G17 居批欽小偷

R語言與資料科學導論

期末專案報告

財金一 B08703045 鄧德齊

工管一 B08701244 蔡銑驊

工管一 B08701228 李羿澄

國企一 B08704040 許嘉仁

前言

Youtube為目前全世界流量第二大的網站，其用戶數與資訊記載量相當龐大，內容擴及各領域、各年齡層，其中「Youtuber」的影響力更是近幾年大家有目共睹的。

為此，我們期望透過youtube頻道、影片當中記載的各項大數據，結合R語言、資料科學等概念，分析當週台灣訂閱人數漲幅前100的頻道主，藉此找出youtuber的黃金成功公式，一方面幫助頻道主優化影片、提高影片質量與粉絲人數，另一方面更希望能以此為基礎，擴大網紅之影響力。

研究分析

研究方法論 —— α值與β值

α值「深度參與率」、β值「既存參與比」，作為一個良好的定量研究專題，關鍵的數值指標是自不可少的，我們以這項專案的基本目標，也就是使Youtuber受到關注，延展出兩項關鍵目標，亦即吸引人的程度以及吸引新人的能力，這兩者的數值化指標，即是α值與β值。以下是兩個數值的定義公式：

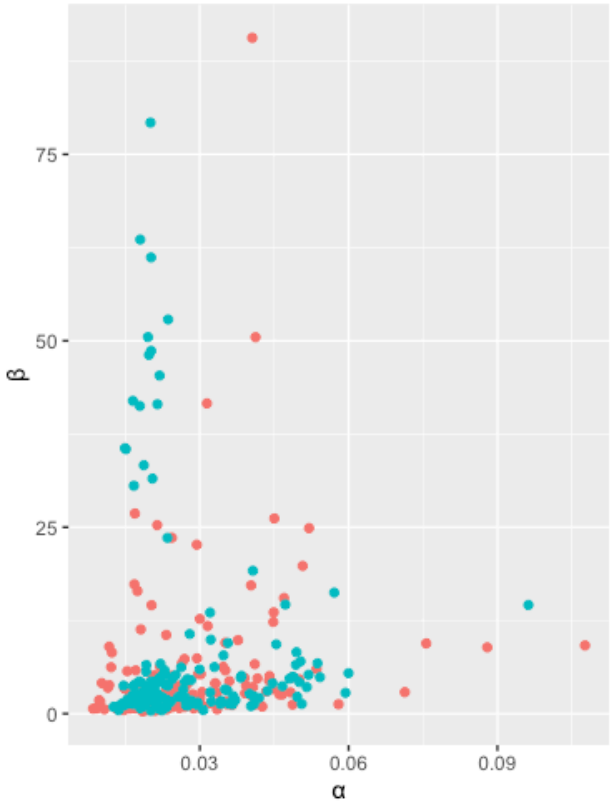
$$\alpha = (likeCount+dislikeCount+commentCount*2)/viewCount$$

