主題：環境與人口密度對登革熱確診數影響

研究變項：

Y（應變數）：每一百萬人中就有幾個人確診登革熱。

$$\text{登革熱確診 ppm} = \frac{\text{登革熱確診數}}{\text{人口數}} * 1000000$$

X（自變數）：空氣品質、人口密度及其他環境因素

資料：

變數名稱	單位	類型	中文說明
station		類別	測站代碼
date	YYYY-MM-DD	datetime	日期
area		類別	北高雄 / 南高雄
season		類別	依月份分類
sum_RAINFALL	mm	連續	當日總降雨量
max_AMB_TEMP	˚C	連續	當日最高溫度
mean_AMB_TEMP	˚C	連續	當日平均溫度
max_PM2_5	μg/m^3	連續	當日最高 PM2.5
mean_PM2_5	μg/m^3	連續	當日平均 PM2.5
max_RH	%	連續	當日最高相對濕度
mean_RH	%	連續	當日平均相對濕度
max_O3	ppb	連續	當日最高臭氧濃度
mean_O3	ppb	連續	當日平均臭氧濃度
max_NO	ppb	連續	當日最高一氧化氮濃度
mean_NO	ppb	連續	當日平均一氧化氮濃度
item_ppm	ppm	連續	登革熱確診率（應變數）

研究目的：

目的	使用變數	統計/圖形
判斷地區與季節是否具關聯性	地區、季節	卡方
判斷不同季節登革熱確診率是否存在差異	季節、確診率 y	ANOVA
判斷哪一季確診情形最嚴重	季節、確診率 y	箱型圖
判斷確診率與多項環境因子是否顯著相關	環境因子、確診率 y	Pearson
確診率與環境因子之間的關係分布	環境因子、確診率 y	散佈圖

壹、 判斷地區與季節是否具關聯性

這裡我們針對地區以及季節去做卡方檢定看他彼此之間是否獨立。假設檢定如下：

H_0：地區、季節兩變數是獨立的

H_1：地區、季節兩變數不是獨立的

FREQ 程序

次數
百分比
列百分比
欄百分比

area * season的表格

area	season			
	冬季	秋季	夏季	總計
北高雄	43 6.25 17.41 30.50	176 25.58 71.26 36.74	28 4.07 11.34 41.18	247 35.90
南高雄	98 14.24 22.22 69.50	303 44.04 68.71 63.26	40 5.81 9.07 58.82	441 64.10
總計	141 20.49	479 69.62	68 9.88	688 100.00

area * season 之表格的統計值

統計值	DF	值	機率
卡方	2	2.7597	0.2516
概度比卡方	2	2.7886	0.2480
Mantel-Haenszel 卡方	1	2.7084	0.0998
Phi 係數		0.0633	
列聯係數		0.0632	
Cramer V		0.0633	

這裡可以看到佔比最高的是秋季的南高雄，佔了比較大的部分。另外可以看到卡方檢定 p value 大於 0.05 不拒絕虛無假設，本研究未能證實地區與季節兩者具有顯著關聯性。

貳、 判斷不同季節登革熱確診率是否存在差異

這裡我們針對季節做 anova 變異數檢定。

H_0：每組(秋季、冬季、夏季)μ 無顯著差異

H_1：每組(秋季、冬季、夏季)至少有一組 μ 存在顯著差異

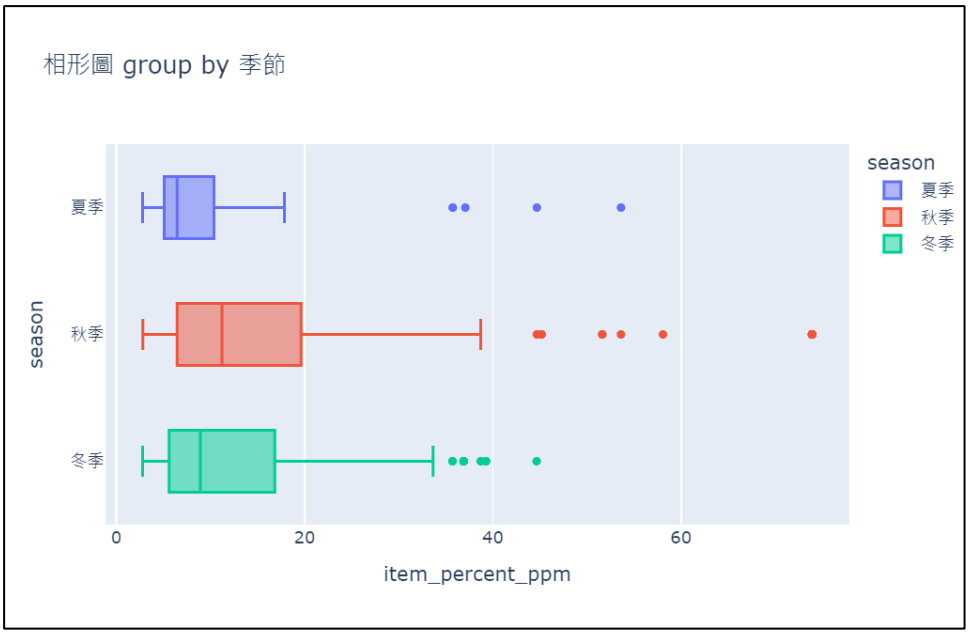
ANOVA 程序

應變數: y

來源	DF	平方和	均方	F 值	Pr > F
模型	2	2447.97592	1223.98796	8.05	0.0004
誤差	585	88918.09291	151.99674		
已校正的總計	587	91366.06884			

