統計應用方法期末報告

阿拉比卡咖啡品質與其生產各項因素之分析

組員：

309551059 羅文慧

309554005 黃睿宇

0511606 陳薀涵

目錄

一、研究動機

目前世界上有超過 100 種咖啡，其中最普遍的是阿拉比卡(Arabica)和羅布斯塔(Robusta)兩大種類，阿拉比卡占全世界咖啡產量的 70～75%，一般認為是較高級的咖啡品種。而咖啡品質的好壞又由許多的因素所影響，包括咖啡豆的生產地、生產高度、栽種方法等等。

而我們想了解在高級的咖啡豆背後，其生產來源與品質究竟有什麼樣的關聯性，不同的栽種方式是否會對應特定的香氣或口味？因此我們蒐集了超過一千筆的資料，在衡量品質的好壞亦選擇許多的變數，包括咖啡的口味、香氣、回甘等等，希望藉由分析這些歷史數據，找出生產環境與品質的關係，如此便能依據自身喜好的變因去選擇適合的咖啡來源。

二、研究內容

本研究欲探討阿拉比卡的咖啡品質與其咖啡豆生產之各項因素是否有關。

1. 資料來源：
   本研究數據取自 Kaggle 上 Coffee Quality database from CQI 之 arabica_data_cleaned.csv 資料，原始資料來自 Coffee Quality Institute 於 2018 年 1 月產生之報告，來源連結如下：來源。

2. 資料簡述：
   原始資料共有 42 個變數與 1311 筆資料，本研究捨棄不感興趣與重複之變數，並刪除不完整的數據後，分析之數據為 1092 筆，變數 18 個，簡介如下：
   (1) 自變數（咖啡豆生產環境及狀態）：
      - 擁有者（Owner）
      - 來源國家（Country of origin）
      - 來源地（Region）
      - 加工方法（Processing method）
      - 顏色（Color）
      - 海拔（Altitude）
      - 收成年份（Harvest year）
   (2) 應變數（品質衡量變數）：
      - 香氣（Aroma）
      - 風味（Flavor）
      - 餘韻（Aftertaste）
      - 酸度（Acidity）
      - 醇厚度（Body）
      - 平衡性（Balance）

- 一致性（Uniformity）
- 乾淨度（Cup cleanliness）
- 甜度（Sweetness）
- 整體性（Cupper points、Total cup points）

3. 資料整理：

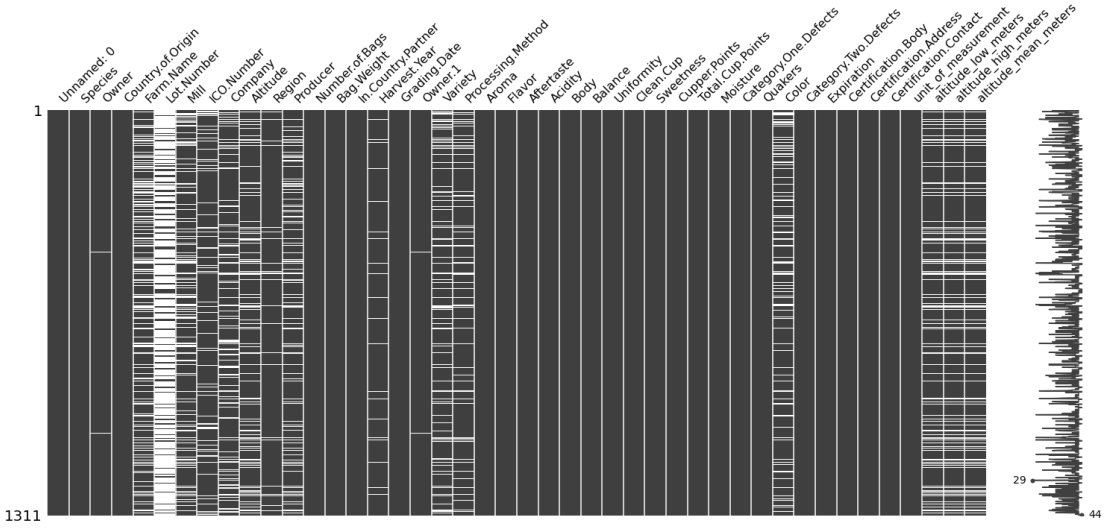
（1）某些變數如 ICO. number 和 lot. number，只是辨別咖啡的序號，在分析上並無需要，故此類變數我們予以刪除。

（2）某些變數如 owner、farm. name、mill、company、producer，因其屬性類似，皆為辨識生產者，故我們選擇最完整亦最適當的 owner 作分析。

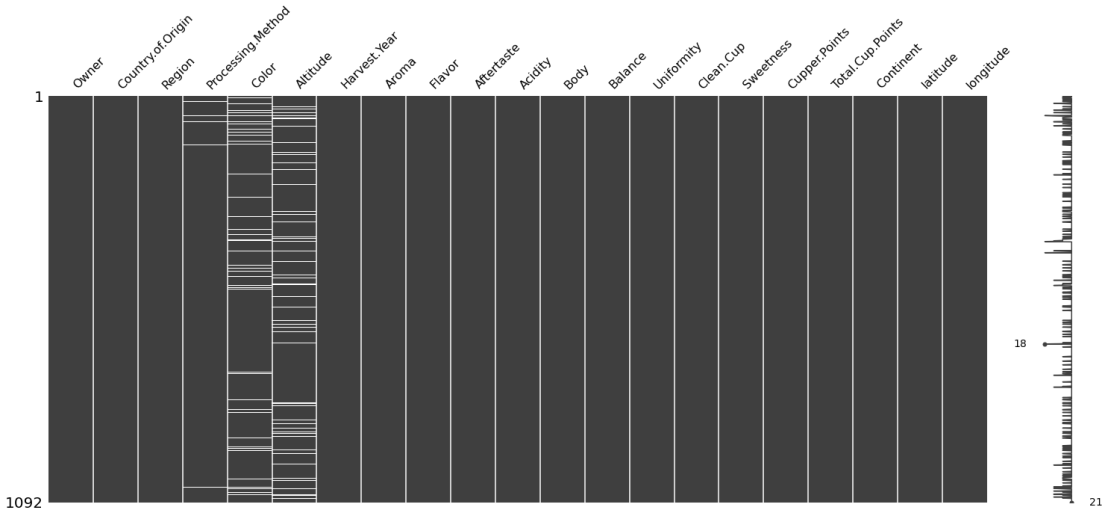
（3）對於某些缺失的數據，我們處理如下：

- Owner 缺失的數據，我們以 farm. name 的資料替代。
- Region 缺失的數據，我們以 country 的資料替代。
- Altitude 缺失的數據，我們以 mean 的資料替代。
- 對於無法找到替代數據的變數如 harvest. year、processing. Method，我們直接刪除該筆資料。

使用 missingno 套件來觀察缺失值的狀況，可以發現原始資料（圖一）在某些變數資料有許多缺失；經過上述處理，以及我們額外針對生產地新增「洲別」的分類後（圖二），數據缺失情況有很好的改善。



