國立成功大學統計系

統計諮詢期末報告

指導教授:蘇佩芳

天氣與商品銷售量關係之研究

組員:

胡博仁

陳柏愷

李茂源

楊庭逸

目錄

圖目錄

表目錄

壹、緒論

一. 研究動機與目的

全球氣候急速變遷，許多企業早已利用準確的氣象預估來增加獲利，因為了解氣象才是掌握商機的關鍵。根據聯合報報導，「世界氣象組織」最新研究顯示，企業每投資一美元在天氣預報，就可以減少十美元的經濟損失。臺灣也有氣象公司專門提供企業準確的氣象預測，冷氣業者、服飾商家、便利商店，甚至清潔用品公司都是客戶。當冬季寒流來襲，百貨公司的大衣專櫃就會生意興隆；夏天熱浪來襲，啤酒與冷飲的業績也會特別出色。由於全球氣候變化愈來愈極端，企業更需要氣象預報來準確計算成本、掌握商機。過去人類無法掌握天候變化，但現在科技進步，人類逐漸能掌握氣象變化因素，預測雨量與氣溫的準確度也隨之提高。純粹的冷暖變化預測不難，但氣溫、下雨機率的準確度，會隨著時間遠近而下降，例如隔天的命中率為八九成，一週後約五六成，兩週後則是五成以下。

因此，雖然氣象資訊本身無法創造價值，但只要人類有效運用，就會產生非常可觀的社會經濟效益，或是避免龐大的損失發生，例如豪雨、暴風雪、濃霧等天災預測。

而接下來在這份報告中，我們也將針對 walmart 提供的資料做深入的探討，期望能給予管理者一些有用的資訊。

貳、前言

一. 背景介紹

Walmart 是一家美國的跨國零售企業，由 Sam Walton 於 1962 年在阿肯色州市成立的，確切註冊日在 1969 年 10 月 31 日。起初，鄉村地區的小城鎮市場顧客量不足，大型百貨連鎖店都往大城市發展，Walmart 卻不斷在城鎮開設零售店，而獲得相當大的擴展機會，而其消費族群從一開始的中下階級到之後發展高級路線，可說是由中下階級包覆了整體。

現今為全球第二大上市公司。也是世界上最大的私人雇主，員工超過兩百萬，是全美最大零售商，同時也是世界上最大的零售商。他們的口號 「Save Money,Live Better」而聞名。

Walmart 於 27 個國家共設立了 11450 間店。如此龐大的店數，Walmart 在控制成本這方面做得十分徹底，故需要找出影響成本的各種因素，也因此會有這個主題－「天氣與銷售量關係之研究」。

只要天氣稍有變化，所有商品銷售都有可能因此產生變化。而在 Walmart 提供的資料中，天氣觀測站共有 20 間，提供各種天氣資訊。並抽出了 45 間店家，對整體做推測。最後提供了 111 種對於天氣較敏感的商品（不提供商品名），好比說：雨傘、雜糧、牛奶……等等，資料中包含各店各個商品每日銷售量。時間達兩年十個月。

接下來，我們將用這些天氣資料，來找出其與銷售量之關係，如此希望能給 Walmart 在進出貨的決策做個參考。

二. 資料與變數解釋

首先，我們介紹資料，會分為「原始資料」及「合併後資料」來看。而我們會想要做合併後資料的原因－**資料筆數過大**。由於總共有超過 400 萬筆的資料，在做統計方法時，會有許多耗時以及資料處理無法完全的問題，因此將資料做合併。而在 2-2

的部分也會詳細講說我們是如何合併資料。

在整份報告裡我們會著重在「**合併後資料**」，因為擁有了所有想要研究的變數，在 R 的操作上也會相當方便。

2.1 原始資料

此資料為 walmart 提供的原始資料，原本為兩個 excel 檔案—weather&train。一為 20 間天氣觀測站各天之天氣資料，另一檔案為產品銷售量的資料中日期卻不完整，少了許多天的資料，故我們決定將其做調整，並把銷售量(train)的檔案合併(merge)到天氣觀測站(weather)之檔案中。

Station 觀測站編號	1	2	3	4	5
Date 日期	2012/1/1	2012/1/1	2012/1/1	2012/1/1	2012/1/1
Tmax 最高溫	52	48	55	63	63
Tmin 最低溫	31	33	34	47	34
Tavg 平均溫度	42	41	45	55	49
Depart 標準溫差	M	16	9	4	0
Dewpoint 露點溫度	36	37	24	28	31
Wetbulb 濕度	40	39	36	43	43
Heat 夏天溫度差	23	24	20	10	16
Cool 冬天溫度差	0	0	0	0	0
Sunrise 日出時間	-	716	735	728	727
Sunset 日落時間	-	1626	1720	1742	1742
Codesum 天氣代碼	RA FZ FG BR	RA			
Snowfall 降雪量	M	0	0	0	0
Preciptotal 降雨量	0.05	0.07	0	0	0
Stnpressure 氣壓	29.78	28.82	29.77	29.79	29.95
Sealevel 海平面氣壓	29.92	29.91	30.47	30.48	30.49
Resultspeed 風速	3.6	9.1	9.9	8	14
Resultdir 風向	20	23	31	35	36
Avgspeed 平均風速	4.6	11.3	10	8.2	13.8

圖 2-1 原始資料(天氣)

上頭為天氣的原始資料，可以看見所有變數都在這裡呈現，大多的變數也是後面研究的重點變數，在 2.3 也會詳細解釋所有變數。

Date 日期	Store 店家編號	Item 商品編號	Units 銷售單位
2012/1/1	1	1	0
2012/1/1	1	2	0
2012/1/1	1	3	0
2012/1/1	1	4	0
2012/1/1	1	5	0
2012/1/1	1	6	0
2012/1/1	1	7	0
2012/1/1	1	8	0
2012/1/1	1	9	29
2012/1/1	1	10	0

圖 2-2 原始資料(每間店各商品銷售量)

圖 2-2 則是每間店的 111 種商品的每日銷售量，而**銷售量**是反應所有研究結果的重要變數，但由於資料量過大，在處理上有一定的程度的困難，故我們將 111 種商品的銷售量加總，成為**總銷售量**。在後面的報告，我們做的各種方法研究也都是以總銷售量為依據。

2.2 Merge 後資料

Date 日期	2012/1/1	2012/1/2	2012/1/3	2012/1/4	2012/1/5
Station 觀測站編號	1	1	1	1	1
Store 店家編號	1	1	1	1	1
Tmax 最高溫	52	50	32	28	38
Tmin 最低溫	31	31	11	9	25
Tavg 平均溫度	42	41	22	19	32
Depart 標準溫差	M	M	M	M	M
Dewpoint 露點溫度	36	26	4	-1	13
Wetbulb 濕度	40	35	18	14	25
Heat 夏天溫度差	23	24	43	46	33
Cool 冬天溫度差	0	0	0	0	0
Sunrise 日出時間	-	-	-	-	-
Sunset 日落時間	-	-	-	-	-
Codesum 天氣代碼	RA FZ FG BR				
Snowfall 降雪量	M	M	M	M	M
Preciptotal 降雨量	0.05	0.01	0	0	0
Stnpressure 氣壓	29.78	29.44	29.67	29.86	29.67
Sealevel 海平面氣壓	29.92	29.62	29.87	30.03	29.84
Resultspeed 風速	3.6	9.8	10.8	6.3	6.9
Resultdir 風向	20	24	31	27	25
Avgspeed 平均風速	4.6	10.3	11.6	8.3	7.8
DU 沙塵暴	0	0	0	0	0
FG 濃霧	0	0	0	0	0
GR 冰雹	0	0	0	0	0
SN 下雪	0	0	0	0	0
SQ 暴風雪	0	0	0	0	0
TS 暴風雨	0	0	0	0	0
01 天氣有或無	1	0	0	0	0
Ext 極端天氣	0	0	0	0	0
Units 總銷售單位	32	66	24	23	17

圖 2-3 合併後資料

合併資料之重點是將主要天氣個別列出以便觀察，以及列出各天每間商店之總銷售量，作為研究其與天氣之關係。

我們感興趣的主要天氣分別為**暴風雪**、**下雪**、**沙塵暴**、**大濃霧** 、**冰雹**及**暴風雨**六種天氣情形，將這六種天氣單獨列為六項變數視為類別變數，有發生記為 1，若沒有則記為 0。

此外，也將有無天氣情形以類別變數處理，只要有天氣情形為 1，完全沒有則為 0。再者，還將上述六種天氣併作為極端天氣類，也同樣視為類別變數，當有發生上述六種天氣至少一種為 1，沒有則為 0。

合併資料之過程及方法：

由於天氣資料(weather)及銷售量資料(train)是分開的，因此我們在資料處理上就花了非常多的力氣。

在天氣資料(weather)的紀錄上面除了 5 號觀測站是從 2012/6/1 開始記錄，此外並沒有任何未登記的日期。而其中在平均氣溫等資料中有一些缺失值，我們採用相同觀測站且類似天氣狀況下的平均值去取代這些缺失值。而像 sunrise 及 sunset 等有些變數我們主觀上認為並不會影響銷售量，因此我們在後面並無使用這些資料。

在銷售量資料(train)中記錄 111 種商品每日每間店的數量，資料量約 400 多萬筆，但是 excel 只能顯示到約 100 萬筆資料，另外我們又發現其中每間不同代號的店，記錄的資料中皆在不同的日期有一定天數資料的缺失，於是我們需要先將銷售量資料與天氣資料的日期對應起來，將銷售量沒有的日期從天氣資料中移除。

我們最剛開始嘗試合併資料是先從 excel 的資料中找尋缺失的日期，可是由於 excel 顯示上限的關係，我們只找到了 2012/12/25 號有缺失值，而其餘的缺失日期無法顯示。接著我們改用 R 將資料讀入(R 在資料量為 400 萬筆時沒有讀取資料的限制)，我們開始嘗試從第一間店的缺失值先開始尋找，發現他們的缺失值在日期上並沒有一定的規律可循，且如果要將 45 間店皆用手動的方式去尋找缺失值的話會花費過於大量的時間，因此我們改嘗試用迴圈的方式去挑出缺失值。

在嘗試用迴圈時，我們先將店家號碼排列，並且將每個店家號碼對應到觀測站的號碼。接著用迴圈將店號和對應的觀測站天氣資料叫出，可是在我們嘗試用此方法去合併資料時，R 卻有最大運算上的限制，所以此方法又宣告失敗。

在嘗試許多次都失敗後請教了許多的同學，然而大家的回答幾乎都是挑出缺失的日期並從天氣資料中將這些對應的日期拿掉，跟我們嘗試的方法都很類似，於是我們又花了很多的時間嘗試用不同的迴圈方式去將資料合併，可是卻又都失敗。在最後，我們詢問老師的建議下，得知了有 merge 這個指令，於是又開始了新的一個路程。

在 merge 時我們用店家號碼(1 到 45)作為第一個迴圈，在第一個迴圈內再放入第二個迴圈，將對應到第一個迴圈號碼的店家銷售量資料取出，也取出對應店家有登記資料的日期，並且將每日 111 種商品的銷售量相加並放入銷售量資料對應的日期，接著再將銷售量資料(將 111 種商品合併放入後)與天氣資料用日期(date)這一行當基準做合併(merge)，然後用 rbind 的方式將第一個迴圈的 45 間店，合併(merge)過後的資料合併(rbind)起來，最後在多次嘗試後終於得出了完整合併後的資料。

2.3、變數名詞解釋

＊紅字為後續章節之重要變數

分類	敘述解釋
Date(日期)	從 2012 年到 2014 年 10 月 31 日，時間達兩年 10 月。
Station(觀測站)	天氣觀測站，記錄各個天氣資訊，共有 1~20 間。
Store(店家)	抽出店家數共有 45 間，記錄各個商品之每日銷售量。
Tmax(最高溫)	當天測量之最高溫度，單位為華氏°F。
Tmin(最低溫)	當天測量之最低溫度，單位為華氏°F。
Tavg(平均溫度)	以最高溫及最低溫做平均溫度，單位為華氏°F。
Depart(標準溫差)	當天均溫是否差往年溫度過多。
Dewpoint(露點溫度)	氣態水凝結到液態水所降到之溫度，單位為華氏°F。
Wetbulb(濕度)	空氣中水蒸氣含量，單位為公克／立方公尺。
Heat(夏天溫度差)	有效時間從七月至一月，以標準均溫 65°F為基準，與標準均溫之差，高於為正低於為負，單位為華氏°F。
Cool(冬天溫度差)	有效時間為一月至七月，標準均溫 65°F為基準，與標準均溫之差，高於為正低於為負，單位為華氏°F。
Sunrise(日出時間)	日出之時間。
Sunset(日落時間)	日落之時間。
Codesum(天氣代碼)	當天出現之天氣情形。下方為天氣總表 FC FUNNEL CLOUD 漏斗雲 **TS THUNDERSTORM 暴風雨** **GR HAIL 冰雹** RA RAIN 下雨 DZ DRIZZLE 毛毛雨 **SN SNOW 下雪** SG SNOW GRAINS 結晶雪 GS SMALL HAIL 小冰雹 PL ICE PELLETS 球狀雪 IC ICE CRYSTALS 雪花 **FG FOG 濃霧** BR MIST 薄霧 VA VOLCANIC ASH 火山灰 +FC TORNADO/WATERSPOUT 龍捲風 **DU WIDESPREAD DUST 沙塵暴(微塵)** DS DUSTSTORM (強烈)沙塵暴(微塵) SA SAND 沙塵(沙粒)

	SS SANDSTORM 沙塵暴(沙粒) PY SPRAY 水霧 **SQ SQUALL 暴風雪** DR LOW DRIFTING 微塵蝻 SH SHOWER 大雨 FZ FREEZING 結冰 MI SHALLOW 淺水 PR PARTIAL 地形雨 BC PATCHES 陰天 BL BLOWING 大風 VC VICINITY 周邊 - LIGHT + HEAVY
Snowfall(降雪量)	當天之積雪累積高度，單位為毫米 mm。
Preciptotal(降雨量)	當天之下雨累積高度，單位為毫米 mm。
Stnpressure(氣壓)	當天大氣層之空氣重力，單位為毫米汞柱 mm-hg。
Sealevel(海平面氣壓)	當天海面之空氣重力，單位為毫米汞柱 mm-hg。
Resultspeed(風速)	當天之風速，單位為公尺／秒(m/s)。
Resultdir(風速)	當天風吹來之方向，北方為 0 度依順時針增加。
Avgspeed(平均風速)	當天之平均風速，單位為公尺／秒(m/s)。
01(有無天氣)	分為有或無天氣狀況，有為 1，無為 0。
ext(極端天氣)	影響力大之天氣情況，包含暴風雨、冰雹、下雪、濃霧、沙塵暴(微塵)、暴風雪。有為 1，無為 0。
Units(總銷售單位)	當天之各間商店的 111 種商品銷售量加總，單位為「個」。

