資料之遺失值機制為 MCAR，從類別型變數的遺失模式（圖 14）可以發現，多數類別變數沒有遺失，而 material、state 與 build_year 的遺失比率接大於 0.2，我們在後續建模的部分過濾掉這些變數。從連續型變數的遺失模型（圖 15）可以發現，有近 1/5 的連續型變數有遺失值，後續建模與降低維度時，過濾掉其遺失比率超過 0.2 的變數。

繪製反應變數 price_doc 的 QQ-Plot（圖 16），曲線向左彎曲，顯示其分布右偏。

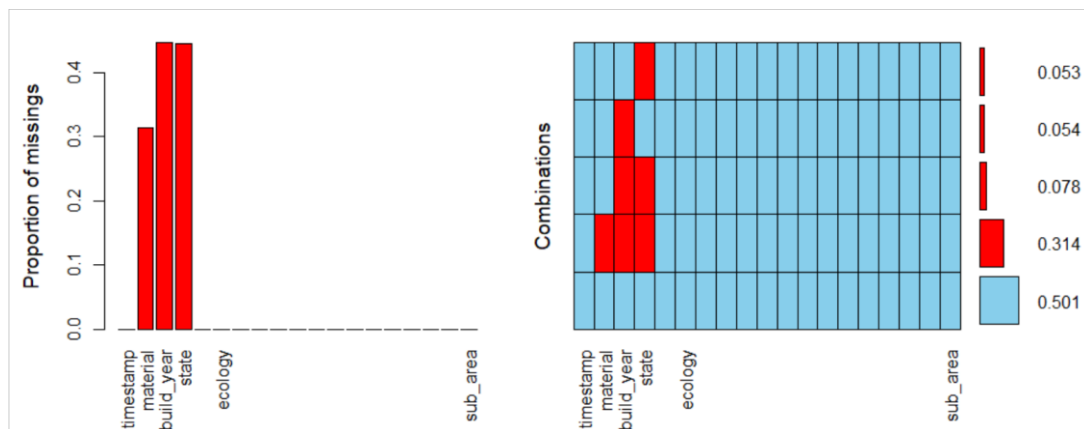


圖 14 Missing Pattern of Categorical Variables

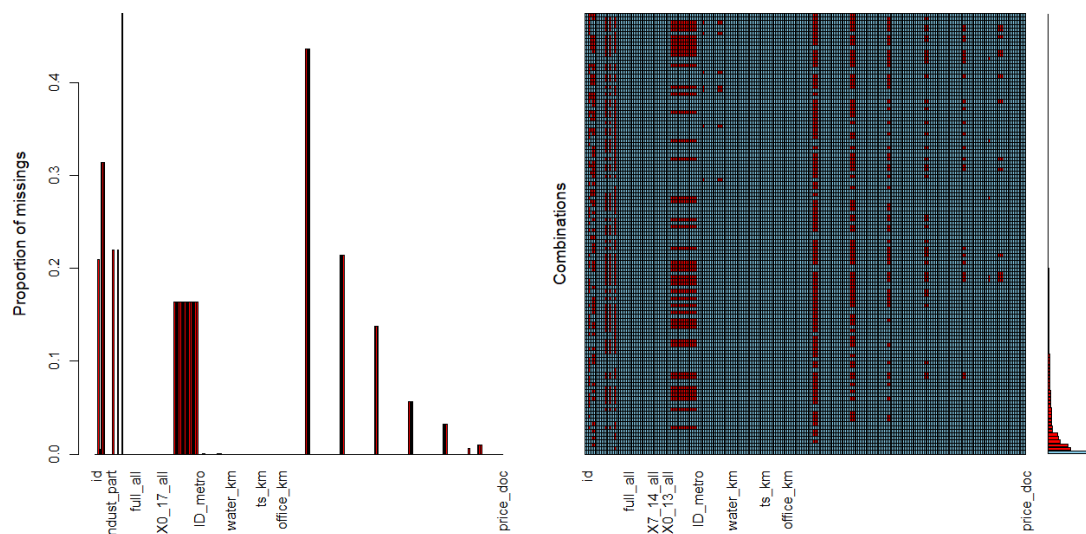


圖 15 Missing Pattern of Continuous Variables

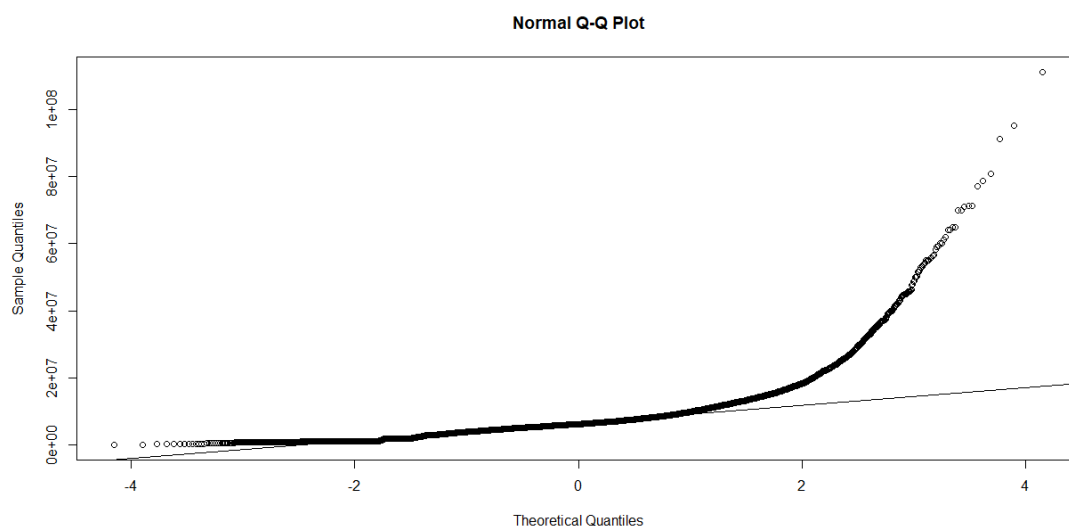


圖 16 QQ-Plot of price_doc

除了過濾掉遺失比率大於 0.2，考慮到含有 ID 的名目型資料提供的資訊與對補值的幫助有限，將其從資料中剷除；而為解決變數之間的共線性，將以男女分類的變數相除（young、work、ekder...etc.），形成男女比之新變數。篩選與過濾變數以後，以 CART 決策樹進行遺失值預測插補（使用 R 語言 mice 套件），產生 20 個個插補資料集，選取第一個插補資料用以進行後續資料分析。

2.5. 資料降維

資料經過插補以後，依 $n \gg p$ 、 $n \approx p$ 、 $n \ll p$ 的資料型態，透過將插補資料隨機抽樣（取 $n = 16037$ ， $n=114$ ， $n=20$ ），產生三個新的資料。對這三個資料皆使用 PCA、ISOMAP、MDS 等降維方法。此外，整個資料在處理過後有 26037 筆觀測值，我們選取 16037 當作 $n \gg p$ 的原因是希望留

10000 筆資料當作 testing data。

2.5.1. $n \gg p$

在 PCA 的部分，陡坡圖（圖 17）紀錄各維度與其解釋資料的變異比例，第四個維度以後解釋的變異程度趨於平緩，而為了達到解釋比例超過 70%，我們最後選擇 10 個維度作為降維後的新線性組合變數，即圖中所顯示的 10 維。圖 18 顯示前 4 個主成分之中，各個變數貢獻的比例，可以發現在第 1 與第 3 主成分之中，解釋變異大多由與各公共設施及消費場所的距離有關的變數貢獻；第 2 主成分由各距離範圍內咖啡廳與餐廳的平均消費有關的變數貢獻；第 4 主成分的貢獻組成則較繁雜。此外，在使用 kmeans 將原始變數分群後，從 Correlation Circle（圖 19）展示之變數組內與主成分之間的關係發現，變數大致分為三大類：1. 文化遺產、高等教育、購物中心等機構數量（綠色區塊）；2. 各距離範圍內咖啡廳與餐廳的平均消費（藍色區塊）；3. 到各公共設施與消費場所的距離（紅色區塊）。

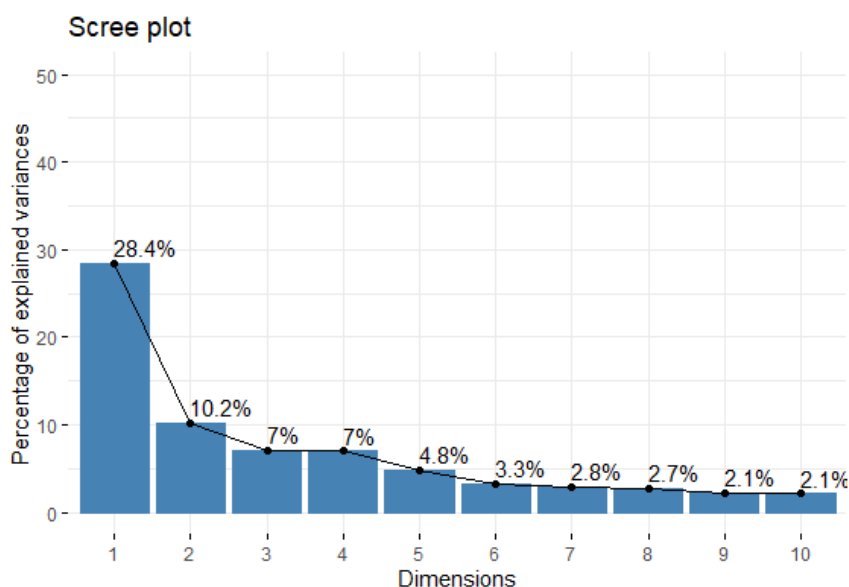


圖 17 Scree Plot of PCA

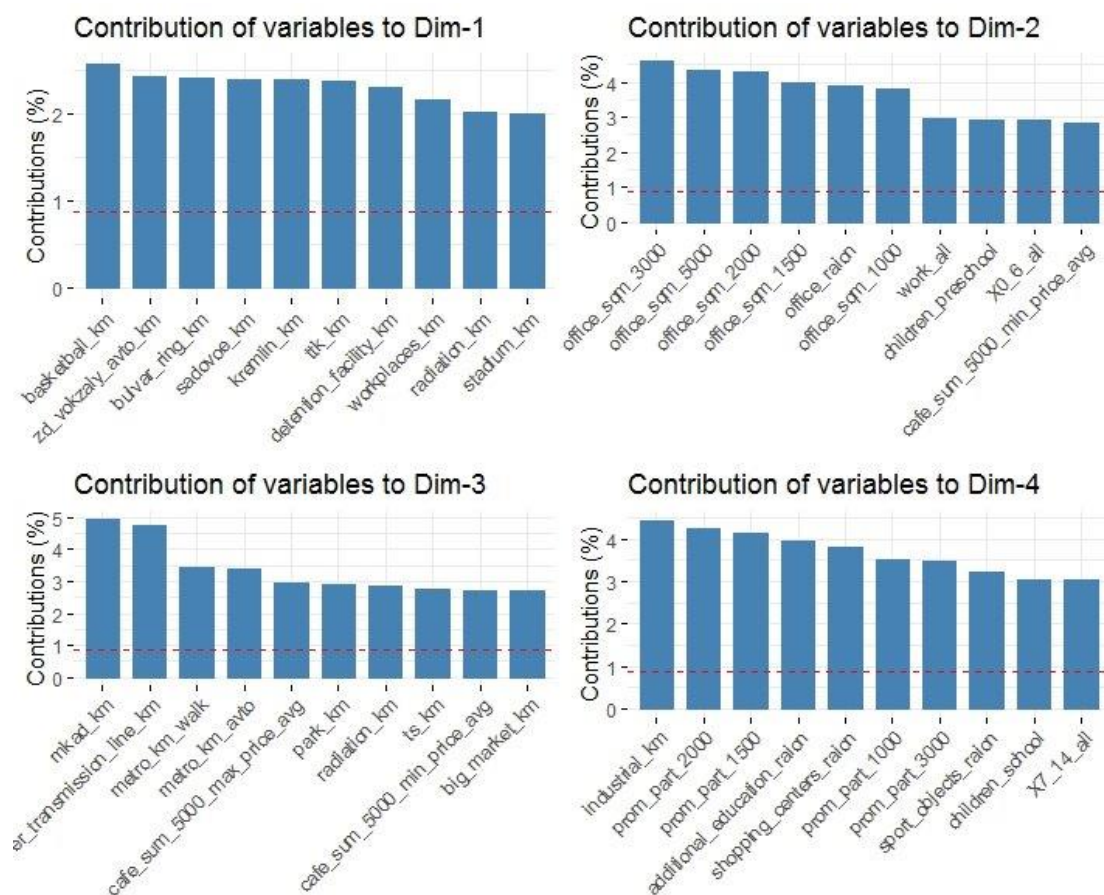


圖 18 變數在各個 PC 貢獻的比例

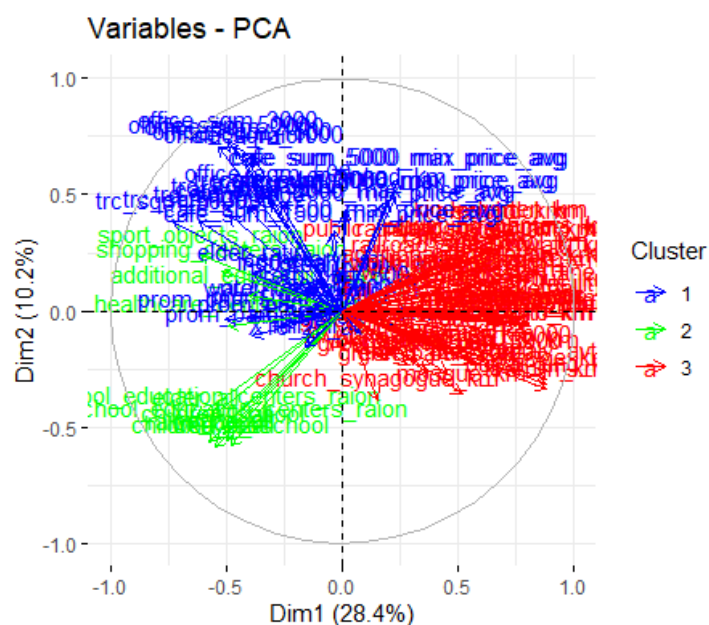


圖 19 Correlation Circle of PCA

若將 PCA 前兩維的座標繪至平面座標可以發現，使用 kmeans 進行分群大

約可以區分成 4 組，其中以黑色、紅色以及綠色的分群較佳，藍色的分群相較之下離散很多，分群效果差。而 MDS 在降維求解的過程中因與 PCA 類似，故我們不難發現在圖 20 中的散佈圖與圖 19 相同，只是左右相反而以，然而這是因為 MDS 只是轉軸不同，故在前兩維的部分結果會呈現相似。