（圖一）

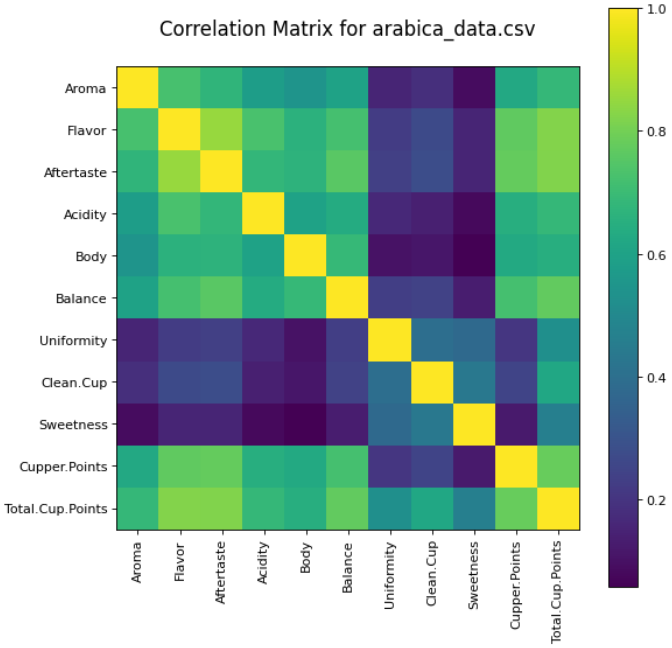


（圖二）

三、資料分析

1. 相關係數矩陣的熱度圖

   由 heatmap 我們可以很快速地察覺出不同變數之間的關係，快速的預覽資料可以幫助我們挑選特徵。



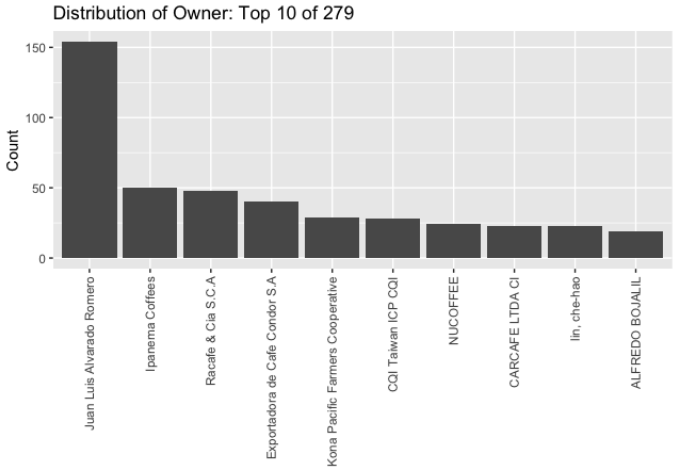
   從上圖可以觀察到整體分數(total cup points)與風味和餘韻有較大的關係，而與一致性和甜度關係較小；此外，一致性、乾淨度和甜度這三個變數與其他變數相關性都很小。

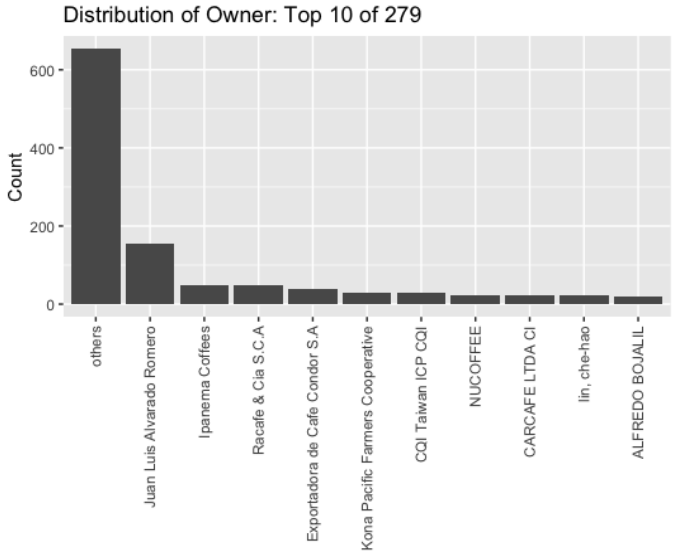
2. 自變數分析

   此次研究分析之咖啡來源，結果如下：

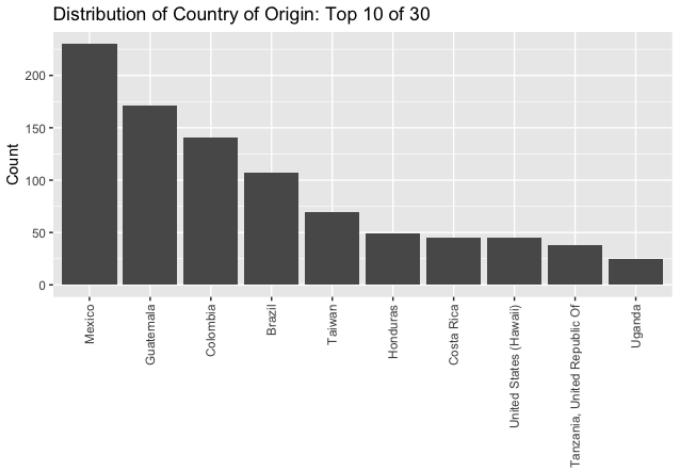
   (1) 第一張圖顯示咖啡擁有者數量的前十名，來自瓜地馬拉 Juan Luis

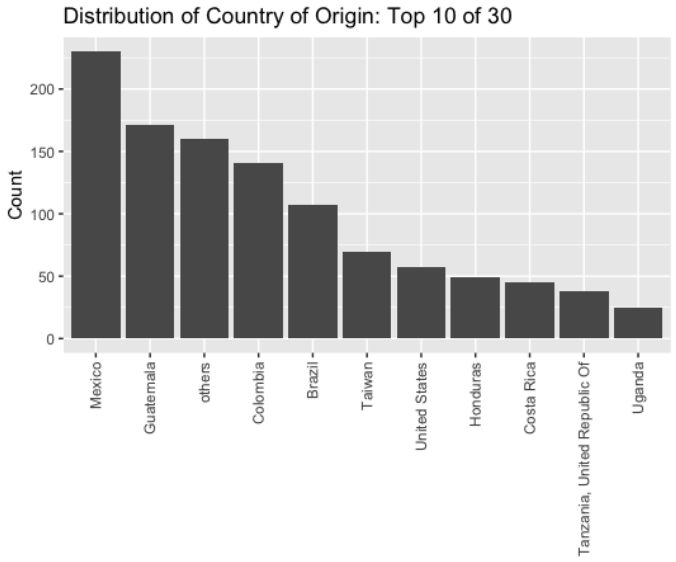
Alvarado Romero 的數量大幅勝過其他人，其餘的九名數量則差不多；第二張圖我們將十名後的所有擁有者加總，得到一個新的「others」，可以發現 others 的數量是遠多於個別前十名的，顯示資料中擁有者是很分散的，有許多規模較小的擁有者。



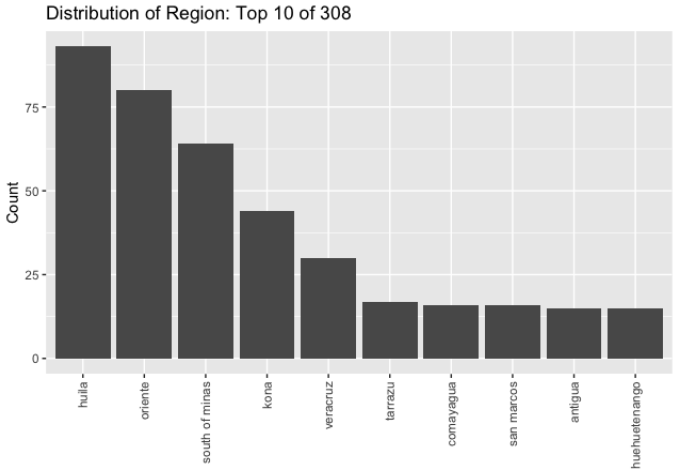


(2) 第一張圖咖啡來源前十名的國家，在 30 個國家中以墨西哥占最多；第二張圖我們一樣將十名後的所有國家加總，可以發現 others 是第三多的，顯示在國家排行裡，墨西哥與瓜地馬拉生產的數量極多，即便是前十名外的加總，依舊是少於這兩個國家。

