# Bike Sharing Demand

廣義線性模式應用分析- 期末報告

2019 年 6 月 21 日

R07849016 黃琪婕

# 目錄

# 表目錄

# 圖目錄

# I. 前言

共享單車是一種新型的單車租賃形式，用戶通過遍布一個城市的自助服務站網絡，便可以從一個地方租用自行車，並根據需要將其返回到不同的地方。此租賃自行車的方式不僅綠色環保，且使用便捷，成為如今解決城市「最後一公里」最重要的交通工具。目前，全世界有超過 500 個自行車共享計劃，而對一個城市而言，單車的投放量至關重要，投放過多，導致資源浪費；投放過少，便無法滿足城市需求，因此準確預算共享單車的投放是關鍵。

而這筆 Bike Sharing Demand 資料，非常特別且具有吸引力，因為它明確記錄了歷史租賃數據和天氣變化。因此，這個自行車共享系統用作傳感器網絡，可用於研究城市中的移動性。在本次分析中，將結合歷史使用模式與天氣數據，以預測華盛頓特區 Capital Bikeshare 計劃中的自行車租賃需求。

# II. 材料與方法

## (I) 研究材料

本報告之資料取自於 Kaggle 平台 4 年前所舉辦的競賽，競賽的名稱同時也是資料名稱為: Bike Sharing Demand。表 1 為本資料前 2 筆數據，共具 12 個欄位資料，每筆欄位為該 1 個小時內的歷史租賃以及天氣變化數據，紀錄時間為 2011-01-01 00 至 2012-12-19 23，共有 10,886 筆，無任何缺失值。

表 1. Bike Sharing Demand 前 2 筆資料。

| | datetime | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2011-01-01 00:00:00 | Spring | 0 | 0 | Clear | 9.84 | 14.395 | 81 | 0.0000 | 3 | 13 | 16 |
| 2 | 2011-01-01 01:00:00 | Spring | 0 | 0 | Clear | 9.02 | 13.635 | 80 | 0.0000 | 8 | 32 | 40 |

表 2 為各個欄位數據的描述，其中有 5 個類別型資料，7 個連續型資料。由於本報告的目的，是要預測華盛頓特區每小時的自行車租賃需求，因此應變數為 count，也就是每小時的單車總租賃數量，而其他欄位則為自變數，用於預測 count。

表 2. 各欄位數據之描述及其類型。

| Column | 描述 | 類型 |
|---|---|---|
| datetime | 小時日期和時間戳 | 類別(時間) |
| season | 1:春天 2:夏天 3:秋天 4:冬天 | 類別 |
| holiday | 當天是否是節假日 | 類別 |
| workingday | 當天是否是工作日 | 類別 |
| weather | 1：晴，少雲，部分多雲，部分多雲.<br>2：薄霧+多雲，薄霧+破碎的雲，薄霧+少量的雲，霧<br>3：小雪，小雨+雷雨+散雲，小雨+散雲<br>4：大雨+冰盤+雷雨+霧，雪+霧 | 類別 |
| temp | 溫度，攝氏度為單位 | 連續 |
| atemp | 體感溫度 | 連續 |
| humidity | 相對濕度 | 連續 |
| windspeed | 風速 | 連續 |
| casual | 未註冊用戶的租賃數量 | 連續 |
| registered | 註冊用戶的租賃數量 | 連續 |
| count | 總租賃數量 | 連續 |

Kaggel 提供的訓練集，也就是我們訓練 model 的資料，為一個月前 19 天的數據和使用情況，而比賽評分的測試集，則是提供一個月 20 號以後的數據，我們主要的任務就是預測 20 號以後的使用量(圖 1)。而兩個資料集皆是 2011-2012 年間收集，所以並無外推年份的問題。

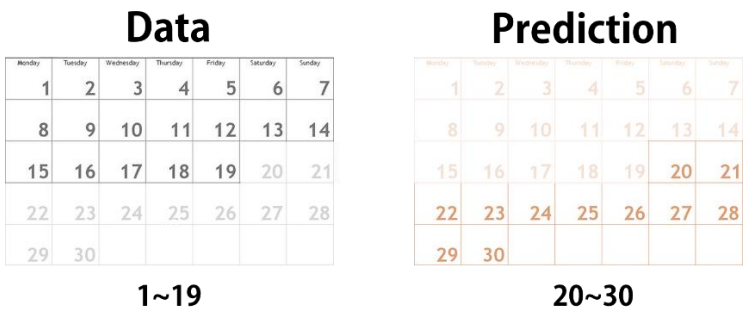


圖 1. 訓練集以及測試集的日期區別

## (II) 研究方法

首先對資料集做探索性分析，找出哪些數據會影響到每小時自行車的租賃數量，count，且由於此資料是以小時為單位，紀錄大約 2 年的時間，因此本報告會特別探討時間與其它變量間的交互作用。

經由探索性分析，找出可用的解釋變數後，還需對這些變數做前處理，如去除極值以及 Transformations 等，才能拿來擬合 GLM。在建模的部分，首先以最簡單的 Normal Regression 建模，可以直覺的察看解釋變數的意義；之後，考量到應變數 count 為計次的每小時單車出租數量資料，因此以 Poisson Regression 進行建模較符合資料的假設，但 Poisson 的假設為期望值等於變異數，但在這篇報告後續分析中，可以看到 Y 並不符合此假設，因此也將用 Negative binomial regression 進行建模，以考量到資料過於離散的狀況。

該報告使用與 Kaggle 一樣的標準，RMSLE，來對模型的好壞進行評估，其公式如如下所示。RMSLE 對預測低估的懲罰大於預測高估，此非常符合這筆資料的需求，因為這筆資料的目的在於預估單車可能的租賃數量，並隨時調放單車的數量以因應城市的需求，因此投放過多並不會造成太大的問題，而投放過少便會無法滿足消費者的需求，低估的後果比起高估來的嚴重，所以使用 RMSLE 來做為模型好壞的評估。

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

在自我評估的部分，我使用 10- fold cross validation ，將訓練集分為 10 份，分別計算 10 次模型 RMSLE，再取平均得到一個平均的 RMSLE，做為評估不同模型間預測能力的好壞的標準。