最後在 ISOMAP 的降維，首先我們需要選擇適當的 k nearest neighbor 數量，然而資料分佈較為分散，為了防止資料粉碎，最終我們選擇 $k=1500, 2500, 5000$ 進行描繪，而從圖 21、圖 22、圖 23 的不同 k 中選擇 2500 作為 k nearest neighbor 的數量。同樣地，對於 ISOMAP 使用 k means 進行分群，發現亦是分成 4 群，其中黑色的分群效果較差。

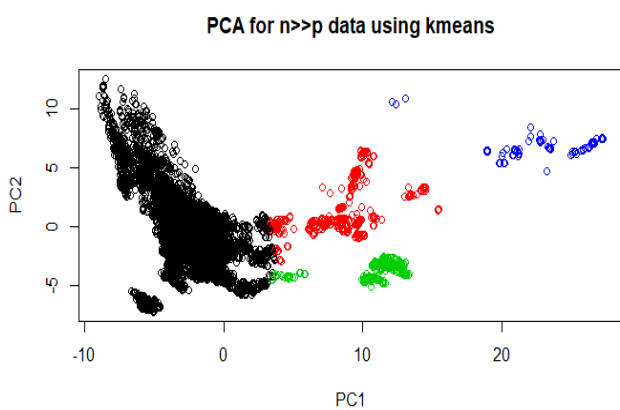


圖 20 使用 K-Means 分群法對前兩個主成分作圖

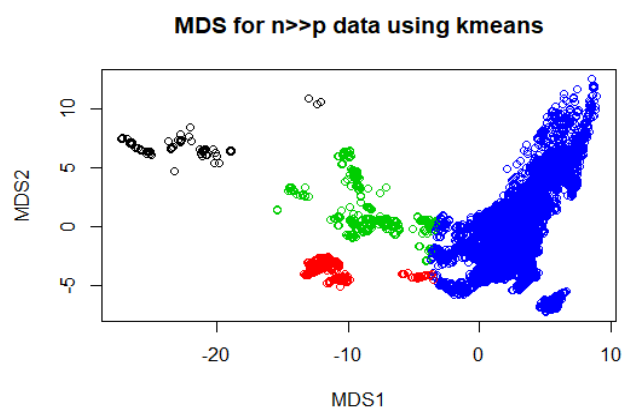


圖 21 使用 K-Means 分群法對 MDS 作圖

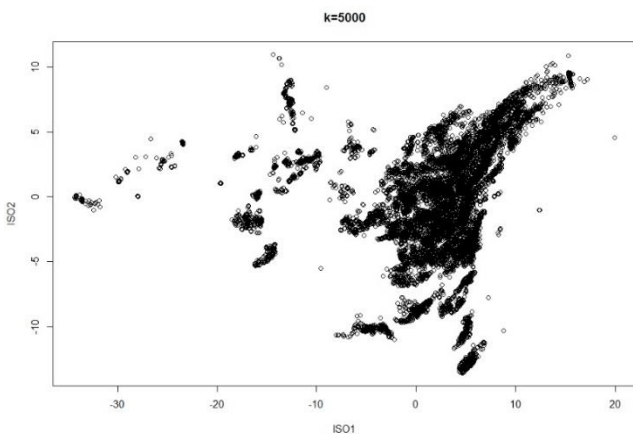


圖 22 ISOMAP of $k=5000$

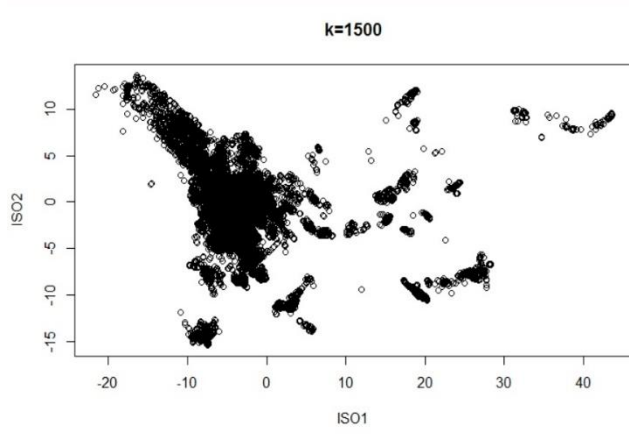


圖 23 ISOMAP of $k=1500$

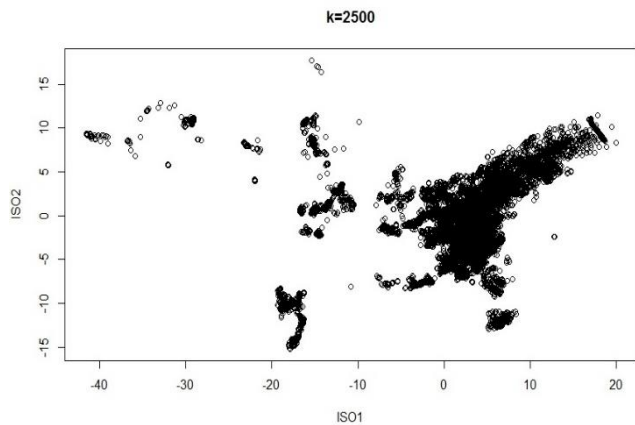


圖 24 ISOMAP of $k=2500$

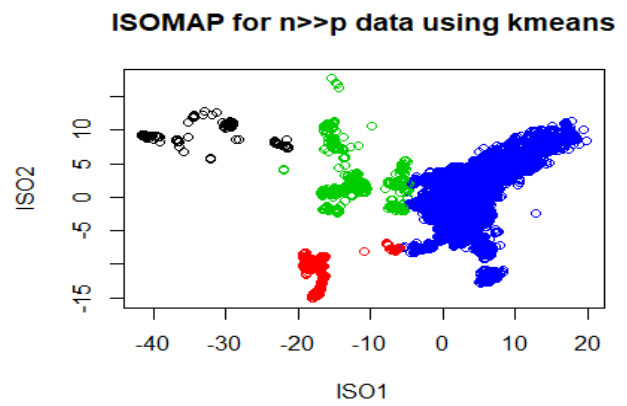


圖 25 使用 K-Means 分群法對 ISOMAP 作圖

最後，在降維效果的比較可以由三種降維方法之 Co-Ranking Matrix 的圖形判斷：發現 PCA 圖形集中於對角線，而相較之下 MDS 與 ISOMAP 較為分散，顯示三種降維方法中，由 PCA 對 $n \gg p$ 做降維的效果較佳。

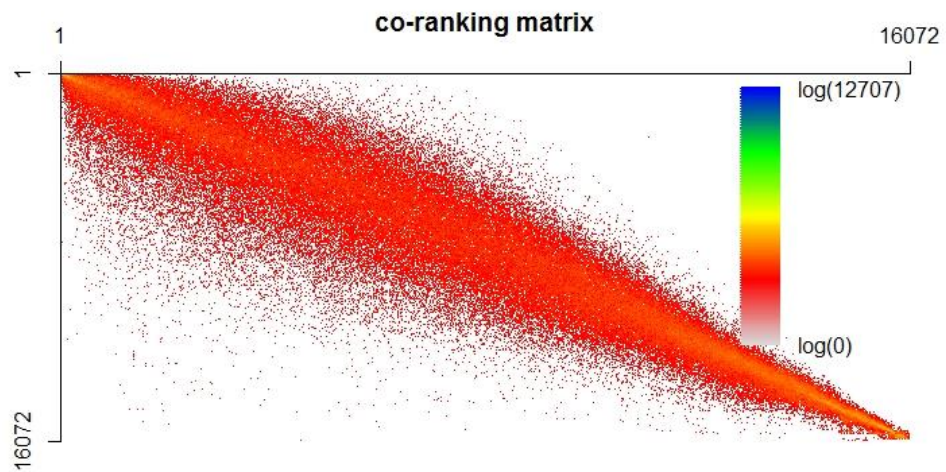


圖 26 CoRanking Matrix of PCA

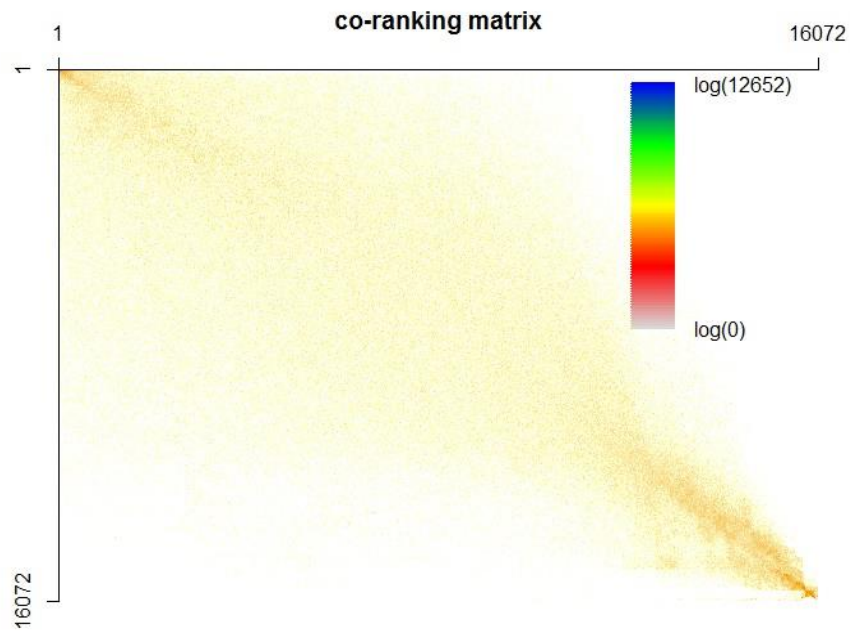


圖 27 Co-Ranking Matrix of MDS

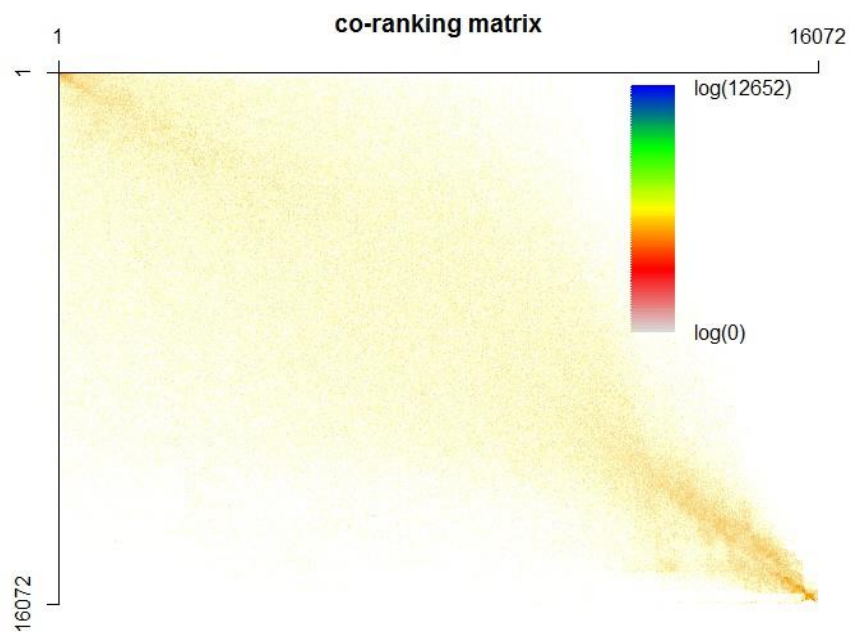


圖 28 Co-Ranking Matrix of ISOMAP

2.5.2. $n == p$

考慮 PCA 的計算使用共變異數矩陣，而 MDS 與 ISOMAP 使用距離矩陣，比較 shrinkage 處理與 empirical 處理之共變異數矩陣，發現 shrinkage 之共變異數矩陣偏離真實共變異數矩陣較多，且分別計算其 Mean Square