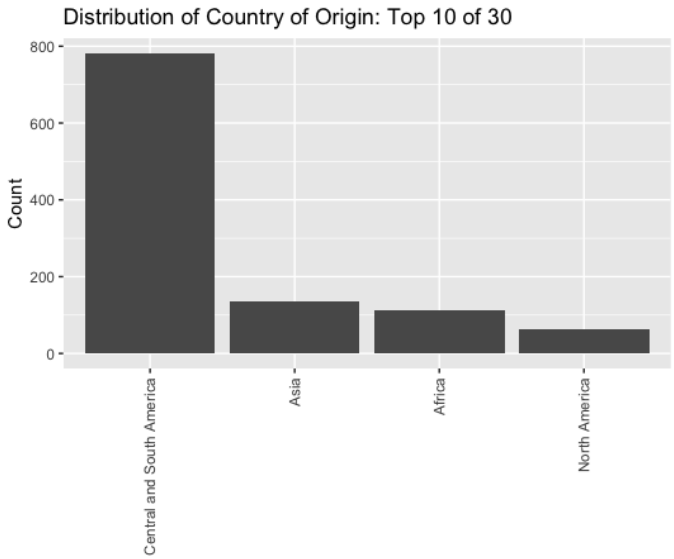




(3) 生產城市前十名所在的國家有哥倫比亞、瓜地馬拉、巴西、夏威夷、墨西哥、哥斯大黎加、宏都拉斯，這些國家也均有出現在上圖的排名中。



(4) 在我們額外新增的「洲別」變數，可以發現蒐集的咖啡來源以中南美洲為最大宗，其次是亞洲與非洲，北美洲則最少。

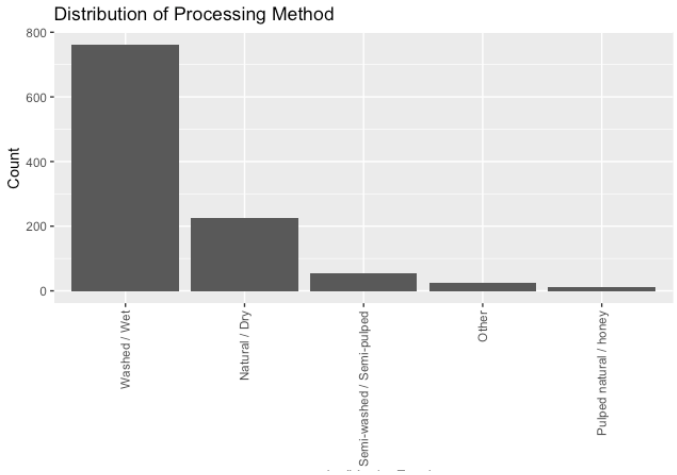


(5) 綜合以上三個變數，我們額外畫出各個國家在地理上的分佈［點下圖可另開連結自由縮放觀察地理位置的分佈］。

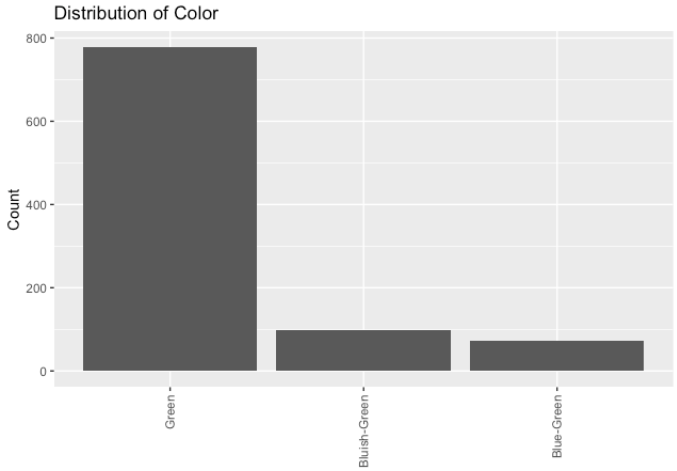
(6) 最常見的加工方法是水洗處理法(washed / wet)，其次是日曬處理法(natural / dry)。

水洗法的過程複雜且繁瑣，但生產出來的咖啡豆雜質較少、外觀較為完整，從數據中可看到使用水洗法的國家中以瓜地馬拉、哥倫比亞與墨西哥占最多數，在這些特別講究生產高品質的咖啡的產地多會利用此方法。

相較於水洗法，日曬法處理的過程較簡單成本也較低，不需要投入太多的工具與設備，且幾乎不需要用到水，在水資源不豐富與較不富裕的地區被廣泛使用，數據中可看到使用日曬法的國家中巴西的占比超過三成，也驗證了因當地山區缺水，故難以使用水洗法。

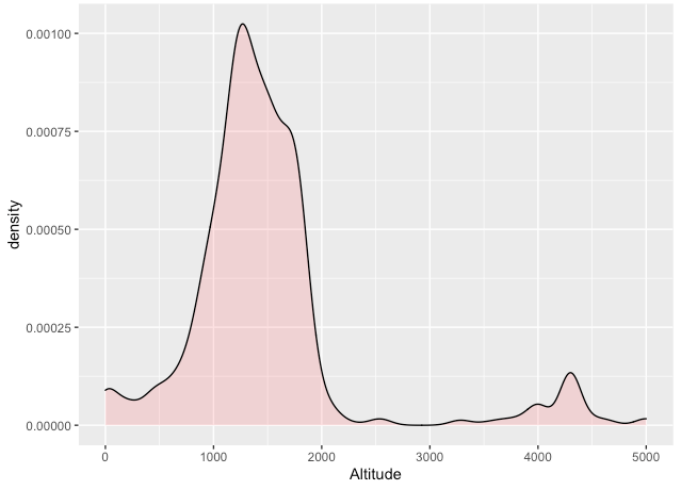


(7) 大部分的咖啡豆都是綠色及藍綠色的，可見蒐集的咖啡豆大多都是新鮮的。

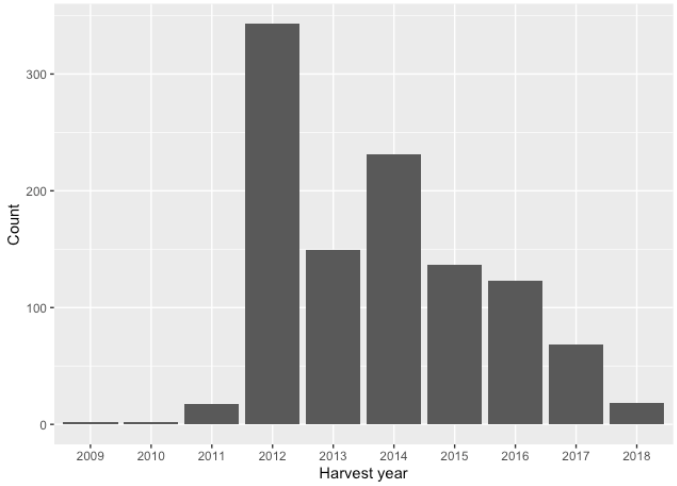


(8) 大部分咖啡都種植在海拔 1000 公尺到 2000 公尺的地方，2000 公尺到 4000 公尺幾乎沒有種植，但是在 4000 公尺以上又有發現少部分的咖啡。

通常海拔越高，因為溫度低所以咖啡生長速度緩慢，使得咖啡酸度更高，香氣和風味也更佳。我們發現在 4000 公尺以上的咖啡是來自於瓜地馬拉的新東方(Oriente)產區，新東方高原咖啡產區內氣候多雨，座落在火山帶，使用水洗法加工，是瓜地馬拉七大產區之一。