# III. 結果

## (I) 探索性分析

### 1. Count

Count 為該報告的應變數 Y，其代表每小時的單車出租數量，因此這是一個連續型的變量，由表 3 可以得知，count 的平均值為 191.57，其標準差為 181.14，因此可以想見此數據的變異非常的大，有過度離散的狀況存在，因此也使用 Negative binomial regression 建模；count 的範圍介於 1 至 977 之間。由圖 2 可以看出 count 是一個右邊的連續分布

表 3. Count 的描述性統計

| mean | sd | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| 191.57 | 181.14 | 1 | 42 | 145 | 284 | 977 |

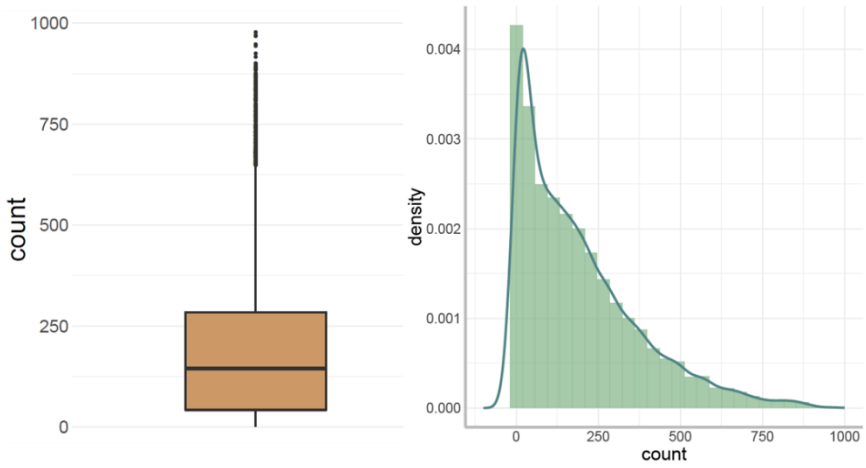


圖 2. Count 的 box plot 與機率分布圖。

### 2. datetime

datetime 紀錄的是時間的數據，以 2011-01-01 00:00:00 的形式記錄，最小的單位為小時，因此可以依照時間特性，將其在分為每小時，星期幾，每月及每年為單位，再令時間作為 X，count 作為 Y，可畫出在時間單位下 count 的變化趨勢，用以探討不同時間單位對於 count 的影響。由圖 3 可以看出，早上 8 時單車的租賃數量達到一個高峰，下午 17~18 時，有另一個高峰。

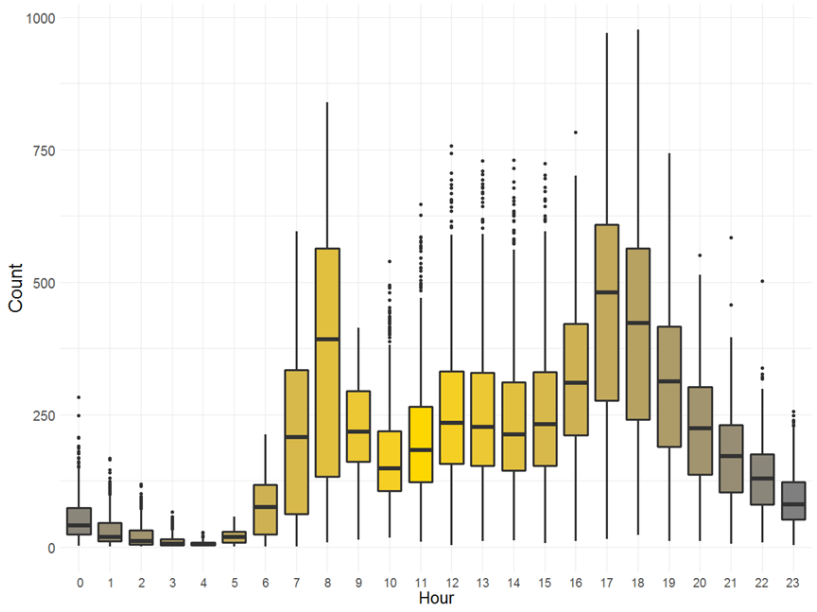


圖 3. 不同小時 count 的分布狀況

圖 4 可以看出，單就周一至周日的單車租賃數量來看，ANOVA 的檢定並不顯著，因此星期幾的平均單車租賃數量並沒有差別。但如果加入小時的影響後，可以發現周六與周日，和周一至五的單車租賃趨勢不太一樣，周六與周日的單車租賃數量由早上 7 時開始穩定上升，直到 12 時達到高峰，16 時才再開始下降，呈現一個單峰的狀況；而周一至五的單車租賃數量則是 7 時達一個高峰後，便下降直到 17-18 時再次達另一個高峰，呈現雙峰的狀況，而這兩個峰剛好分別對應到早上上班以及下午下班的人潮數。

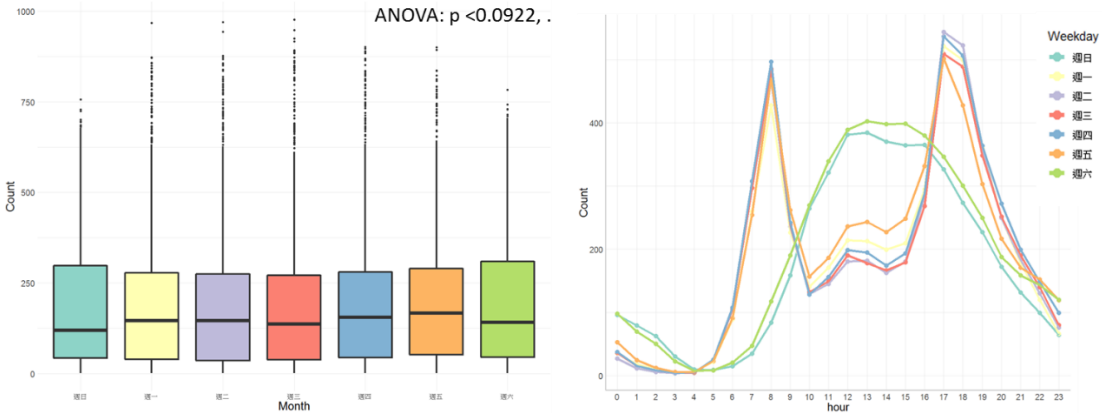


圖 4. 星期幾與 count 的分布狀況，左圖為周一至周日單車租賃情況；右圖為周一至周日不同小時單車租賃情況。

圖 5 可以看出，單車的租賃數量由一月為最低，之後開始上升直到 7 月夏天達到最高，後續開始下降，因此可以說秋夏的租賃數量較多，春冬較少。而在加入不同小時的影響後，可以發現到，每個月分不同小時的單車出租趨勢近乎一致，與單看不同小時單車租賃數量的趨勢一致(圖 3)，皆是早上 8 時單車的租賃數量達到一個高峰，下午 17~18 時，有另一個高峰。