表 2-4 變數名詞解釋

參、一般天氣有無情形

所謂一般天氣有無，意指只要出現任何天氣情況則為有，完全沒有天氣情況則為無，依此來做區分。

針對一般天氣情況的有無，我們進行初步平均銷售量分析，探討有一般天氣情況下及無一般天氣情況下平均銷售量是否存在差異。首先我們將資料分為“有一般天氣狀況”及“無一般天氣狀況”兩類樣本，分別檢驗其是否成常態，在此我們選用*nortest* package 中ad.test()，其主要是透過critical value是否遵循於某一分配進行檢定，後來亦常用來檢定是否為常態分配，而不選用一般常見shapiro test是因為此檢定方法在R程式當中，只能檢驗樣本數小於 5000 的資料。 若資料都為常態，則採用t.test檢定，若有一為非常態，我們即用無母數統

計方法 Wixconon rank sum test 來檢驗平均銷售量是否有差異。

一、敘述統計

◎從下表 3-1，我們可以得知：

1. 樣本數：沒有相差至太多。

2. 最小值：一樣，皆有當日無銷售情形。

3. 四分位距：都有些微量之差距。由此觀察出，在無天氣狀況下，銷售量會較有天氣狀況來的高。

4. 最大值：有天氣狀況高出無天氣狀況許多，推論有可能是為變天前的準備。

5. 平均數：無天氣狀況較有天氣狀況高一些。

表 3-1　有無天氣狀況之敘述統計量

	n	min	1st Qu	Median	Mean	3rd Qu	Max	sd	trimmed	se
有天氣狀況	18458	0	51	89	105	139	903	78.74	95	0.58
無天氣狀況	23140	0	57	96	112	145	819	80.65	102	0.53

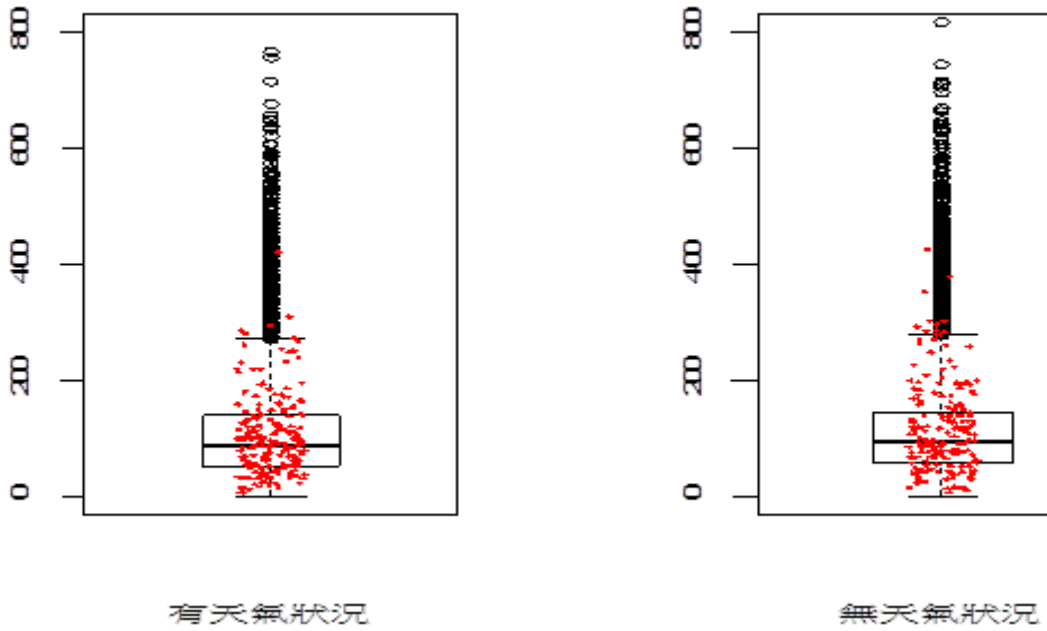


圖 3-2　有無一般天氣狀況之盒鬚圖

◎從圖 3-2 得知，兩者之盒鬚圖沒有太大之差異。但仔細看可發現，無天氣狀況較有天氣壯況高出一些。

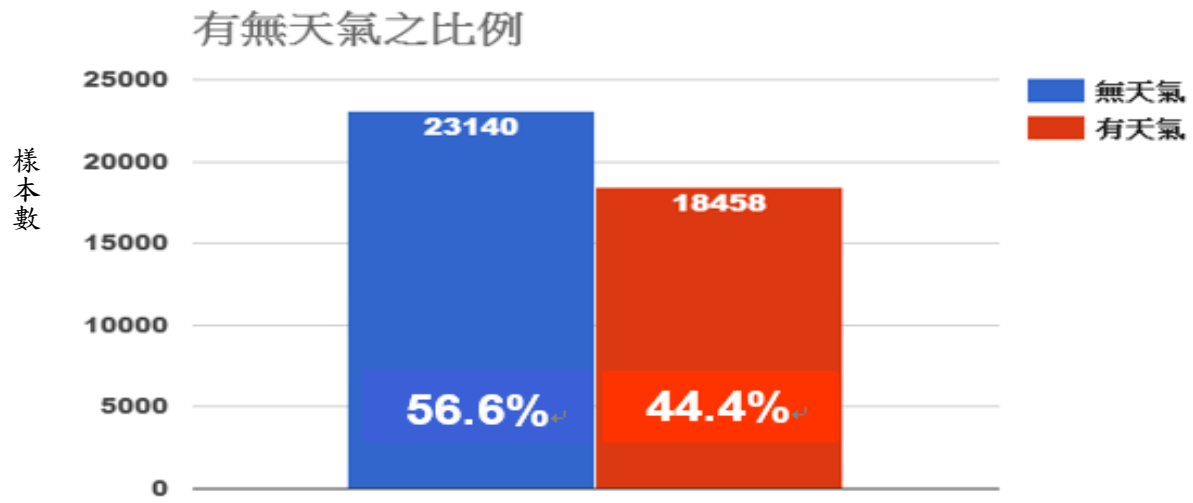


圖 3-3　有無天氣狀況之天數及比例

◎在圖 3-3 能看出無天氣況之樣本數較多，相差約 5000 筆左右。

二、Anderson-Darling 檢定法

在此部分，首先我們將資料分為“有一般天氣狀況”及“無一般天氣狀況”兩類樣本，檢驗兩者是否成**常態**，如此才能選出正確的檢定方法確認兩者的銷售量差異。

在此我們選用 nortest package 中 Anderson-Darling.test(ad.test)，這裡我們必須額外安裝一個軟體套件”nortest ”，才能讓 Anderson-Darling test 在 R 中運作。由於其是透過 crtical value 進行檢驗，故也能推廣運用到檢定資料是否為其他的特定分配，只要更改內定的虛無假設即可，是方便的檢定方法。

其檢定統計量為$A^2 = -N - S$ ， $S = \sum_{i=1}^{n} \frac{(2i-1)}{N} [\ln F(Y_i) + \ln(1 - F(Y_{N+1-i}))]$

F 是常態分配的 cumulative distribution function，Y_i是資料的順序

接者再透過查表，看其是否落在拒絕域中。然而，我們不選用一般常見 shapiro test 是因為此檢定方法在 R 程式當中，只能檢驗樣本數小於 5000 的資料。

接下來，我們要先檢定兩者資料是否為常態，故先設立虛無假設 H_0 及對立假設 H_a ：

H_0 ： *樣本資料呈常態*

H_a ： *樣本資料非常態*

α *=0.05*

接著，將有無天氣情況兩者資料，各別做 ad.test，得出各自的 P-value：

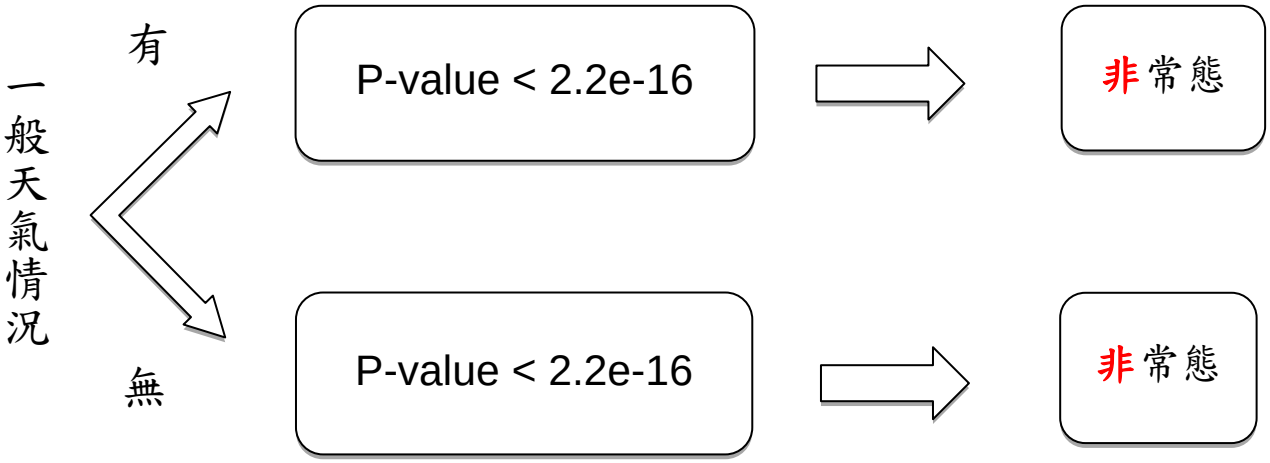


圖 3-4　有無天氣狀況之常態判斷

◎ 做完 ad.test 後，得出兩者皆非常態。無法使用常態之檢定方法，故做無母數檢定。

三、Wilcoxon rank sum test

由於在檢定常態性的部分，得出兩者皆為**非常態**。故在此我們使用無母數檢定之方法，而方法為：Wilcoxon rank sum test。我們這裡選擇不用一般最常用來檢定兩平均

數是否相同的 t-test，原因是 t-test 必須假定兩母體是成常態的，這裡明顯兩母體違反此一假設，所以我們選擇透過無母數中的 Wilcoxon rank sum test 來進行兩母體平均數檢定，此一方法不需假設母體是否成常態。Wilcoxon rank sum test 主要是透過排序編號等級來比較兩母體之間的平均數是否有差異。$W_{1=}\sum_{j=1}^{n_1} R_{1j}$　R_{1j} 為其編號等級、而檢定統計量:

$u = W_1 * \frac{n(n+1)}{2}$，n 為總樣本數，接者我們再透過查 Wilcoxon rank sum 表，找出其 critical value 看是否落在其拒絕域中

現在我們對兩者平均銷售量的差別設立虛無假設 H_0 及對立假設 H_a ：

H_0 ： *一般天氣情況不會影響平均銷售量*

$(\mu_{有一般天氣} = \mu_{無一般天氣})$

H_a ： *一般天氣情況會影響平均銷售量*

$(\mu_{有一般天氣} \neq \mu_{無一般天氣})$

$\alpha = 0.05$

對其做 wilcoxon rank sum test，求其檢定統計量，得出 P-value：

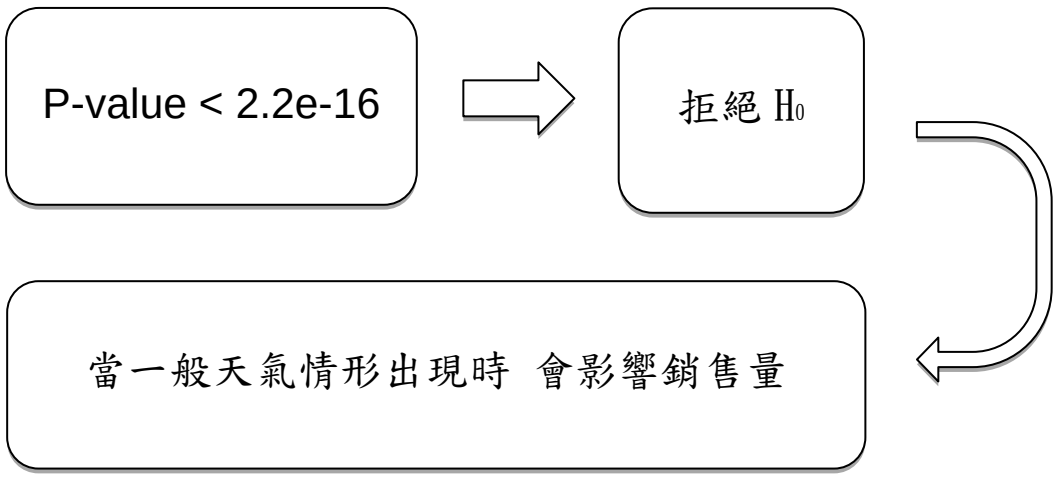


圖 3-5　一般天氣情況是否影響銷售量結論

◎ 由上面的檢定方法可得知，拒絕虛無假設 H_0。故兩者之間存在差異為顯著的。

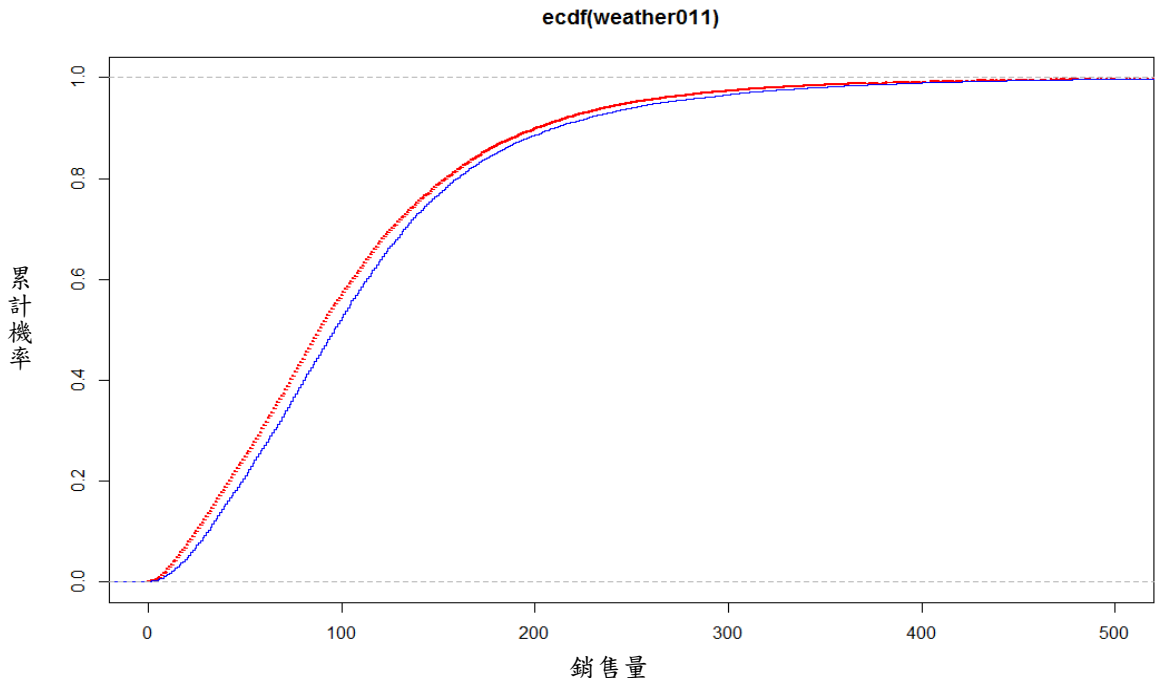


圖 3-6　有無一般天氣情形之 cdf 圖

另外我們由累積相對次數圖來佐證上述之結果（圖 3-6 紅線為有一般天氣情形、藍線則無一般天氣情形），紅線整體上普遍落於藍線左方，得知當有天氣情形發生時，銷售量會降低，與我們檢定相符，我們也可以推估：當有一般天氣情形發生時，人們多會懶得出門，選擇待在家裡。