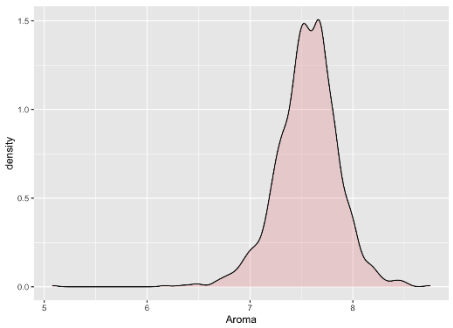
(9) 此次分析之咖啡多是收成於 2012 到 2016 年。

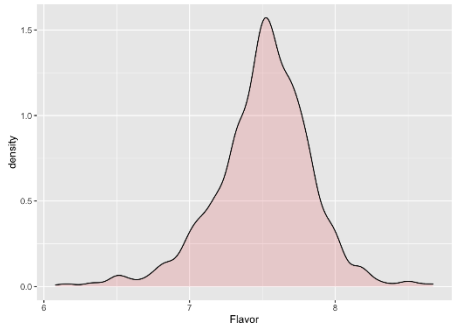


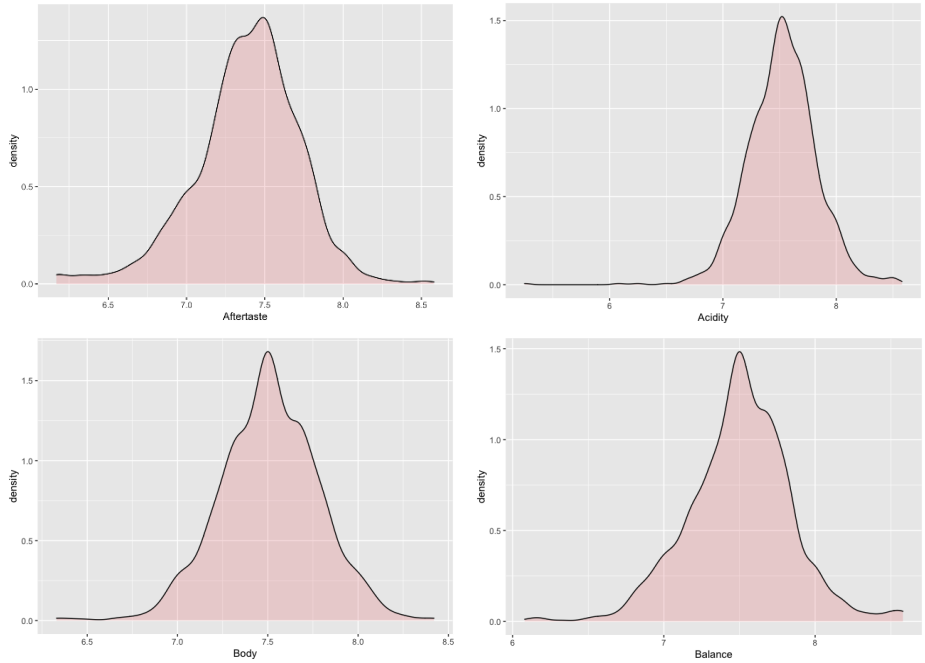
3. 應變數分析

此次研究分析之咖啡品質評分，結果如下：

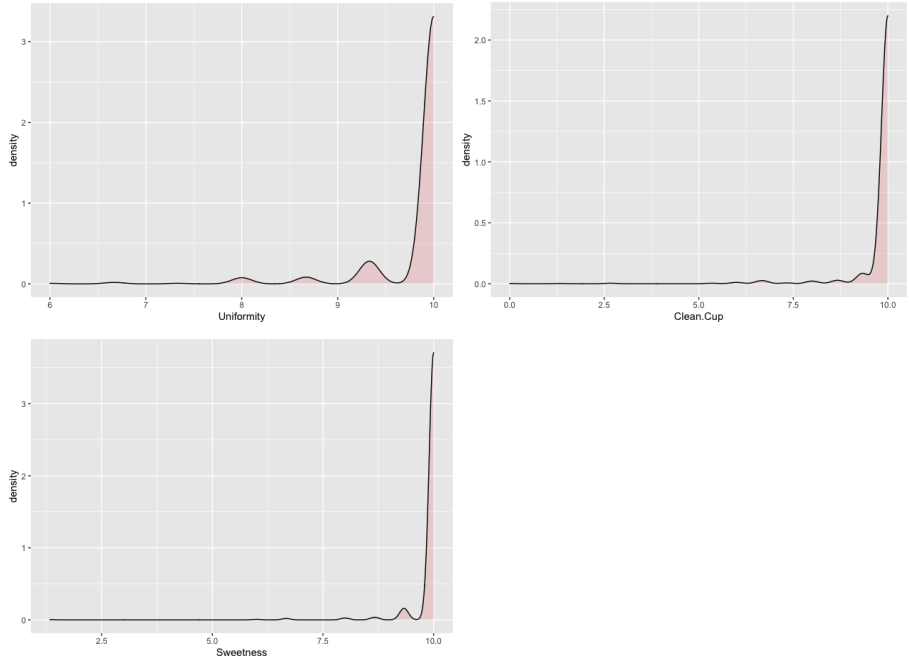
(1) 在香氣、風味、餘韻、酸度、醇厚度和平衡性的分數，大部份咖啡都落在 6.5～8.5，平均分數則是 7.5 左右。