$$\beta = viewCount/subscriberCount$$

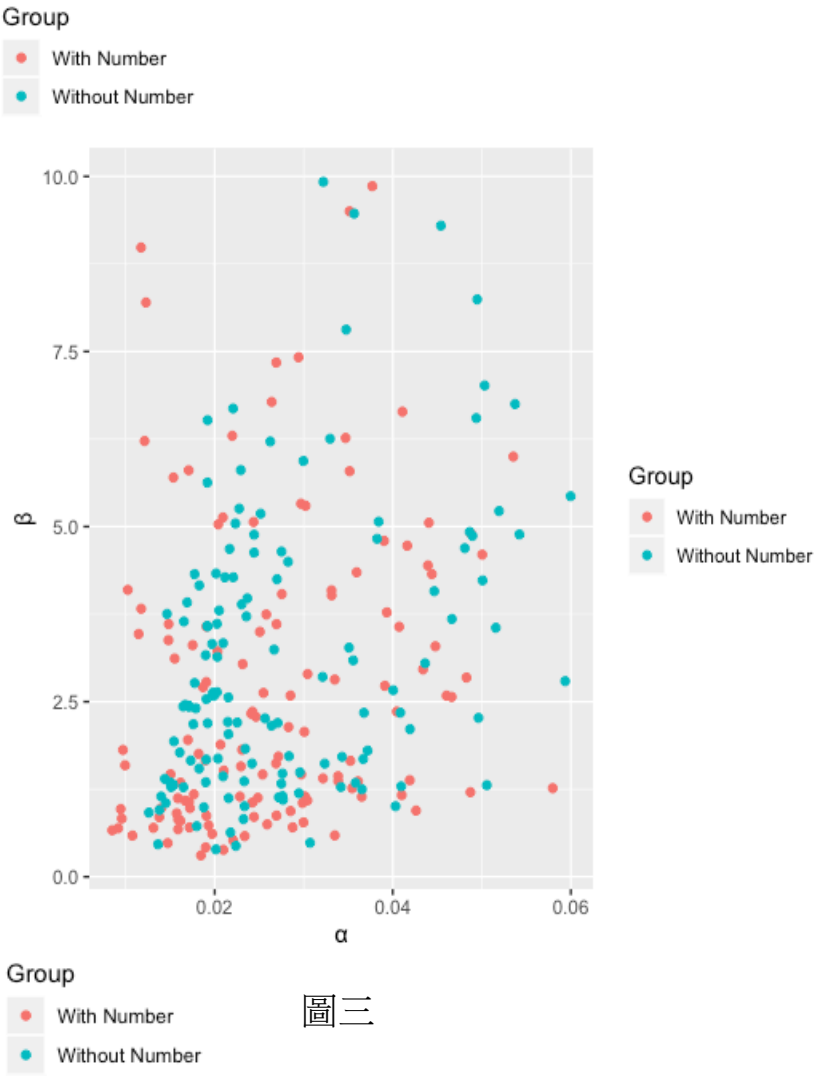
影片標題與頻道成長之關係 許嘉仁

試問，「台大你不可不知的真相」和「台大你不可不知的五大真相」你想看哪個？
此部分探討中文標題與α值和β值的相關性，具體作法如下：在利用tuber套件取得數據後，使用str_match()函數配合is.na()邏輯判斷作篩選分群，之後逐項資料計算α值和β值，最後，以ggplot2套件將結果作散佈圖呈現。

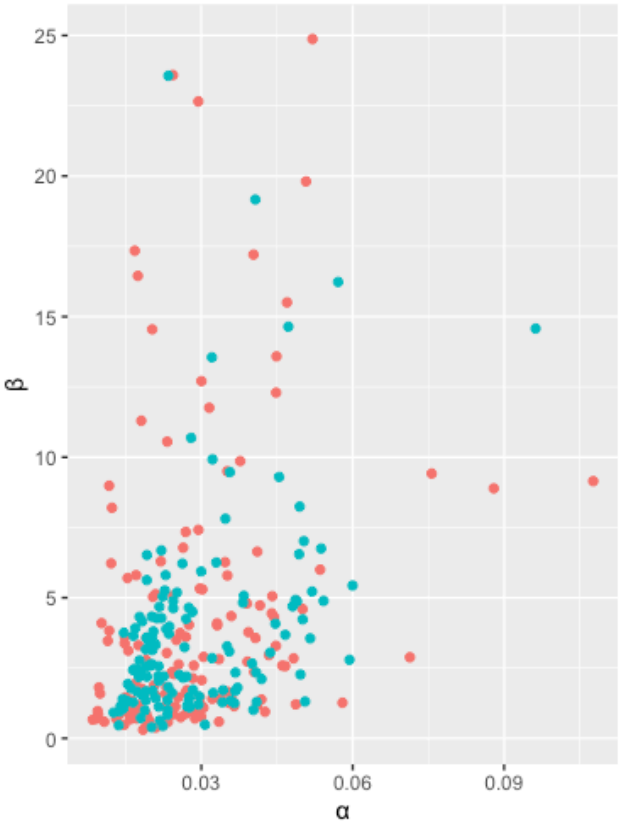
將最初的數據分佈（圖一）作了洽當的偏差值處理後，所產出的分佈圖（圖二、圖三）中可以看出，基本上數字是否存在並不影響α值和β值，但具備內含數字標題的影片卻在圖的左側產生了一條點狀帶，其α值低下的特色，正洽證明了標題的數字不但未帶來預期的正面效果，反而可能造就異常低落的深度參與率。


