倫敦共享單車租借量迴歸分析

- 動機與資料來源
- 資料預處理
- 相關與共線性
- 迴歸分析
- 準確率分析
- 討論與結論

112701006 吳昭泓

動機與資料來源

- 動機:因為我是運管系學生，因此找尋運輸相關主題
  希望能將所學與數據分析連結

- 資料來源:kaggle 資料集 ─ London bike sharing dataset
  2015/01/04~ 2017/01/03 每小時共享單車租借量
  與當時天氣、溫度、風速、濕度，共17414筆資料

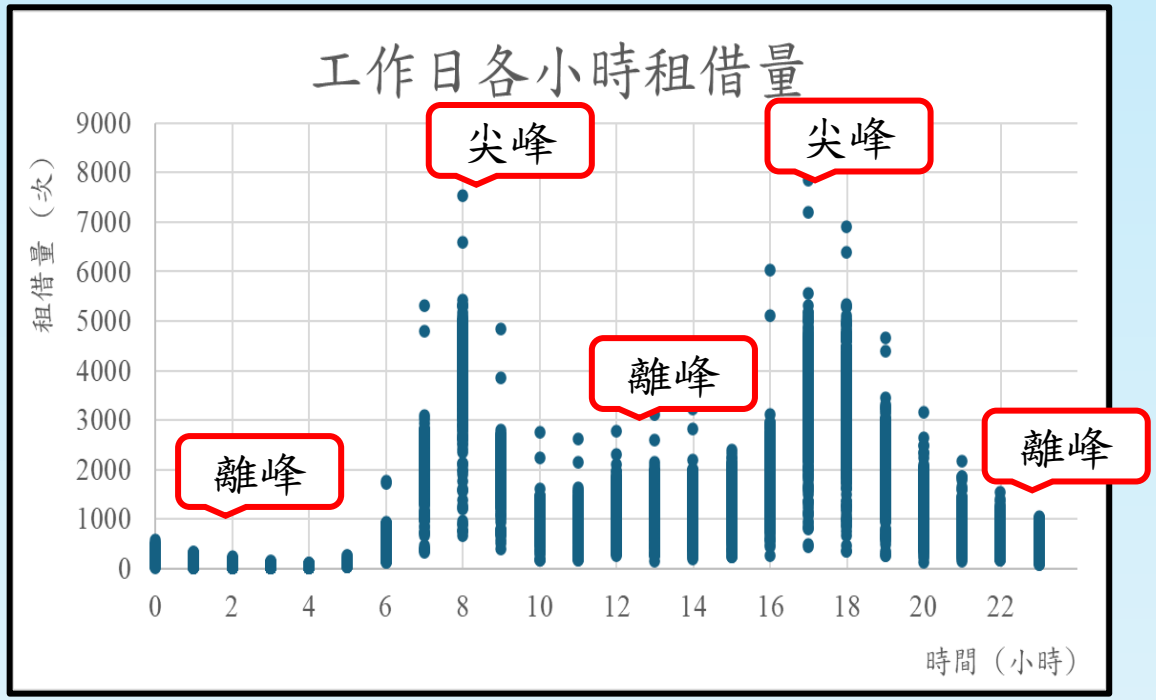
原始資料

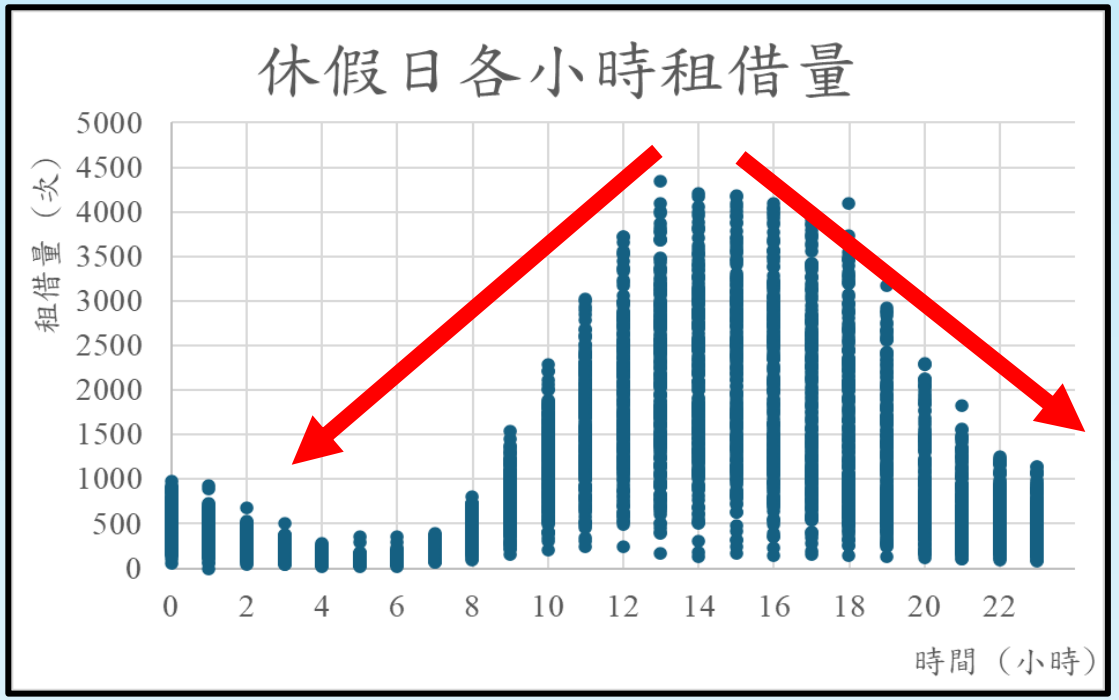
	A	B	C	D	E	F	G	H	I	J	K	L
1	日期	時間 (整點)	租借量 (次)	時間	溫度 (°C)	濕度 (%)	風速 (公里/小時)	是否為假日 (0:是 1: 否)	是否為周末 (0:是 1: 否)	天氣 (如右)		註: 天氣代碼
2	2015/1/4	0:00	182	0:00	3	93	6	0	1	3		1: 晴朗
3	2015/1/4	1:00	138	1:00	3	93	5	0	1	1		2: 少雲
4	2015/1/4	2:00	134	2:00	2.5	96.5	0	0	1	1		3: 碎雲
5	2015/1/4	3:00	72	3:00	2	100	0	0	1	1		4: 多雲
6	2015/1/4	4:00	47	4:00	2	93	6.5	0	1	1		7: 小雨
7	2015/1/4	5:00	46	5:00	2	93	4	0	1	1		10: 雷雨
8	2015/1/4	6:00	51	6:00	1	100	7	0	1	4		26: 雪
9	2015/1/4	7:00	75	7:00	1	100	7	0	1	4		94: 霧

~中間省略~

17412	2017/1/3	20:00	541	20:00	5	81	21	0	0	4
17413	2017/1/3	21:00	337	21:00	5.5	78.5	24	0	0	4
17414	2017/1/3	22:00	224	22:00	5.5	76	23	0	0	4
17415	2017/1/3	23:00	139	23:00	5	76	22	0	0	2

工作/休假日差異