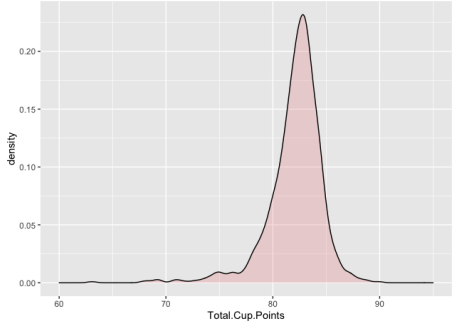




(2) 一致性、乾淨度和甜度的分數則是幾乎都是 10 分。



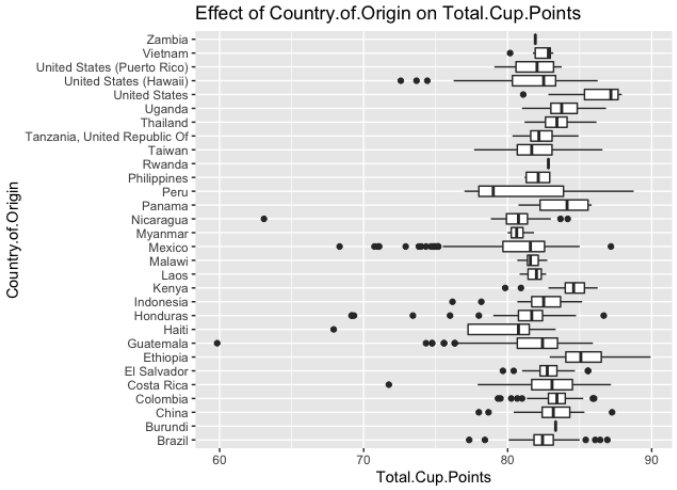
(3) 整體分數則是落在 77～88 之間。



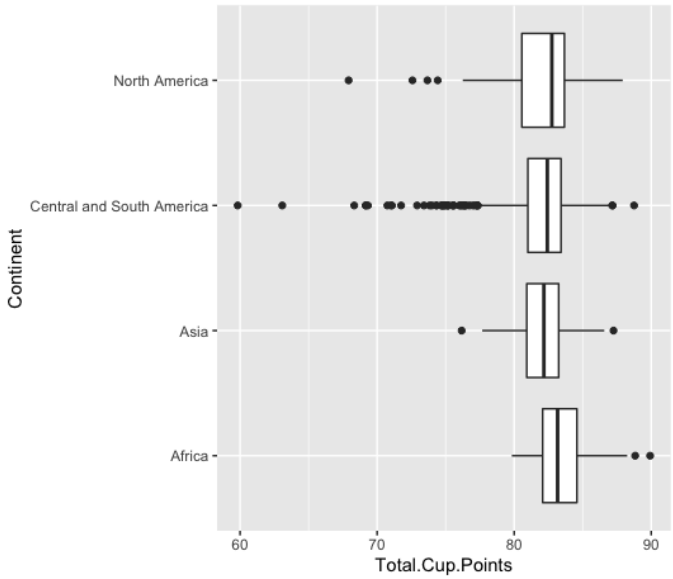
4. 關聯性分析

(1) 由整體分數與自變數的圖，可以發現：

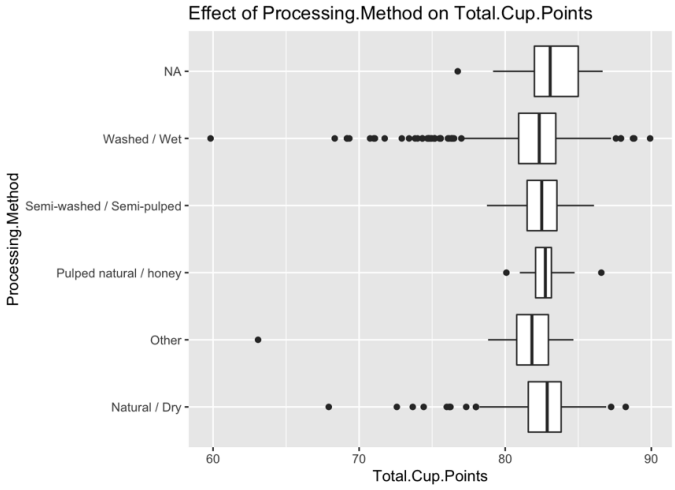
- 來自衣索比亞(Ethiopia)的咖啡分數較其他國家高。

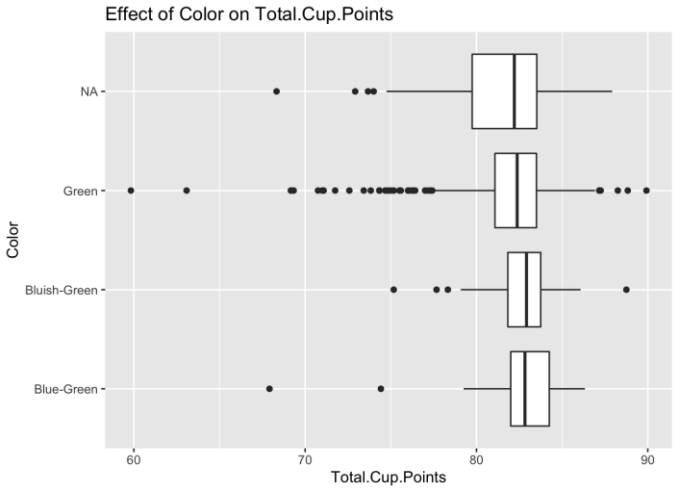


- 改用洲別來觀察，可以發現非洲的整體分數較其他洲高，中南美洲則是有較多低分的離群值，推估可能是因為數量較多的關係。

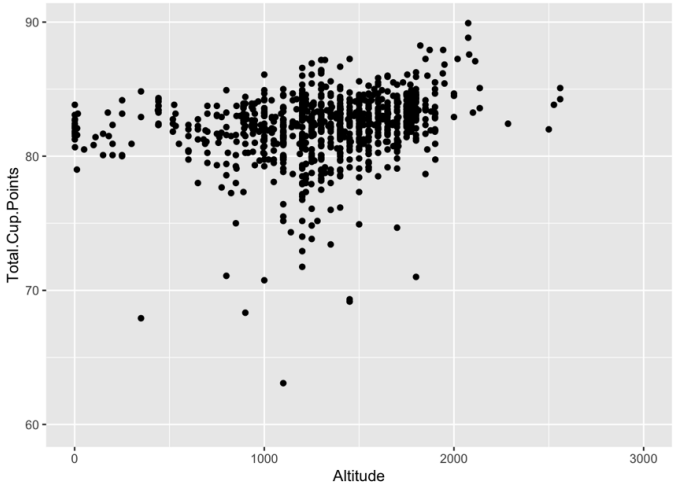


- 不同的加工方法和咖啡豆顏色在分數上相近，可得知這兩個變數對於分數影響不大。

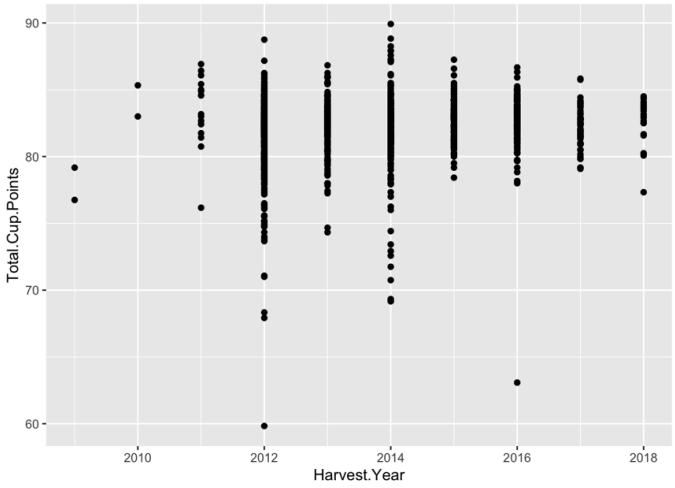




- 在海拔 2000 公尺以下，栽種高度越高整體分數也越高，符合前面查詢的資料結果。

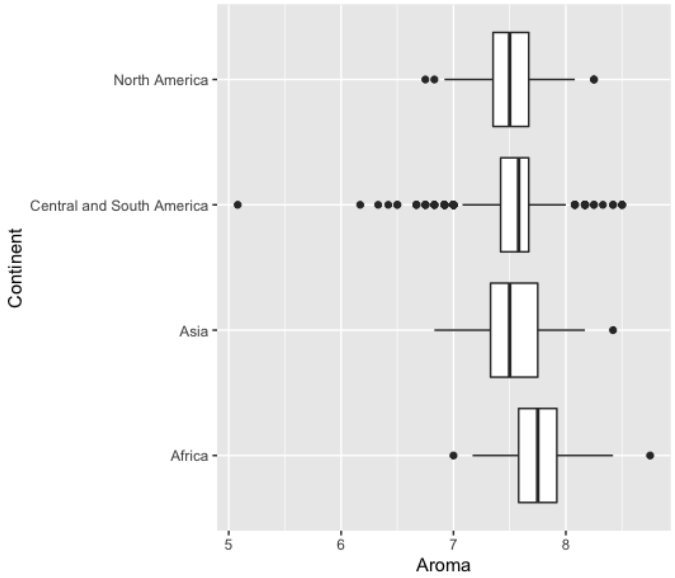


- 收成年份與分數並無太大的關聯性。

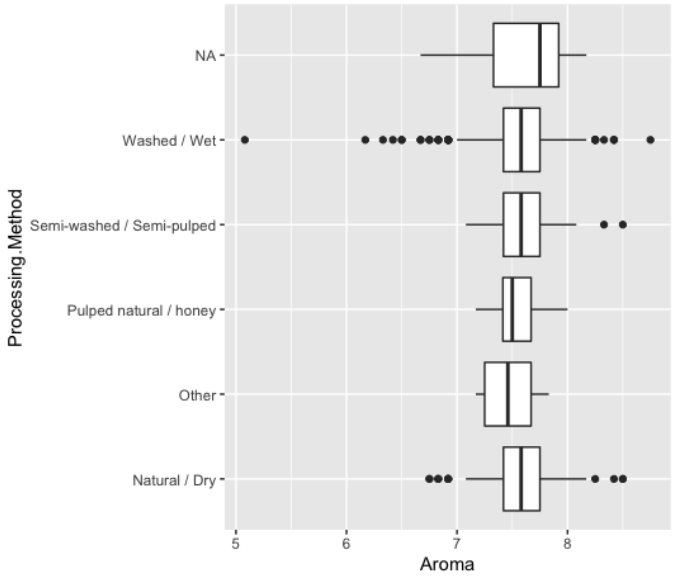


(2) 由香氣與自變數的圖，可以發現：

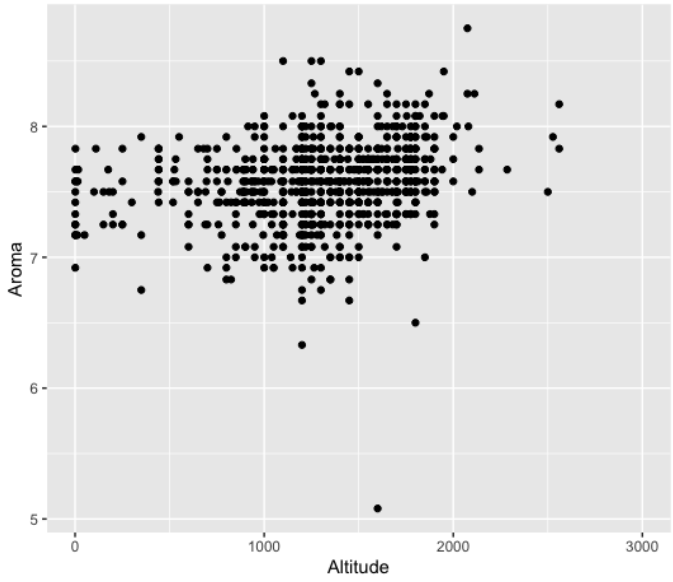
- 香氣的分數非洲最高，其餘三個洲則相差不大。



- 水洗法在風味評比差距極大，其餘方法分數則相差不大。

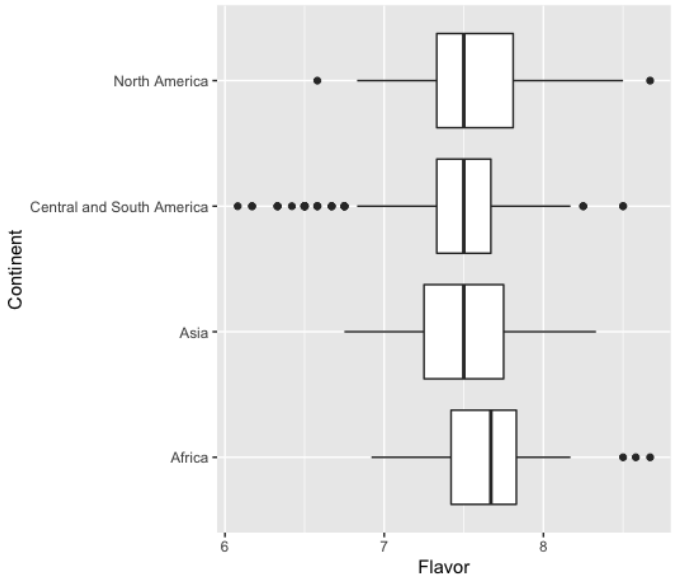


- 大部份分數集中在 7～8 分之間，在海拔 2000 公尺左右的地方有零星幾筆高於 8 分的數據。

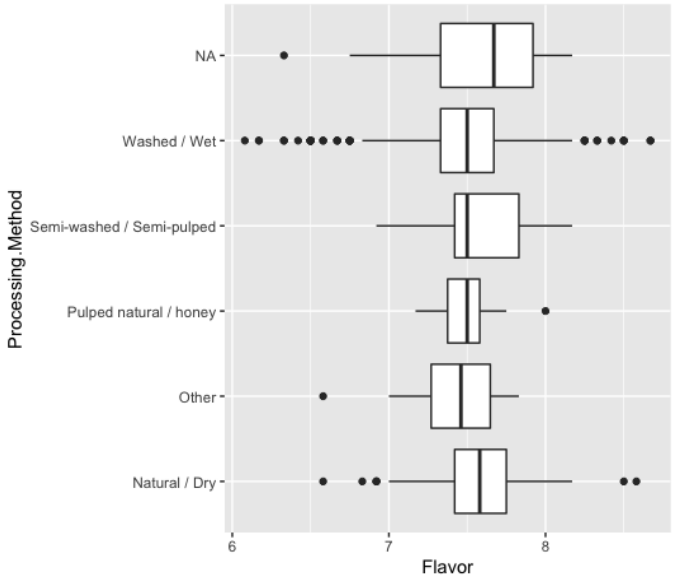


(3) 由風味與自變數的圖，可以發現：

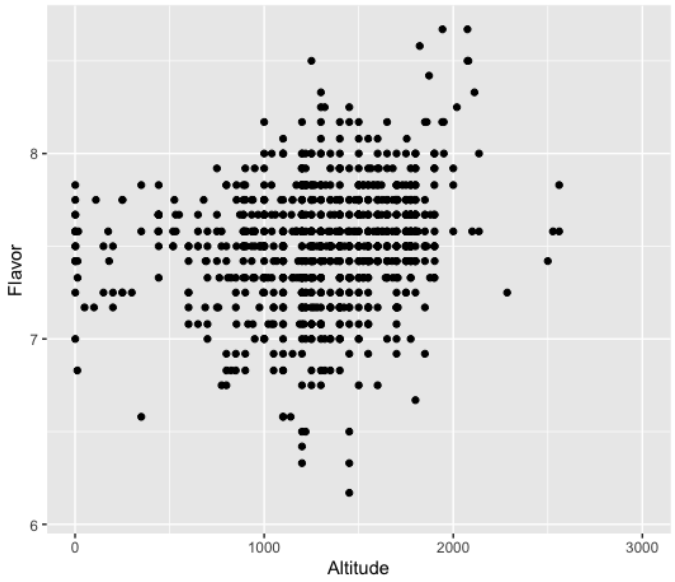
- 風味的分數非洲的中位數略高於其餘三洲，但差距不大。



- 水洗法在風味評比差距極大，日曬法也有一些離群值但中位數是最大的，其餘方法分數則相差不大。

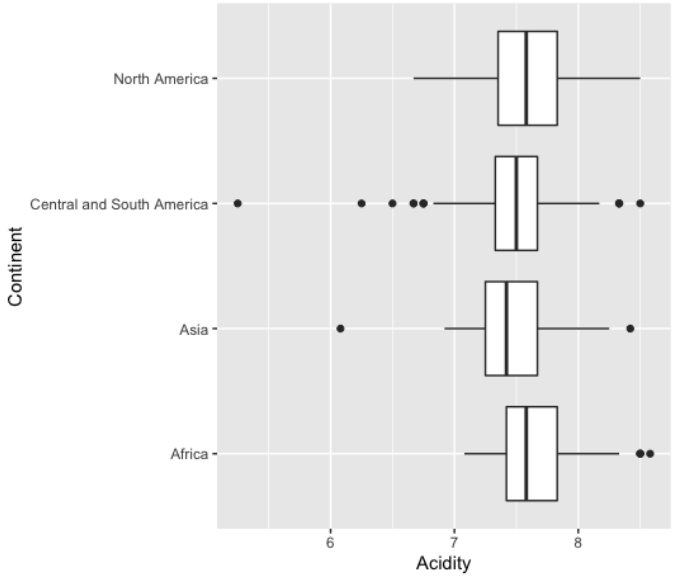


- 大部份分數集中在 7～8 分之間，在海拔 2000 公尺左右的地方有些高於 8 分的數據，低於 1500 公尺有較多低於 7 分的分數。

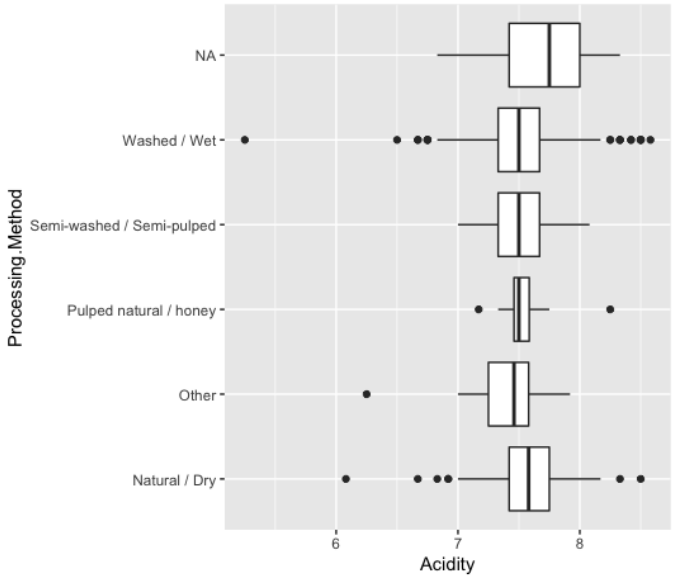


(4) 由酸度與自變數的圖，可以發現：

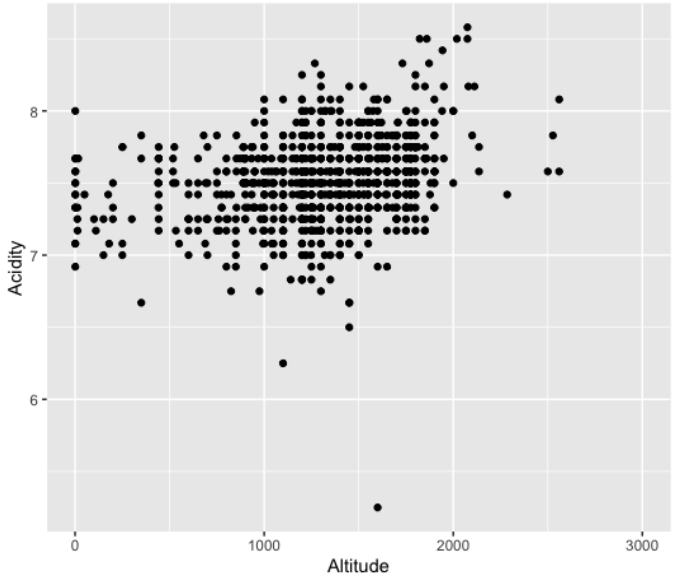
- 酸度分數四個洲差距不大，只有北美洲的分佈級距較大。



- 日曬法中位數最大，整體分佈也較其他方法高。



- 大部份分數集中在 7～8 分之間，在海拔 2000 公尺左右的地方有零星幾筆高於 8 分的數據，分佈和香氣圖極相似。



四、模型預測

1. 在模型預測，我們觀察了以下兩個模型：

（1）Full model：

針對建設模型，因 Owner、Country of origin 及 Region 的變數數量太多且都是離散變數，因此我們先做了以下處理：

- Owner 變數裡取前十大主要擁有者，其他設為 Others。
- 自變數裡的來源國家(Country of origin)與來源地(Region)以新增的洲別(Continent)變數取代，避免變數裡的分類過多，導致 overfitting，模型沒有意義。

```
> lm1 = lm(Total.Cup.Points ~ Owner + Continent +
   Processing.Method + Color + Altitude + Harvest.Year, data =
   data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-21.5945  -0.9872   0.2115   1.3450   6.9314

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                -2.346e+02  1.097e+02  -2.140 0.032648 *