75
50
25
0
β
0.03
0.06
0.09
α

圖一


Group
With Number
Without Number
10.0
7.5
5.0
2.5
0.0
β
0.02
0.04
0.06
α
Group
With Number
Without Number

圖三


25
20
15
10
5
0
β
0.03
0.06
0.09
α

圖二

影片上傳頻率與頻道成長之關係 蔡鈺驊

上傳頻率對觀看人數、alpha值、beta值、訂閱人數的影響

之前滿多人說日更是衝訂閱數很好的方法，因此想看看更新頻率對一個頻道的影響

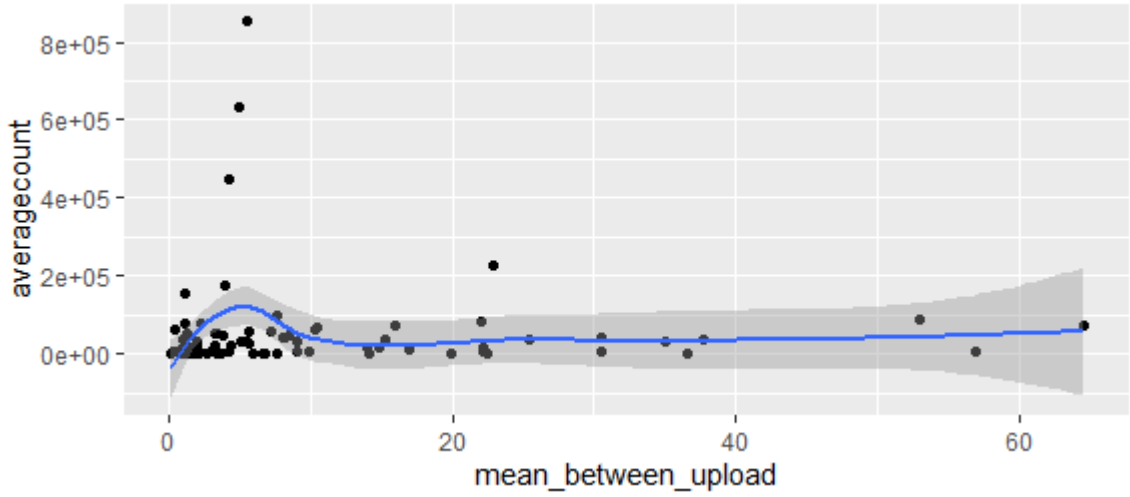
上傳頻率是以上傳影片的間隔天數為依據，因此數字越大，頻率越小

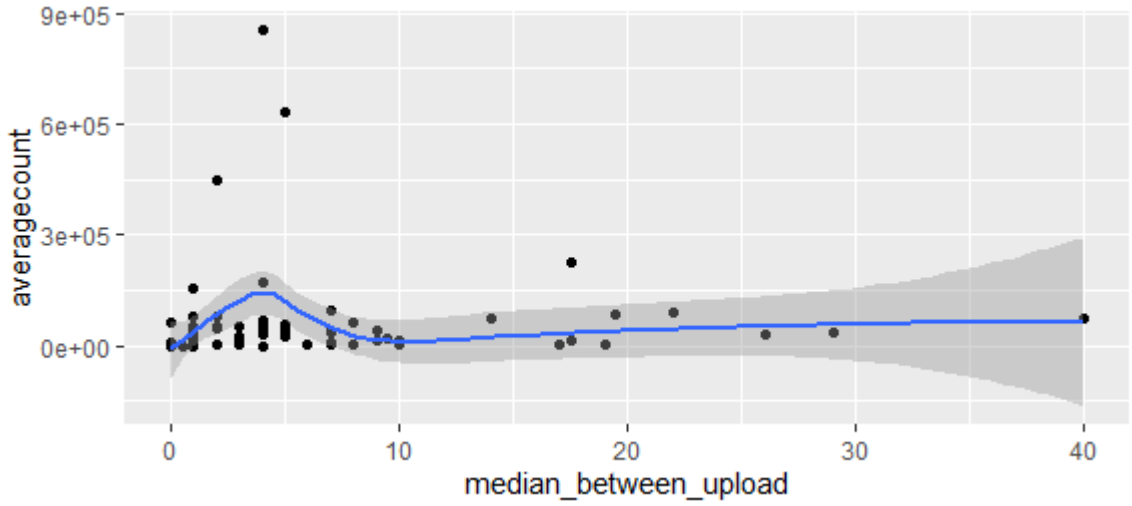
但怕某些頻道的上傳頻率起伏很大，因此做了平均值跟中位數作為參考

觀看人數:

這方面平均值跟中位數顯示出來的趨勢差不多，大約每隔4、5天上

傳一次影片會得到最好的效果。

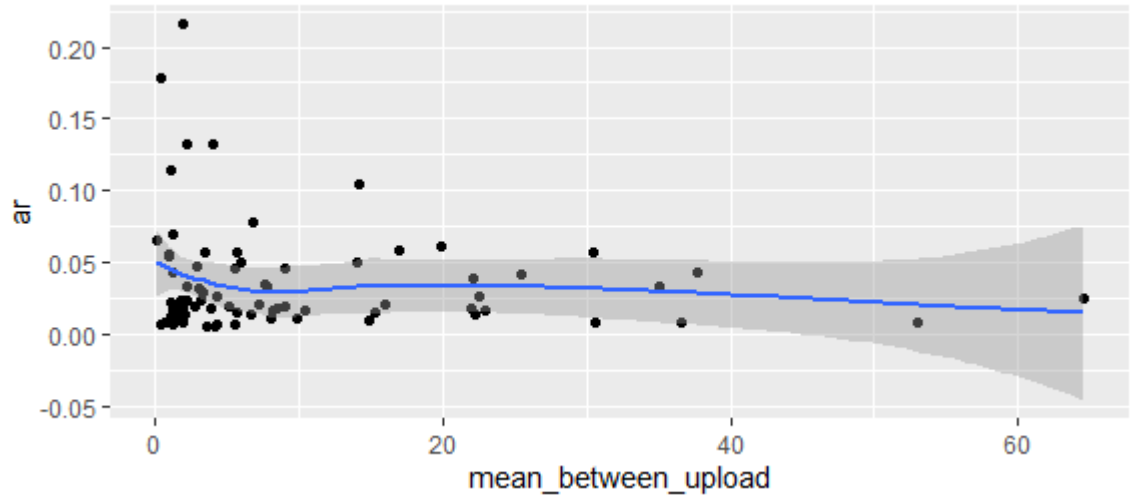


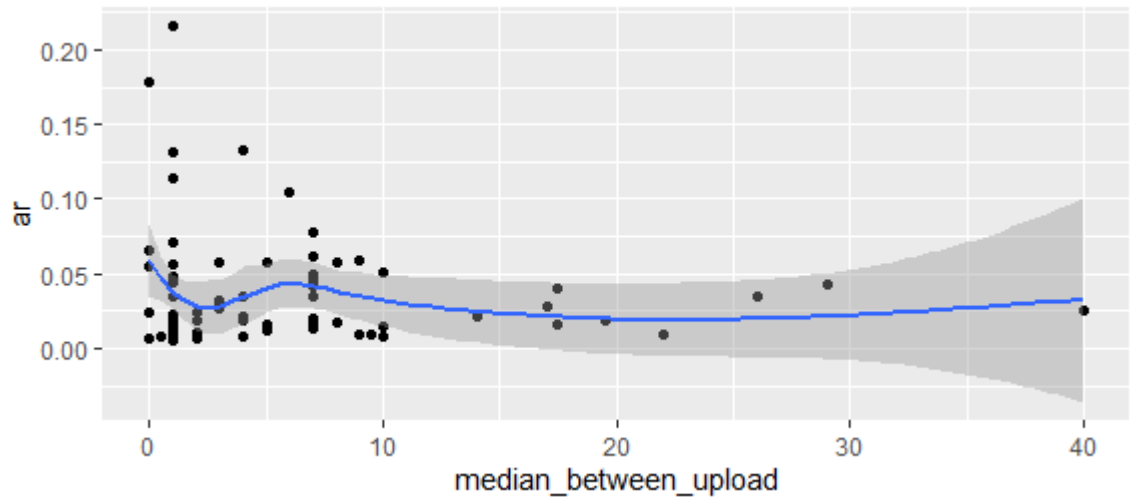


Alpha值:

這方面的平均值跟中位數值在2~7天之間的趨勢有點不同，應該代表