Error，發現 empirical (MSE=125.9974) 的誤差相較 shrinkage (MSE=145.9453) 小，表示沒有做 shrinkage 之必要性。在此以 empirical 處理之共變異數矩陣進行 PCA 的計算。

繪製 $n=p$ 時 PCA 之陡坡圖 (圖 30)，發現第 3 個主成分以後解釋變異的程度趨於平緩，而為了取累積解釋百分比超過 70%，故我們與 $n \gg p$ 相同取 10 個維度 (如圖 30)。Correlation Circle (圖 31) 顯示與 $n \gg p$ 之 PCA 相同，變數大致分為機構數量、平均消費與距離三大類。

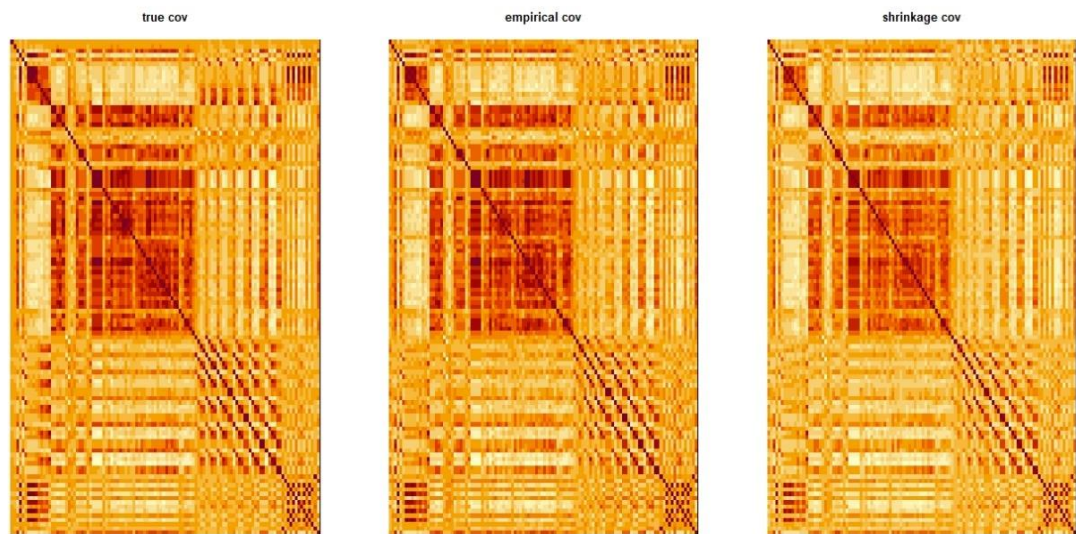


圖 29 Comparison of Covariance Matrix

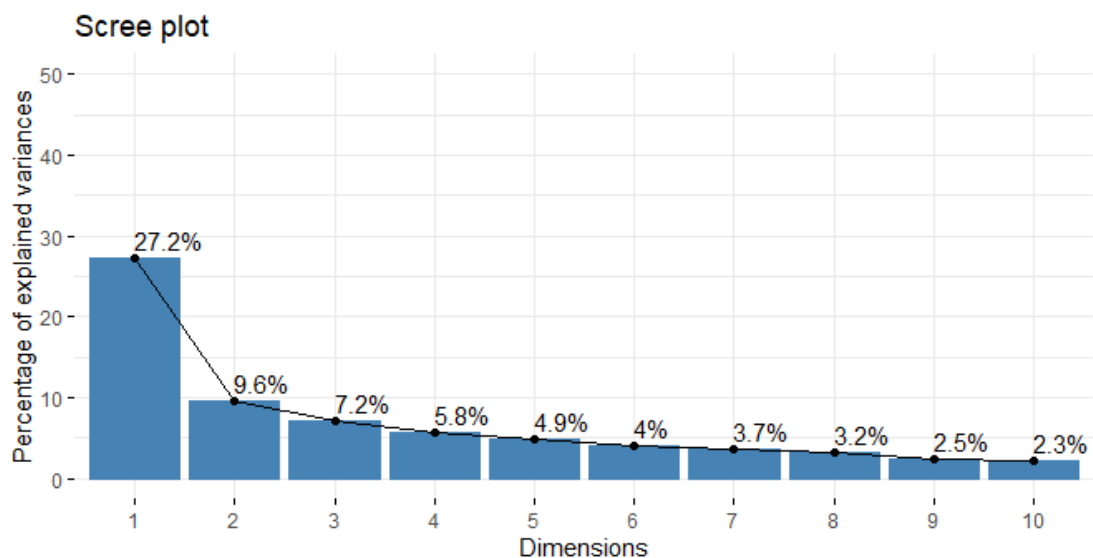


圖 30 Scree Plot of PCA

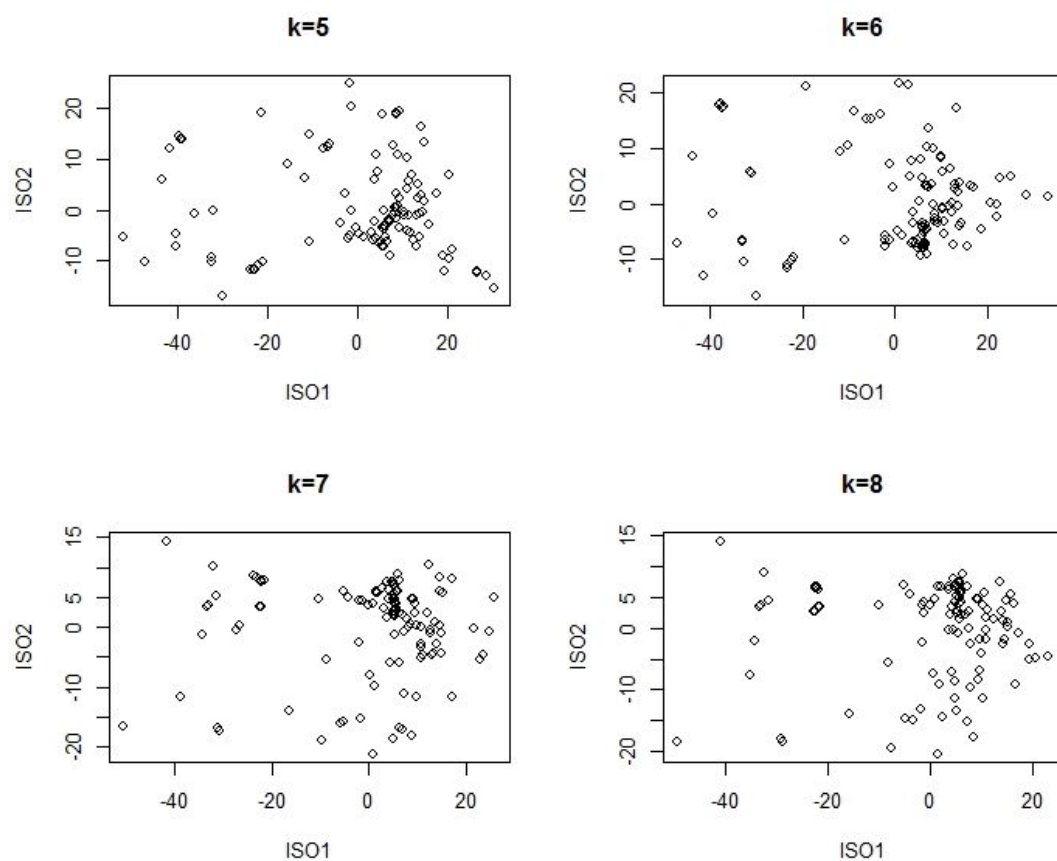


圖 32 Comparison of Different K of ISOMAP

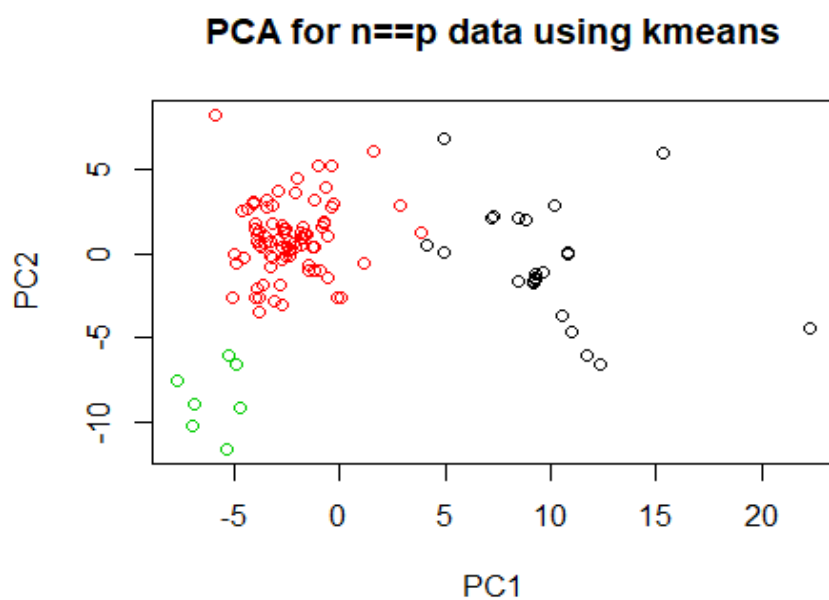


圖 33 使用 K-Means 分群法對前兩個主成分作圖

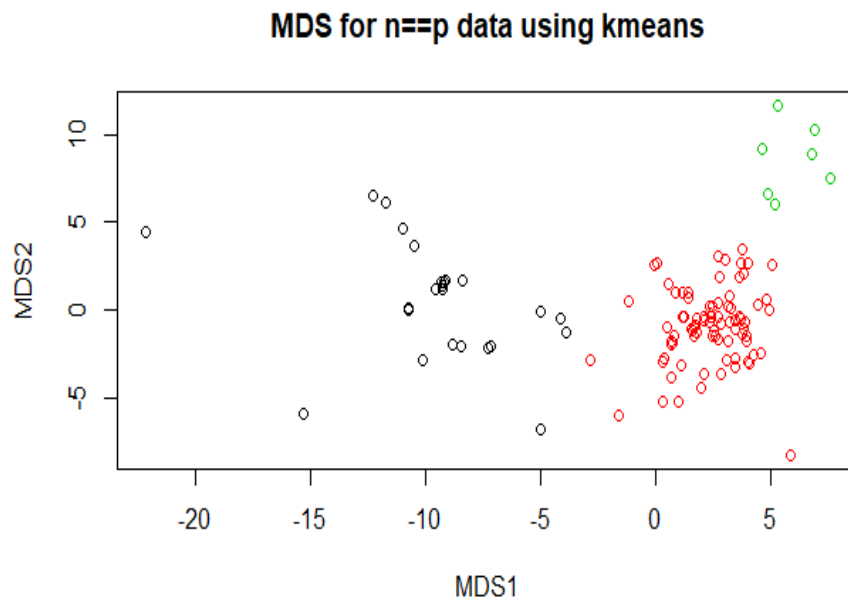


圖 34 使用 K-Means 分群法對 MDS 作圖

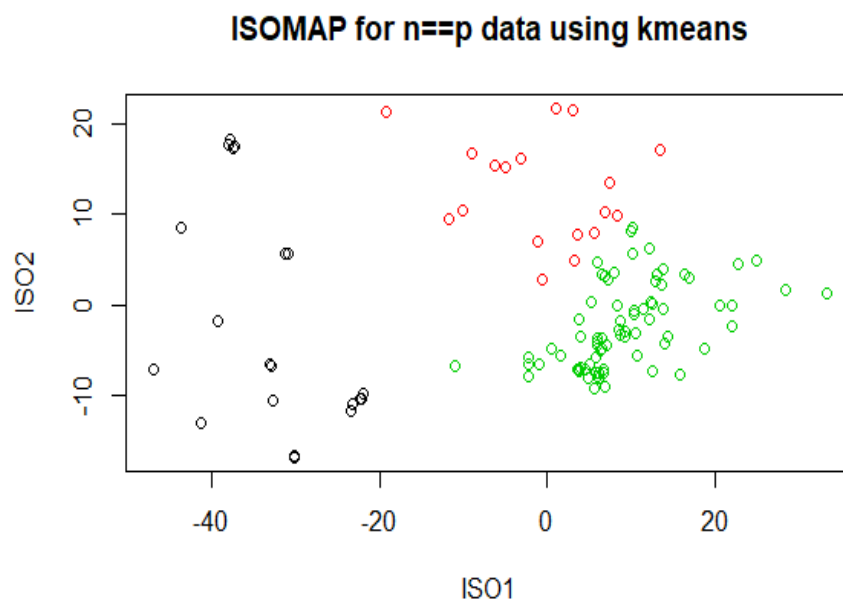


圖 35 使用 K-Means 分群法對 ISOMAP 作圖

最後同樣以 Co-Ranking Matrix 的圖形（圖 36、圖 37 與圖 38）判斷：相比之下 MDS 與 ISOMAP 較 PCA 為分散，顯示三種降維方法中，由 PCA 對 $n == p$ 做降維的效果較佳，但與 $n \gg p$ 的情形相比顯然不是那麼理想。