OwnerCARCAFE LTDA CI                        2.201e-01  8.686e-01   0.253 0.800054
OwnerCQI Taiwan ICP CQI                    -1.311e+00  8.025e-01  -1.633 0.102733
OwnerExportadora de Cafe Condor S.A         4.434e-01  7.925e-01   0.560 0.575928
OwnerIpanema Coffees                       -1.231e+00  7.931e-01  -1.552 0.121033
OwnerJuan Luis Alvarado Romero             -2.207e-01  6.265e-01  -0.352 0.724703
Ownerlin, che-hao                          -7.366e-01  8.324e-01  -0.885 0.376441
OwnerNUCOFFEE                               4.677e-01  8.804e-01   0.531 0.595385
OwnerOthers                                -7.407e-01  5.904e-01  -1.255 0.209943
OwnerRacafe & Cia S.C.A                     4.345e-01  8.010e-01   0.542 0.587696
ContinentAsia                              -1.271e+00  3.764e-01  -3.376 0.000764 ***
ContinentNorth America                     -1.934e+00  2.776e-01  -6.966 6.06e-12 ***
ContinentSouth America                     -1.077e+00  4.281e-01  -2.515 0.012053 *
Processing.MethodNatural / Dry             -6.392e-01  1.053e+00  -0.607 0.544166
Processing.MethodOther                     -1.678e+00  1.149e+00  -1.460 0.144730
Processing.MethodPulped natural / honey    -4.322e-01  1.334e+00  -0.324 0.745998
Processing.MethodSemi-washed / Semi-pulped -5.309e-01  1.091e+00  -0.487 0.626649
Processing.MethodWashed / Wet              -1.146e+00  1.031e+00  -1.111 0.266716
ColorBlue-Green                             1.291e-01  4.325e-01   0.298 0.765395
ColorBluish-Green                          -7.086e-02  4.138e-01  -0.171 0.864072
ColorGreen                                 -5.581e-01  3.066e-01  -1.820 0.069068 .
ColorNone                                  -1.469e+00  4.788e-01  -3.069 0.002209 **
Altitude                                   -2.775e-06  1.135e-05  -0.244 0.806961
Harvest.Year                                1.590e-01  5.452e-02   2.916 0.003626 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.461 on 955 degrees of freedom
