大數據分析方法
期末書面報告

吃貨就是要了解美食

組員：

B0544203 石家安

B0544215 王姵云

B0544224 樊　驊

B0544225 陳勗恩

指導老師： 曾意儒 老師

一、動機

在長庚待了三年，即將邁入第四年了，不僅學校裡面沒什麼東西吃，連外面的「桃園市」都被大家稱之為：「美食沙漠」，因此我們想要知道，桃園的美食數量是不是真的相較於其他5個直轄市較少，同時也想知道桃園和臺北市的熱門美食店家有哪些。

再者，我們也好奇當上網打出「熱門美食」的關鍵字時，就會跑出各種看起來非常好吃的食物的照片，但那真的是受到大家喜愛的「熱門」的美食嗎？還是那只是透過修圖技巧拍出來的「照騙」或一種網路行銷宣傳的手法呢？

此外，當晚上 10 點一過，只要 PO 出任何跟食物有關的照片，就一定會被身邊的好友撻伐說：「怎麼可以發消夜文？！」因此我們還想知道，PO 出美食文的時間點，會不會間接地影響到該篇文的按讚數量。

最後，當我們追蹤某一個專門分享美食的 Instagram 帳號時，有一天突然因為葉配的關係，PO 出一張自己戴眼鏡的照片，又或是平常 PO 出各種生活自拍照，心裡就會跑出一種「恩？我追蹤的不應該是推薦美食的帳號嗎？怎麼變成個版了呢？」的想法，因此我們也想要知道，PO 出自拍照，和其他的因素，會不會影響到該位美食 Instagram 經營者 PO 文的讚數。

二、資料介紹

Insta stalker： https：//www.instastalker.net/（代替 Instagram 網站），並利用爬蟲技術將資料抓下來。

1. 選擇原因：
   A. Instagram 的 API 是要申請且需經過審核，因此我們無法在短時間內使用 Instagram 來分析，所以我們找了相似的網站，該網站會儲存所有公開帳號的貼文。
   B. Insta stalker 會顯示詳細的時間，IG 只會出現日期，而且 Insta stalker 的貼文會按照時間排序從新的到舊的排序。

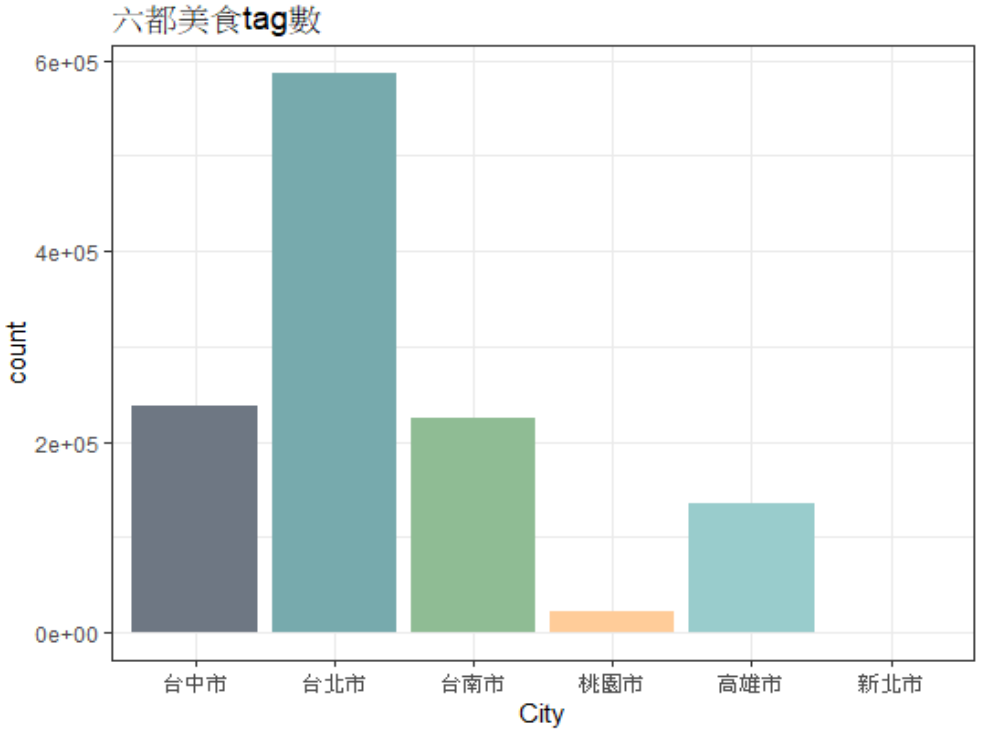
三、分析議題

1. 大家都說桃園市美食沙漠，真的是這樣嗎？
   A. 分析過程：

      此分析將以台灣 6 都來做比較，若桃園的美食 tag 數在六都最後一名，那桃園就是美食沙漠。首先搜尋六都個別的美食 tag 後，爬取所有的美食相關 hashtag 並加總算出每一個直轄市的所有美食 tag 數。(分析一)

      但由於各縣市的人口數也可能會影響到個別的美食 tag 數，例如：台北市人口最多所以可能造成台北市被 tag 較多，所以我們進一步再將所有跟六都相關的 hashtag 總數算出後當作分母，各都美食 tag 數當作分子來算出個個比例再來做一次比較。(分析二)