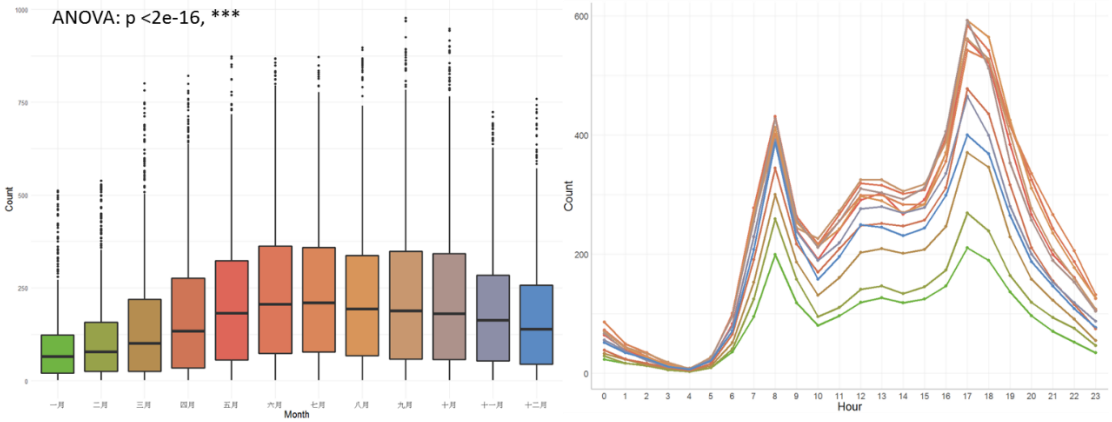


圖 5. 不同月分與 count 的分布狀況，左圖為不同月份單車租賃情況；右圖為不同月份不同小時單車租賃情況。

該資料總共橫跨 2 年的時間 2011-2012，而從圖 6 可以看出，2012 年的平均單車出租數顯著大於 2011 年(p <2e-16)。而當另外再加入不同月份的影響時，可以看出這 2 年的月趨勢大致相同，與圖 5 一致。

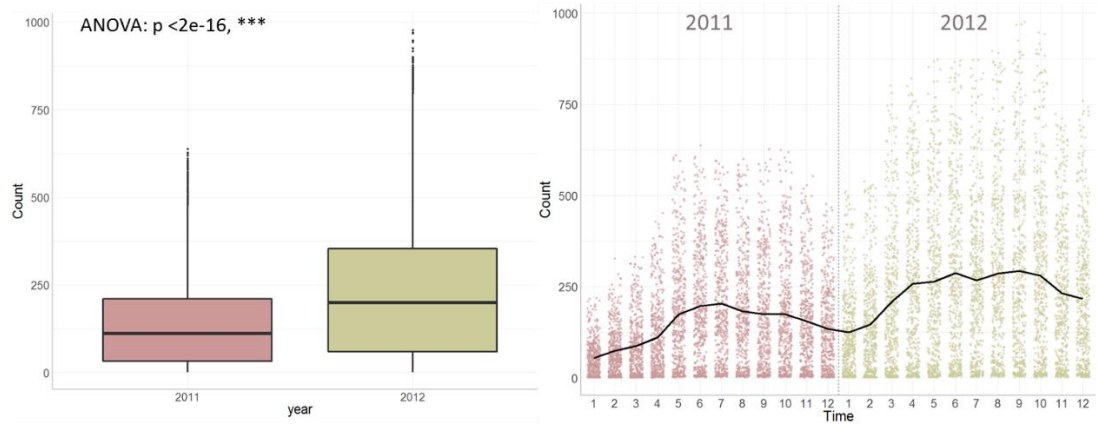


圖 6. 不同月分與 count 的分布狀況，左圖為不同年份單車租賃情況；右圖為不同年份不同月份單車租賃情況。

### 3. season

season 為類別資料，1:春天 2:夏天 3:秋天 4:冬天。由圖 7 可以看到，4 個季節分布的筆數非常平均；而不同季節對於 count 的影響則不太一樣(圖 8)，與圖 5 的發現一樣，秋夏的租賃數量較多，春冬較少，不同季節的平均單車租賃數量具有顯著性差異(p <2e-16)，而表 4 則為其詳細的敘述性統計。一樣加入小時的影響，來看不同季節不同小時的單車租賃趨勢，發現到這與單看不同小時單車租賃數量的趨勢一致(圖 3)。