  (113 observations deleted due to missingness)
Multiple R-squared:  0.1353,    Adjusted R-squared:  0.1145
F-statistic: 6.498 on 23 and 955 DF,  p-value: < 2.2e-16
```

```
> anova(lm)
Analysis of Variance Table

Response: Total.Cup.Points
                   Df Sum Sq Mean Sq F value    Pr(>F)
Owner               9  273.3  30.362  5.0126 1.274e-06 ***
Continent           3  418.3 139.436 23.0199 2.159e-14 ***
Processing.Method   5   43.9   8.781  1.4497 0.2039146
Color               4  118.3  29.564  4.8808 0.0006718 ***
Altitude            1    0.0   0.036  0.0059 0.9387374
Harvest.Year        1   51.5  51.514  8.5045 0.0036256 **
Residuals         955 5784.6   6.057
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

（2）Reduced model：

我們從加入所有自變數的模型裡刪去不顯著的變數(Owner、Processing. method 與 Altitude)，想要了解 reduced model 相比 full model 是否較值得，因為基於模型的簡約原則，如果增加一大堆解釋變數後的模式 fit 只進步一點點，則不值得用複雜的模式。

```
> lm2 = lm(Total.Cup.Points ~ Continent + Color + Harvest.Year,
   data = data)
```

```
> summary(lm)

Call:
lm(formula = Total.Cup.Points ~ Continent + Color + Harvest.Year,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.3903  -1.0502   0.2797   1.4191   6.8824

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)            -127.22519   92.76075  -1.372  0.17049
ContinentAsia            -1.30579    0.32135  -4.063 5.19e-05 ***
ContinentNorth America   -1.84595    0.25962  -7.110 2.10e-12 ***
ContinentSouth America   -0.40424    0.28538  -1.417  0.15692
ColorBlue-Green           0.22497    0.39306   0.572  0.56721
ColorBluish-Green         0.21344    0.36493   0.585  0.55874
ColorGreen               -0.36187    0.27585  -1.312  0.18986
ColorNone                -1.42397    0.44646  -3.189  0.00147 **
Harvest.Year              0.10470    0.04607   2.272  0.02326 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.503 on 1081 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:   0.11,     Adjusted R-squared:  0.1034
F-statistic:  16.7 on 8 and 1081 DF,  p-value: < 2.2e-16
```