肆、主要天氣有無情形

所謂主要天氣有無，意指將沙塵暴、濃霧、冰雹、下雪、暴風雪、暴風雨這六種天氣情況視為主要天氣情形，而沒有上述六種主要天氣情況則為無主要天氣情形，依此來做區分。

針對主要天氣情況的有無，我們會同第參章－一般天氣有無情形，進行初步平均銷售量分析，探討有主要天氣情況下及無主要天氣情況下平均銷售量是否存在差異。步驟為，將資料分為“有主要天氣狀況”及“無主要天氣狀況”兩類樣本，分別檢驗其是否成常態，使用 ad.test 以檢測（此方法原理及選用理由可參造第參章）。 若資料都為常態，則採用 t.test 檢定，若有一為非常態，我們即用無母數統計方法 Wixconon rank sum test 來檢驗平均銷售量是否有差異。

一、敘述統計

◎從下表 4-1，我們可以得知：

1.樣本數：相差極多，在檢定方面需留意。

2.最小值：一樣，皆有當日無銷售情形。

3.四分位距：都有些微量之差距。然而，此與第參章－一般有無天氣情況完全相反，在有主要天氣狀況下，銷售量反而會較無主要天氣狀況來的高。

4.最大值：有主要天氣狀況略高於無天氣狀況，在此相差不多。

5.平均數：無天氣狀況較有天氣狀況高一些。

	n	min	1st Qu	Median	Mean	3rd Qu	Max	sd	trimmed	se
有主要天氣	7111	0	57	94	111	148	903	79.94	101	0.95
無主要天氣	34487	0	54	93	109	142	873	79.87	98.11	0.43

表 4-1　主要天氣情銷售量結論

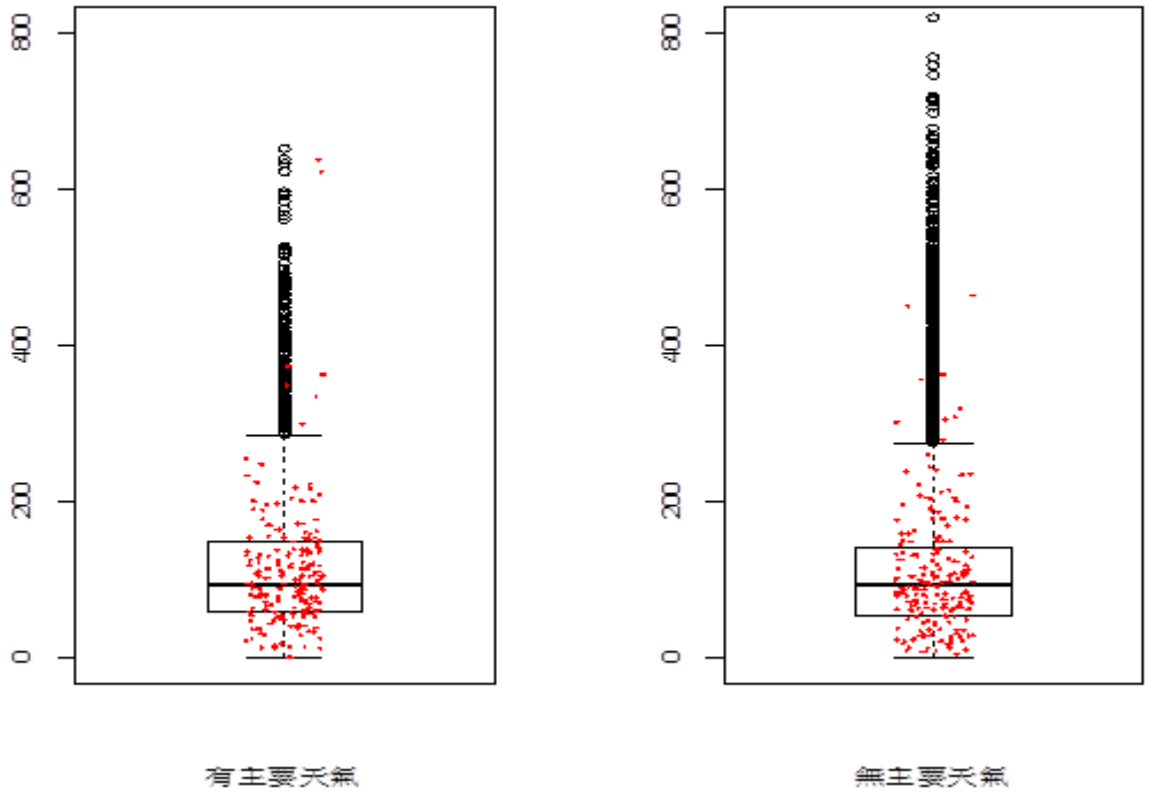


圖 4-2 有無主要天氣狀況之盒鬚圖

◎從圖 4-2 得知，兩者之盒鬚圖沒有太大之差異。但仔細看可發現，有主要天氣情況較有主要天氣壯況高出一些。

	有	無
一般天氣型態	105	112
主要天氣型態	111	109

表 4-3 一般天氣型態有無 及 主要天氣型態有無之平均銷售量比較

◎我們從表 4-3 可以發現，當我們只把天氣區分為有無一般天氣情況時，有天氣狀況時銷售量比無天氣狀況時銷售量少，然而，當我們把區分天氣的條件設定為有無”**主要**”天氣情形時，有主要天氣狀況時銷售量反而會比無天氣狀況時來的高。

二、Anderson-Darling 檢定法

H_0 ：樣本資料呈常態

H_a ：樣本資料非常態

$\alpha = 0.05$

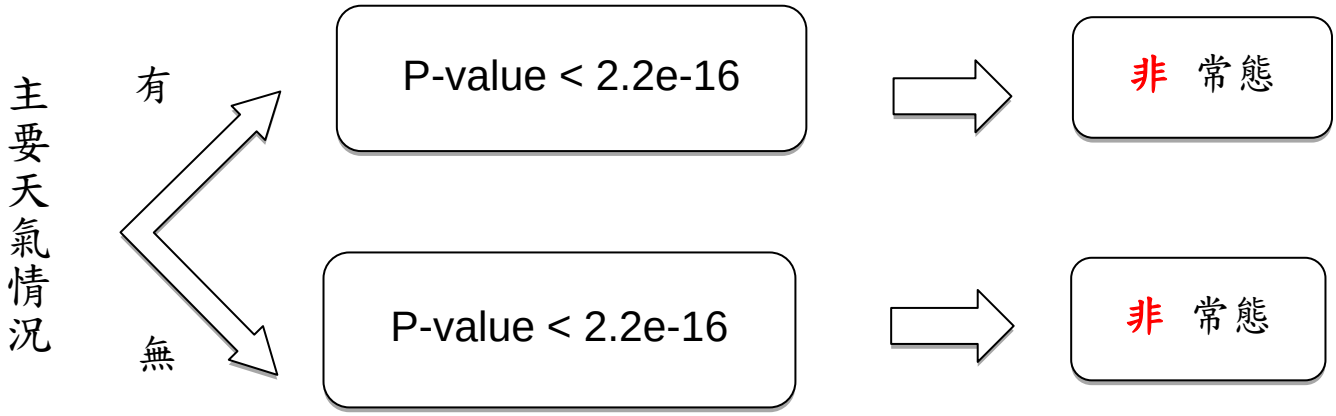


圖 4-4 主要天氣型態有無之常態檢定

◎ 兩者皆非常態，故做無母數檢定 Wilcoxon rank sum test

三、Wilcoxon rank sum test

H_0 ：主要天氣情況不會影響平均銷售量

（$\mu_{有主要天氣} = \mu_{無主要天氣}$）

H_a ：主要天氣情況會影響平均銷售量

（$\mu_{有主要天氣} \neq \mu_{無主要天氣}$）

$\alpha = 0.05$

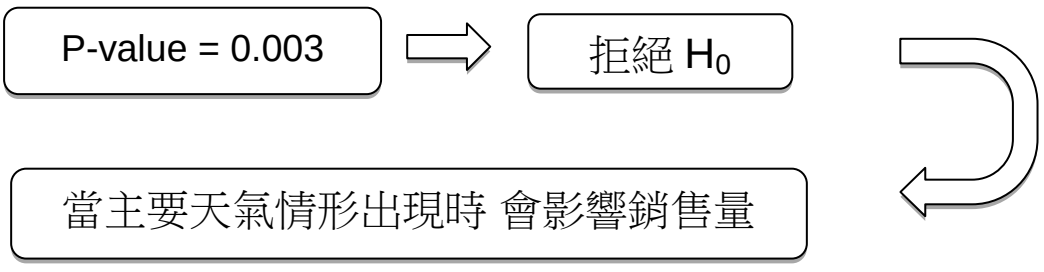


圖 4-5 主要天氣情況是否影響銷售量結論

由上面的檢定方法可得知，兩者之間存在差異，另外我們以累積相對次數圖來觀察（下圖 4-6 紅線為有主要天氣情形、藍線則無無天氣情形），紅線整體上普遍落於藍線左方，故可得知，當有主要情形發生時，銷售量會提高，與我們檢定相符，這代表當有主要天氣即將出現時，人們可能會傾向先去大採購一番，以備不時之需，未雨綢繆，補足家內之必需品。

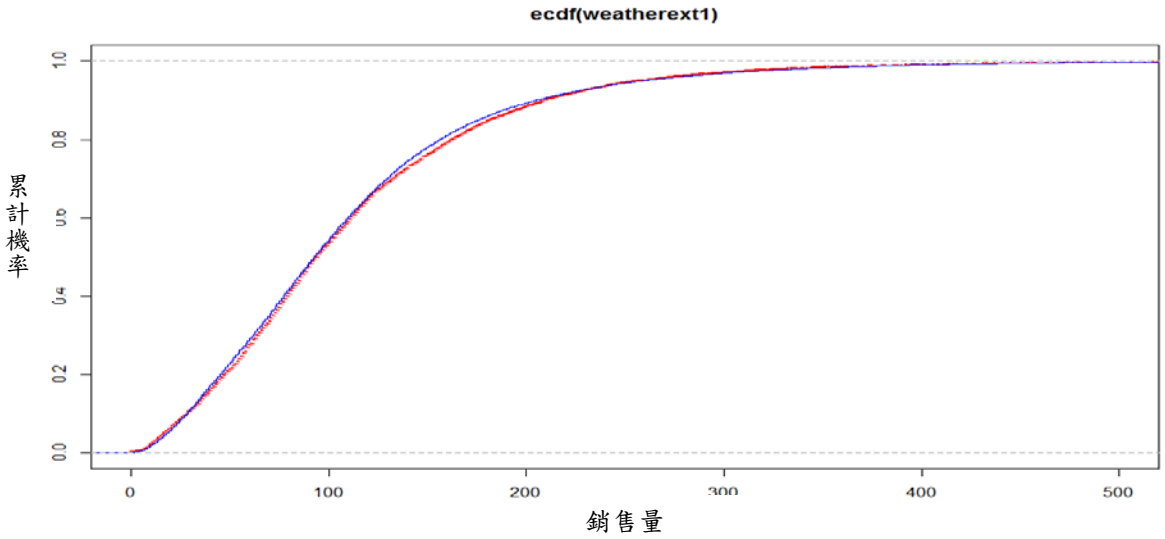


圖 4-6　有無主要天氣情形之 cdf 圖

四、各主要天氣型態之比較

以下我們把各個主要天氣型態都分別抓出來看，以沙塵暴為例，我們選取的準則為把當日有出現沙塵暴歸為一類，沒有出現沙塵暴則皆歸為另一類。其他五種皆是如此。

下表 4-7 為六個主要天氣之敘述統計。從各個情況來看，各個有無主要天氣情況之平均數以及四分位距，除了暴風雨之外，其他五種的差距皆頗大。而冰雹的部分，可以看到樣本數只有 3 筆，很明顯的資訊較不足，在後面檢定的部分會做完整解釋。

	n	min	1st Qu	Median	Mean	3rd Qu	Max	sd	trimmed	se
有沙塵暴	64	24	93	128	143	177	453	77	134	9.63
無沙塵暴	41434	0	54	93	110	142	903	79.88	99	0.39
有濃霧	1446	0	41	84	102	142	584	80.85	91	2.13
無濃霧	40152	0	55	93	110	142	903	79.84	99	0.4
有冰雹	3	76	78	79	81	83	87	5.69	81	3.28
無冰雹	41595	0	54	93	109	142	903	79.89	99	0.39
有下雪	1899	0	65	112	126	172	848	88.06	117	2.02
無下雪	39699	0	54	92	109	141	903	79.38	98	0.4
有暴風雪	26	27	90	113	128	140	323	62.54	121	12.26
無暴風雪	41572	0	54	93	109	142	903	79.89	99	0.39
有暴風雨	3990	1	59	91	107	136	903	74.69	97	1.18
無暴風雨	37608	0	54	93	110	143	873	80.42	99	0.41