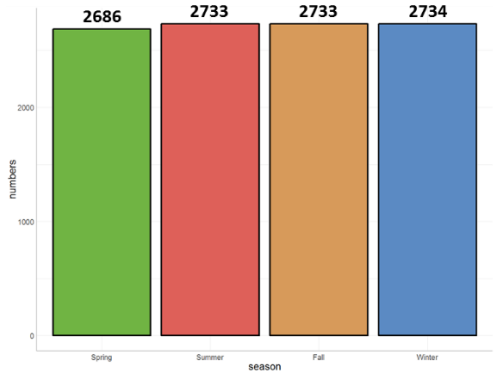


圖 7. 不同季節的資料筆數。

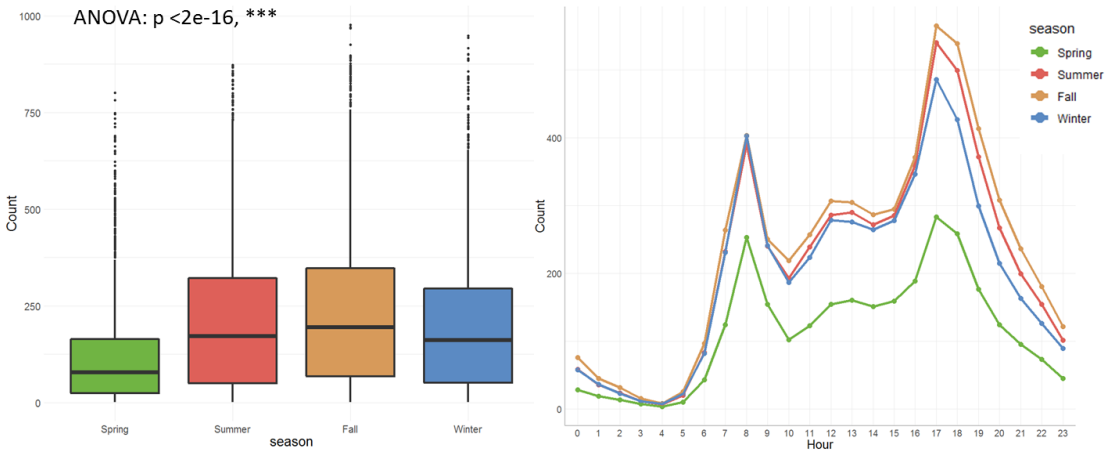


圖 8. 左圖為不同季節單車租賃情況；右圖為不同季節不同小時單車租賃情況。

表 4. 不同季節單車租賃數量的敘述性統計

| | mean | sd | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|---|
| Spring | 116.3433 | 125.2740 | 1 | 24 | 78 | 164 | 801 |
| Summer | 215.2514 | 192.0078 | 1 | 49 | 172 | 321 | 873 |
| Fall | 234.4171 | 197.1510 | 1 | 68 | 195 | 347 | 977 |
| Winter | 198.9883 | 177.6224 | 1 | 51 | 161 | 294 | 948 |

## 4. holiday

holiday 為類別資料，表示當天是否為假日，0 表否，1 表是。由圖 9 可以看到，這筆資料大部分的小時皆不是假日。圖 10 可以看到，假日以及非假日的平均單車租賃數量並沒有顯著性差異(p <0.574)，表 5 為其詳細的敘述性統計；但是加入小時的影響後，可以發現到假日以及非假日的租賃趨勢不太一樣假日較像單峰，非假日則呈現雙峰，結果似圖 4 所示。

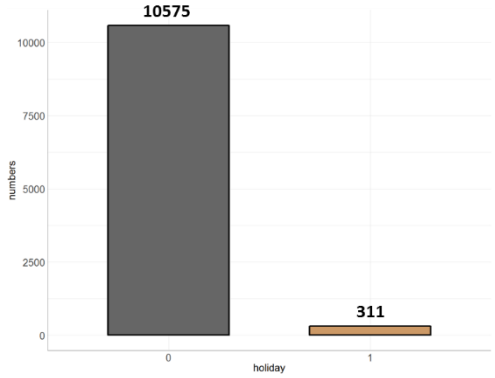


圖 9. 假日以及非假日的資料筆數。

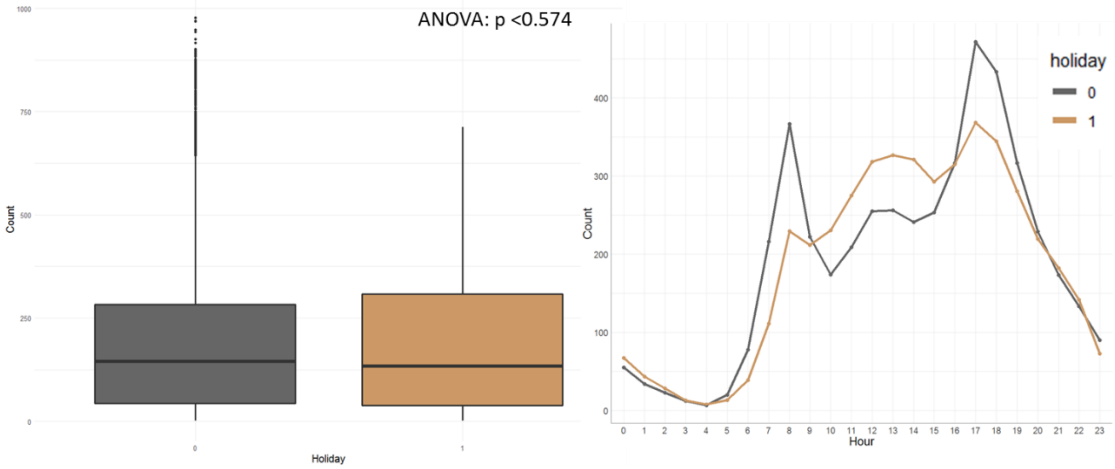


圖 10. 左圖為假日以及非假日單車租賃情況；右圖為假日以及非假日不同小時單車租賃情況。

表 5. 假日以及非假日單車租賃數量的敘述性統計

| holiday | mean | sd | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|---|
| 0 | 191.74 | 181.51 | 1 | 43.0 | 145 | 283 | 977 |
| 1 | 185.88 | 168.30 | 1 | 38.5 | 133 | 308 | 712 |

## 5. workingday

workingday 為類別資料，表示當天是否為工作日，0 表否，1 表是。由圖 11 可以看到，這筆資料大部分的小時皆是工作日。圖 12 可以看到，非工作日以及工作日的平均單車租賃數量並沒有顯著性差異(p <0.226)，表 6 為其詳細的敘述性統計；但是加入小時的影響後，可以發現到非工作日以及工作日的租賃趨勢並不一樣，工作日為雙峰，非工作日則呈現雙峰，結果與圖 4 所示一致。