B. 分析結果：



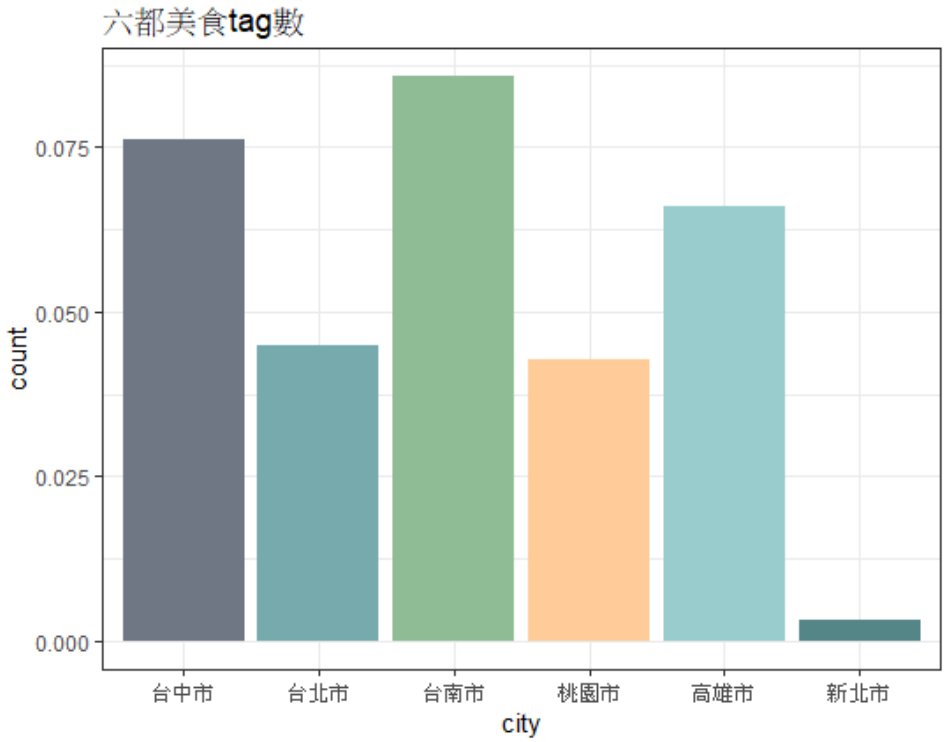
圖一 六都美食 tag 數

在第一個分析階段的結果可以從圖一得知，台北市的美食 tag 數量是明顯比其他五都高出許多，而新北市則落在最低的位置，而此次我們要探討的桃園市則是落在第二低的。

六都	美食 tag	總 tag
台中市	237,009	3111,223
台北市	586,390	13040,209
台南市	223,840	2610,502
桃園市	21,422	501,713
高雄市	135,444	2052,156
新北市	500	162,697

表一 六都總相關 tag 數與美食 tag 數

從表一中可以看出新北市的總 tag 數在六都中視最低的，而桃園市則位在第二低，台北市的相關總 tag 數則是在六都中最多的，被 tag 的數量遠遠高於其他直轄市。



圖二 六都美食與總相關 tag 數比例圖

而從表一中的數據來繼續繪製我們於第二個階段的長條圖，可以從圖二中清楚看到台北市的比例算出來後變成明顯低於台中、台南、高雄等直轄市，而桃園與其他直轄市的差距則是明顯縮小，但還是處在第二低的位置，則新北市則是和上一張長條圖一樣占有最少的比例。

C. 分析限制：

從圖一中，得到的結果是新北市最少 tag 數，但是我們認為新北市會是最少的原因為大部分的使用者在用 hashtag 時並不會特別標註「新北市」而是只會單純標註「台北市」，所以造成新北市分析結果比實際上的美食數量來要低。

而圖二中台北市的比例明顯降低的原因可能為會標註台北地區的文章總數量太多，造成比例降低；桃園地區明顯降低和其他縣市的差距的原因則是比較少人會去標住桃園的其他東西，所以造成算出來的比例提高。整體來說，除了剛剛提到的新北市 tag 問題外，桃園市還是在最後的分析中佔有最低的比例，這也代表桃園在六都中可以說是真正的美食沙漠。

2. 台北及桃園的美食景點趨勢

A. 分析過程：

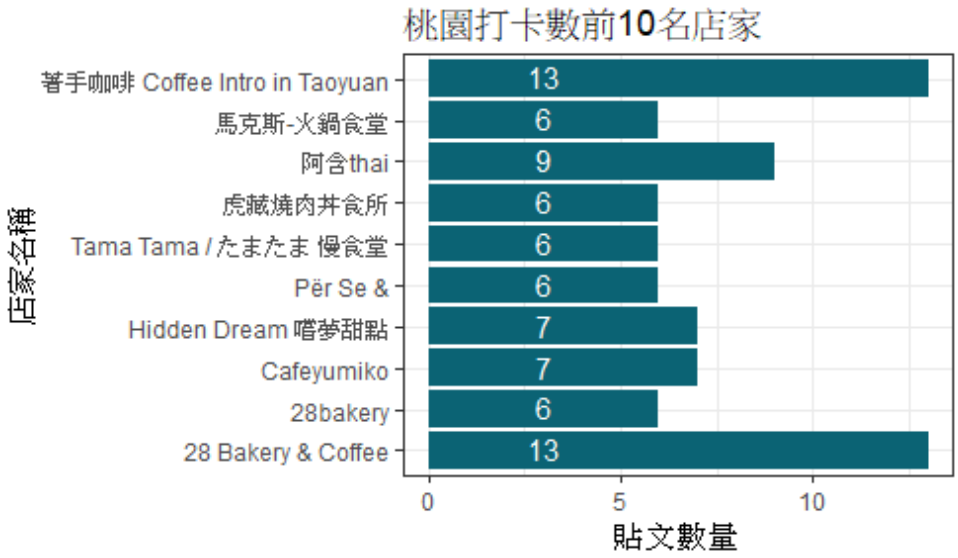
在 hashtag 關鍵字中找哪先是跟 taipei 以及 taoyuan 有關，會看到像是 taoyuanfood、taoyuan_food、taoyuanfoodie、taipeicofe 等等的