表 4-7 各個主要天氣情況之敘述統計

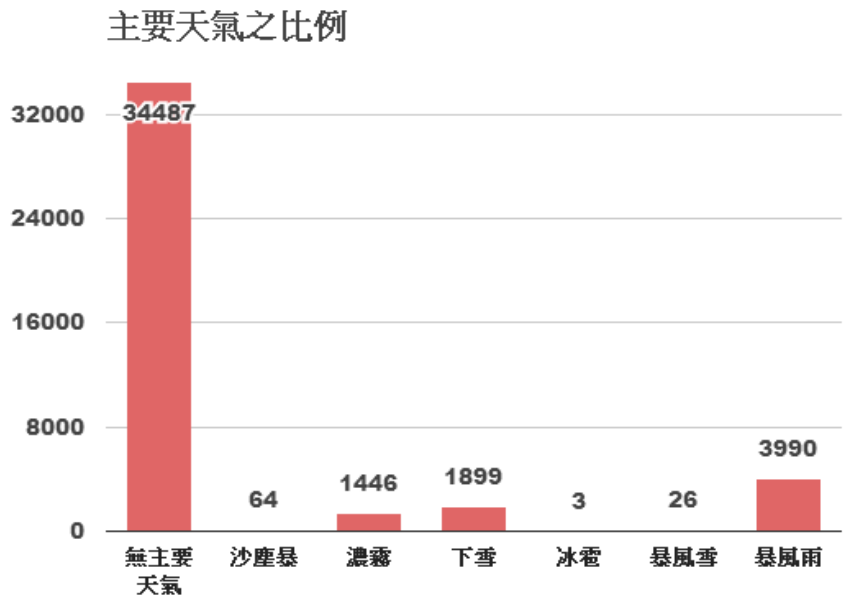
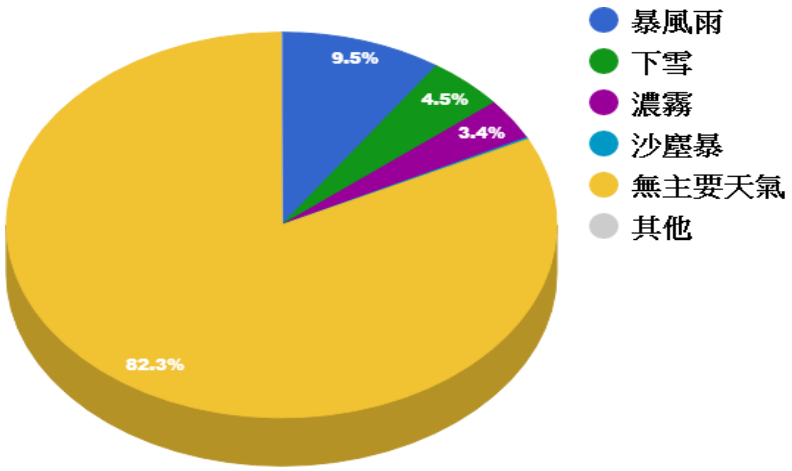


圖 4-8(上) 圖 4-9(下) 各個主要天氣情況之比例



◎由表 4-7 中可清楚發現，有冰雹的天數僅有 3 天，故我們之後將用 shapiro test 來檢其是否為常態。而其他五種主要天氣及無主要天氣，還是使用 ad test 來檢測。

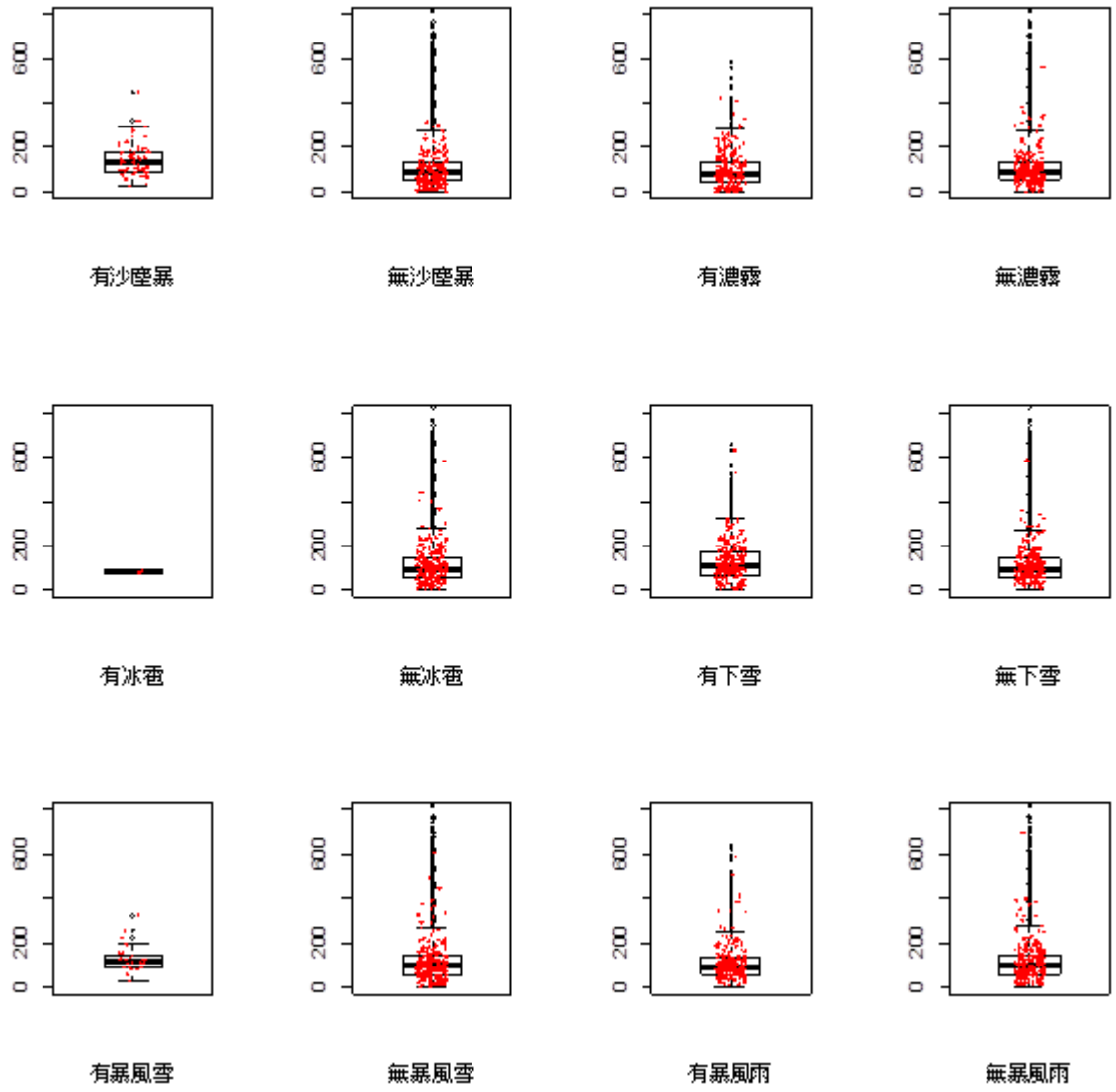


圖 4-10 各個主要天氣情況之盒鬚圖

圖 4-10 為我們將各個主要天氣一個一個來看的狀況下之盒鬚圖，第一三行為有出現該天氣狀況時的銷售量盒鬚圖，反之則為二四行。而相關的比較我們會在本章節最後面總結。

4.1 檢驗各天氣型態是否符合常態

	檢定方法	p-value	是否符合常態
有沙塵暴	Anderson-Darling	0.0005	否
無沙塵暴	Anderson-Darling	< 2.2e-16	否
有濃霧	Anderson-Darling	< 2.2e-16	否
無濃霧	Anderson-Darling	< 2.2e-16	否
有冰雹	Shapiro test	0.5098	是
無冰雹	Anderson-Darling	< 2.2e-16	否
有下雪	Anderson-Darling	< 2.2e-16	否
無下雪	Anderson-Darling	< 2.2e-16	否
有暴風雪	Anderson-Darling	0.00297	否
無暴風雪	Anderson-Darling	< 2.2e-16	否
有暴風雨	Anderson-Darling	< 2.2e-16	否
無暴風雨	Anderson-Darling	< 2.2e-16	否

表 4-11 各個主要天氣情況之常態檢定

◎由表 4-1 我們可以看出每個天氣型態除了有冰雹之外皆符合常態，又因為我們是要對於有無某一天氣型態下做檢定，所以當有無天氣型態其中有一方呈現**非**常態，我們即須將其整個視為非常態的資料來處理，亦即對於有無冰雹是否影響平均銷售量，我們也需使用 Wilcoxon rank sum test。

4.2 Wilcoxon rank sum test

H_0 ：此一天氣情況不會影響平均銷售量

（$\mu_{有主要天氣} = \mu_{有主要天氣}$）

H_a ：此一天氣情況會影響平均銷售量

（$\mu_{有主要天氣} \neq \mu_{有主要天氣}$）

$\alpha = 0.05$

	檢定方法	P-value	此一天氣情況是否影響銷售量
濃霧	Wilcoxon rank sum	2.441e-08	是
冰雹	Wilcoxon rank sum	0.6282	否
沙塵暴	Wilcoxon rank sum	3.213e-05	是
下雪	Wilcoxon rank sum	< 2.2e-16	是
暴風雪	Wilcoxon rank sum	0.0386	是
暴風雨	Wilcoxon rank sum	0.7792	否

表 4-12 各個主要天氣情況是否影響銷售量

從上表可得知，濃霧、沙塵暴、下雪、暴風雪對於銷售量的影響是顯著的，而冰雹及暴風雨則是不顯著。冰雹不顯著原因主要是樣本數不足使得 P-value 過大。而暴風雨的原因，從前面敘述統計可看出，其與無暴風雨天氣情況之銷售量沒有什麼差異。

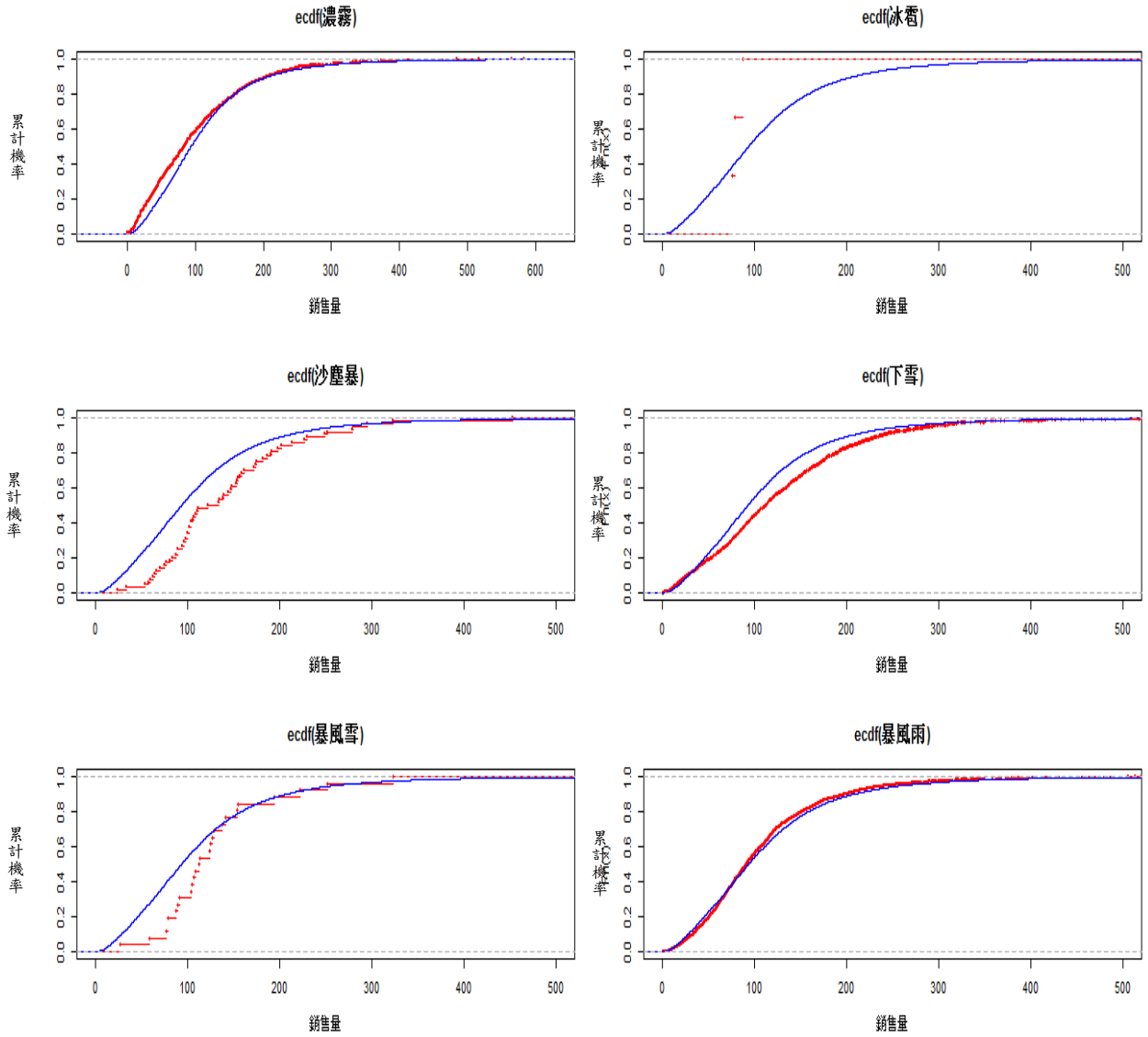


表 4-13 各個主要天氣情況之 CDF（紅線為有此天氣情形，藍線則無）

4.3 各天氣情形比較之結論

濃霧：不需特別購買商品，故有濃霧時會稍微降低人們購買的慾望。

冰雹：這三天中的銷售量都偏低，都不超過 100 單位，但由於兩樣本數量差異太大，故此檢定之結果可能無法完全表達真實狀況。

沙塵暴：對於交通影響較小，但可能需要購買大量口罩，導致銷售量提高。

下雪、暴風雪：銷售量皆較高，可能是需要出門買些保暖衣物、鏟雪用具或是暖爐等用具。

暴風雨：不需特別購買商品，且在美國大多數家庭都有汽車，故對銷售量影響不大。

伍、主要天氣之負二項迴歸

在此我們亦將資料先做整理，在上述的方法中，我們感興趣的主要天氣分別為**暴風雪**、**下雪** 、**沙塵暴**、**大濃霧** 、**冰雹**及**暴風雨**六種，我們將這六種天氣單獨列為六項變數視為類別變數，當其發生時，我們記為 1，若沒有則記為 0，再來進行多方面的迴歸分析

這裡我們選用的是負二項迴歸(Negative Binomial Regression)，負二項迴歸的形式如下：

$$y_i \sim NB(u_i, k)$$

運用 log link,

$$\log(u_i) = x_i^T \beta$$

擁有 offset 項

$$\log(u_i) = \alpha + x_i^T \beta$$

這裡我們選用的是負二項迴歸(Negative Binomial Regression)，值得注意的部分是，我們放進去模型的不止上述六種主要天氣，我們把平均氣溫以及平均風速一併放入模型進行負二項迴歸。

再來我們觀察資料，大致歸類出以下三種現象:

1.暴風雨會伴隨著較強的平均風速

2.暴風雪會伴隨著低溫及較強的平均風速

3.下雪會與低溫同時發生

為了因應上述三種現象，我們會在模型中多放入以下三點變數:

1. 暴風雨天氣與平均風速的相互作用

2. 暴風雪天氣與平均氣溫與平均風速之間的相互作用

3. 下雪天氣與平均氣溫的相互作用

為了要驗證我們的模型配適的好壞與否，我們做 cross validation，將隨機將資料切成 10 等份，並拿其中的九份作模型配適，一份拿來套入模型，檢驗配適程度。並運用 MSE 進行推估及大概的誤差

$$\text{MSE} = \frac{(\text{預測值} - \text{實際值})}{\text{資料數}}$$

以下是我們配適的模型中顯著的變數項及其係數，可以從中得知變數項之重要性：

項目	係數	Std. error	z-value	p-value
常數項	4.818	9.059 e-01	53.1791	< 2e-16
主要天氣-濃霧	-9.805e-02	1.953 e-02	-5.02	2.46e-07
主要天氣-沙塵暴	2.145e-01	1.031 e-01	2.07893	0.018812
主要天氣-下雪	3.828e-01	1.537 e-02	2.48995	0.006388
主要天氣-暴風雨	1.238e-01	3.703 e-02	3.34259	0.000415
平均氣溫	-1.052e-01	3.245 e-02	-3.24175	0.000594
平均風速與暴風雨交互作用	-5.671e-04	1.6965 e-04	-3.34259	0.000415

表 5-1 負二項模型迴歸係數

我們再將剩下 1/10 的資料帶入，並求出預測值，並與實際值求出 MSE 求出 MSE= 6342.453，大約有 80 個銷售量的預測誤差。

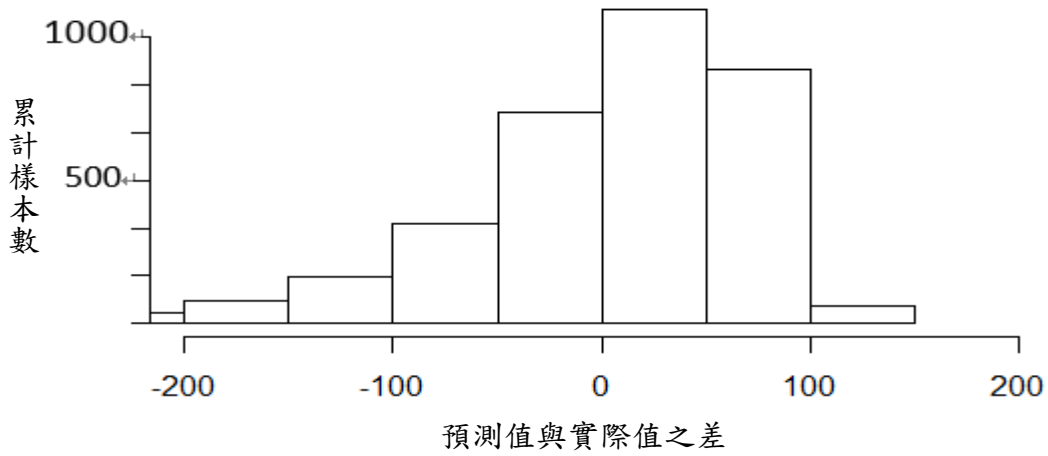


表 5-2 預測值與實際值之差異累積次數圖

上圖為橫軸預測值與實際值之間的差距，縱軸是累計樣本，其預測值與實際值之差值大約還是落在 0~50 這個區間。

◎ 綜合上述內容中可以看出，濃霧及平均氣溫升高會讓店家減少銷售量，暴風雨與沙塵暴及下雪的發生，可能會增加銷售量;由上述兩點稍作推測，我們可以看出其濃霧的發生可能會造成交通上的阻礙，而消費者在這情況下則無法出門，造成消費量的降低。主要天氣中，沙塵暴及下雪及暴風雨可能是該商品對於其有相關關係，而導致消費者在這三種天氣來臨時，增加消費，而導致店家的消費量增加，我們推測其可能為一些災後復原家園的用品或是工具，也有可能是一些保暖用品，用來因應這三種主要天氣的來臨。

陸、其他變數之決策樹

此章節我們使用決策樹方法找出能明顯區分銷售量的變數，並藉由觀測節點找出變數介於哪個臨界值時能將銷售量區分出來。

在裡頭我們將使用三種方法－C5.0、Rpart、RandomForest。以此三種方法分別作出決策樹，找出具有影響力的變數，並對各個決策樹檢測分類能力及預測能力之正確率。

在章節最後我們將比較三種方法的優缺點，並選出一個最適合 Walmart 資料型態的方法。

一、C5.0

最一開始，我們先隨機抽取 80%的資料來培育決策樹，留 20%的資料來觀測分類後的準確度。

首先將每筆總銷售量用四分位距分為四個等分：

Q1 代表小於四分位距中的 Q1

Q2 代表介於 Q1～Q2

Q3 表示介於 Q2～Q3，

Q4 代表大於 Q3

在最一開始我們使用變數，每日平均溫度(tavg)、平均風速(avgspeed)、降雨量(preciptotal)、降雪量(snowfall)、是否有主要天氣狀況(codesumext)，並且在不修改內設參數的情況下先培育出第一棵決策樹。

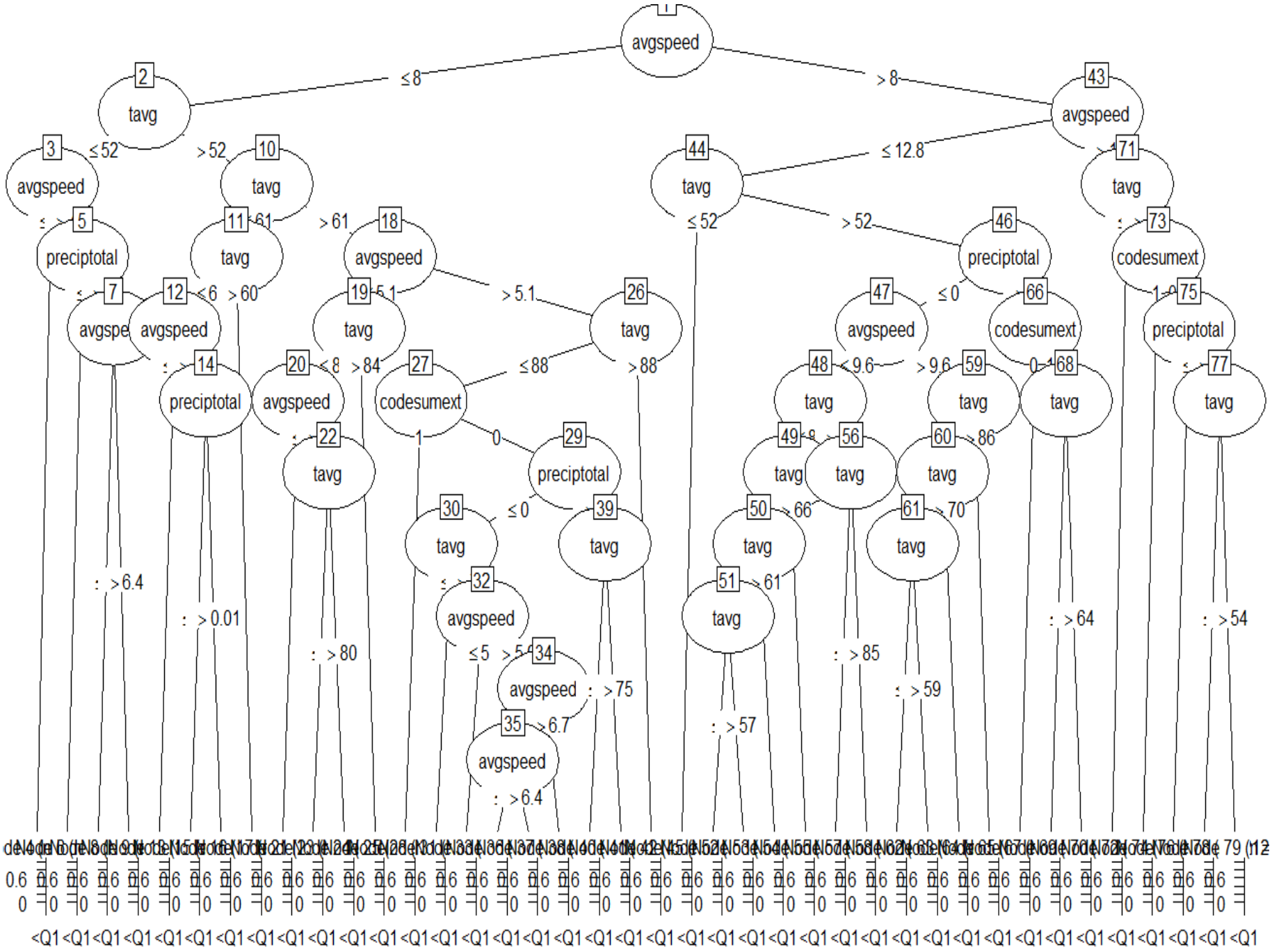

avgspeed
≤8
>8
2
tavg
43
avgspeed
3
≤52
>52
10
tavg
44
tavg
≤12.8
71
tavg
avgspeed
5
preciptotal
11
tavg
≤52
>52
46
preciptotal
73
codesumext
>61
18
avgspeed
47
≤0
66
codesumext
75
preciptotal
7
12
<6
>60
19
tavg
>5.1
26
tavg
avgspeed
avgspeed
48
<9.6
>9.6
59
0
68
77
tavg
tavg
tavg
tavg
14
20
<8
>84
27
≤88
>88
49
56
60
>86
preciptotal
avgspeed
codesumext
tavg
tavg
tavg
22
1
0
29
50
>66
61
>70
tavg
preciptotal
tavg
tavg
30
≤0
39
51
>61
tavg
tavg
32
avgspeed
≤5
>5
34
avgspeed
35
>6.7
avgspeed
: >6.4
: >75
: >57
: >85
: >59
: >64
: >54
: >80
: >0.01
: >6.4

圖 6-1 初始分類樹

從圖(6-1)我們可以看出分類樹的枝葉太繁雜，節點的分類過於細瑣，所以我們希望藉由調整每次分割最小所需分配到的數量，達到簡化分類樹的效果。

在簡化的過程中，我們會觀測分類樹的分類及預測能力，如果與修剪前比較並無顯著的差異，我們將傾向使用修剪過後的分類樹。

首先我們對未修剪的決策樹做分類及預測能力的檢測：

使用資料	正確率
分類能力的檢測（運用建立決策樹 80%資料）	34.83879%
預測能力(剩餘 20%資料)	33.29727%

表 6-2 分類及預測能力檢測

接著我們用 summary 得到各變數的貢獻程度：

Attribute usage: 100.00% tavg 100.00% avgspeed 19.82% preciptotal	從資料中可看出平均氣溫(tavg)和風速(avgspeed)是分類的重要依據，而降雨量的貢獻度相較起來不明顯。

表 6-3 變數貢獻度

而第二棵樹的培育，在經過許多嘗試後，我們決定將最小所需分配到的數量調為1500，可以得到一棵修剪後的決策樹，如圖(6-2)。

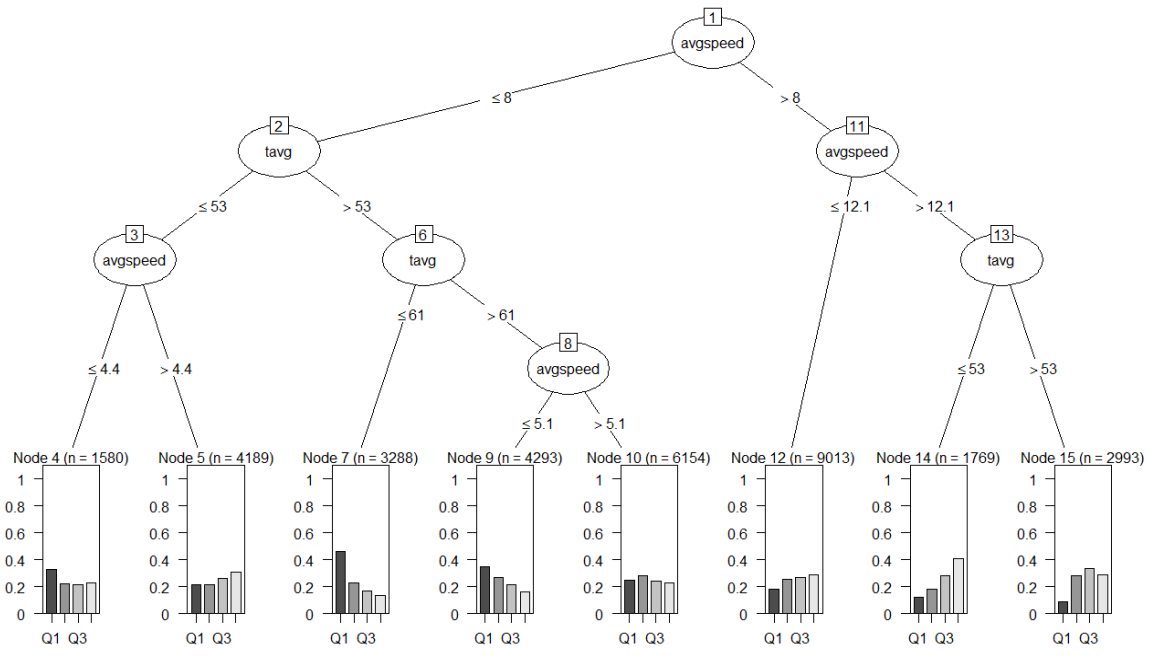


圖 6-4 C5.0 修剪後決策樹

使用資料	正確率
分類能力的檢測（運用建立決策樹 80%資料）	33.50161%
預測能力(剩餘 20%資料)	32.84049%

表 6-5 分類及預測能力檢測

對修剪後決策樹做變數的檢測：

接著我們用 summary 得到各變數的貢獻程度：

Attribute usage: 100% avgspeed 79.55% tavg 29.58% preciptotal	從資料看出平均氣溫(tavg)、風速(avgspeed)依舊是分類的重要依據，而降雨量(preciptotal)也占了一定的比例。

表 6-6 變數貢獻度

另外我們比較修剪前後的決策樹，可以看出分類及預測能力在兩者之間並無明顯的差異，因此我們較建議採用修剪過後的分類樹。

二、Rpart

接著，我們安裝 package”rpart”來培育另一棵決策樹。在 Rpart 的部份，我們跟 C5.0 一樣，抽取 80%培育分類樹，20%觀測準確度。

有一點差異，C5.0 的使用，其觀測值必須是**類別資料**。而 Rpart 沒有此限制，因此我們在 Rpart 的第一部分先試著用每日的銷售量培育分類樹。

由於 Rpart 的內設 cp(complexity parameter)值為 0.1，會使樹枝太過於簡單，導致不能準確分類資料。故我們將 cp 值設為 0.001。