密集上傳的重要性比穩定上傳的重要性還大。

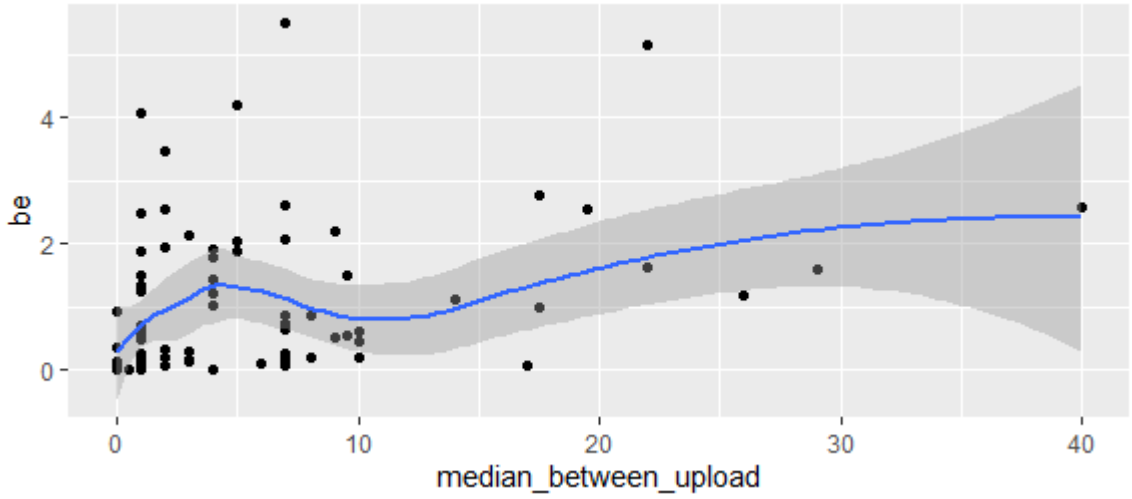




Beta值:

在這方面獲得較高分數的大多是控制在10天至少上傳一次，但也有少數頻道是久久上傳一

次卻也獲得很高分，那可能就是其他因素影響的了。



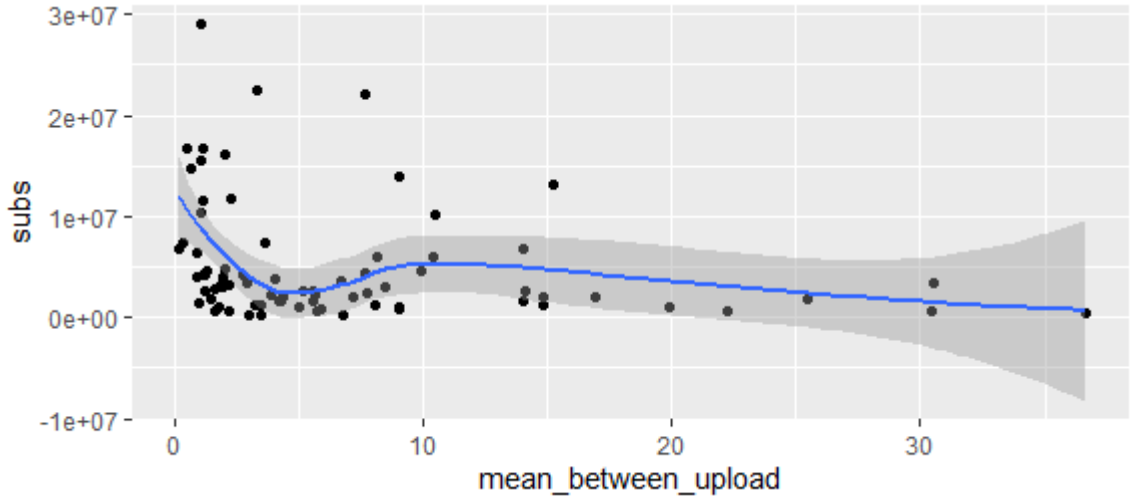
訂閱數:

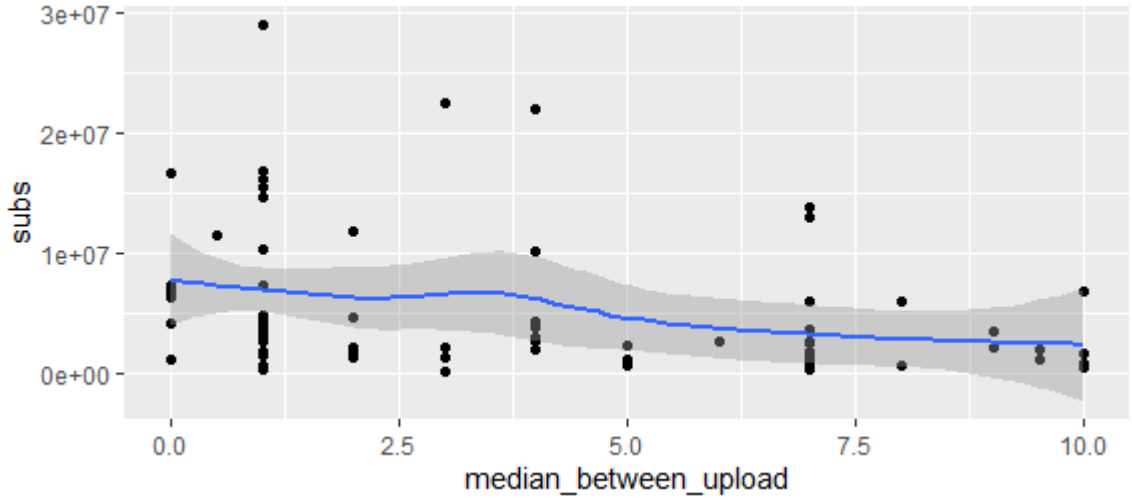
以平均數來看1~4及8~10附近的頻道有最多訂閱數，而以中位數來看

的最高點是3、4天，日更反而不一定最好，但影響似乎不大。大致可

以得出的結果為平均上傳頻率需在10天以內。這樣就可以兼顧成長率

以及訂閱數量。





留言與頻道成長之關係 鄧德齊

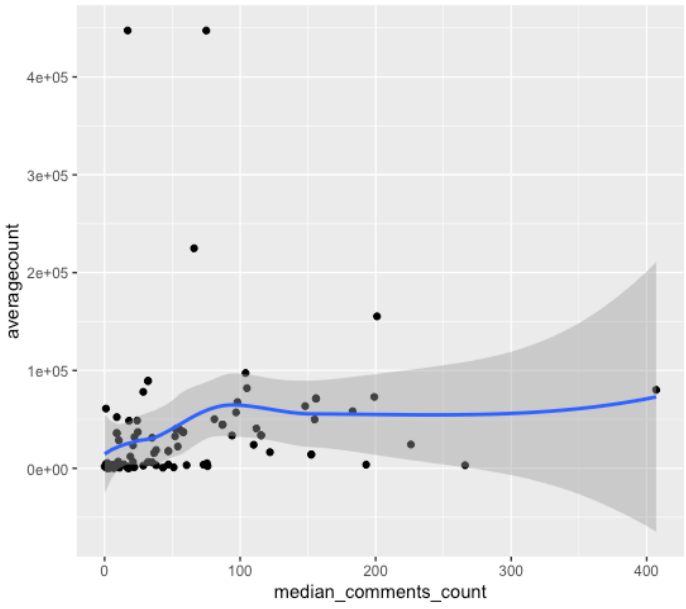
我們還想知道，留言與頻道成長是否有相關？「頻道主回覆留言的比率」和「總留言數」是否與頻道成長有關係？