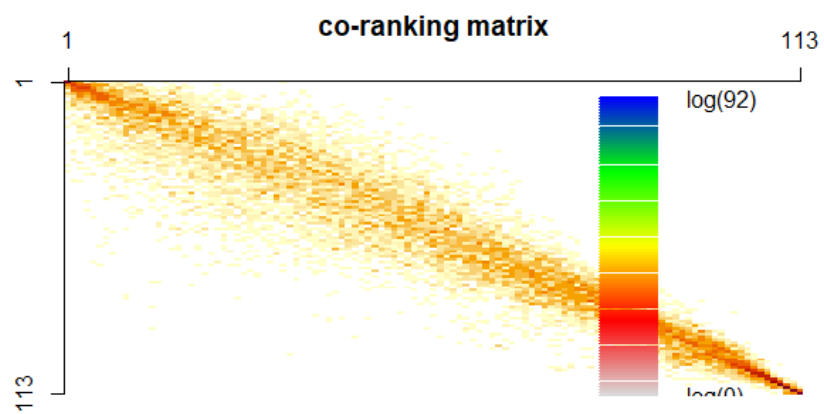


圖 36 Co-Ranking Matrix of PCA

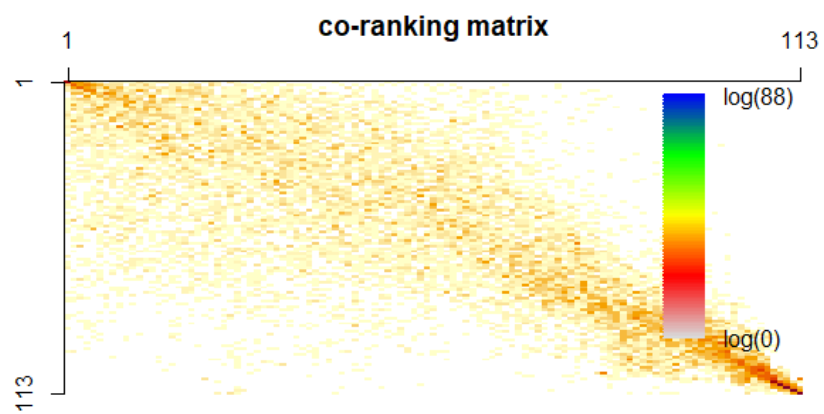


圖 37 Co-Ranking Matrix of MDS

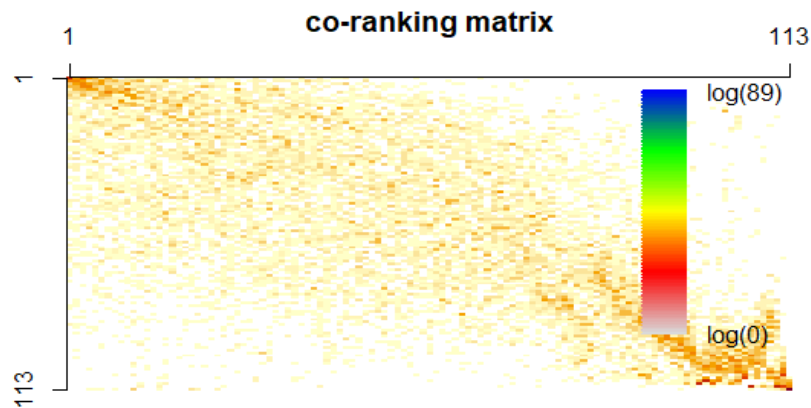


圖 38 Co-Ranking Matrix of ISOMAP

2.5.3. $n \ll p$

在 $n \ll p$ 的資料中，首先勢必要觀察共變異數矩陣圖，從圖 39 中可以看到，若不處理共變異數矩陣，在 empirical 的部分似乎是高估其相關性（深色部分），然而，在使用 shrinkage 後反而變成過度修正，整個共變異數矩陣都下修（變淺），因此我們懷疑在選取維度時會有問題發生。接著藉由圖 40 可以發現，在 5 維時 shrinkage 的特徵值稍微低估，但是到了接近 10 維時卻變成正常的 empirical 特徵值高估。最後，在圖 41 我們將 $n \gg p$ 的累積解釋百分比與 $n \ll p$ 的進行比對，發現共變異數在經過縮水後竟累積極慢，同樣取 70% 的解釋百分比來說在 $n \gg p$ 只需取 10 維，然而，在 $n \ll p$ 卻須超過 50 個維度才可以達成，因此我們最後決定取 16 個維度總計 55% 的解釋百分比做為降維後的資料。但可以預期的是，若拿此降維後資料進行預測，效果一定不會太好。

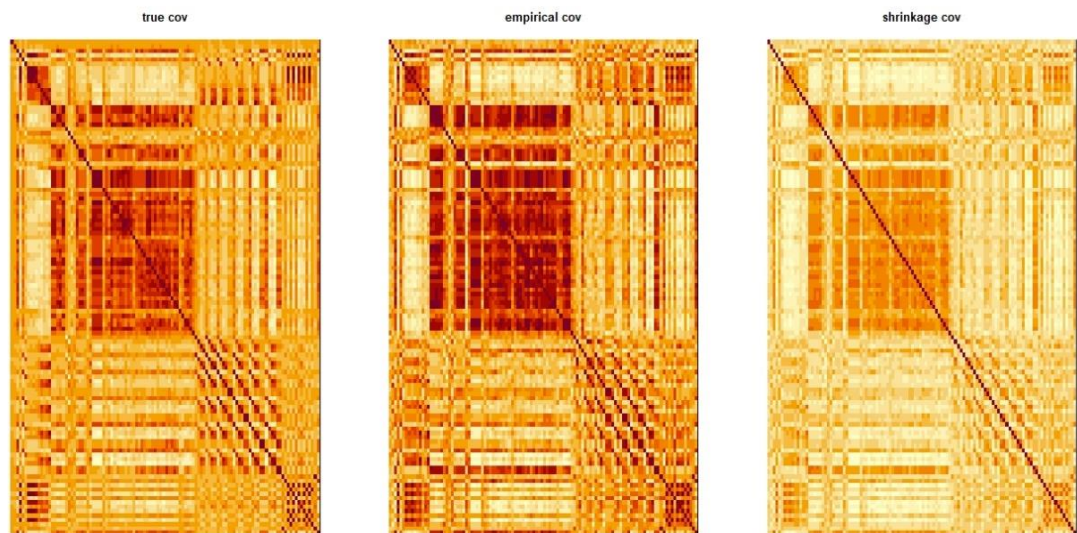


圖 39 Comparison of Covariance Matrix

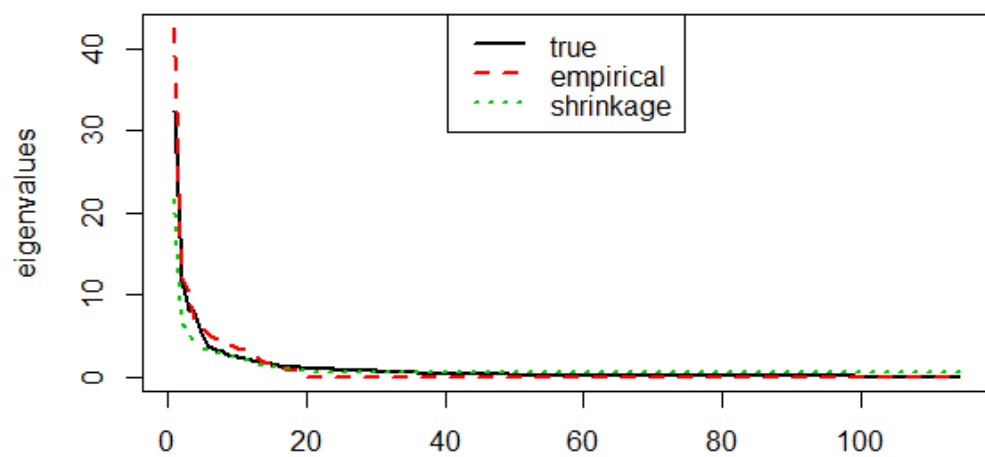


圖 40 Eigenvalue between result treatment and untreatment result

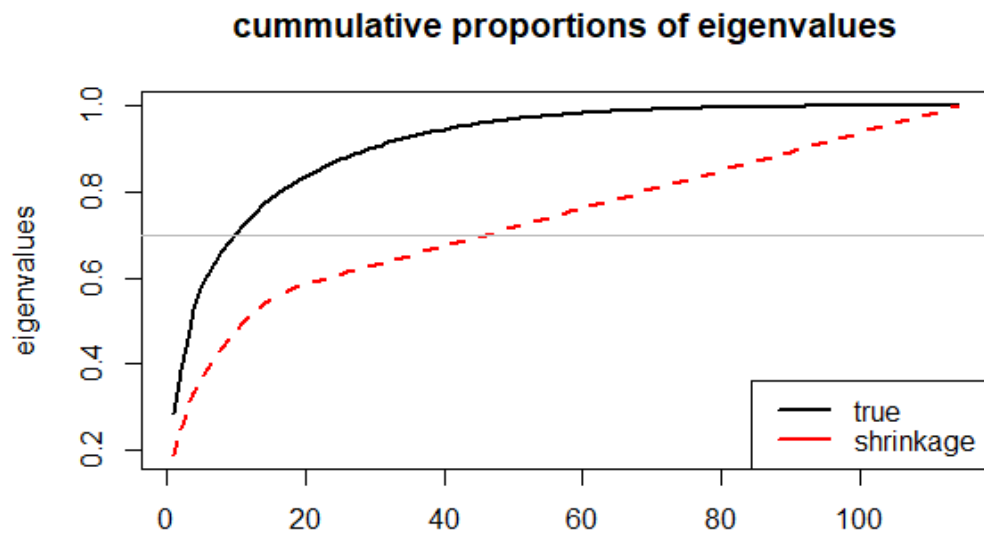


圖 41 Cumulative Proportions of eigenvalues between True & Shrinkage

分群的部分，或許是樣本數過少的情況，導致在兩維的散佈情況都無法至少分成兩群，因此不使用 kmeans 觀察分群結果。

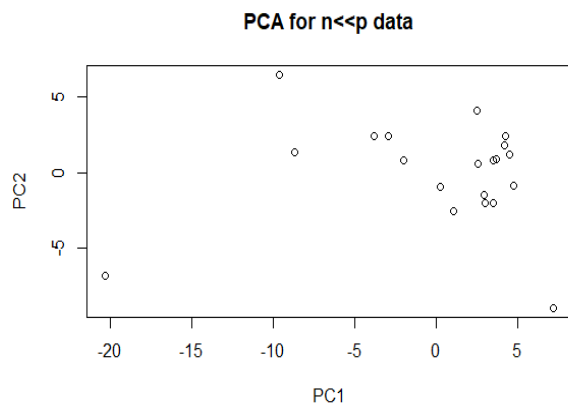


圖 42 PCA

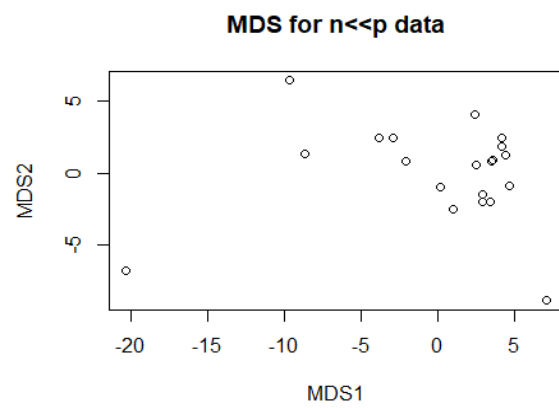


圖 43 MDS

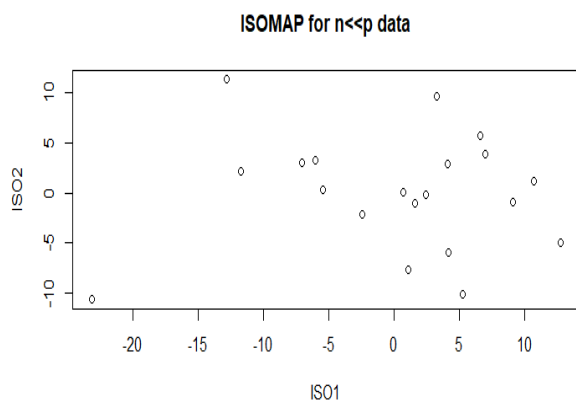


圖 44 ISOMAP

在 coranking matrix (圖 45、圖 46 以及圖 47) 佐證下，MDS 與 ISOMAP 的效果最差，在不同的資料大小亦是如此。然而，PCA 的結果呈現極佳，有高達 80% 的情形，但畢竟只抽樣一次，無法直接斷定 PCA 在小樣本下降維較好，須重複測試驗證。