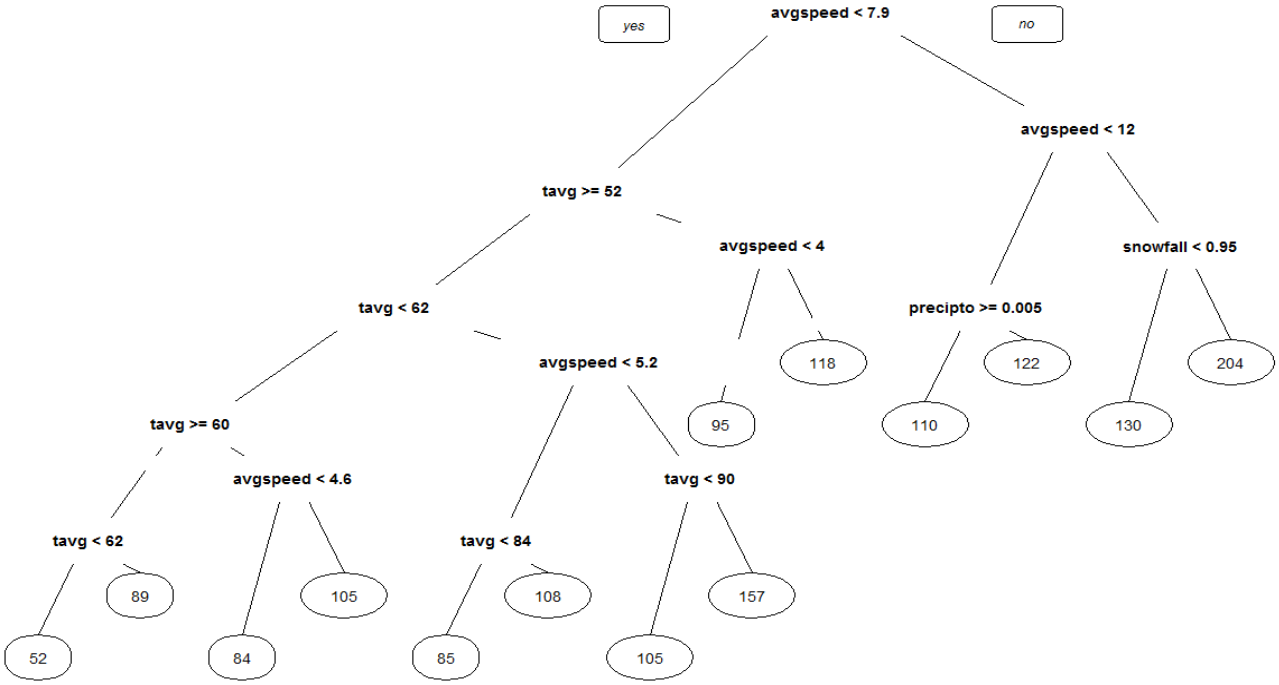


圖 6-7 Rpart 決策樹(cp=0.001)

上圖的樹枝顯得有點繁雜，在這我們將 complexity parameter(cp)設為 0.0025，試著修剪樹枝。

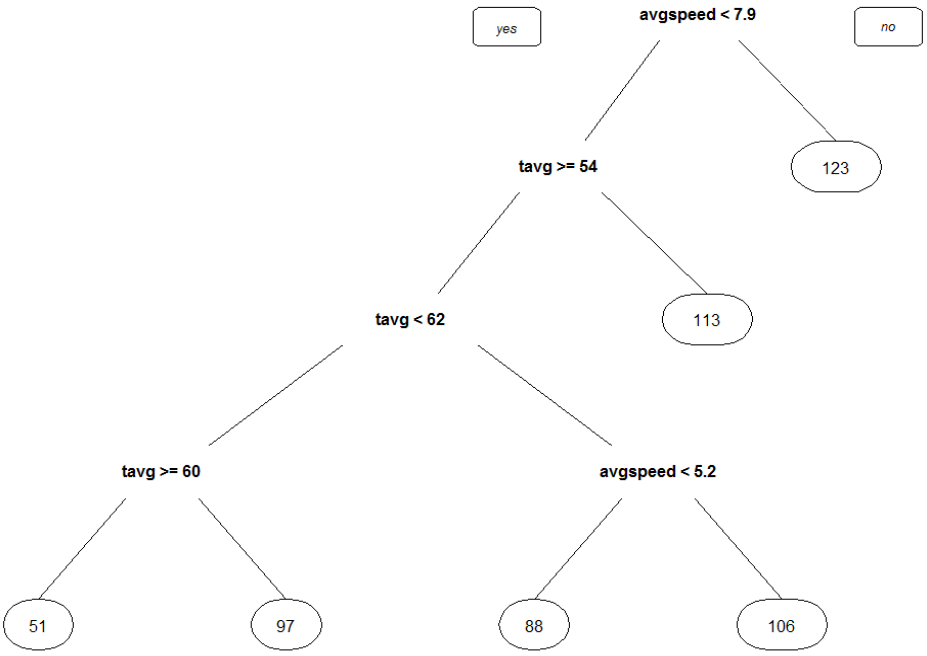


圖 6-8 Rpart 決策樹(cp=0.0025)

我們決定用風速(avgspeed)7.9 為分界，對銷售量做 summary，觀察兩者差異。

	min	1st Qu	median	mean	3st Qu	max
風速<7.9	0	69	105	122.9	157	873
風速>7.9	0	45	82	99.03	131	903

表 6-9 以風速 7.9 為界之敘述統計

由上圖可知，可以看出其實兩者的平均銷售量有明顯的差異。

接著再用風速(avgspeed)小於 7.9 的節點以平均氣溫(tavg)62°F為分界，對銷售量做 summary，觀察兩者差異。

風速<7.9	min	1st Qu	median	mean	3st Qu	max
均溫<62°F	0	54	98	112.4	150	757
均溫>62°F	0	43	76	93.4	122	903

表 6-10 以均溫 62°F為界之敘述統計(given 風速>7.9)

◎由上圖可知，可以看出兩者的平均銷售量依然有明顯的差異。將每個節點的分類以敘述統計量的資料觀察，發現皆有顯著的差異，我們可以得知 Rpart 的決策樹效果不錯。

接著我們採用如同 C5.0 的方式，用四分位距分將銷售量為四個等分，採用相同的變數，培育出第一棵分類數。

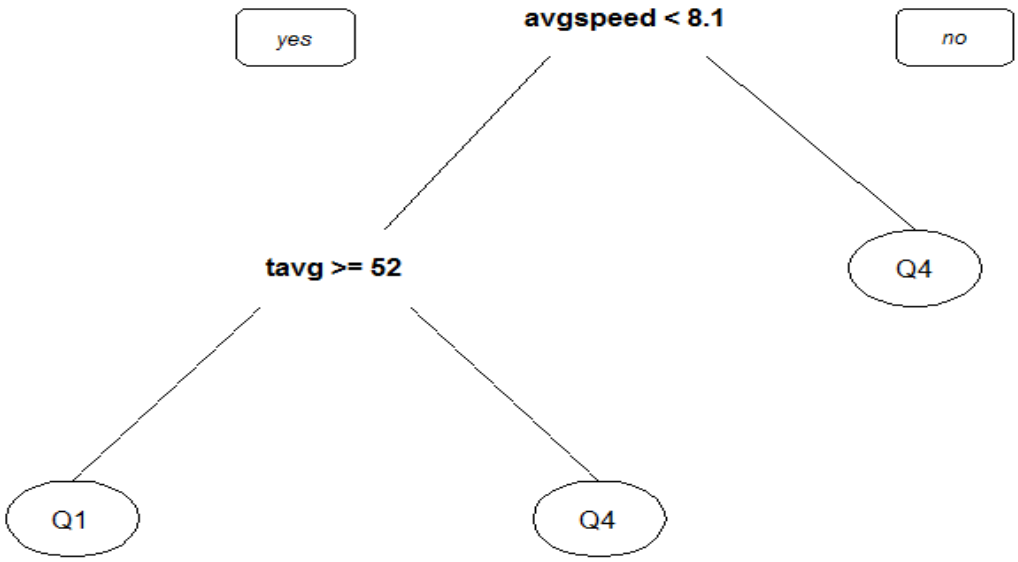


圖 6-11 Rpart 決策樹(以四分位距區分)

在 Rpart 裡面，與 C5.0 不同的是，用 summary 不會有各變數的貢獻度。故我們使用”variable importance”來得知。

風速	平均溫度	降雪量	降雨量
274.801	87.401	12.425	1.196

表 6-12 變數重要程度

◎從上面資料可看出風速(avgspeed)、平均溫度(tavg)皆是重要的變數。

接下來，一樣對決策樹做變數的檢測：

分類能力的檢測(80%資料)	預測能力(20%資料)
CORRECTION RATIO(test)=27.90949%	CORRECTION RATIO(test)=27.14269%

表 6-13 分類及預測能力檢測

◎從上圖及資料是使用內設 cp 的情況，做出來的決策樹顯得過於簡單，導致準確性不高，因此我們將 complexity parameter(cp)設為 0.005。

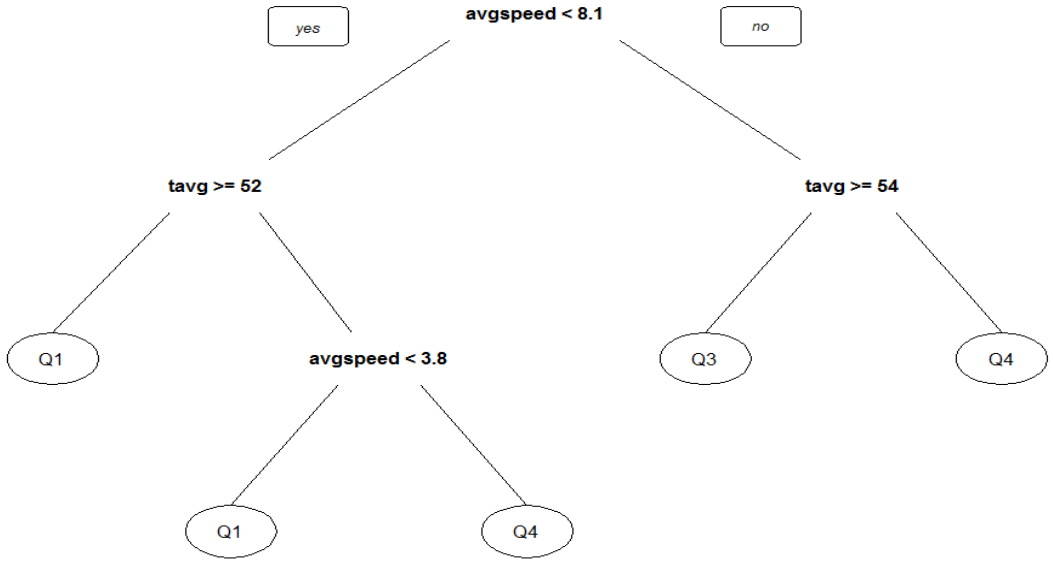


圖 6-14 Rpart 決策樹(以四分位距區分、cp=0.005)

◎我們從圖可看出將 cp 改為 0.005 的情況下決策樹會變的有許多的枝葉，但不算是太繁雜。

一樣對決策樹做變數的檢測：

分類能力的檢測(80%資料)	預測能力(20%資料)
CORRECTION RATIO(test)=33.85018%	CORRECTION RATIO(test)=33.30929%

表 6-15 分類及預測能力檢測

風速	平均溫度	降雪量	降雨量	主要天氣
372.106	245.621	18.873	16.739	11.032

表 6-16 變數重要程度

◎而從上表可看出降雪量(snowfall)、降雨量(preciptotal)、是否有主要天氣(codesumext)對於分類的貢獻度不大，於是我們將他們拿掉再做一次分類樹看看是否預測能力會受到影響。

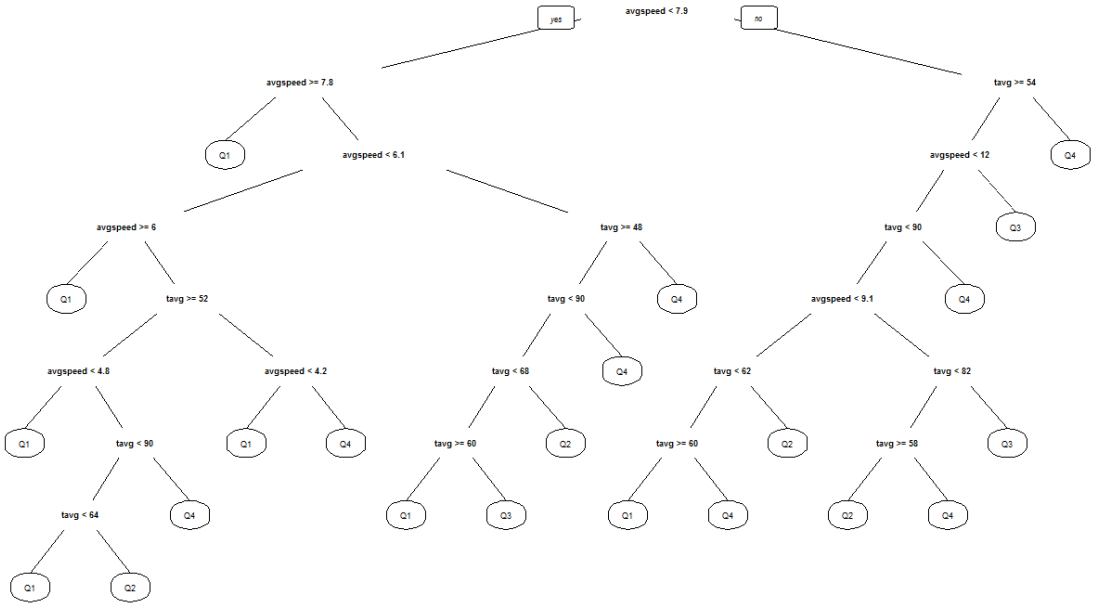


圖 6-17 Rpart 決策樹(以四分位距區分、cp=0.005、只留平均溫度及風速)

一樣對決策樹做變數的檢測：

分類能力的檢測(80%資料)	預測能力(20%資料)
CORRECTION RATIO(test)=33.9854%	CORRECTION RATIO(test)=33.9854%

表 6-18 分類及預測能力檢測

平均溫度	風速
163.879	611.778

表 6-19 變數重要程度

◎從上圖和資料可看出預測能力跟拿掉變數前相近，但分類樹卻變得複雜。我們也嘗試只拿掉降雪量或降雨量單一變數，可是預測準確度皆不如完整五個變數的決策樹，因此我們不建議將變數拿掉。