hashtag 關鍵字。

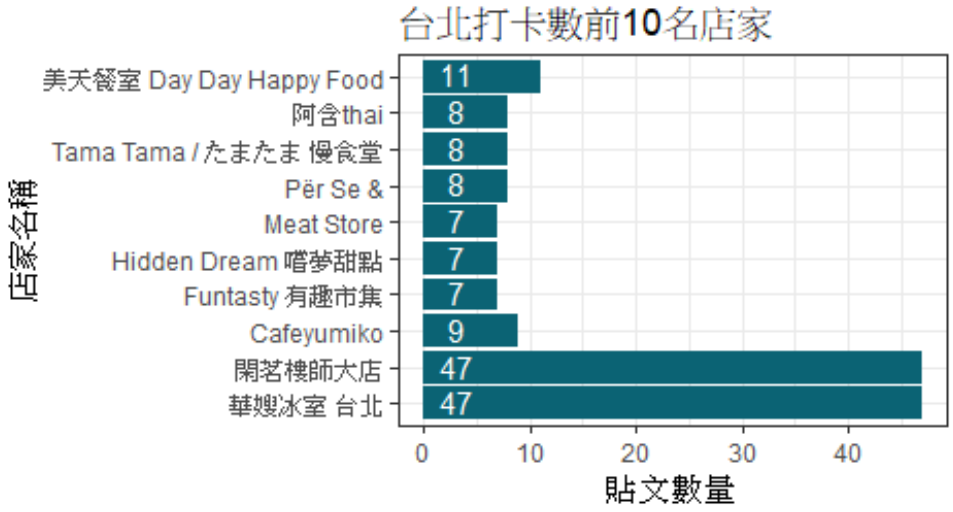
將這些關鍵字相關的貼文裡的數據擷取下來，有發文帳號、時間日期、地點、hashtag，放入陣列裡面。

利用 table 以及 arrange 將擷取到的地點排序，看那一個地點的被發文數最多。最後在手動刪除商業帳號。

B. 分析結果：



圖三 桃園打卡店家前 10 名



圖四 台北打卡店家前 10 名

C. 分析限制：

- hashtag 會被亂用：有些貼文根本與美食不相關，但是卻 tag taipeifood，或是地點明明是在桃園的店家卻 tag 台北。
- 有許多店家有自己的帳號，會一天發很多篇文章來增加曝光度，只能用手動的方式刪除。
- 如果有一篇貼文用了所有 hashtag，店家就會被重複很多次。像是一

篇文章若有兩個跟美食有關的 hashtag 那他就會被重複計算到。每篇文章的相關 hashtag 數量又不同，很難用平均去計算這篇貼文到底出現了多少次。

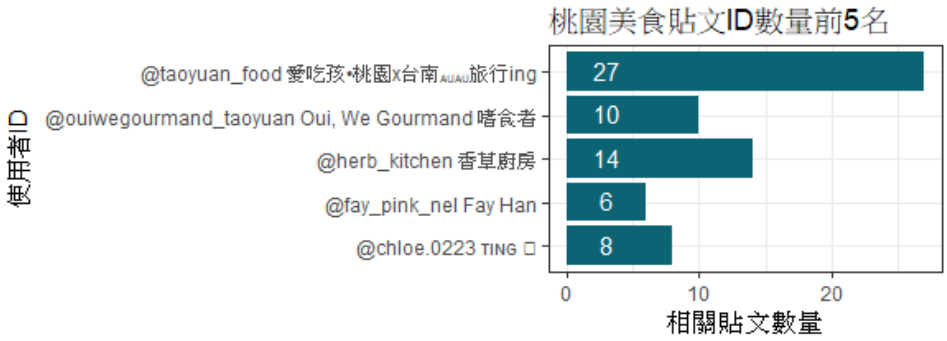
3. 判斷 po 出美食文章的使用者，是不是專業的 Instagram 美食平台經營者