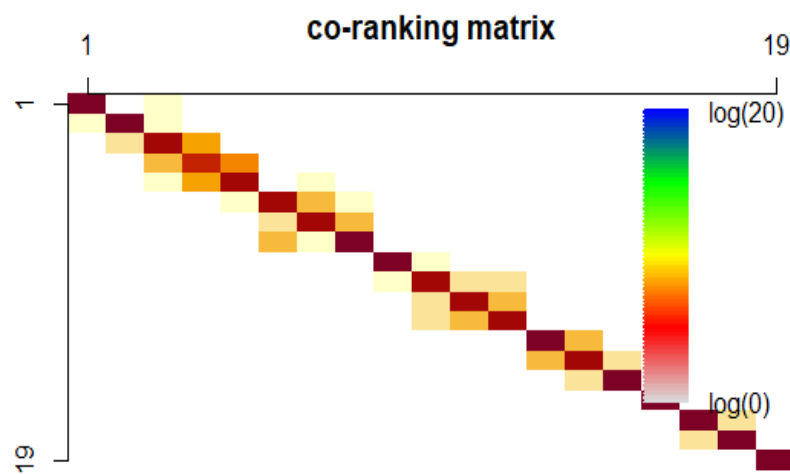


圖 45 Co-Ranking Matrix of PCA

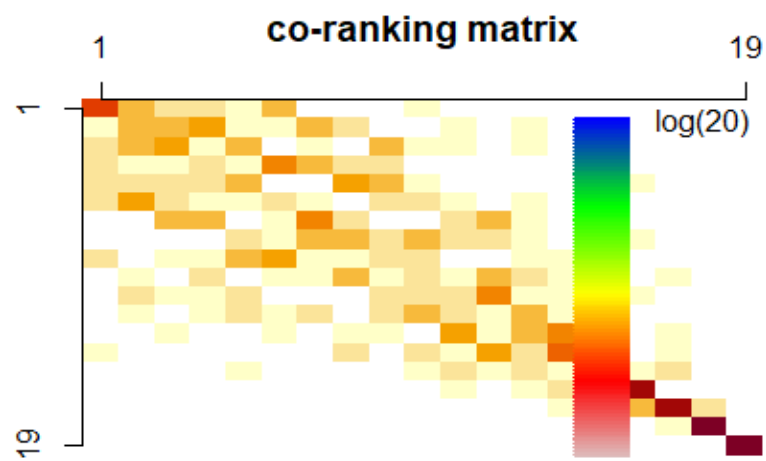


圖 46 Co-Rnaking Matrix of MDS

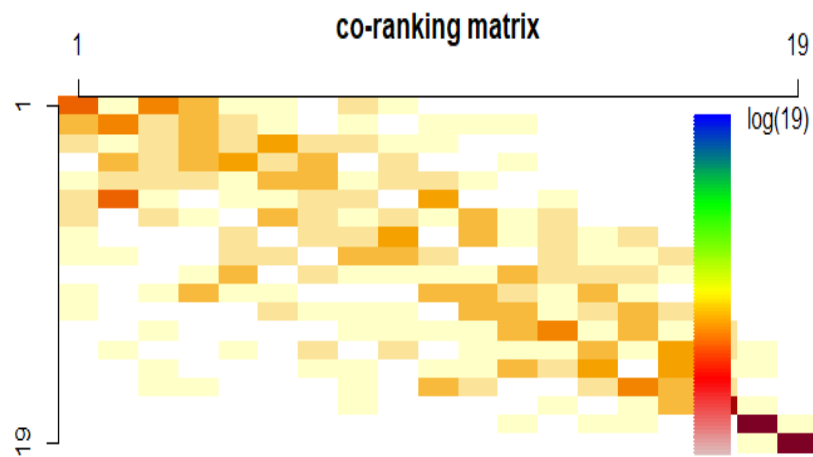


圖 47 Co-Ranking Matrix of ISOMAP

2.5.4. 小結

綜上所述，以透過對 LCMC 做圖佐證：無論資料大小，PCA 皆為最佳降維方法。當取的 k 越大時，PCA 在 $n \ll p$ 的表現就越差，反之在 $n \gg p$ 的表現最佳；而 k 值與資料大小對 MDS 與 ISOMAP 的影響不大，唯其在 $n \gg p$ 時表現略佳。