結果顯示 P value = 0.0004 小於 0.05，拒絕虛無假設，表示在統計上組別之間至少有一組的 μ 不相等，後續也可以進一步探討每兩組之間的 μ 是否存在顯著差異。

參、 判斷哪一季確診情形最嚴重



此圖為用以季節作分組的箱型圖，可以看得出來它們彼此之間的變異數、μ都有顯著差異。其中**秋季確診情形最為嚴重**。

肆、 判斷確診率與多項環境因子是否顯著相關

這裡我們針對所有連續的解釋變數對反應變數做相關係數檢定，假設檢定如下：

H_0：該環境因子與登革熱確診率之間沒有線性相關

H_1：該環境因子與登革熱確診率之間具有線性相關

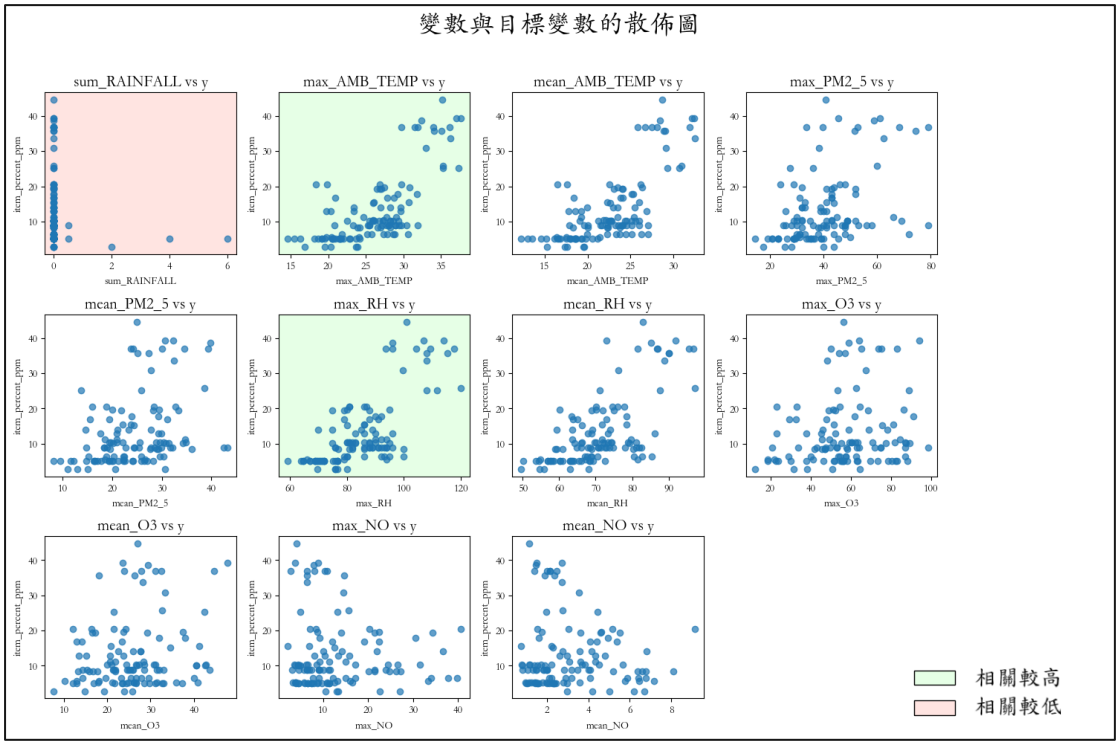
Pearson 相關統計值 (Fisher z 轉換)

變數	常變數	N	樣本相關	Fisher z	偏差調整	相關估計值
sum_RAINFALL	y	588	-0.12029	-0.12088	-0.0001025	-0.12019
max_AMB_TEMP	y	585	0.57707	0.65805	0.0004941	0.57674
mean_AMB_TEMP	y	585	0.54287	0.60822	0.0004648	0.54254
max_PM2_5	y	588	0.40060	0.42437	0.0003412	0.40032
mean_PM2_5	y	588	0.39135	0.41339	0.0003333	0.39106
max_RH	y	587	0.58958	0.67702	0.0005031	0.58925
mean_RH	y	587	0.48545	0.53009	0.0004142	0.48513
max_O3	y	588	0.27048	0.27739	0.0002304	0.27027
mean_O3	y	588	0.32026	0.33194	0.0002728	0.32002
max_NO	y	588	-0.08569	-0.08590	-0.0000730	-0.08561
mean_NO	y	588	-0.12297	-0.12359	-0.0001047	-0.12286

變數名稱	中文說明	相關方向	強度評等*
sum_RAINFALL	日累積降雨量	負	低度
max_AMB_TEMP	當日最高溫度	正	**中度**
mean_AMB_TEMP	當日平均溫度	正	**中度**
max_PM2_5	當日最高 PM2.5	正	**中度**
mean_PM2_5	當日平均 PM2.5	正	低度
max_RH	當日最高相對濕度	正	**中度**
mean_RH	當日平均相對濕度	正	**中度**
max_O3	當日最高臭氧濃度	正	低度
mean_O3	當日平均臭氧濃度	正	低度
max_NO	當日最高一氧化氮濃度	負	低度
mean_NO	當日平均一氧化氮濃度	負	低度

** 0.4 以上視為中度相關*

伍、 確診率與環境因子之間的關係分布



由散佈圖結果可觀察到，各環境變數與登革熱確診率之間多呈現分散分布，整體線性關係較弱。其中，以當日最高溫度與當日最高相對濕度之分布較具明顯正向趨勢，與前述相關係數分析相符，代表氣溫及濕度越高，確診率有上升之傾向。其餘變數則未顯示明確線性關係。