A. 分析過程：

以美食相關的 hashtag 為依據搜尋貼文，找尋發相關貼文的帳號，將發文帳號計算次數後由大到小排序，並且將商業帳號去除。

B. 分析結果：

圖五 台北-前 5 名經常發美食貼文的 ID



圖六 桃園-前 5 名經常發美食貼文的 ID

4. PO 文的時間和按讚數有無相關

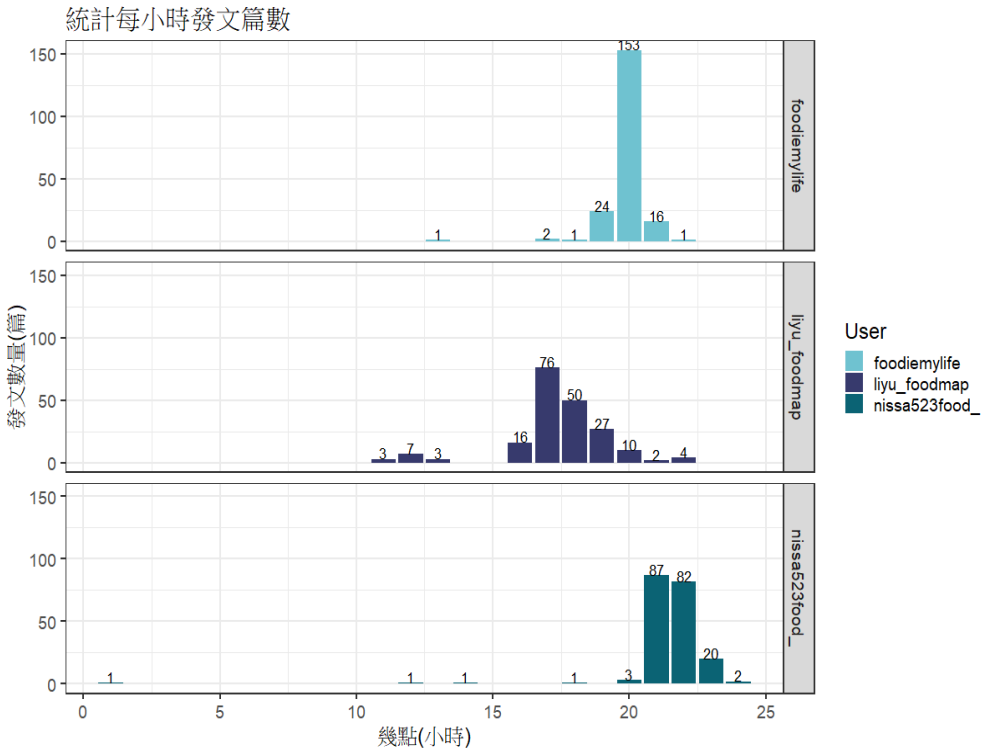
A. 分析過程：

首先，我們選取了三位粉絲人數介於 23K 至 25K 的 Instagram 美食平台經營者，分別是「老爺說半夜不要 foodiemylife」、「護理系吃貨 liyu_foodmap」和「CHIEN 簡 nissa523food_」。從這三位 Instagram 美食平台經營者的發文中，按照 PO 文時間順序，各擷取 200 篇貼文，也就是總資料數共 600 篇貼文。

再從每一篇貼文中抓取「發文日期」、「發文時間」及「按讚數」做此題的分析。

B. 分析結果：

我們統計了三位 Instagram 美食平台經營者每個小時發文的數量，以一個小時為單位做統計，如圖七所示。



圖七 統計每小時發文篇數

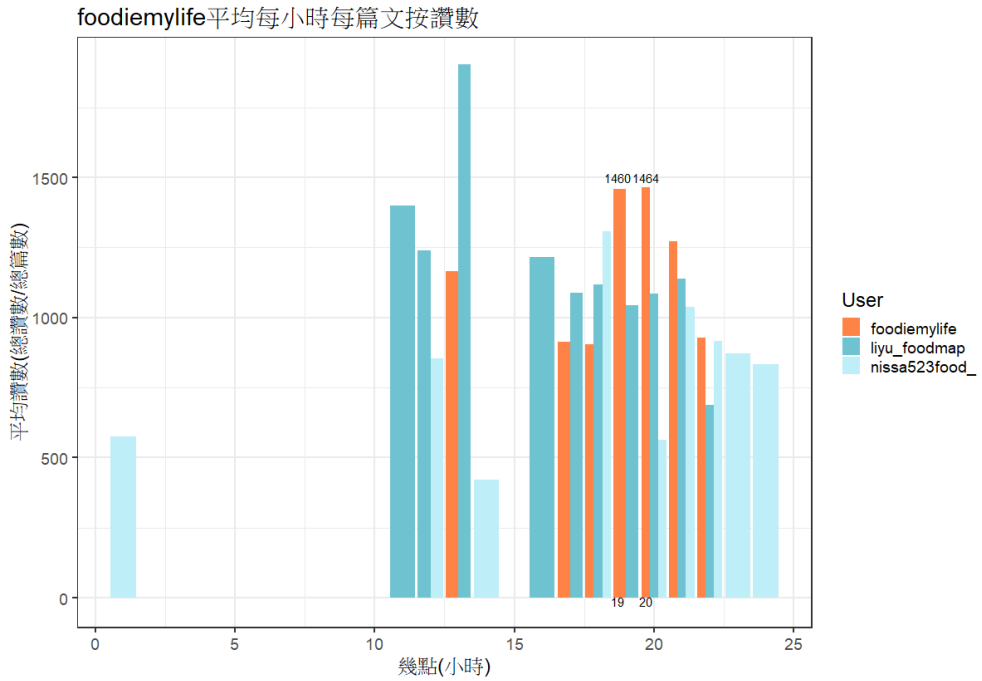
由圖七可知，第一位 Instagram 美食平台經營者「老爺說半夜不要 foodiemylife」，主要在 20 至 21 點間發文。

第二位 Instagram 美食平台經營者「護理系吃貨 liyu_foodmap」，最常在晚上 17 點至 18 點發文，而其他文章的發文時間大概落在 11 點到 22 點之間，而沒有在 14 點和 15 點發文。