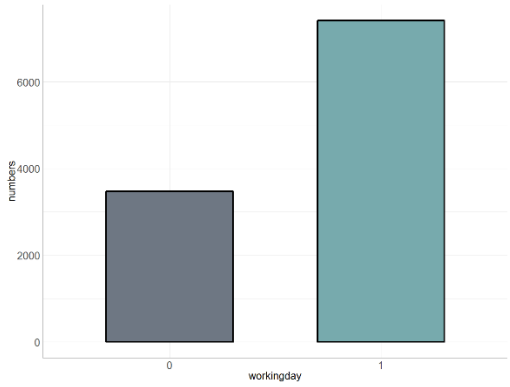


圖 11. 非工作日以及工作日的資料筆數。

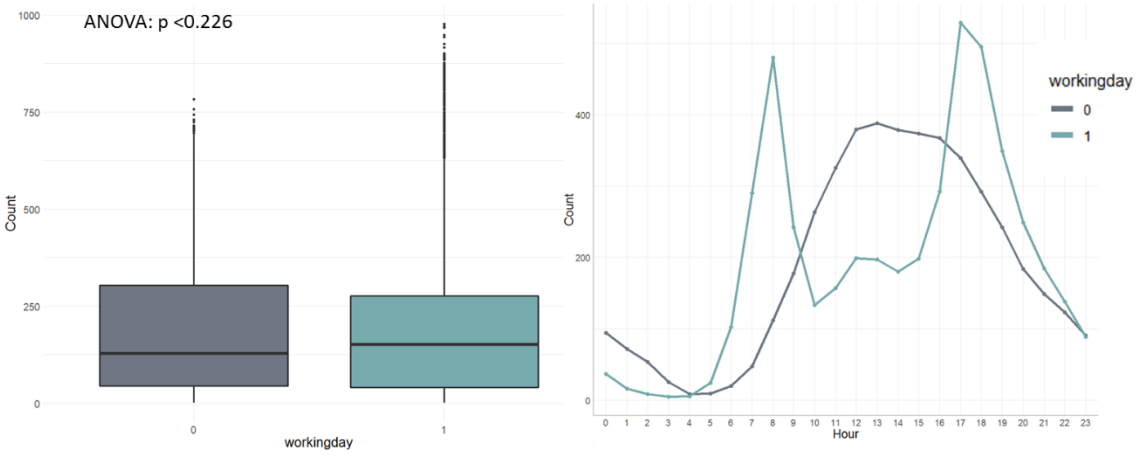


圖 12. 左圖為非工作日以及工作日單車租賃情況；右圖為非工作日以及工作日不同小時單車租賃情況。

表 6. 非工作日以及工作日單車租賃數量的敘述性統計

| workingday | mean | sd | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|---|
| 0 | 188.51 | 173.72 | 1 | 44 | 128 | 304 | 783 |
| 1 | 193.01 | 184.51 | 1 | 41 | 151 | 277 | 977 |

## 6. weather

weather 為類別資料，表示天氣狀況，詳情請查看表 2。圖 13 可看出，大部分的小時天氣狀況都不錯，以晴天及霧居多，下暴雨的情況在 10,886 小時中僅出現 1 小時；而 4 種不同的天氣狀況其單車平均租賃數量有顯著性差異(p <2e-16)，天氣越好，單車出租數越多，表 7 為其詳細的敘述性統計。

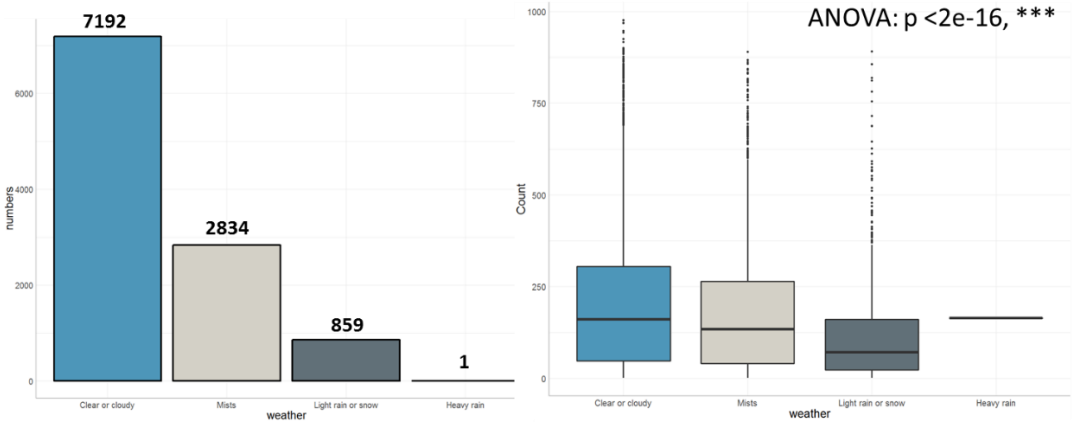


圖 13. 左圖為不同天氣狀況資料筆數；右圖為不同天氣狀況單車租賃情況。

表 7. 不同天氣狀況單車租賃數量的敘述性統計

| weather | mean | sd | 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|---|
| Clear or cloudy | 205.24 | 187.96 | 1 | 48 | 161 | 305 | 977 |
| Mists | 178.95 | 168.37 | 1 | 41 | 134 | 264 | 890 |
| Light rain or snow | 118.85 | 138.58 | 1 | 23 | 71 | 161 | 891 |
| Heavy rain | 164.00 | NaN | 164 | 164 | 164 | 164 | 164 |

## 7. temp, atemp, humidity, windspeed

溫度、體感溫度、相對濕度及風速等天氣資料皆為連續型資料，因此可以對 count 做散佈圖，算出天氣資料與 count 的相關性，在此以 person correlation 為主。

由圖 14 可得知，溫度、體感溫度與相對濕度的自身分布較似常態分佈，而風速則明顯是一個右偏的分布，且風速為 0 的數據非常多，推測可能是因為缺

失值太多，以 0 取代。在天氣資料各自的相關性則可以發現到，溫度與體感溫度呈現高度相關(0.985)，因此後續模型擬合時應只擇一放入，以避免高度共線性的問題。而在天氣狀況與 count 的相關性方面，則可以發現溫度與 count 呈現正相關(0.394)，也就是溫度越高，單車租賃數量越多；而濕度則與 count 呈現負相關(-0.317)，濕度越高，單車租賃數量越少；最後，風速與 count 呈現輕微的正相關(0.101)，但由於將 0 當 null 填充的問題，推測會低估其相關性。