三、隨機森林(RandomForest)

最後一個方法為 RandomForest，我們使用隨機森林，以內設參數 CP 的情況下，用五個變數，在每個節點隨機取出三個變數來決定此節點的變數為何，而培育一棵決

策樹。重複上述的過程，培育出 500 棵樹。在這裡我們是採取全部銷售量的四分位距來觀察銷售量變化。

做完之後我們用這 500 棵樹，plot 出 Q1 到 Q4 預測的錯誤率。

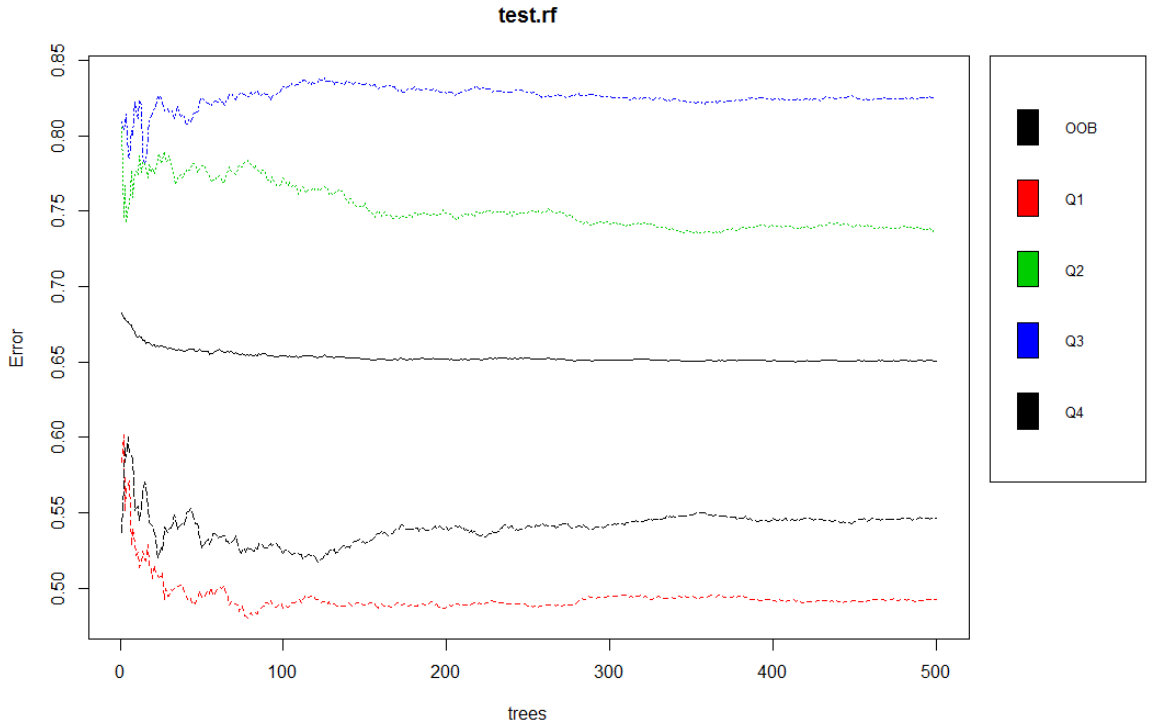


圖 6-20 RandomForest Q1~Q4 之錯誤率

◎由上圖可以看出明顯在 Q1 及 Q4 的分類較為準確，在 Q3 及 Q2 的分類上錯誤率則相對高出許多。

接著，我們使用”MeanDecreaseGini”來觀察各個變數的重要性

	MeanDecreaseGini
Tavg	1428.36
Snowfall	94.82
Preciptotal	876.46
Avgspeed	1751.69
codesumext	139.27

表 6-21 各變數重要性

◎從 RandomForest 中我們可以得知平均氣溫和風速仍然是分類的最主要依據，而降雨量為次要的依據。與前述兩個方法做出的結果是大同小異。

最後，依然對這些決策樹做變數的檢測：

分類能力的檢測(80%資料)	預測能力(20%資料)
CORRECTION RATIO(test)=45.07647%	CORRECTION RATIO(test)=44.66454%

表 6-22 分類及預測能力檢測

◎由上表，我們可以明顯的看出正確率比前面用 C5.0 和 Rpart 高出許多。因此我們認為 RandomForest 與 C5.0 和 Rpart 相比下是較好的選擇。

四、分類樹三種方法之比較:

	特性	預測準確率	時間
C5.0	只能用於類別觀測值	32.84%	1.46sec
Rpart	連續類別觀測值皆可用	33.31%	1.39sec
RandomForest	運算時間較久	44.66%	21.19sec

表 6-23 三方法之比較

◎從以上三種分類樹的比較，C5.0 只能用於類別的觀測值，在連續型觀測值的分類上必須先分為數個類別，因此在分類樹的培育上有一定的限制，且不管在運算時間和預測準確率接 Rpart 沒有太大的差異。

而 RandomForest 雖然在運算時間上明顯比 C5.0 和 Rpart 久，可是在預測的準確率上卻也有明顯的上升，因此我們認為在這三種方法當中 RandomForest 是最適合此筆資料的分類方式。

柒、假日與平日之差別

這一部分我們也想探討平日與假日對銷售量有無影響，雖然這一部分與我們的主題”天氣”較無關係，且直觀上而言，假日銷售量一定會比平日來得高，可是我們還是以簡單保守的統計方法做此次檢定。首先，我們一樣把資料分成兩部分，假日與平日，分別檢驗其是否成常態，因為兩類都超過 5000 筆資料，故我們一樣選用 nortest package 中的 ad.test 來判斷。

一. 敘述統計

◎從下表 7-1，我們可以得知當假日來臨時，不論是四分位數、平均數亦或是截尾平均數，平均銷售量都高出平日許多，此與我們上面推論相符，

	n	min	1st Qu	Median	Mean	3rd Qu	Max	sd	trimmed	se
平日	29715	0	50	86	100	131	819	72.96	91	0.42
假日	11883	0	69	115	132	171	903	91.25	120	0.84

表 7-1 平日與假日之敘述統計

二、Anderson-Darling 檢定法

H_0 ： 樣本資料呈常態

H_a ： 樣本資料非常態

$\alpha = 0.05$

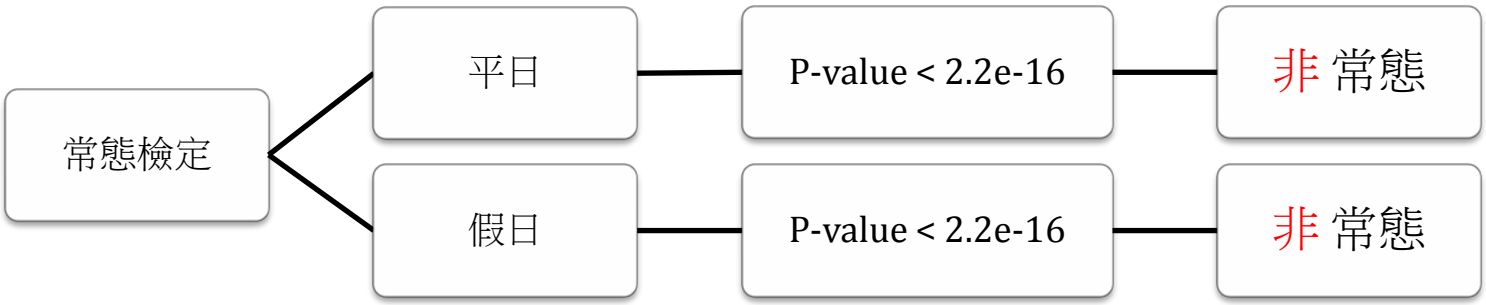


表 7-2 平日與假日之常態檢定

◎ 不管平日或假日，皆呈非常態，故我們對其做無母數檢定 Wilcoxon rank sum test

三、Wilcoxon rank sum test

H_0 ： 平日與假日不會影響銷售量

$(\mu_{平日} = \mu_{假日})$

H_a ： 平日與假日會影響平均銷售量

$(\mu_{平日} \neq \mu_{假日})$

$\alpha = 0.05$

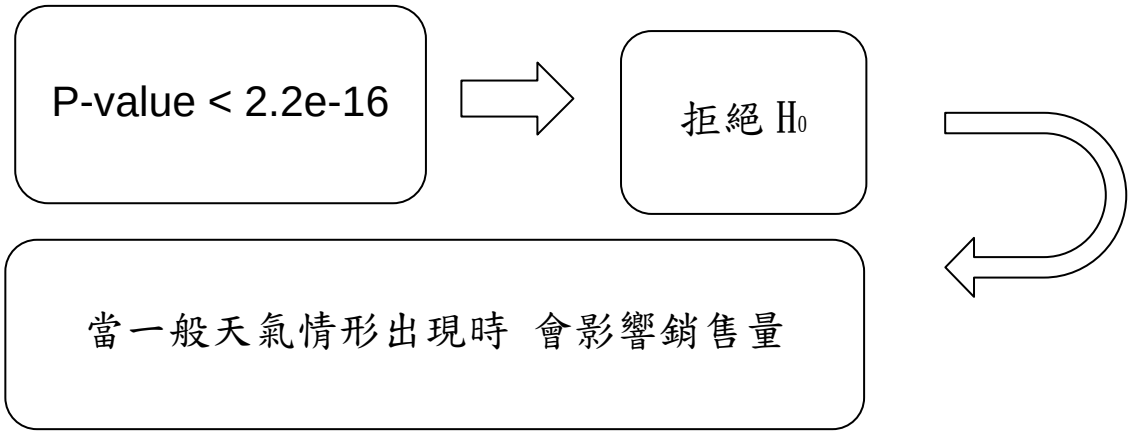


圖 7-3 平日與假日是否影響銷售量

◎ 由上面的檢定方法可得知，兩者之間存在差異

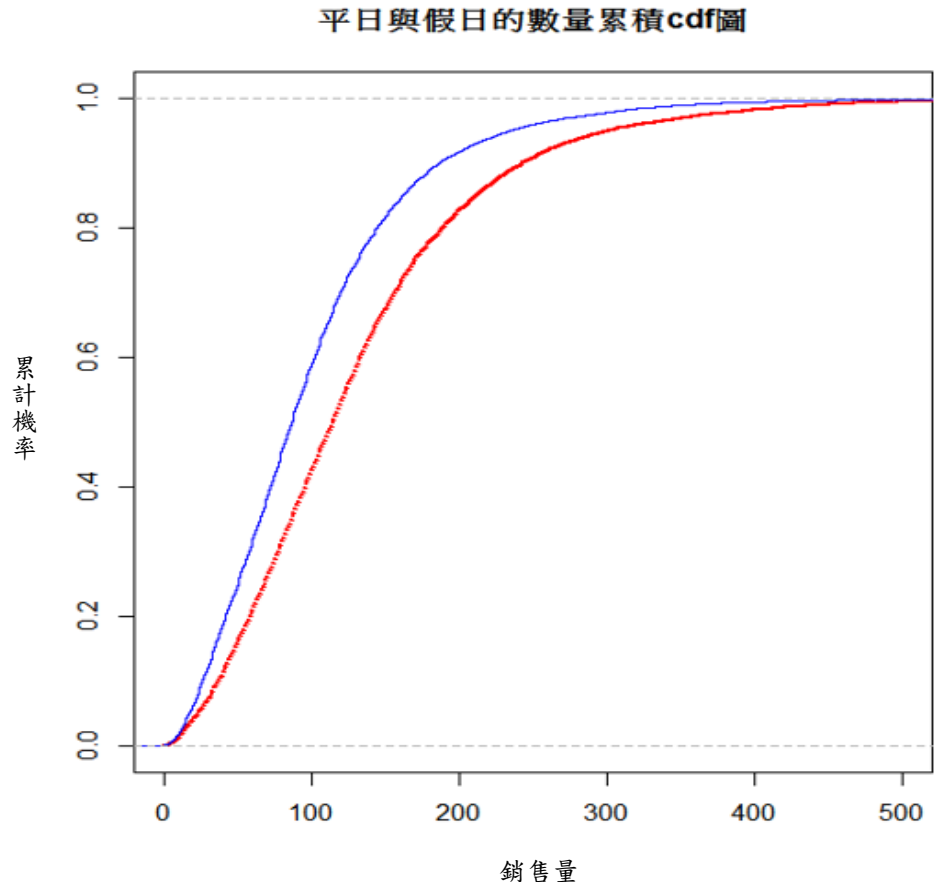


圖 7-4 平日與假日之 cdf

◎此外我們亦可由累積相對次數圖來佐證上述之結果（圖 7-4 紅線為有假日、藍線則平日），紅線整體上普遍落於藍線右方。

我們可從此得知假日銷售量會提高許多，與我們檢定以及假設皆相符。因而得出結論：大多數人平日都要上班，較無空閒時間能去大賣場。到了假日，成年人與小孩有空閒的比例比較高，多數家庭便會較傾向於假日在去大賣場購物。

捌、結論與建議

在前面章節裡，我們再重新彙整一次重點結論：

第參章－一般天氣有無：

先使用了常態性檢定(Anderson Darling test)確認其是否為常態。確認後，皆為非常態，故使用無母數檢定(Wilcoxon rank sum test)，檢測"有天氣情況"及"完全沒有天氣情況"之間的銷售量有無差異。而檢定後得知兩者銷售量是有顯著差異的，且"完全沒有天氣情況"之銷售量是高於"有天氣情況"。故推論出－「當有天氣情況發生時，人們大多會選擇待在家中，而不出門購物」。

第肆章－主要天氣有無：

在本章分為兩個重點，第一，有主要天氣情況的有無其銷售量之差異。第二，針對個別的主要天氣情況與沒有主要天氣情況發生時的銷售量差異。

在此使用的方法亦是先使用了常態性檢定(Anderson Darling test)確認其是否為常態。確認後，皆為非常態，故使用無母數檢定(Wilcoxon rank sum test)。

第一部分，在有無主要天氣情況的銷售量得出是有顯著的差異，而有主要天氣情況的銷售量比無主要天氣情況來的高。故推論出－「當主要天氣情況要出現時，人們反而會傾向去採購，未雨綢繆而以備不時之需，補足必需品等。」