最後一位 Instagram 美食平台經營者「CHIEN 簡 nissa523food_」的主要發文時間是晚上 21 點至 22 點，其次為 22 點至 23 點和 23 點至 24 點，比起前兩位 Instagram 美食平台經營者的發文時間更晚。

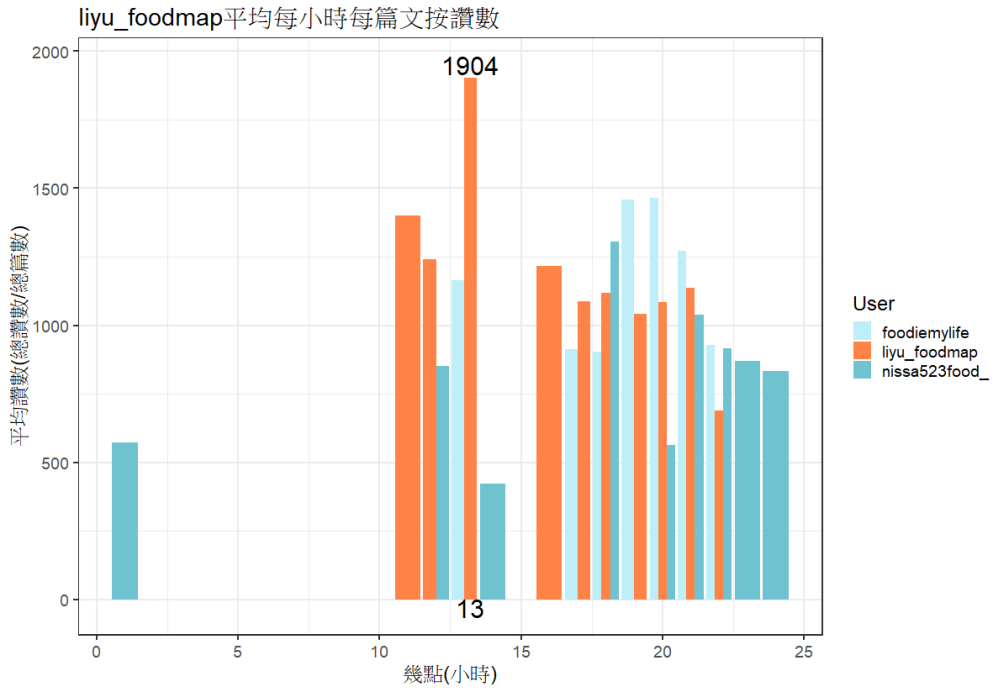
此外，從圖七我們可以看出 Instagram 美食平台經營者的發文時間大多數落在晚上，比較少會在中午 12 點前 po 文。

再者，我們將三位 Instagram 美食平台經營者每篇文章獲得的讚數加總，並除以發文的篇數，以每小時平均每篇文章獲得的按讚數量繪製成圖八、圖九和圖十。



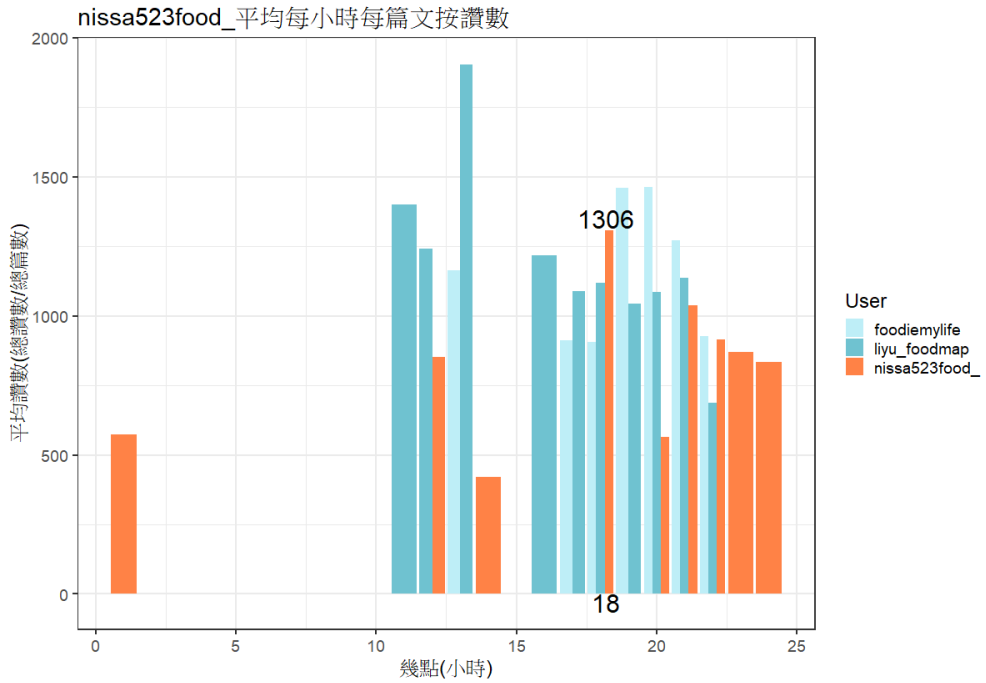
圖八 foodiemylife 平均每小時每篇文按讚數

由圖八可知，第一位 Instagram 美食平台經營者「老爺說半夜不要 foodiemylife」平均每篇文章獲得的按讚數量最高的是 20 點，平均一篇文章有 1464 個讚，再來是 19 點的 1460 個讚。