工作日: 上午與下午尖峰
與離峰差距明顯

本專題聚焦於工作日模型的建立

休假日 (周末與假日):
離中午越遠租借量大致越低
可能須多項式(二次式)迴歸

資料預處理

因租借量與時間非呈線性相關

不宜將時間視為一變數，故區分

虛擬變數1: 尖峰 ─ 07、17~18點

虛擬變數2: 晚間 ─ 21~05點

是:1, 否: 0

因天氣類別過多，因此重分類為兩種天氣

0: 壞 (小雨、雷雨、雪、霧)

1: 好 (晴朗、少雲、碎雲、多雲)

藉此定義晴天一個虛擬變數

時間	0~5	6	7	8~16	17~18	19~20	21~23
尖峰	0	0	1	0	1	0	0
晚間	1	0	0	0	0	0	1

拆分資料集

測試資料集: 2016年各月份日期整除4的各小時資料，共1295筆

訓練資料集: 其他資料，共10622筆

目的: 藉由2015年全部資料, 預測次年租借量

並藉由間隔取樣以達成亂數

舉例說明	
日期	用途
1/1	訓練
1/2	訓練
1/3	訓練
1/4	測試
1/5	訓練
1/6	訓練
1/7	訓練
1/8	測試

註: 訓練資料集扣除離群值 — 2015/7/9資料

經查證原因為倫敦地鐵罷工，使單車租借量異常增加

預處理後資料

	A	B	C	D	E	F	G	H	I	J	K
1	年	月	日	時間 (整點)	租借量 (次)	尖峰	晚間	溫度 (°C)	濕度 (%)	風速 (公里/小時)	晴天
2	2015	1	5	0	83	0	1	4	93	6	1
3	2015	1	5	1	67	0	1	4	93	5	1
4	2015	1	5	2	32	0	1	5	87	6	1
5	2015	1	5	3	22	0	1	6	84	7.5	1

~中間省略~

10617	2016	12	30	17	824	1	0	4	100	11	1
10618	2016	12	30	18	674	1	0	4	100	9	1
10619	2016	12	30	19	482	0	0	4	100	11	1
10620	2016	12	30	20	367	0	0	4	100	10	1
10621	2016	12	30	21	251	0	1	4	100	12	1
10622	2016	12	30	22	287	0	1	5	93	12	1
10623	2016	12	30	23	225	0	1	5	93	11	1