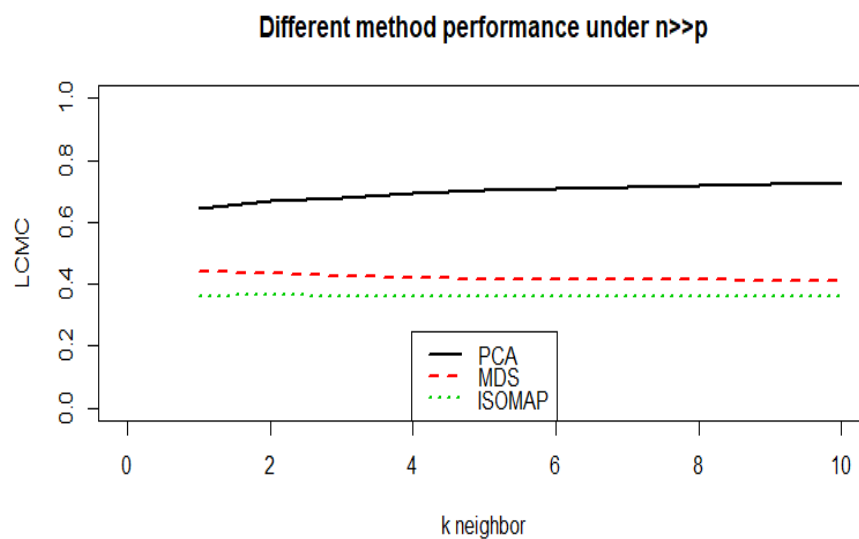


圖 48 各維度縮減方法在 $n \gg p$ 之 LCMC

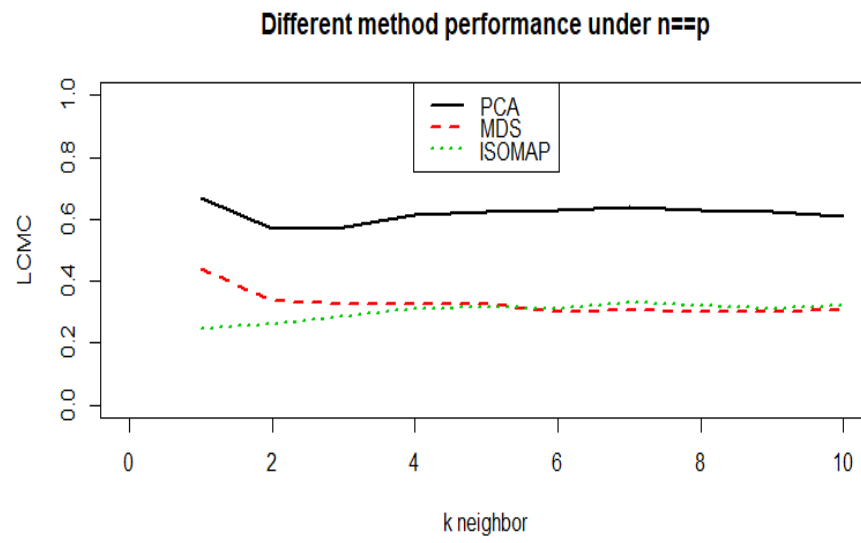


圖 49 各維度縮減方法在 $n = p$ 之 LCMC

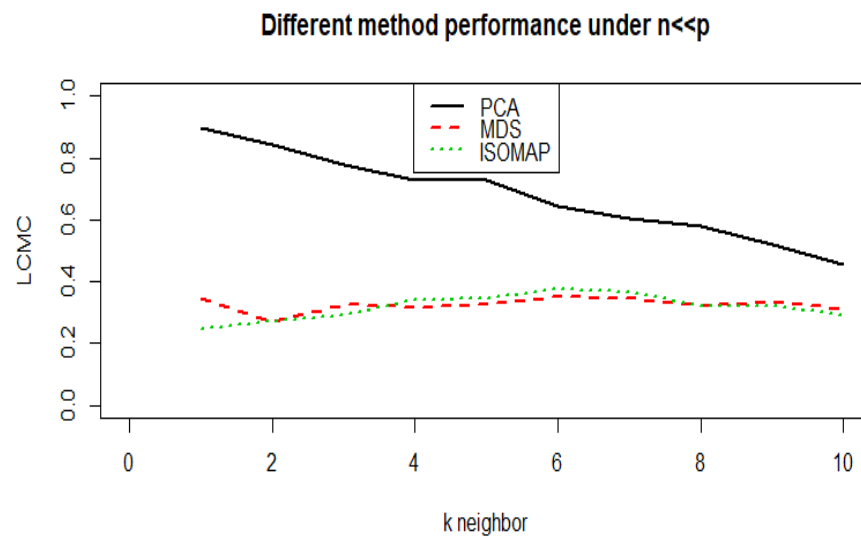


圖 50 各維度縮減方法在 $n \ll p$ 之 LCMC

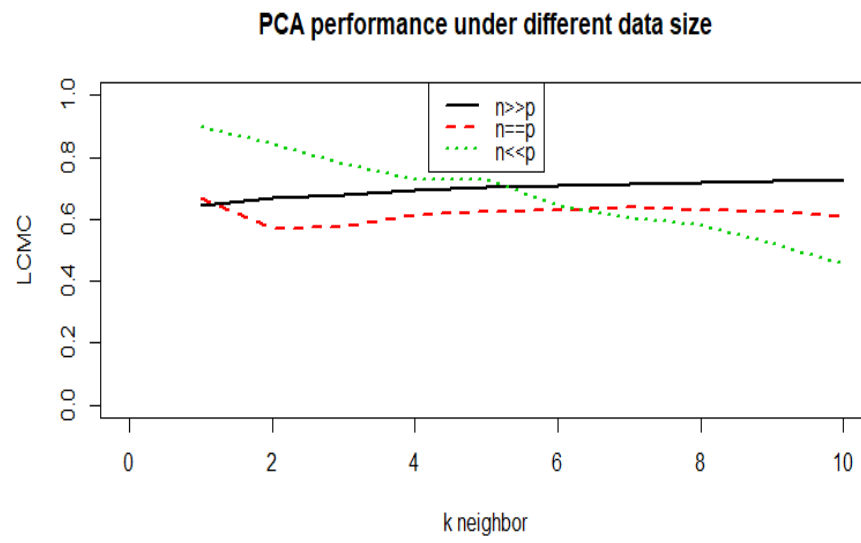


圖 51 不同資料大小進行 PCA 之 LCMC

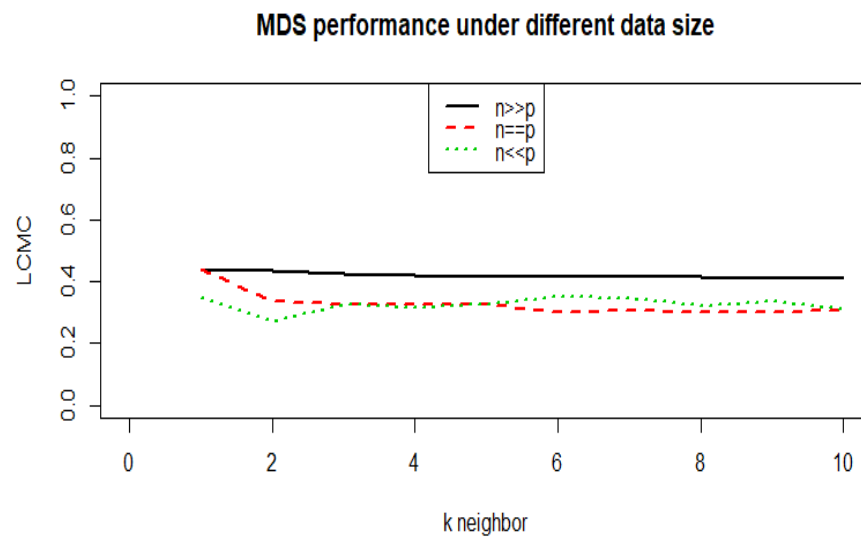


圖 52 不同資料大小進行 MDS 之 LCMC

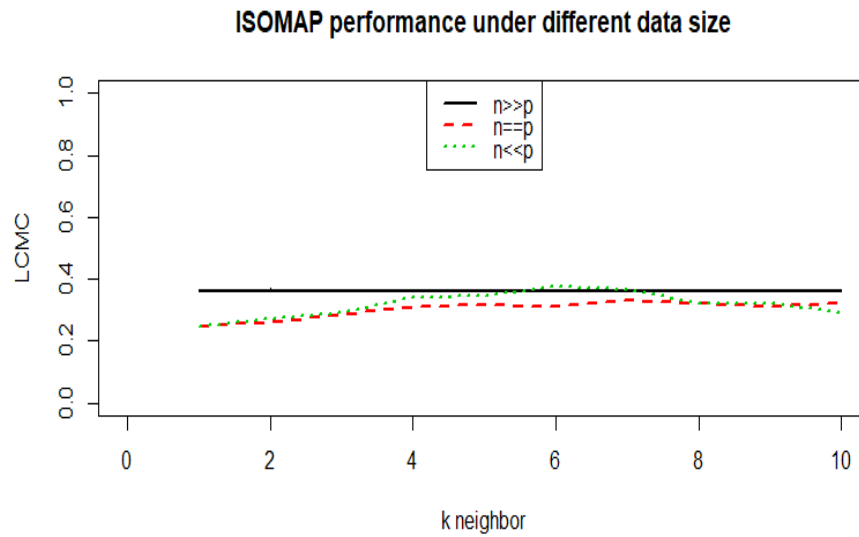


圖 53 不同資料大小進行 ISOMAP 之 LCMC

2.6. 模型預測

2.6.5. $n \gg p$

模型預測中，首先繪製以 MDS 降維後的資料審視房屋價格與降維後前兩維的趨勢（圖 54）。然而，從圖 54 明顯可以看出：若以房價高低表示顏色呈現，明顯顏色是雜亂無章的，亦即不管以 MDS1 或 MDS2 的方向檢視均可以發現房價的高低沒有固定的趨勢，所以可以預期在線性迴歸模型配適的結果不會太好，若改用 pairwise 圖呈現（圖 55），可以了解到 MDS1 與 MDS2 分別對房屋價格的相關係數為 0.219 與 0.104，相關程度低，接著實際以 MDS 降維後資料配適線性迴歸，結果如表 5 所示：

線型迴歸模型 R-square 為 0.06，可見使用此模型效果不佳。同理在使用 PCA 或 ISOMAP 的降維後資料進行配適會得到相同極差的結果，因此不再多加贅述。

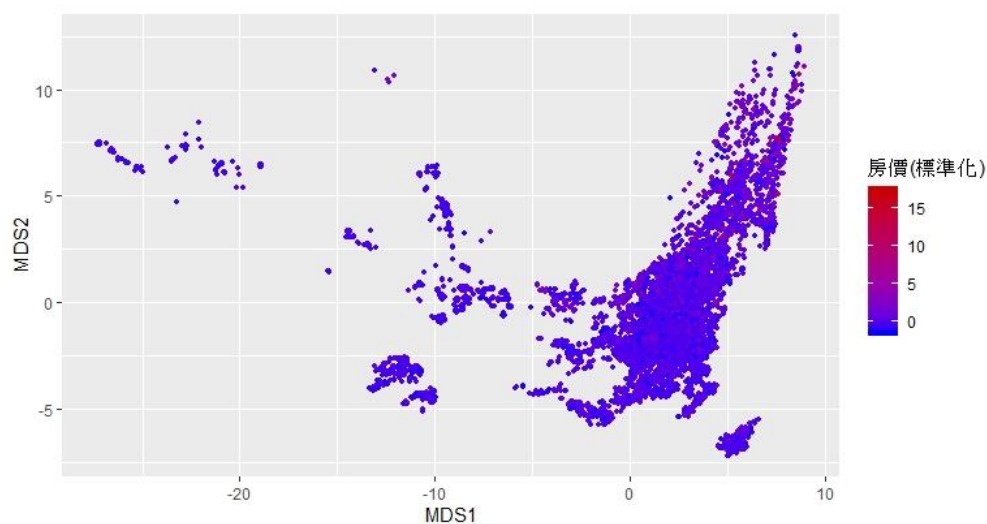


圖 54 房屋價格與以 MDS 降維後前兩維的關係圖

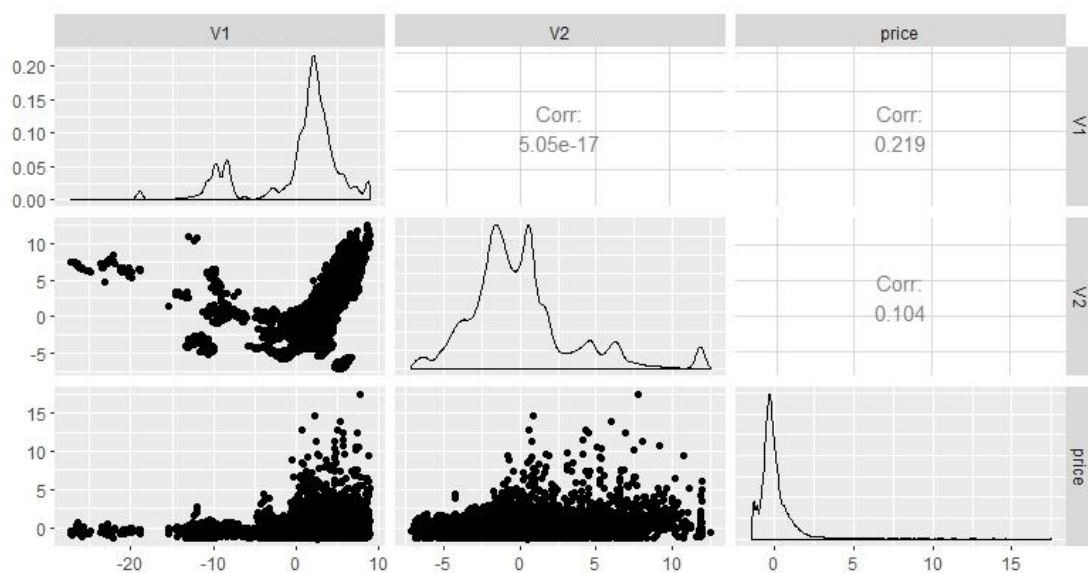


圖 55 Pairwise of first 2 MDS & price

表 5 以 MDS 降維後線性回歸之參數

Coefficients:	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.251e-16	7.653e-03	0.00	1
V1	3.845e-02	1.345e-03	28.58	<2e-16 ***
V2	3.055e-02	2.247e-03	13.60	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '.' 0.1 ' ' 1				
Residual standard error: 0.9703 on 16070 degrees of freedom				
Multiple R-squared: 0.05869, Adjusted R-squared: 0.05857				
F-statistic: 501 on 2 and 16070 DF, p-value: < 2.2e-16				

2.6.6. $n \ll p$

在觀測值遠小於變數個數的狹長情形中會因非 full rank 導致傳統線性回歸有共線性的問題。然而，因線型迴歸無法求得，本資料使用向後消去 (backward elimination) 也無法配適，導致在變數選擇 (自由度為 0，R 運行結果如表 6)，因此考慮 lasso regression。

表 6 regression of backward elimination

```
We are eliminating variables based on p value
Variables Removed:
Error in Anova.lm(m):residual df=0
```

在建立 lasso regression 前，首先以 cross validation 選擇最適當的懲罰項，因為 $n \ll p$ 的觀測值只有 20 筆，所以以 Leave-One-Out 以及 MSE 最小的方式進行選擇 λ ，結果顯示取 $\exp(-2.64)=0.07$ 即可 (圖 56)。

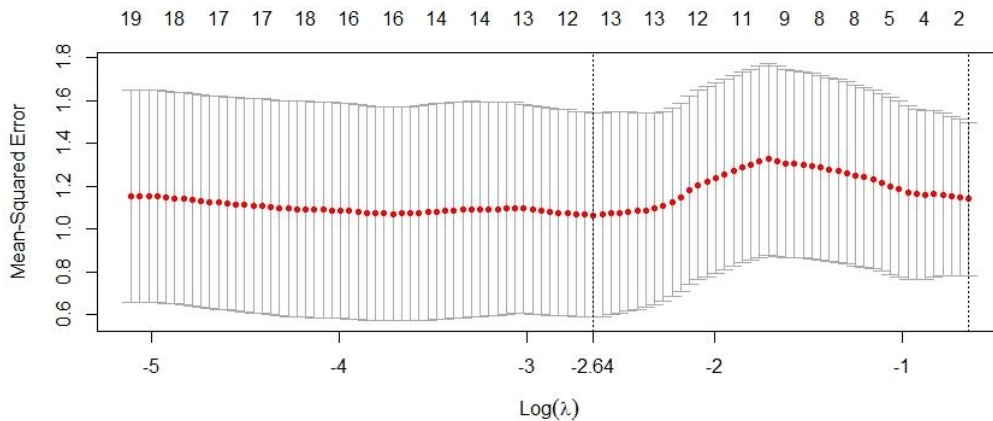


圖 56 Mean Square Error of $\text{Log}(\lambda)$

然而，在建立 lasso 模型的最後顯示警告消息，呈現出來的係數貌似沒問題，但在信賴區間卻出現無窮大以及 p-value 皆為 0 或 1：

表 7 lasso regression

Var	Coef	Z-score	P-value	LowConfPt	UpConfPt	LowTail Area	UpTail Area
1	0.528	1.855	0	28.490	Inf	0	0
4	0.099	0.067	0	148.934	Inf	0	0
7	-0.011	-0.009	1	-Inf	-118.613	0	0
12	0.142	0.184	1	-Inf	-77.140	0	0
21	-0.349	-1.058	1	32.987	Inf	0	0
22	0.284	1.035	0	27.384	Inf	0	0
23	0.340	0.936	0	36.278	Inf	0	0
29	0.066	0.204	1	-Inf	-32.456	0	NaN

63	-0.227	-0.741	1	30.652	Inf	0	0
91	0.025	0.039	0	64.092	Inf	0	0
107	0.119	0.220	1	-Inf	53.931	0	0
110	-0.378	-1.378	0	-Inf	-27.434	0	0

而根據 **Gabriel Vasconcelos** 於 **insightr** 中《**When the lasso fails?**》¹中提及之資訊，當資料中的共變異數矩陣可以被區分成明顯區集如下：

$$\Sigma = \begin{pmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{pmatrix}$$

$C_{1,1}$ 為資料中的重要變數， $C_{2,2}$ 為不重要的變數， $C_{1,2}$ 與 $C_{2,1}$ 為重要與不重要變數的共變異數矩陣，當 Σ 明顯被分割的情形下且滿足以下條件：

$$|C_{2,1}C_{1,1}^{-1}\text{sign}(\beta)| < 1$$

則 lasso 係數的估計將出現非嚴格遞減的情形（如圖 58 中紅色線所示）。

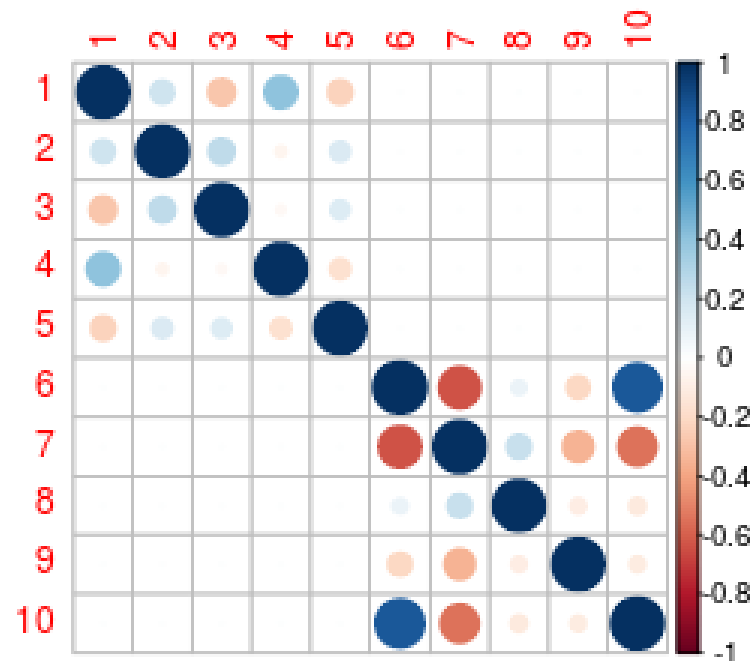


圖 57

¹ 《When the lasso fails?》

https://insightr.wordpress.com/2017/06/14/when-the-lasso-fails/?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com