圖九 liyu_foodmap 平均每小時每篇文按讚數

由圖九可知，第二位 Instagram 美食平台經營者「護理系吃貨 liyu_foodmap」平均每篇文章獲得的按讚數量最高的是 13 點，平均一篇文章有 1904 個讚，甚至比 20 點和 21 點獲得的讚數還高。

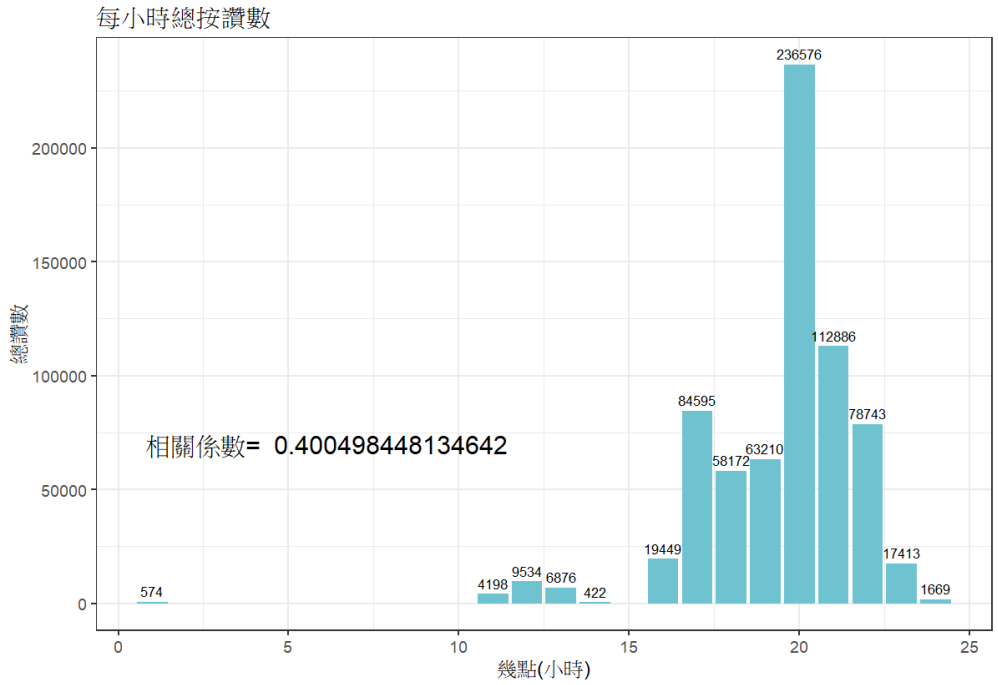


圖十 nissa523food_平均每小時每篇文按讚數

由圖十可知，最後一位 Instagram 美食平台經營者「CHIEN 簡 nissa523food_」平均每篇文章獲得的按讚數量最高的是 18 點，平均一篇文章有 1306 個讚。

因此，如果三位 Instagram 美食平台經營者(「老爺說半夜不要 foodiemylife」、「護理系吃貨 liyu_foodmap」和「CHIEN 簡 nissa523food_」)想要在貼文中獲得較多的按讚數，建議可於 20 點、13 點和 18 點 PO 出文章。

最後，我們統計了所有貼文的按讚數，以每小時做區分，繪製成成圖十一。



圖十一 每小時總按讚數

我們可以看到在晚上 20 點獲得的總按讚數是最多的，高達 236576 個讚。並以小時和獲得的讚數算取相關係數，為 0.400498448134642，證明 PO 文時間和按讚數為中度相關。

5. 讚數和露臉有沒有關係

A. 分析過程：

抓取每篇文章的按讚數，使用 magick 和 image.libfacedetection 套件，偵測照片中有沒有露臉，挑選一位 Instagram(raymond.hou)美食平台經營者進行分析，再者，為考慮到一些可能影響讚數之因子，如：