```
> anova(lm)
Analysis of Variance Table

Response: Total.Cup.Points
              Df Sum Sq Mean Sq F value    Pr(>F)
Continent      3  680.1 226.691 36.1957 < 2.2e-16 ***
Color          4  124.5  31.129  4.9704 0.0005673 ***
Harvest.Year   1   32.3  32.340  5.1637 0.0232600 *
Residuals   1081 6770.2   6.263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. 從上面結果可以發現兩個模型的 p-value 跟 Adjusted R-squared 相差都不大，因此我們改做 partial F-test 來觀察兩個模型是否有顯著差異。

```
> state.lm1 = dyn$lm(Total.Cup.Points ~ Owner + Continent +
  Processing.Method + Color + Altitude + Harvest.Year, data =
  data)
> state.lm2 = dyn$lm(Total.Cup.Points ~ Continent + Color +
  Harvest.Year, data = data)
> anova(state.lm1, state.lm2)
```

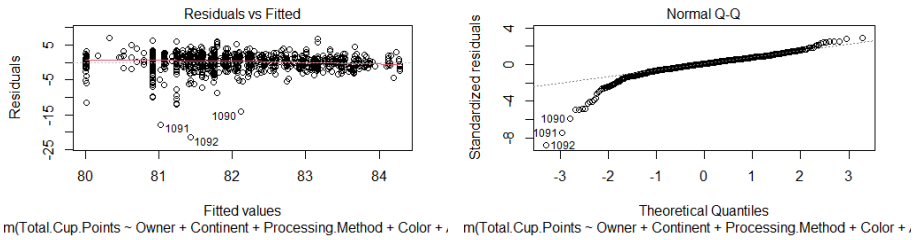
```
Analysis of Variance Table

Model 1: Total.Cup.Points ~ Owner + Continent + Processing.Method + Color +
    Altitude + Harvest.Year
Model 2: Total.Cup.Points ~ Continent + Color + Harvest.Year
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    955 5784.6
2    970 5942.9 -15   -158.22 1.7414 0.0386 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

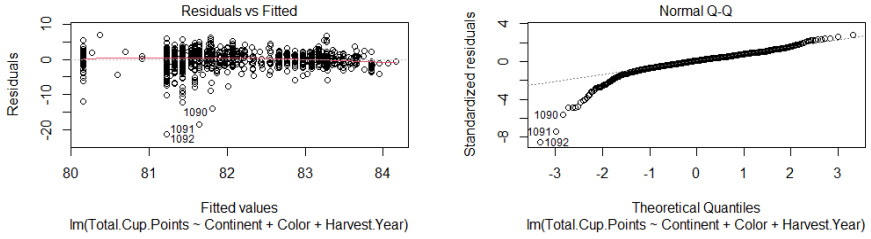
從結果可以觀察到 full model 的殘差平方和(RSS)較小，且 p-value＝0.0386＜0.05，因此拒絕虛無假說（H_0:假設從完整模型中刪除的所有係數都為零），故結果顯示 full model 在模型的擬合度是較好的。

3. Plot

（1）Full model：



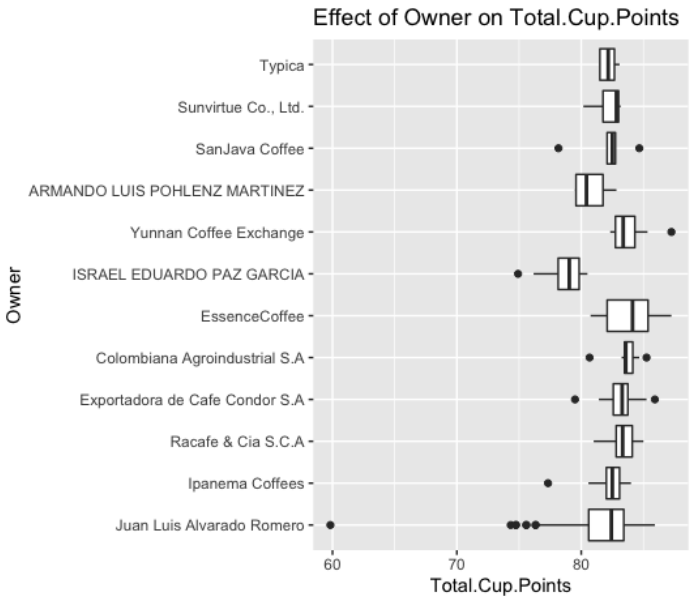
（2）Reduced model：



兩個模型的殘差圖和 Normal Q-Q plot 相似，都沒有很符合常態分佈，看不太出哪一個模型比較好。

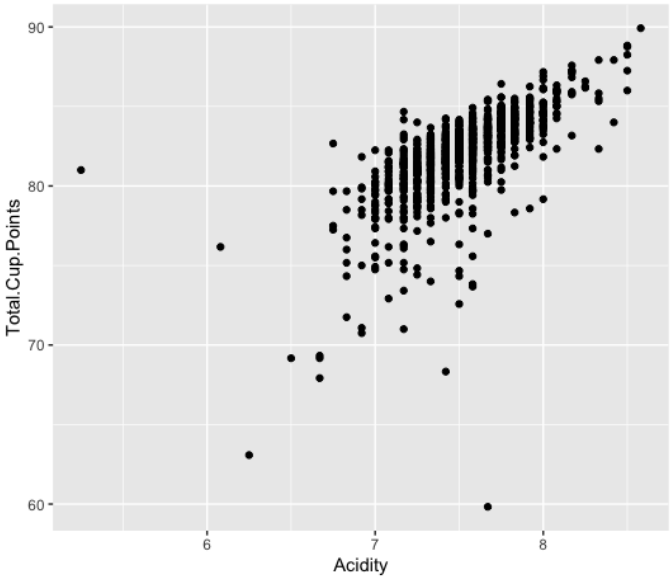
五、討論與結論

1. 我們額外分析了 owner 數量與整體分數的盒狀圖：盒狀圖最下面的四筆為擁有咖啡數量前四多的人，中間四筆為擁有咖啡數量中段的四人(個數 12 個)，最上面四筆為擁有咖啡數量較少的四人(個數 5 個)；圖中 owner 的分數差距其實不大，顯示擁有咖啡數量最多的 owner 其咖啡分數並沒有明顯高於咖啡數量少的人。



2. 除了海拔高度外，原先我們想加入溫度、雨量、日照分析關聯性，但因region裡許多資料是小城鎮，難以搜尋到天氣資訊，若用國家去分析又有失精準，因此最後沒有更深入探討天氣與咖啡品質的關聯性。

3. 我們十分好奇酸度與整體分數的相關性，究竟越酸的咖啡品質是否也越好呢？因此我們額外分析了酸度與整體分數的散佈圖，發現兩者呈現正相關，也驗證了整體分數越高的咖啡，通常其香氣、風味與酸度也會越高。



4. 本研究所蒐集數據的產地數量以中南美洲為最多，也符合阿拉比卡咖啡產量的排名，因此在分析上與實際的落差推估應不會太大。

5. 在國家部份，我們發現雖然台灣數量是第五多的，但在城市的前十名卻

沒有出現台灣的城市，重新觀察數據後，發現在資料中台灣的區域很分散，在嘉義、彰化、雲林等都只有零星幾筆資料；但我們搜尋資料發現雲林古坑及嘉義阿里山都算是台灣咖啡主要的產區，推估可能是原資料來源蒐集數據沒有很完整的緣故。

6. 加工方法部份，一般而言使用水洗法處理的咖啡豆品質會優於使用日曬法處理的，但在我們分析的整體分數上發現兩個的分數差距不大，然而個別分析香氣、風味及酸度時，又顯示水洗法分數是較優的，因此我們推測可能是有其他的因素影響了水洗法的整體分數。

7. 海拔高度部份，一般而言海拔越高，咖啡酸度會越高，香氣和風味也更佳，而我們的分析結果亦符合此結論。

8. 模型預測部份，使用改良過的全部變數的模型(full model)優於刪減掉不顯著變數的模型(reduced model)，但兩個模型結果並沒有差距太大，推測可能是因為我們有先對變數做過處理導致結果不明顯，惟若不事先處理，因自變數數量均為離散型且數量太多，分析結果會無意義。