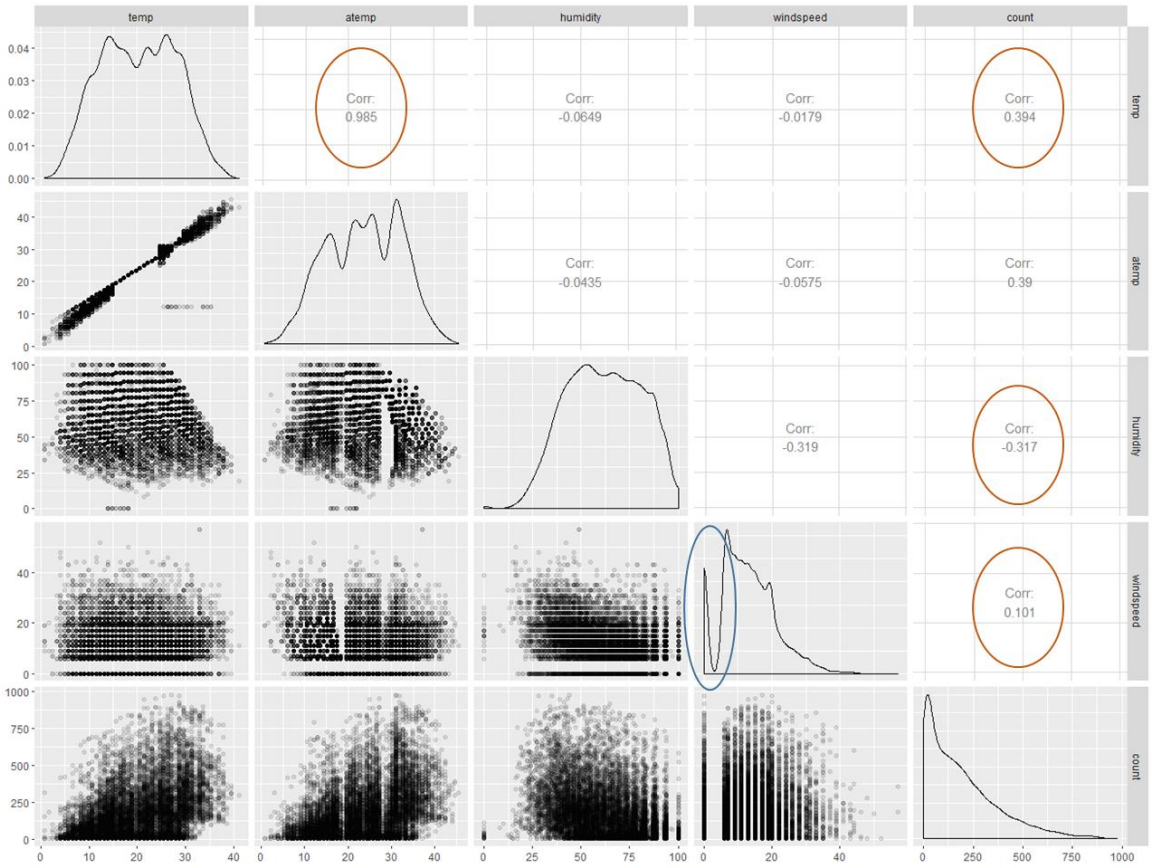


圖 14. 溫度、體感溫度、相對濕度及風速的分布及其與 count 的相關性。

## 8. 探索性分析之結論

由上述探索性分析可以發現到，時間的因素非常重要，因此會放入每小時，星期幾，每月及每年作為解釋變數。另外若加入小時為單位來看不同變量與 count 的關係時，更可以發現到租賃數量不同的趨勢。如時間與周一至周日

應具有一個交互作用項，才能反映出其趨勢的差別，而 workingday 的解釋與周一至周日過於重複，因此不將其加入模型中。另外 holiday 中假日及非假日在不同小時的趨勢雖然不太一致，但在初步的模型中，僅先把他單獨放入模型中，不考慮其與小時的交互作用。而在天氣變量方面，則選用溫度、相對濕度加入模型中。

**(II) 資料前處理**

模型的選擇標準是看 RMSLE，其對預測低估的懲罰大於預測高估，因此在模型擬合前，先將過大的極值去除掉，期望模型能對較低的租賃數掌握準確。首先嘗試使用 Boxplot 法去除極值，該報告選用的標準是 Q3+ 2*IQR，並非一般常用的 1.5*IQR，因為我們的目的是將極值去除掉的同時，模型還擁有預測較大租賃量的能力，而經由此方法共去除掉 117 筆資料。另外也使用另一個方法，標準化法，來看其去除極值的結果。該報告的標準是先將 count 標準化後，大於 3 的值去除掉，共刪去 147 筆資料。若仔細查看，可以發現 2 個方法刪去的資料皆差不多，因此選用標準化法作為去除極值的標準，而之後的分析中，皆是使用經過標準化法處理過的資料。

由於之後有使用到 Normal Regression，因此在對資料進行擬合之前，先對應變數 count 進行 Transformations，可降低殘差值不符常態的影響。在此本報告使用了 2 種方法，Log 以及 Box cox，2 者方法皆會在之後分析中出現。

**(III) 模型擬合**

首先以探索性分析得到的結論，將解釋變數放入 3 個模型中進行擬合，這種模型非常直覺且好解釋，符合我們探索到的變量特性。以下依序以 Normal Regression、Poisson Regression 及 Negative binomial regression 進行建模，並進行比較。

1. 探索性分析結果直接建模

Normal Regression 的 Y 是使用過 log 或 box cox 轉換的，以下為 log 轉換過後的模型。解釋變數的選擇是依據探索性分析而來。該模型使用 Log Transformation 的 RMSLE 為 0.3521，而使用 Box cox 的 RMSLE 則為 0.3633。因此就模型預測能力來說，使用 Log Transformation 較好。另外使用 Log Transformation 的模型其 $R^2$ 為 0.9347，表示這些解釋變數經由線性迴歸，可以解釋 count 變異量的比例為 0.9347，擬合的不錯。

$$\log(count) = \beta_0 + \beta_1 weather + \beta_2 temp + \beta_3 humidity + \beta_4 month + \beta_5 year + \beta_6 holiday + day * hour$$

而 Poisson Regression 與 Negative binomial regression 的模型長的一模一樣，僅在於 Dispersion parameter 的假設不一，Poisson 假設其為 1，Negative binomial 則會去估計離散的狀況，而用以下的模型進行擬合後，Negative binomial regression 估出的 Dispersion parameter 為 14.251，更加確認了該本資料的 Y 確實有過度離散的狀況。在這個模型中，Poisson Regression 的 RMSLE 為 0.3696；而 Negative binomial regression 的 RMSLE 則為 0.3485，因此可以發現 Negative binomial regression 在這 4 個模型中，預測能力是表現最好的。

$$\log(E(count)) = \beta_0 + \beta_1 weather + \beta_2 temp + \beta_3 humidity + \beta_4 month + \beta_5 year + \beta_6 holiday + day * hour$$

2. 加入其他解釋變數以建模

由於此資料的目的在於預測，因此想更進一步的找出預測能力更好的模型，我們進一步將風速放進模型，也放入 holiday 與小時的交互作用項，來看模型的擬合狀況。首先在 Normal Regression 的部分，加入這些解釋變數後，確實使 Log Transformation 的 RMSLE 從 0.3521 下降到 0.3382，甚至比起之前模型