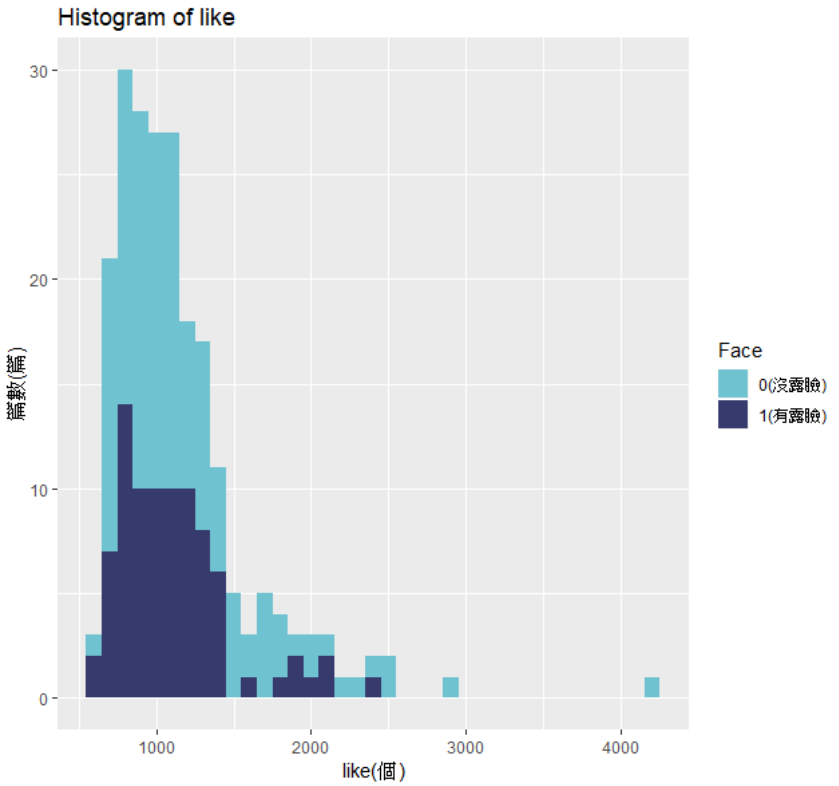
- 發文日：

  考慮到發文者若於資料蒐集日前後發文，讚數可能會因時間不夠長而使此篇文章的按讚數特別少，意為讚數的次數還未穩定，可能仍有許多追蹤者不是因為不喜歡未按讚，而是因為還沒看到此篇文章，所以尚未按讚；此外，考慮到若爬取較早期的文章，可能會因為發文者的追蹤人數有大幅度的不同而影響讚數的多寡，因此決定爬取 2018/09/06~2019/5/29 的文章。

- 抽獎文：

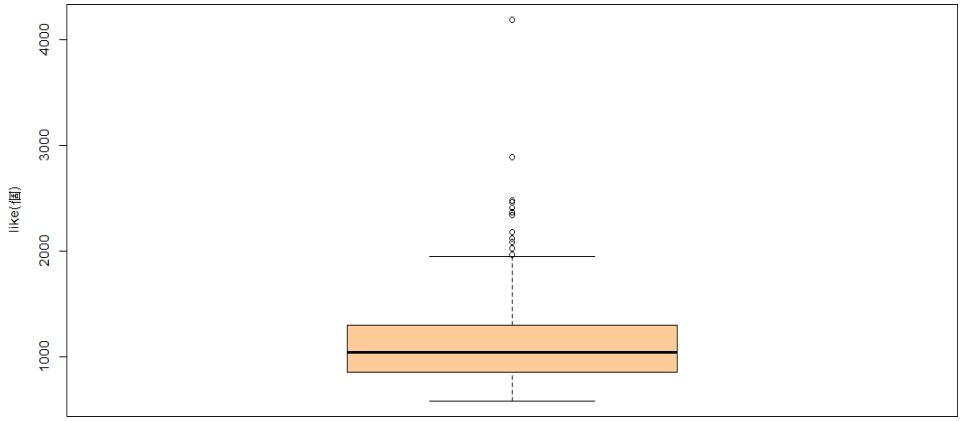
  通常抽獎文章的內容會要求參加抽獎的人，一定要按該文章讚，故抽獎文的讚數通常都會異常的高，再者，我們發現此發文者的抽獎文皆標有『#抽獎』的 hashtag，故為在資料蒐集時，將標有此 hashtag 之資料去除。

由圖十二中看到呈現資料分布，為想確認資料是否為常態分佈，因此使用 shapiro.test function 去判斷，檢驗出 p value = 2.139e-15，小於 0.05 不為常態分佈。但由於我們蒐集的母體數有 216 筆資料，根據統計只要樣本數夠大，樣本平均數的抽樣分佈也會非常接近 t 分佈，因此使用 t 檢定。



圖十二 讚數的分布圖

再者，為精準判斷此資料，故須先了解資料是否有偏離值存在，由圖十三所示，發現有大量的偏離值，故刪除 17 筆偏離值，最後在使用 t 檢定判斷露臉與讚數是否有關係。



圖十三 讚數盒鬚圖

B. 分析結果：

由圖十四可得知，p value 小於 0.05，具有顯著差異，表兩者具有關係。

```
        Two Sample t-test

data:  Ttest$like and Ttest$Followers
t = -1192.2, df = 396, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23468.44 -23391.17
sample estimates:
mean of x mean of y
 1056.196 24486.000
```

圖十四 T 檢定-讚數與露臉與否

C. 分析限制：

由上圖十三可得知，就算以扣除可能影響讚數之因子(抽獎文)，仍有許多偏離值，故仍有許多可能因子存在，為想知道是否有其他變數會影響到讚數，將由下題深入探討。

再者，此臉部分析套件並不能準確辨識，時常會把食物或玩偶...等物體辨識成人，經由我們實際使用後，被偵測有人臉的照片，約有 4 呈並非真正人臉，因此，此 4 成資料是在此分析中，我們尚未無法找到可解決方法改善。

6. 預測新文章的讚數

A. 分析構想：

由上述題目可得知，若把抽獎文此影響因子去除，仍有許多偏離值出現，為此我們另外列出四個可能因子：

- 是否有露臉
- 發文者是男性還是女性
- 追蹤人數為多少人
- 發文時間為平日還是假日

判斷上述四個因子是否能影響讚數，以及算出一最佳公式以預測，若要提升讚數，需改變哪些因子。因此，我們找尋四個 Instagram 美食平台經營者，以兩男兩女進行分析(男：raymond.hou、ryan.food；女：jojoxdaily、duffy_lifediary)。