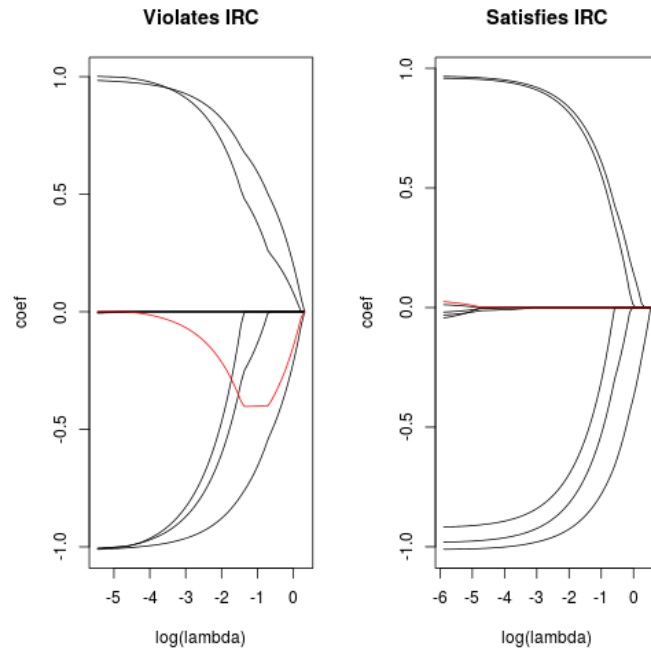


圖 58

基於以上原則，檢視本資料的相關係數矩陣及係數懲罰的情形：

把變數整理成矩陣熱圖（圖 57）的形式，檢視係數懲罰，發現有部分的變數有非嚴格遞減的情形（圖 60），因此，可以考慮使用 adaptive lasso regression 或者其他修正之 lasso regression 進行配適。

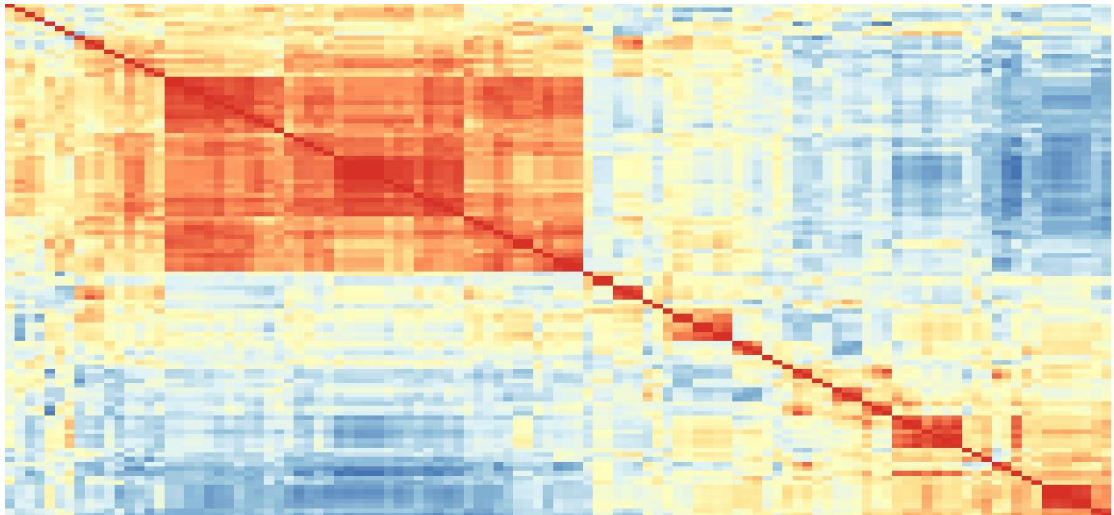


圖 59 heatmap of Correlation Matrix

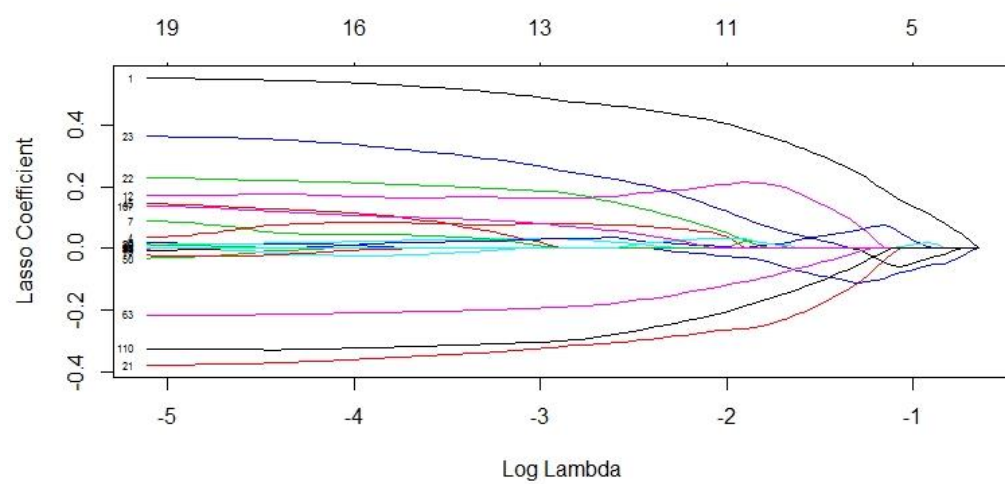


圖 60 $\text{Log}(\lambda)$ v.s Lasso Coefficient

3. 結論

綜合以上的參考文獻、EDA 以及降維，我們可以合理推論俄羅斯房價無法配適良好的原因：

1. 以俄羅斯來說，或許距離以及性別人數並不是影響房價最主要的原因，也就是說俄羅斯與臺灣影響房價的原因不同，因此不適用於俄羅斯。或許 Sberbank 須重新考量與俄羅斯房屋價格之變數。

2. 根據 EDA 及其他圖例顯示，在不同類別變數的標籤區分下，無法明顯顯示價格上有差異，以及最後在 MDS 降維後的散佈圖使用價格去標示仍無法看出 pattern。

3. 從參考文獻或許也可以推論，因俄羅斯房價在近年來炒作嚴重，不管房屋所在地區以及生活便利性已無法反映在價格上，也就是說，實際價值較低的房屋很有可能因為透過上漲導致價格誤解，在模型配適時無法得到理想的結果。

4. 參考書目與文章

1. 《R Graphics Cookbook》 *Winston Chang*
2. 《When the lasso fails?》 *Gabriel Vasconcelos*
3. 《俄羅斯 2001 年前房價趨勢》, 盛世良 , 中國房地產業, 1999
4. 《俄羅斯的房價上漲》 聶書嶺, 中亞信息, 2004
5. 《俄羅斯房價飛漲》 岳萍, 中亞信息, 2006

附錄 I、資料結構

附錄表 1 資料結構表

類型	變數名稱	
連續 numeric		
	additional_education_km	area_m
	big_church_km	big_market_km
	big_road2_km	bulvar_ring_km
	cafe_avg_price_1000	cafe_avg_price_1500
	cafe_avg_price_3000	cafe_avg_price_500
	cafe_sum_1000_max_price_avg	cafe_sum_1000_min_price_avg
	cafe_sum_1500_min_price_avg	cafe_sum_2000_max_price_avg
	cafe_sum_3000_max_price_avg	cafe_sum_3000_min_price_avg
	cafe_sum_500_min_price_avg	cafe_sum_5000_max_price_avg
	catering_km	cemetery_km
	detention_facility_km	exhibition_km
	green_part_1000	green_part_1500
	green_part_3000	green_part_500
	green_zone_km	green_zone_part
	ice_rink_km	ID_railroad_station_avto
	incineration_km	indust_part
	kindergarten_km	kremlin_km
	metro_km_avto	metro_km_walk
	metro_min_walk	mkad_km
	museum_km	nuclear_reactor_km
	oil_chemistry_km	park_km
	preschool_km	prom_part_1000
	prom_part_2000	prom_part_3000
	prom_part_5000	public_healthcare_km
	public_transport_station_min_walk	radiation_km
	railroad_station_avto_km	railroad_station_avto_min
	railroad_station_walk_min	sadovoe_km
	shopping_centers_km	stadium_km
	theater_km	thermal_power_plant_km
	ttk_km	university_km
	water_treatment_km	workplaces_km
		basketball_km
		big_road1_km
		bus_terminal_avto_km
		cafe_avg_price_2000
		cafe_avg_price_5000
		cafe_sum_1500_max_price_avg
		cafe_sum_2000_min_price_avg
		cafe_sum_500_max_price_avg
		cafe_sum_5000_min_price_avg
		church_synagogue_km
		fitness_km
		green_part_2000
		green_part_5000
		hospice_morgue_km
		ID_railroad_station_walk
		industrial_km
		market_shop_km
		metro_min_avto
		mosque_km
		office_km
		power_transmission_line_km
		prom_part_1500
		prom_part_500
		public_transport_station_km
		railroad_km
		railroad_station_walk_km
		school_km
		swim_pool_km
		ts_km
		water_km
		zd_vokzaly_avto_km

計數 integer

additional_education_raion	afe_count_3000_price_2500	afe_count_5000_na_price
big_church_count_1000	big_church_count_1500	big_church_count_2000
big_church_count_3000	big_church_count_500	big_church_count_5000
build_count_1921.1945	build_count_1946.1970	build_count_1971.1995
build_count_after_1995	build_count_before_1920	build_count_block
build_count_brick	build_count_foam	build_count_frame
build_count_mix	build_count_monolith	build_count_panel
build_count_slag	build_count_wood	cafe_count_1000
cafe_count_1000_na_price	cafe_count_1000_price_1000	cafe_count_1000_price_1500
cafe_count_1000_price_2500	cafe_count_1000_price_4000	cafe_count_1000_price_500
cafe_count_1000_price_high	cafe_count_1500	cafe_count_1500_na_price
cafe_count_1500_price_1000	cafe_count_1500_price_1500	cafe_count_1500_price_2500
cafe_count_1500_price_4000	cafe_count_1500_price_500	cafe_count_1500_price_high
cafe_count_2000	cafe_count_2000_na_price	cafe_count_2000_price_1000
cafe_count_2000_price_1500	cafe_count_2000_price_2500	cafe_count_2000_price_4000
cafe_count_2000_price_500	cafe_count_2000_price_high	cafe_count_3000
cafe_count_3000_na_price	cafe_count_3000_price_1000	cafe_count_3000_price_1500
cafe_count_3000_price_4000	cafe_count_3000_price_500	cafe_count_3000_price_high
cafe_count_500	cafe_count_500_na_price	cafe_count_500_price_1000
cafe_count_500_price_1500	cafe_count_500_price_2500	cafe_count_500_price_4000
cafe_count_500_price_500	cafe_count_500_price_high	cafe_count_5000
cafe_count_5000_price_1000	cafe_count_5000_price_1500	cafe_count_5000_price_2500
cafe_count_5000_price_4000	cafe_count_5000_price_500	cafe_count_5000_price_high
children_preschool	children_school	church_count_1000
church_count_1500	church_count_2000	church_count_3000
church_count_500	church_count_5000	culture_objects_top_25_raion
ekder_all	ekder_female	ekder_male
female_f	floor	full_all
full_sq	healthcare_centers_raion	hospital_beds_raion
id	ID_big_road1	ID_big_road2
ID_bus_terminal	ID_metro	ID_railroad_terminal
kitch_sq	leisure_count_1000	leisure_count_1500
leisure_count_2000	leisure_count_3000	leisure_count_500
leisure_count_5000	life_sq	male_f
market_count_1000	market_count_1500	market_count_2000
market_count_3000	market_count_500	market_count_5000
max_floor	mosque_count_1000	mosque_count_1500

mosque_count_2000	mosque_count_3000	mosque_count_500
mosque_count_5000	num_room	office_count_1000
office_count_1500	office_count_2000	office_count_3000
office_count_500	office_count_5000	office_raion
office_sqm_1000	office_sqm_1500	office_sqm_2000
office_sqm_3000	office_sqm_500	office_sqm_5000
preschool_education_centers_raion	preschool_quota	price_doc
raion_build_count_with_builddate_info	raion_build_count_with_material_info	raion_popul
school_education_centers_raion	school_education_centers_top_20_raion	school_quota
shopping_centers_raion	sport_count_1000	sport_count_1500
sport_count_2000	sport_count_3000	sport_count_500
sport_count_5000	sport_objects_raion	trc_count_1000
trc_count_1500	trc_count_2000	trc_count_3000
trc_count_500	trc_count_5000	trc_sqm_1000
trc_sqm_1500	trc_sqm_2000	trc_sqm_3000
trc_sqm_500	trc_sqm_5000	university_top_20_raion
work_all	work_female	work_male
X0_13_all	X0_13_female	X0_13_male
X0_17_all	X0_17_female	X0_17_male
X0_6_all	X0_6_female	X0_6_male
X16_29_all	X16_29_female	X16_29_male
X7_14_all	X7_14_female	X7_14_male
young_all	young_female	young_male
Date		
build_year	timestamp	
Factor		
big_market_raion	big_road_line	culture_objects_top_
detention_facility_raion	incineration_raion	nuclear_reactor_raion
oil_chemistry_raion	radiation_raion	railroad_line
railroad_terminal_raion	sub_area	ecology
material	product_type	state
thermal_power_plant_raion	water_line	