預測最好的 Negative binomial regression 的 RMSLE 還低。

$$\log(count) = \beta_0 + \beta_1 weather + \beta_2 temp + \beta_3 humidity + \beta_4 windspeed + \beta_5 month + \beta_6 year + holiday * hour + day * hour$$

而在 Poisson Regression 與 Negative binomial regression 的部分，模型預測力也都獲得了改善，Poisson Regression 的 RMSLE 由 0.3696 降為 0.3552；而 Negative binomial regression 的 RMSLE 則從 0.3485 降為 0.3473。而 Negative binomial 估出來的 Dispersion parameter 為 14.312，與之前模型相差不遠。

$$\log(E(count)) = \beta_0 + \beta_1 weather + \beta_2 temp + \beta_3 humidity + \beta_4 month + \beta_5 year + \beta_6 holiday + day * hour$$

由以上可以發現到，在新模型中，Normal Regression 的 RMSLE 是最低的，預測能力最好，因此本報告使用該模型去對比賽評分的測試集，也就是一一個月 20 號以後的數據去進行擬合，並上傳到 Kaggle 平台，由 Kaggle 算出測試集的 RMSLE，而這個 Normal Regression 的新模型其測試集的 RMSLE 為 0.4597，略高於訓練集的 RMSLE，因為測試集的資料並沒有拿來訓練模型，因此較高的 RMSLE 是合理的。而這個成績在 Kaggle 競賽大概可以拿前 27%左右。

## IV. 討論

由結果可知，經由探索性分析找到的解釋變數，確實能有效的對本資料進行模型擬合，不但具有解釋意義，同時也具備預測能力，在此我們先對探索性分析結果直接建模的結果中，最好的模型 Negative binomial regression 來詳細探討，其每一個解釋變數對於 Y 的解釋意義，模型如以下所示：

$$\log(E(count)) = \beta_0 + \beta_1 weather + \beta_2 temp + \beta_3 humidity + \beta_4 month + \beta_5 year + \beta_6 holiday + day * hour$$

Negative binomial regression 是對機率建模，解釋與 Poisson Regression 一樣，首先看到 day*hour 的部分，它具有 6 + 23 + 6*23 共 167 個係數，將其算出後加減一下以用於探討這個交互作用項，是否可以準確的捕捉到周末與周一至五不同小時單車租賃數量不同的趨勢。圖 15 可以明確地看出，加入這個交互作用項，確實可以幫助我們把周末與周一至五不同小時單車租賃數量的趨勢捕捉到模型中，增進模型的預測能力。

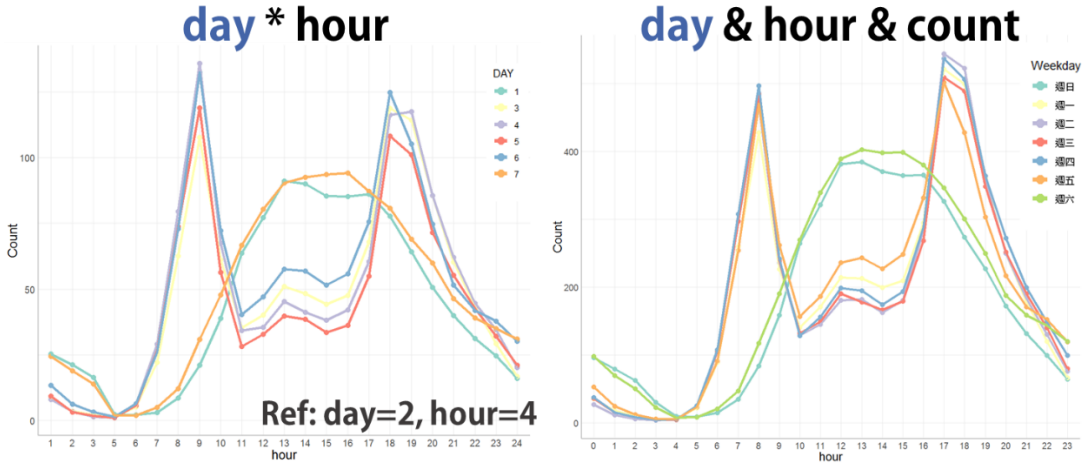


圖 15. 左圖為 day*hour 交互作用向所算出的趨勢；右圖為探索性分析所探討的趨勢。

而在 holiday 的部分，只有將其單放進解釋變數中，並沒有另外加入交互作用項，因此其係數 0.064 僅能解釋為，在控制其他解釋變數的情況下，假日的平均單車出租率是非假日 exp(0.064)= 1.159 倍，達顯著上的統計意義，在此參考組為非假日(0)。

而 weather 共有 4 個類別，其中我們以晴天作為參考組，其他 3 類之係數個別取過 exp 後，分別為 Mists: 0.928，Light rain or snow: 0.596 及 Heavy rain: 0.704，前兩者達統計顯著，後面因為只有 1 筆數據，因此未達統計顯著。從這些數據可以看出，天氣越晴，單車的出租率越高。

temp 與 humidity 的部分則與探索性分析的結果一致，溫度與 count 呈現正相關，其係數取過 exp 後為 1.026，也就是在控制其他解釋變數的情況下，溫度每上升一度，單車的出租率會變為原本的 1.026 倍；humidity 與 count 呈現的是負相關，其係數取過 exp 後為 0.998，也就是濕度每上升一度，單車的出租率會變為原本的 0.998 倍，在控制其他解釋變數的情況下。兩者天氣變量的變數皆達顯著統計。

時間變量的方面還有不同月份及不同年沒討論，表 8 可以看出，month 這個解釋變數的確有抓到不同月份的單車租賃數量趨勢(圖 5 左)，但是其在 9-12 月呈現一個高估的狀況，因此這個解釋變數並沒有很有效，且完整的掌握到不同月份的單車租賃數量趨勢。而在 year 的方面，以 2011 年為參考組後，其係數取過 exp 後為 1.645，也就是在控制其他解釋變數的情況下，2012 年的平均單車出租率是 2011 的 1.645 倍，達顯著統計，而在探索性分析也可以看到差不多的結果(圖 6)。