第二部分，在針對各主要天氣情況與無主要天氣情況的銷售量差異上，濃霧、沙塵暴、下雪、暴風雪是顯著影響銷售量的，而冰雹及暴風雨則無。故推論出－「

濃霧→不需特別購買商品防範，且對於交通有影響不適合出門，故有濃霧時會稍微降低人們購買的慾望。

冰雹→由於只有三筆資料，而這三天中的銷售量都偏低，都不超過100單位，且有冰雹及無主要天氣情況之兩樣本數量差異太大，故此檢定之結果可能無法完全表達真實狀況。

沙塵暴→雖可能對於交通方面有影響，但可能需要購買大量口罩及阻擋沙塵之器具，導致銷售量提高。

下雪、暴風雪→銷售量皆較高，可能是需要出門買些保暖衣物、鏟雪用具或是暖爐等用具。

暴風雨→不需特別購買商品，且在美國大多數家庭都有汽車，故對銷售量影響不大」

第伍章－主要天氣之負二項迴歸：

在本章要配適負二項迴歸模型，而丟入的變數包括：暴風雪、下雪 、沙塵暴、大濃霧 、冰雹及暴風雨，外加交互作用：暴風雨天氣與平均風速、暴風雪天氣與平均氣溫與平均風速、下雪天氣與平均氣溫的交互作用。

接著驗證模型配適的好壞，使用 cross validation，將資料切成十份，九份作模型配適，一份套入模型，檢驗配適程度。

得出重要變數：

濃霧、沙塵暴、下雪、暴風雨、平均氣溫、平均風速與暴風雨之交互作用

並推論出：濃霧及平均氣溫升高會讓店家減少消售量，暴風雨與沙塵暴及下雪的發生，可能會增加銷售量;由上述兩點稍作推測，我們可以看出其濃霧的發生可能會造成交通上的阻礙，而消費者在這情況下則無法出門，造成消費量的降低。主要天氣中，沙塵暴及下雪及暴風雨可能是該商品對於其有相關關係，而導致消費者在這三種天氣來臨時，增加消費，而導致店家的消費量增加，我們推測其可能為一些災後復原家園的用品或是工具，也有可能是一些保暖用品，用來因應這三種主要天氣的來臨。

第陸章－其他變數之決策樹：

在此章使用了三種決策樹找出還有哪些變數對於銷售量有影響，分別使用的是 C5.0、Rpart、Randomforest。幸運的是，使用這三種方法皆找出了相同的主要影響變數－平均氣溫(Tavg)、平均風速(Avgspeed)。

而也從這三種決策樹的特性及比較找出最適合 Walmart 資料的方法－Randomforest，由於其預測準確率高出其他兩者頗多，儘管運算時間較長，卻也是最有說服力的決策樹。

	特性	預測準確率	時間
C5.0	只能用於類別觀測值	32.84%	1.46sec
Rpart	連續類別觀測值皆可用	33.31%	1.39sec
RandomForest	運算時間較久	44.66%	21.19sec

第柒章－假日與平日之差別：

在本章，主要是探討假日與平日之銷售量有無差異，首先，使用與先前一樣的常態性檢定－ad.test。確定假日與平日的銷售量皆為非常態後，使用無母數檢定－Wilcoxon rank sum test 來檢定是否銷售量有差。檢定後為有顯著差異，假日之銷售量大於平日之銷售量。

因而得出結論：假日銷售量較平日來的高，這與現實邏輯上判斷相符，因為大多數人平日都要上班，而多數家庭可能會較傾向於假日在去大賣場購物，我們認為此為主要影響銷售量的因素。

在整份報告裡頭，我們使用了三種完全不同的方法，來找出哪些天氣型態為主要變數。我們比較這三種方法：

比較無母數統計方法(Wilcoxon rank sum test)以及負二項回歸，都判斷濃霧的發生成銷售量的減少，我們合理推測是因為路上交通的影響，濃霧會造成汽車駕駛的路況不佳，漸少出門意願，降低銷售量；兩種方法亦同時評斷，沙塵暴發生及下雪的發生時會造成店家銷售數量的上升，我們這裡是以推測這與其銷售商品有關，可能是一些如口罩或是手套…等等相關商品。讓消費者會在這種天氣狀況下，有未雨綢繆，多買一些備用的心理，造成銷售量的增加。

再者，比較負二項迴歸以及決策樹，則是同時認為氣溫視為重要變數，當氣溫越低時，銷售量會增加，我們認為這與下雪等主要天氣有連帶關係，因為下雪會造成銷售量的增加，並會一併帶來低溫，所以會造成低溫時，銷售量也會增加。

綜括整體來說，當一些主要天氣的發生時會造成消費者有未雨綢繆的現象而造成銷售量的增加，我們亦推測這 111 種商品亦跟天氣有關，如口罩或雪鏟…等這類防護或是清掃工具。在綜合報告的所有章節後，以下是我們建議店家可以增加存貨以因應增加銷售量的幾個指標:

1.當下雪或是沙塵暴這兩種主要天氣情況發生時，這兩種天氣相關的商品需增加存貨，如：剷雪工具、口罩、暖爐…等商品。

2.低溫發生時，銷售量與低溫為正相關，店家需特別留意「保暖」商品或者「熟食」類商品。

3.當假日來臨時，銷售量與假日為正相關，這已不只是天氣敏感商品的部分，所有商品都需注意，店家可以先找出假日時熱賣商品，並在禮拜五就將熱賣商品之存貨增加。

下次店家在營業時，需特別注意天氣預報，並針對哪種天氣型態做出相對的因應措施，控管好存貨，店家本身也能自己觀察以往資料，推估今日可能的銷售量，對其做準備。

玖、參考文獻

1. 蘇佩芳統計諮詢上課講義,2015 年

2. 中華 R 軟體協會

(http://www.r-software.org/movielist)

3. 免費電子書 -- 機率與統計（使用 R 軟體）

(http://ccckmit.wikidot.com/st:main)

4. 提綱挈領學統計 張翔

心得

陳柏愷：

從大一修課到現在，無母數、抽樣、類別等課學了很多的觀念和知識，可是運用的範圍永遠只限於作業題目。而這次算是第一次從資料選取、處理、分析、到結論都自己做的完整統計分析報告，其實十分的具有挑戰性。

這次的報告主題我們選擇了 walmart，希望能和商業領域做一個結合。資料處理方面，其實我是第一次遇到資料分成兩個 csv 檔，而且也是第一次處理 400 萬筆這麼龐大的資料，所以剛開始對於如何處理這些資料毫無頭緒，要怎麼去分析更是遙不可及的事情，於是決定先從資料的合併開始慢慢一步一步做。

而剛開始在資料合併就遇到了問題，在嘗試了很多方式後才找到了用 merge 這個 code 去解決。再接下來的分類樹等方式其實也都是之前課堂上所沒有教過的，於是就慢慢翻書、上網慢慢學。而在合併和分析的過程中其實也激盪了很多之前所學的知識，突然有一種頓悟的感覺。

在上這堂課之前我的學習幾乎都是老師出甚麼作業，教甚麼課程才去被動的吸收知識。可是這堂課除了將許多重要的課程複習過，也在寫作業和報告的過程中自己去學習了很多的 packages 和很多的統計方法，把之前所學的東西和觀念串聯起來，讓自己感覺在統計分析的運用上有了很大的進步，Rcode 方面也是有很大的突破。

這堂課給我最重要的東西其實我覺得不是知識，而是處理問題的能力。對於一個沒處理過的資料型態，沒用過的統計方法，其實這些在網路上都是有許多資料可以搜尋的到，Rcode 的使用也有很多的教學影片可以讓人學習。而統計方法真的是太多種了，課堂上所學其實只是基礎，感覺很像 “師父引進門，修行在個人”的感覺，想要精進自己的能力就必須學會自己學習。

最後很感謝老師開這堂課，不管是從前面複習各科目的報告，到最後 kaggle 資料的分析，過程中讓我們自己摸索到了很多的東西，也比較具有自己解決問題的能力。

胡博仁：

這學期的統諮課就像是趟雲霄飛車一樣，有點難形容這種感覺，就好像原本很順利，突然急轉直下盪到谷底，之後又一步一步爬起來，這樣的感覺滿令人難忘的，我們四個沒有一個格外會寫程式，所以有時就是遇到問題，可能就會卡很久，網路上也盡量再找解決的方法，但感覺就是 R 程式一直無法讀懂我們的語言，再加上研究所的考試也是迎面而來，因為六日都在補習班上課，補習班也會模擬考，其實一考出來就大概知道自己的程度大約在哪，離自己的目標還差多遠，或許在學期初的時候還不是特別明顯，但從 100 天到 50 天到現在的 20 天，其實說不害怕是騙人的，所以我們一開始上台的報告統計檢定那次準備的比較充分，到後面的時候有要分析大數據，再加上另外一邊研究所考試的壓力，真的快爆了阿，所以在最後一次上台的報告，真的也很慘，不過接下來的兩個禮拜，我們重新又做了一次，幾乎重頭翻新了一次報告，回想起那兩個禮拜的時光，真的滿地獄的，早上打到晚上的程式碼，晚上回家繼續念著研究所考試，或許還是滿希望可以獲得老師的肯定，所以我付出滿大的心力，雖然修報告的時候，還是很多地方要改，不過至少我們真的進步，統計方法也不像之前報告那次空泛。

其實我本身不是一個成績好的學生，大一大二的成績其實也滿慘的，不過大三大四開始爬起來，其實我現在也還在重修著大二的程式設計，不過托統諮這堂課的福，我現在還滿熟練的，所以程設那邊的考試跟作業其實都很簡單，也算是個可以 HIGH PASS 的科目，一切的一切感覺都還是息息相關，統諮這堂課學到很多，也要學會與自己的組員溝通，當發生意見相左時，也顯得格外重要，最後也很開心自己還是有努力完成報告，或許報告的某些手法或地方還是相較之下有些拙劣，不過我覺得我還滿努力的，至於之後的研究所考試，也就盡力啦!希望可以完成自己的目標!!!

感謝老師這學期的教導，老師的上次講義做的很精美，總是可以比對著參考，那之後就可能有什麼需要老師幫忙，就在麻煩老師了，我們這組的報告每次都比較有創意一點，希望給予老師的驚喜大於驚嚇，謝謝老師。

李茂源：

在大四學期末了，回過頭來看看自己以前對統計方面的認識，從統計學到數統再到迴歸、抽樣、無母數、類別、實設等。我們總是在書本與考題間盤旋，學習著你覺得不一定用得到卻必須學起來的知識。回頭看看，驚覺那時的我很迷惘，自己像是被嵌入了「被動」的晶片。

然而，在這門課－統計諮詢，讓我看到在大三以前不曾看過的東西－「自我觀感」，在每一次的作業中，雖然老師會給予我們方向，但我們是隨著那方向要做出自己的答案，而非以往那種考試既定模式。在最後的報告中，連題目的方向都由我們自己定，從資料處理、整合、分析、解釋到最後的結論都是由自己一手打造的，這相信是每位修這堂課的同學，在大學時期本系最意義非凡的一堂課了。

在一開始，我們選了 Walmart 這份資料，主要是因為我們四個組員未來的方面都想往商業領域發展，這對未來我們的幫助也匪淺。我們相信有了這個概念之後，未來出社會一定也有相當大的幫助。

然而，我們開始了名為天堂卻是以地域之路。在討論時，我們投注了很多想法以及不同層面，希望能將這份報告做出非常全面性的分析。然而，也許這就是現實的差別，我們需要處理不少 missing value 以及要合併兩份資料才方便我們做分析，又或者因為商業關係而匿名的商品也迫使我們放棄了滿多的想法。這期間，光是在合併 400 多萬筆資料就讓我們焦頭爛額了，我們就像原本在溫室的花朵被放逐到外面的現實世界一樣，非常的徬徨失措。

但在後來，我們一步步將問題列出，再上網翻書慢慢解答，讓自己對於 R 又有重新一份認識，也覺得在分析資料及 R 的處理方面上都有慢慢的在進步。接著再慢慢補充慢慢豐富整份報告，讓他盡量接近我們原本想像的畫面。也許做這份報告也像是估計吧，他的完全體就是在一個確定的位子，我們永遠不知道他在哪裡，只能一步步的找出他「可能」在哪裡。

我覺得這份報告下來，我學到的不只是統計方面的知識，也學到了很多處理問題和在團隊合作時的意見溝通。不管是約時間或者是分配工作還是問題如何解決，又或者是當意見不同時要怎麼做出 win-win 的最佳選擇。其實都讓我收穫很大，我非常感謝我的組員能讓我有這麼多的經驗以及收穫。最重要的，也謝謝老師給了我們這個機會以及總是在旁協助讓我們一步步慢慢將統計精華的部分給吸收。

我知道在這門課期間，我們給了老師很多的期望，同時也給了很多失望。這讓我們體悟到了很多也覺得需要改變，也因為這樣我們努力將整份報告慢慢填補慢慢豐富，希望老師對我們的期望不會落空。最後，謝謝老師，因為這堂課，我對統計的方法重新複習並更新了不同的東西，在組員間也得到了很多收穫。

楊庭逸：

這一次的期末報告我們選用的資料是 kaggle 上 walmart 銷售量的資料，會選用 walmart 的資料最主要是因為我們這組四個同學未來都想往商發展，而這次資料中考慮的變數是各種天氣型態，其中包含一些我們平時認為與銷售量不會扯上關係的變數，譬如說：大氣壓力、日出日落時間和風速。一開始我們就遇到了一個很大的困難，兩筆資料日期並不完全一樣，而且其中一筆資料在 excel 無法完全打開，所以一開始這一部分就讓我們一度想換主題，後來在我們嘗試了一些方法後發現了 merge 這個方法，我們總算成功將兩筆資料合在一起，但這已經是我們期末報告的前兩三天了，所以即便我們報告前兩天都熬夜趕工，但依舊與我們想像中的完整期末報告有一段落差，對於糟糕的期末報告表現真的深感抱歉。也因此在這一次期末書面報告中，雖然小組中有些人快要考研究所、其他人忙著補習準備出國的考試，但我們依然花了很多時間努力補足我們之前缺的統計方法。這一次我深刻體會到，在一個團隊中，向心力是很重要的一件事情，像是一開始我們小組約討論時常會因為有人缺席導致整個團隊很鬆散、氣氛不佳，直到經過了期末上台報告老師嚴厲的點評，不服輸的個性把渴望成功的態度內化成積極向上的動力，最後在我們不眠不休奮鬥三天三夜後，終於完成了這用我們一部分肝換來的完整報告。

進來成大統計系之後，從大一的計概、大二的程設跟迴歸、大三的統模以及現在大四的統諮，基本上都與 code 脫不了關係，一開始其實覺得打程式很不親切、很排斥，但是經過了這四年的訓練，現在已經很能敞開心胸跟它做朋友。

我覺得統計諮詢這一門課毫無疑問的是統計系四年來最重要的一門課，前三年我們其實都只是在打基礎就像是準備食材，而這一門課教我們的是實務上面的應用，也是最能讓我們與職場接軌的一門課，就像是教我們如何把這些我們準備好的食材炒成滿漢全席，很謝謝老師上課的用心，教了很多做諮詢時會應用到的方法，這些都是老師好幾年來的心血啊！！也謝謝老師給了我們一個機會能夠完整的分析一個真實的資料，讓我們體會到這真是一份不輕鬆的工作阿！老師辛苦了！！！