附錄 II、連續變數摘要表

附錄表 2 連續變數摘要表

Continuous Variables	min	1 st Qu.	Median	Mean	3 rd Qu.	Max	NMiss	MissRate
full_sq	0	38	49	54.21	63	5326		
life_sq	0	20	30	34.4	43	7478	6383	0.2095
floor	0	3	6.5	7.671	11	77	167	0.0055
max_floor	0	9	12	12.56	17	117	9572	0.3141
num_room	0	1	2	1.91	2	19	9572	0.3141
kitch_sq	0	1	6	6.399	9	2014	9572	0.3141
area_m	2081628	7307411	10508030	17657051	18036437	206071809		
raion_popul	2546	21819	83502	84056	122862	247469		
green_zone_part	0.0019	0.0638	0.1675	0.218922	0.336177	0.852923		
indust_part	0	0.01951	0.0722	0.11887	0.19578	0.52187		
children_preschool	175	1706	4857	5140	7103	19223		
preschool_quota	0	1874	2854	3271	4050	11926	6688	0.2195
preschool_education_centers_raion	0	2	4	4.065	6	13		
children_school	168	1564	5261	5354	7227	19083		
school_quota	1012	5782	7377	8325	9891	24750	6685	0.2194
school_education_centers_raion	0	2	5	4.705	7	14		
school_education_centers_top_20_raion	0	0	0	0.1097	0	2		
hospital_beds_raion	0	520	990	1191	1786	4849	14441	0.4739
healthcare_centers_raion	0	0	1	1.321	2	6		
university_top_20_raion	0	0	0	0.1383	0	3		
sport_objects_raion	0	1	5	6.635	10	29		
additional_education_raion	0	1	2	2.896	4	16		
culture_objects_top_25_raion	0	0	0	0.2867	0	10		
shopping_centers_raion	0	1	3	4.201	6	23		
office_raion	0	0	2	8.253	5	141		
full_all	2546	28179	85219	146306	125111	1716730		
male_f	1208	13522	39261	67208	58226	774585		
female_f	1341	15031	45729	79099	67872	942145		
young_all	365	3459	10988	11179	14906	40692		

young_male	189	1782	5470	5724	7597	20977		
young_female	177	1677	5333	5455	7617	19715		
work_all	1633	13996	52030	53668	77612	161290		
work_male	863	7394	26382	27254	38841	79622		
work_female	771	6661	26092	26414	37942	81668		
ekder_all	548	4695	20036	19210	29172	57086		
ekder_male	156	1331	6180	5812	8563	19275		
ekder_female	393	3365	13540	13398	20165	37811		
X0_6_all	175	1706	4857	5140	7103	19223		
X0_6_male	91	862	2435	2631	3523	9987		
X0_6_female	85	844	2390	2509	3455	9236		
X7_14_all	168	1564	5261	5354	7227	19083		
X7_14_male	87	821	2693	2743	3585	9761		
X7_14_female	82	743	2535	2611	3534	9322		
X0_17_all	411	3831	12508	12541	16727	45170		
X0_17_male	214	1973	6085	6423	8599	23233		
X0_17_female	198	1858	6185	6118	8549	21937		
X16_29_all	575	5829	17864	31316	27194	367659		
X16_29_male	308	2955	8896	15369	13683	172958		
X16_29_female	253	2874	9353	15947	14184	194701		
X0_13_all	322	3112	9633	9841	13121	36035		
X0_13_male	166	1600	4835	5037	6684	18574		
X0_13_female	156	1512	4667	4804	6699	17461		
metro_min_avto	0	1.721	2.803	4.961	4.832	61.438		
metro_km_avto	0	1.037	1.784	3.701	3.777	74.906		
metro_min_walk	0	11.48	20.45	42.74	45.32	711.22	25	0.0008
metro_km_walk	0	0.957	1.704	3.561	3.777	59.268	25	0.0008
kindergarten_km	0.00047	0.2000	0.3538	0.98168	0.97142	29.08577		
school_km	0	0.2697	0.4749	1.324	0.8865	47.3947		
park_km	0.00374	0.9733	1.8039	3.09994	3.40479	47.35154		
green_zone_km	0	0.101	0.2143	0.3005	0.4155	1.9824		
industrial_km	0	0.2883	0.5765	0.7688	1.0411	14.0482		
water_treatment_km	0.2741	5.3046	10.378	11.1676	16.7914	47.5912		
cemetery_km	0	1.335	1.969	2.315	3.089	15.779		
incineration_km	0.1981	6.2219	10.3242	10.8846	13.3938	58.632		
railroad_station_walk_km	0.0282	1.9314	3.23554	4.38694	5.14764	24.65304	25	0.0008
railroad_station_walk_min	0.3378	23.1765	38.82650	52.6433	61.7717	295.8365	25	0.0008
railroad_station_avto_km	0.0282	2.1169	3.42832	4.58728	5.39143	24.65398		

railroad_station_avto_min	0.0352	3.2355	4.94456	6.08661	7.30357	38.69192
public_transport_station_k m	0.0028	0.1013	0.16028	0.414136	0.278403	17.413002
public_transport_station_ min_walk	0.0337	1.2158	1.9233	4.96963	3.34084	208.95602
water_km	0.0067	0.3396	0.6212	0.690947	0.963865	2.827709
mkad_km	0.0136	2.6334	5.4675	6.27476	8.18475	53.27783
ttk_km	0.0019	5.3398	9.8426	11.31815	15.67545	66.0332
sadovoe_km	0.0004	8.3463	12.7487	14.05672	18.71662	68.85305
bulvar_ring_km	0.0020	9.2567	13.6115	15.02334	19.94519	69.98487
kremlin_km	0.0729	10.4605	14.8792	16.0448	20.6668	70.7388
big_road1_km	0.0004	0.7790	1.72412	1.881276	2.806196	6.995416
big_road2_km	0.0019	2.1034	3.21197	3.396649	4.316292	13.798346
railroad_km	0.0023	0.6550	1.23836	1.88938	2.520431	17.387119
zd_vokzaly_avto_km	0.1367	9.992	14.7576	17.2148	24.0612	91.2151
bus_terminal_avto_km	0.0620	5.214	7.4545	9.99245	13.28391	74.79611
oil_chemistry_km	0.5107	8.721	16.6985	17.4016	23.4245	70.4134
nuclear_reactor_km	0.3098	5.238	8.9653	10.9453	16.3725	64.257
radiation_km	0.0047	1.232	2.4352	4.41078	4.68705	53.89016
power_transmission_line_ km	0.0303	0.976	1.8957	3.49223	4.92655	43.32437
thermal_power_plant_km	0.4006	3.770	5.8924	7.3401	9.8187	56.8561
ts_km	0	2.057	3.972	4.931	5.552	54.081
big_market_km	0.6614	7.530	11.9104	13.2839	16.5602	59.5016
market_shop_km	0.0039	1.5436	2.92742	3.95888	5.48542	41.10365
fitness_km	0	0.3612	0.6563	1.1546	1.334	26.6525
swim_pool_km	0	1.7090	2.877	4.232	5.37	53.359
ice_rink_km	0	3.0440	5.547	6.124	7.957	46.037
stadium_km	0.1148	4.0182	6.9692	9.4367	13.5918	83.3985
basketball_km	0.0055	1.3082	2.8780	4.78764	6.36452	56.70379
hospice_morgue_km	0.0025	1.1182	1.8957	2.64649	3.29732	43.69464
detention_facility_km	0.0412	5.6697	11.31144	14.5511	24.88321	89.37137
public_healthcare_km	0	1.279	2.342	3.357	3.984	76.055
university_km	0.00031	2.2012	4.3376	6.85589	9.38027	84.86215
workplaces_km	0	1.017	2.032	3.927	5.416	55.278
shopping_centers_km	0	0.4838	0.8396	1.5058	1.5495	26.2595
office_km	0	0.5552	1.053	2.011	3.0467	18.9589
additional_education_km	0	0.4748	0.899	1.3285	1.5711	24.2682

preschool_km	0	0.2851	0.493	1.3452	0.9363	47.3947		
big_church_km	0.0041	0.8605	1.4908	2.33005	2.92226	45.66906		
church_synagogue_km	0	0.5325	0.86	0.972	1.2485	15.6157		
mosque_km	0.0055	3.7661	6.5436	7.73924	10.04705	44.84983		
theater_km	0.0268	4.2253	8.6120	9.63807	13.45959	87.60069		
museum_km	0.0079	2.8794	5.6435	7.0632	10.3286	59.2032		
exhibition_km	0.0090	2.2438	4.1067	5.55226	6.9687	54.43124		
catering_km	0.0004	0.2086	0.4127	0.687988	0.841418	12.162697		
green_part_500	0	1.48	8.38	13.38	19.92	100		
prom_part_500	0	0	0	5.718	5.76	98.77		
office_sqm_500	0	0	0	13983	0	611015		
trc_sqm_500	0	0	0	21797	120	1500000		
cafe_sum_500_min_price_avg	300	500	666.7	741.3	954.8	4000	13281	0.4359
cafe_sum_500_max_price_avg	500	1000	1167	1247	1500	6000	13281	0.4359
cafe_avg_price_500	400	750	916.7	994.2	1250	5000	13281	0.4359
green_part_1000	0	6.31	13.04	16.96	24.18	100		
prom_part_1000	0	0	4.02	8.783	12.62	72.2		
office_sqm_1000	0	0	0	62267	54500	2244723		
trc_sqm_1000	0	0	7800	65881	67183	1500000		
cafe_count_1000	0	1	4	15.41	11	449		
cafe_sum_1000_min_price_avg	300	543.2	669.2	710.9	839.3	2500	6524	0.2141
cafe_sum_1000_max_price_avg	500	1000	1143	1207	1400	4000	6524	0.2141
cafe_avg_price_1000	400	750	912.5	958.8	1120	3250	6524	0.2141
green_part_1500	0	8.47	14.95	19.2	26.69	90.41		
prom_part_1500	0	1.52	7.82	10.6	15.34	63		
office_sqm_1500	0	0	16650	140371	117300	2908344		
trc_sqm_1500	0	0	49410	127715	154590	1533000		
cafe_sum_1500_min_price_avg	300	585.7	692.3	714.1	821.4	2500	4199	0.1378
cafe_sum_1500_max_price_avg	500	1000	1167	1206	1367	4000	4199	0.1378
cafe_avg_price_1500	400	795	926.3	960	1093.8	3250	4199	0.1378
green_part_2000	0.01	10.16	17.63	20.84	28.33	75.3		
prom_part_2000	0	3.12	8.8	11.22	16.21	56.1		

office_sqm_2000	0	0	58411	246497	207193	3602982		
trc_sqm_2000	0	12065	117300	212759	286681	2448300		
cafe_sum_2000_min_price_avg	300	607.7	683.3	720	791.7	2166.7	1725	0.0566
cafe_sum_2000_max_price_avg	500	1000	1156	1211	1322	3500	1725	0.0566
cafe_avg_price_2000	400	823.5	919.2	965.4	1057.2	2833.3	1725	0.0566
green_part_3000	0.31	12.15	20.26	22.73	30.36	74.02		
prom_part_3000	0	4.24	9.66	10.98	15.73	45.1		
office_sqm_3000	0	0	130303	543262	494706	6106112		
trc_sqm_3000	0	41100	294350	438132	659453	2654102		
cafe_sum_3000_min_price_avg	300	650	711.1	765.9	815.6	1833.3	991	0.0325
cafe_sum_3000_max_price_avg	500	1102	1212	1283	1333	3000	991	0.0325
cafe_avg_price_3000	400	875.8	961.1	1024.6	1083.3	2416.7	991	0.0325
green_part_5000	3.52	14.78	19.76	22.77	31.41	75.46		
prom_part_5000	0.21	6.05	8.98	10.35	14	28.56	178	0.0058
office_sqm_5000	0	85159	432438	1401057	1433847	12702114		
trc_sqm_5000	0	262000	1075495	1173871	1683836	4585477		
cafe_sum_5000_min_price_avg	300	670.9	721.7	765.1	816.7	1875	297	0.0097
cafe_sum_5000_max_price_avg	500	1144	1212	1278	1346	3000	297	0.0097
cafe_avg_price_5000	400	909.4	966.7	1021.7	1091.7	2437.5	297	0.0097
price_doc	100000	4740002	6274411	7123035	8300000	111111112		

附錄 III、使用的 R 套件

附錄表 3 使用 R 套件一覽表

用途	套件
遺失值處理	mice、VIM
處理資料運算	magittr
繪圖	scales、ggplot2、grid、patchwork、plotrix、 plotly、pheatmap
資料處理	dplyr、caret
模型配適	olsrr、glmnet、selectiveInference、 cmprsk、parcor

附錄 iv、工作分配表

附錄表 4 工作分配表

姓名	工作分配
張馨云	書面報告，EDA
羅季安	EDA，影片剪輯，PPT
王勃淵	資料前處理
李承志	降維與配適模型