表 8. 以 1 月為參考組之不同月份的係數

| 月份 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 係數 | 1.195 | 1.311 | 1.645 | 2.058 | 1.947 | 1.755 | 1.788 | 1.955 | 2.225 | 2.146 | 2.040 |

加入 windspeed 以及 holiday 與 hour 的交互作用後，模型的預測力改進很多，其中 windspeed 的係數為 0.0044，跟探索性分析得到的結果一樣，但由於沒有去處理 null 填充為 0 的問題，加入 windspeed 只能幫助模型改善模型一點預測能力。而加入 holiday 與 hour 的交互作用則代表，星期一至日不同小時的單車租賃數量可以不一樣，是否為假日又可以再不一樣，增進了不少模型的預測能力。

上述兩個模型都有很嚴重的共線性問題存在，可以看到最後的模型的膨脹係數非常大，其原因主要是因為設了許多類別變項之間的交互作用，因此估出

的參數非常多，且大多所解釋的意義相似。因此我在想若要解決這個問題，我可以將小時分組，將該小時間租賃數量相近的放在一組，共分為 10 組(圖 16)，而這個方法確實可以使膨脹係數下降，但仍未達到標準(圖 17)。不過換個方向想，我的模型重點在於預測，高度的共線性並不會侷限我模型的預測能力，它影響的是模型的收斂能力，因此在這個前提下，是否真的要去解決這個問題，是值得探討的。

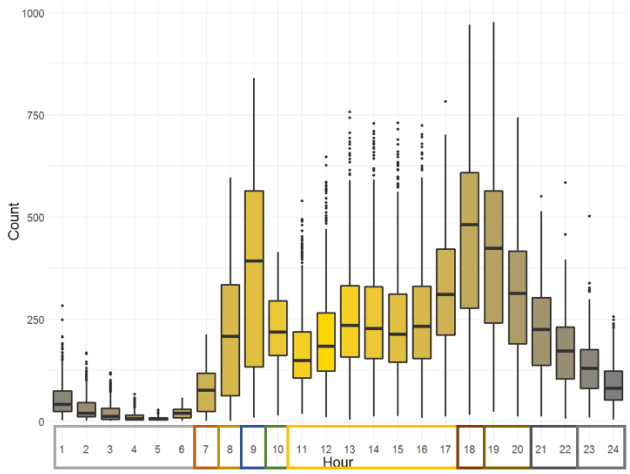


圖 16. 將 24 小時分為 10 類

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| holiday | 2.512016e+01 | 1 | 5.012002 |
| hour | 1.512965e+19 | 23 | 2.611878 |
| weather | 1.401154e+00 | 2 | 1.087981 |
| temp | 5.804497e+00 | 1 | 2.409252 |
| humidity | 1.977891e+00 | 1 | 1.406375 |
| month | 6.724367e+00 | 11 | 1.090487 |
| year | 1.042752e+00 | 1 | 1.021152 |
| day | 1.999233e+08 | 6 | 4.917435 |
| windspeed | 1.209517e+00 | 1 | 1.099780 |
| holiday:hour | 5.236103e+02 | 23 | 1.145800 |
| hour:day | 5.450832e+27 | 138 | 1.260359 |

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| holiday | 2.722304e+01 | 1 | 5.217571 |
| hour_lasso | 4.783282e+06 | 8 | 2.615067 |
| weather | 1.361985e+00 | 2 | 1.080297 |
| humidity | 1.905116e+00 | 1 | 1.380259 |
| month | 1.235760e+00 | 11 | 1.009669 |
| year | 1.015255e+00 | 1 | 1.007599 |
| windspeed | 1.201006e+00 | 1 | 1.095904 |
| day | 1.666729e+08 | 6 | 4.843457 |
| holiday:hour_lasso | 7.420769e+01 | 8 | 1.308890 |
| hour_lasso:day | 3.035249e+14 | 48 | 1.415325 |

圖 17. 左圖為最終模型的膨脹係數；右圖為將小時分組後之模型的膨脹係數。

另外該報告也測試了一個做法，因為不在本次報告的範圍中，因此僅以文字簡單描述。以上模型中放入許多交互作用項，會導致高度共線的問題存在，然而其交互作用項都是在 2 個類別變項之間，所以簡單來說只是算出在許多不同 x 狀況下的平均值，例如星期一的 24 小時都分別有一個平均值，出現在模型的系數中；以此類推，星期二的 24 小時也都分別以一個系數表示。因此我在想

可以直接把這些平均值算出來，放入模型中當解釋變數，然後使用 Ridge Regression，其所加入的懲罰項能使模型能穩定收斂，不被共線性問題影響，但同時會使的模型的 bias 增加。不過如此一來我便能算出各種 x 狀況下的平均值，如以上模型以用到的 hour*day、hour*holiday 等，還可以再將 weather*weekday*hour 等平均值的資訊算出放入解釋變數中，交由 Ridge 算出。如此一來模型訓練集 CV 的 RMSLE 為 0.3162，而測試集的 RMSLE 在 Kaggle 上的評分為 0.40401，大約為前 8%的程度。但這麼做有一個缺點，就是它幾乎捨去了模型被推論的能力，以換取預測力，與機器學習大部分的方法類似，是以電腦幫忙挑變數，而不是經由研究者確認過，具因果關係的解釋變數，因為在這種情況下，該模型只能解釋為，過去資料中的星期幾的第幾個小時，對未來模型的預測很有影響等推論，比起前面所述的模型，其推論能力不佳。

## V. 結論

這筆 Kaggle 的競賽資料，Bike Sharing Demand 非常適合使用 GLM 進行擬合，雖然預測力比起其他的機器學習的方法，如 Random Forest 或 XGBooting 等模型來說，還是低了點，但是卻有著其他方法無可取代的優點，就是模型的可被推論的能力，例如經由模型我們可以更加確認天氣越晴，平均而言單車租賃數量越多等趨勢。另外這筆資料有一個缺點，就是它並沒有進行外推的動作，並沒有用於預測未來，也就是說用測試集來評估模型好壞的標準，與建模的目的不太一樣，對該資料建模的目的，應是想預測未來在某些情況下，可能會有的租賃數量，進而調動單車的配置，達到消費著的需求，但 Kaggle 只局限於在內推的部分，較沒有注重於未來預測的方面。因此這個議題應建立一個動態的模型，隨時更新，用以預測明天，甚至是未來一周的情況，如此才能幫助公司進行決策，進而增加單車的租賃數以及營收。