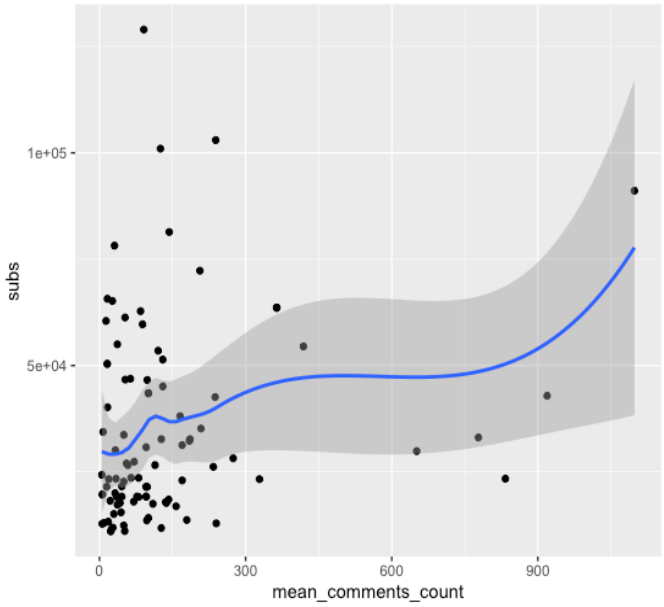
因此，我們將 get_all_channel_video_stats() 中獲得的影片id放入迴圈中，使用 get_all_comments() 取得影片留言資料，並依序擷取