相關與共線性

相關係數	租借量(次)	尖峰	晚間	溫度(°C)	濕度(%)	風速(公里/小時)	晴天
租借量(次)	1.00						
尖峰	0.76	1.00					
晚間	-0.67	-0.54	1.00				
溫度 (°C)	0.32	0.07	-0.24	1.00			
濕度 (%)	-0.39	-0.12	0.40	-0.43	1.00		
風速(公里/小時)	0.10	0.06	-0.24	0.17	-0.30	1.00	
晴天	0.15	0.02	-0.04	0.02	-0.30	-0.10	1.00

與租借量的相關係數

尖峰、晚間: 高度相關

溫度、濕度: 中度相關

風速、晴天: 低度相關

濕度與晚間、溫度、風速、晴天呈中度相關

可能有共線性

迴歸試算一 — 加入所有變數

迴歸統計	
R 的倍數	0.91
R 平方	0.82
調整的 R 平方	0.82
標準誤	475.27
觀察值個數	10622

模型解釋力: 0.82

P值: 皆通過

	係數	標準誤	t 統計	P-值
截距	1127.52	44.68	25.24	<0.001
尖峰	2019.77	14.73	137.17	<0.001
晚間	-995.66	10.87	-91.56	<0.001
溫度 (°C)	34.58	0.93	37.29	<0.001
風速 (公里/小時)	-8.66	0.64	-13.53	<0.001
濕度 (%)	-4.68	0.41	-11.41	<0.001
晴天	327.77	14.87	22.04	<0.001

迴歸試算二 — 扣除濕度變數

迴歸統計	
R 的倍數	0.91
R 平方	0.82
調整的 R 平方	0.82
標準誤	478.16
觀察值個數	10622

扣除原因:

與其餘變數有共線性

模型解釋力: 0.82

P值: 皆通過

	係數	標準誤	t 統計	P-值
截距	679.99	21.55	31.55	<0.001
尖峰	2015.76	14.81	136.11	<0.001
晚間	-1033.54	10.42	-99.21	<0.001
溫度 (°C)	38.57	0.86	44.63	<0.001
風速 (公里/小時)	-6.84	0.62	-10.97	<0.001
晴天	383.09	14.15	27.08	<0.001

迴歸試算三 — 再扣除晴天虛擬變數

迴歸統計	
R 的倍數	0.90
R 平方	0.81
調整的 R 平方	0.81
標準誤	494.37
觀察值個數	10622

扣除原因:

與租借量低度相關

模型解釋力: 0.81

P值: 皆通過

	係數	標準誤	t 統計	P-值	下限 95%	上限 95%
截距	1045.12	17.39	60.11	<0.001	1011.04	1079.20
尖峰	2016.37	15.31	131.68	<0.001	1986.36	2046.39
晚間	-1046.73	10.76	-97.28	<0.001	-1067.82	-1025.64
溫度 (°C)	39.28	0.89	43.98	<0.001	37.53	41.03
風速 (公里/小時)	-8.94	0.64	-13.97	<0.001	-10.19	-7.69

迴歸試算四 — 再扣除風速變數

迴歸統計	
R 的倍數	0.90
R 平方	0.80
調整的 R 平方	0.80
標準誤	498.87
觀察值個數	10622

扣除原因:

與租借量低度相關

模型解釋力: 0.80

P值: 皆通過

	係數	標準誤	t 統計	P-值	下限 95%	上限 95%
截距	902.11	14.18	63.62	<0.001	874.31	929.90
尖峰	2022.64	15.45	130.96	<0.001	1992.36	2052.91
晚間	-1015.65	10.62	-95.61	<0.001	-1036.47	-994.82
溫度 (°C)	38.23	0.90	42.57	<0.001	36.47	39.99

準確率評估

針對模型解釋力 (R平方) 相同者，再以平均絕對誤差 (MAE) 評估

試算一 : 425.53　試算二 : 390.39

因此，試算二為目前最佳模型，公式:

*小時租借量(次) = 679.99 + 2015.76 尖峰 (7, 17~18 時)**

*– 1033.54 晚間 (21~5 時) * + 38.57 溫度 (℃)*

*– 6.84 風速 (公里/小時) + 383.09 晴天**

*標示*者為虛擬變數: 是為 1, 否為 0*

討論

- 風速、晴天與租借量相關係數低，但 p 檢定<0.001 仍通過

  → 因相關係數矩陣僅討論單獨因素對租借量的影響

  但迴歸探討固定其他因素下，某因素對租借量影響

- 試算二扣除濕度後，使準確率較高

  → 因溼度與其他因素有共線性

  → 或試算一變數過多導致過度擬合 (overfitting)

結論

- 尖峰、晚間與晴天迴歸直線係數大
  → 時間與天氣為影響租借量的主要因素

- 溫度與風速影響租借量有限，但共享單車公司若可取得資料仍可增加模型解釋力與預估準確率

- 濕度與其餘變數共線性高，從模型中移除可增加預估準確率

- 此模型可供共享單車公司做為車輛購置與調度的參考依據