B. 分析結果：

使用帕松分佈的原因為：Poisson 適合描述單位時間內隨機事件發生的次數的機率分佈，而讚數正為在時間單位內，隨機點擊按讚的次數機率分佈。

由下圖十五，變數 Face 1 為有露臉為基準、Weekday 1 為平日發文

為基準，以及 Gender 1 為發文者為男性當基準。

```
Call:  glm(formula = like ~ Face + Weekday + Followers + Gender, family = "poisson",
    data = AllBind2)

Coefficients:
(Intercept)        Face1     Weekday1    Followers      Gender1
  6.111e+00   -4.324e-03   -6.224e-02    2.504e-05    2.796e-01

Degrees of Freedom: 783 Total (i.e. Null);  779 Residual
Null Deviance:      233400
Residual Deviance: 139900       AIC: 146800
```

圖十五 讚數預測回歸模型

為探討是否所有變數皆對讚數有顯著的影響，故逐步驗證模型查看是否有更好的模型可得到最準確的結果。由圖十六，可得知所有變數皆對讚數有影響。

```
Degrees of Freedom: 783 Total (i.e. Null);  779 Residual
Null Deviance:      233400
Residual Deviance: 139900       AIC: 146800
> summary(finalModel_B)$coefficient
                 Estimate   Std. Error     z value      Pr(>|z|)
(Intercept)  6.110899e+00 4.728195e-03 1292.438059  0.000000e+00
Face1       -4.324412e-03 2.184614e-03   -1.979486  4.776134e-02
Weekday1    -6.223672e-02 2.352273e-03  -26.458120 2.942491e-154
Followers    2.503623e-05 8.555586e-08  292.630186  0.000000e+00
Gender1      2.795676e-01 2.615199e-03  106.901069  0.000000e+00
```

圖十六 逐步驗證讚數預測模型

由上述可得知，『 log(like) = -0.004324*(Face1) - 0.06224*(Weekday1) -0.00002504*(Followers)+0.2796*(Gender1) +0.6111 』為最佳迴歸公式，即當發文者露臉的話，該篇文章的讚數會減少約 4%的 log(like)，其中較為顯著的為發文者若是男性，即可增加約 28%的 log(like)。

C. 分析限制：

此分析的限制如同上題分析所述，臉部辨識套件並非 100%準確，故此分析可能會有些許的不精確。再者，仍有可能有其他的因子會影響，例如： 發文者的長相、照片的角度、濾鏡、照片的真實度和文章內容的敘述...等，但上述之其他因子可能會產生每個觀看者的主觀因素，例如：發文者的長相是每位觀看者主觀的看法，可能有部分人認為此長相是順眼的而紛紛點讚，然而也會有部分人認為此長相沒什麼吸引人之處，故不會特別點讚，故我們目前無法控制此主觀因素。

四、 能解決的問題：

1. 提供給找尋美食的使用者之幫助
   A. 可以節省民眾從網路上各個平台搜尋近期趨勢美食的時間成本。
   B. 可以了解不同地區所流行的不同餐廳美食。
   C. 可以了解哪些地區可以吃到最多美食。
2. 提供給 ig 美食經營者之幫助
   A. 可以幫助美食家分析出在哪個時間點所 po 出的文章可以接觸到最多人。
   B. 可以幫助美食家分析哪些因素可以提高文章按讚的機率。

五、 課程建議相關修改：

1. 後面的課程(資料視覺化到資料探勘)內容較深，使同學在學習上可能會需要花更多時間理解，故於後幾個禮拜因為學生在做題時需要較長的思考時間，導致在課堂上仍有很多練習題不能實作，覺得有點可惜。建議可以重新安排影片課程分配，前面的課程內容可以上的快一點，雖然知道老師是想大概固定每個禮拜的影片總時數，但由於後面的課程內容較深，才更需要大量的練習。
2. datacamp 是個很好的練習平台，但建議可以當作加分使用，因為好像有可以刷分的功能(？)。
3. 其餘都超完美，如果之後有類似這樣的課程安排(翻轉教學)，仍然會想去上的!!!

五、 分析心得：

經過我們這次的期末專題，發現了原來桃園真的是美食沙漠啊！難怪我們每次想要吃甜點或是好吃的東西的時候都只能往台北跑。

再者，這個專題最困難的地方就是：每個人對好吃的定義太主觀了，趨勢美食不一定就是好吃的東西。雖然我們的目標是推薦趨勢餐廳，所以只要注意那家餐廳有沒有造成趨勢就好了，但是像同學們問的問題一樣，不好吃的時候就會覺得這個分析不可靠，無法使人信服，因為大家還是想要被推薦好吃的餐廳，所以如果有辦法能在最後多一個可以把關的標準，就會有更多人相信我們做出來的數據。

此外，我們這次的分析也有很多限制，除了上述所說的限制外，經由同學對報告的提問後，發現仍然有很多不足，例如：使用 Hashtag 真的是能準確判斷美食的好吃與否嗎？是否文章內容有 hashtag 美食但實際卻是在批評此美食呢？所以目前想到最好的解決方法是利用文字探勘去判斷。

最後!我們也瞭解到「統計」對數據分析的重要性，分析資料除了要學習獲得資料的方法、資料的清洗和數據的呈現外，要用什麼樣的檢定方式來證明我們的假設理論，就要運用到「統計」的知識，因此進而顯現出我們在大一學了統計就忘了的事實和不足之處。

六、 參考資料：

無。(參考曾意儒老師厲害的建議，與江彥逸老師專業的統計建議)

八、 分工：

石家安：分析一的程式碼；分析五與六的程式碼及書面

王姵云：分析四的程式碼及書面；動機書面

樊驊　：分析五與六的程式碼；分析一的程式碼及書面

陳勗恩：分析二與三的程式碼及書面；簡報美編