1. $parentID 不為NA者（有受到回覆的留言串）
2. $authorChannelId.value 和頻道主的id一樣者（是頻道主回覆的留言），

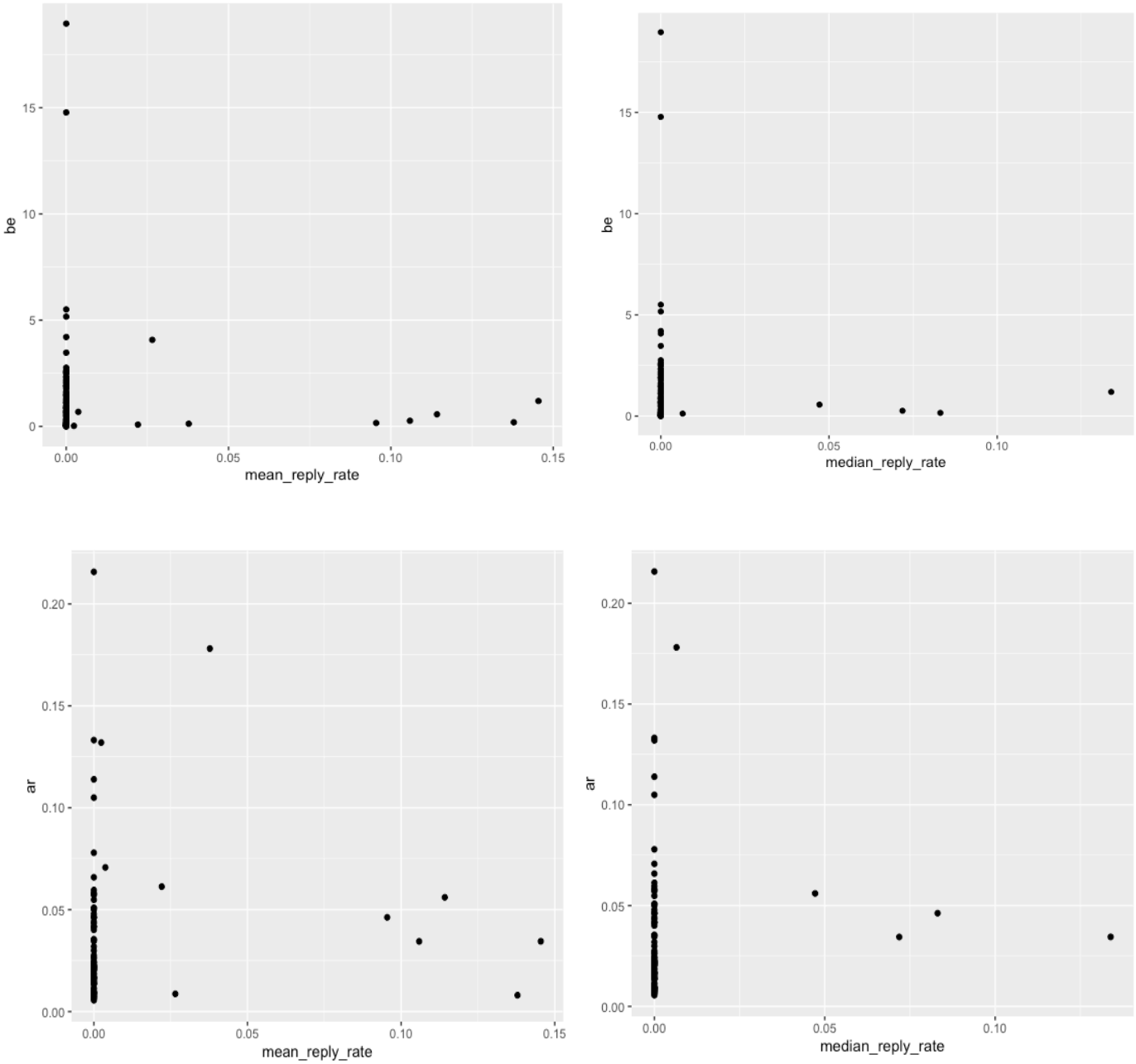
加總後除以總留言數量即得頻道主之回覆率。

我們觀察留言數量、留言回覆率之平均數與中位數，並將之與α值與β值、訂閱數、平均觀看數做比對，並將結果利用 ggplot 套件製圖如下：





由上二張圖可發現，總留言數與平均觀看次數、訂閱人數僅在前期（剛發展之階段）呈正相關，達到一定水準後便呈漸近水平。



（上二圖為回覆率平均和中位數與β值之關係，下二圖則為與α值之關係）

反觀回覆率與α、β值之比較則可發現，兩者並無顯著關係，甚至許多頻道的回覆率甚至是掛零。這是一項驚人的發現，顯示以往我們所認為的用心經營頻道包含認真回覆粉絲的留言和頻道成長並無顯著相關，因此未來若有心經營新創Youtube頻道者可能可以朝其他方面投資更多資源努力。

結論與未來願景

關於這次資料科學專案的未來發展，大致可分為三個方向。

第一，由於youtube api配額的限制，使我們不能爬取影片數量較多，也就是所謂知名頻道主的相關數據，為此，我們計畫透過書面申請的方式，分享專案內容並向youtube要求提高配額上限，藉此使我們數據的收集得以擴大，更加全面。

第二，我們認為「文字情緒詞分析」會是相當有趣的一個面向，而影片中標題、字幕與影片簡介等等當中的字詞，皆能夠判斷出為正面、負面等各式情緒用詞類別，目的是希望了解是否高度情緒化的鋪陳、描述更能吸引觀眾，創造更多觀看、點閱人數，然而現階段由於技術方面限制，我們能做的僅僅只是分析單一影片，未來要解決程式迴圈使其更完善。

第三，我們希望做出數據回歸模型，甚至開發出「網紅分析工具」，用戶可以在我們的應用程式中輸入其經營youtube頻道之策略，或將現有影片放進去做分析，一方面能夠預測用戶未來成長潛能與軌跡，另一方面，我們同時提供最新百大網紅各項數據給用戶參考。

最終，我們期望youtube不單單只是頻道主用來創造收益、分享生活的管道，也可以是政府單位、各類公益團體，用以向社會大眾宣傳理念、教育民眾的工具，如此一來好的理念能夠擴展，人類社會也因而受益，創造雙贏局面。