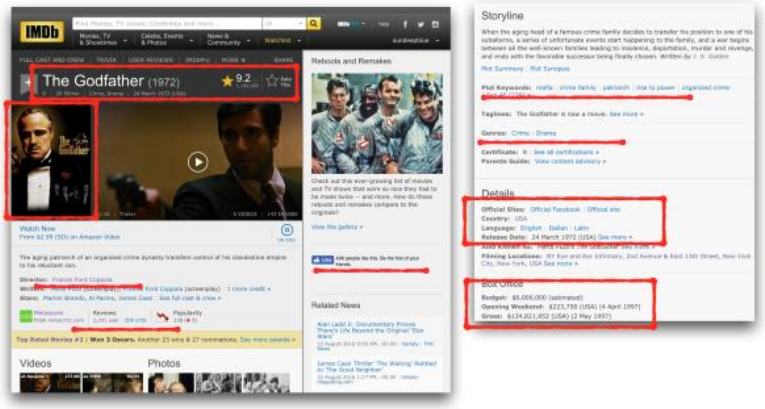
資料集內容描述

資料集：[IMDB 5000 Movie Dataset]
資料來源：[Kaggle]



由一名網友分享，內容為 IMDb(Internet Movie Database)從 1916 年到 2016 年歷時 100 年共超過 5,000 部電影的資料，目的在於預測電影的 IMDB 評分。

原資料集共 28 個欄位，包含重點演員及導演和演員分別擁有的 FB 粉絲數，這次分析只擷取其中的部份欄位作練習，篩選後共 11 個欄位，欄位名稱及資料型態如下表：

表一、IMDB 資料集之欄位列表

欄位名稱	資料型態
導演名稱	String
評論	Numeric
票房	Numeric
電影名稱	String
IMDB 用戶投票數量	Numeric
預算	Numeric
上線日期	Numeric
IMDB 評分	Numeric
電影 FB 粉絲數	Numeric
語言	String
國家	String

	導演名稱	評論	票房	電影名稱	IMDB用戶投票數量	預算	上線日期	IMDB評分	電影FB粉絲數	語言	國家
0	James Cameron	723.0	760505847.0	Avatar?	886204	237000000.0	2009.0	7.9	33000	English	USA
1	Gore Verbinski	302.0	309404152.0	Pirates of the Caribbean: At World's End?	471220	300000000.0	2007.0	7.1	0	English	USA
2	Sam Mendes	602.0	200074175.0	Spectre?	275868	245000000.0	2015.0	6.8	85000	English	UK
3	Christopher Nolan	813.0	448130642.0	The Dark Knight Rises?	1144337	250000000.0	2012.0	8.5	164000	English	USA
4	Doug Walker	NaN	NaN	Star Wars: Episode VII - The Force Awakens? ...	8	NaN	NaN	7.1	0	NaN	NaN

圖一、IMDB 資料集之部份資料截圖

分析方法

一、資料預處理

包含去除遺漏值和更改欄位型態，確保應為數值的欄位不會被視為字串。

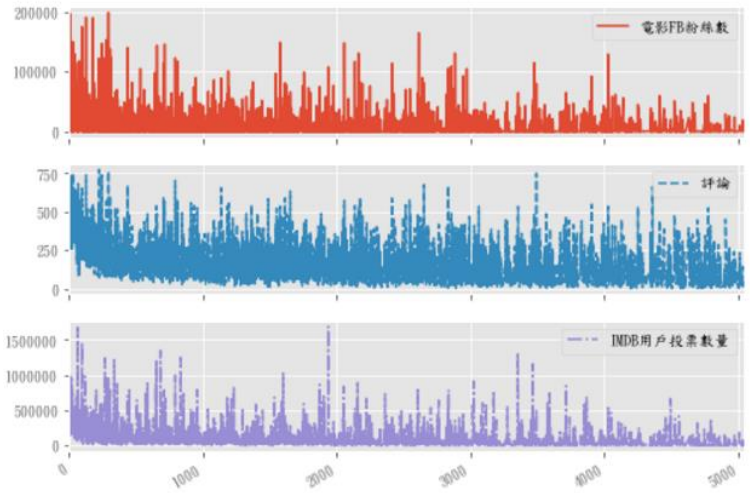
二、描述性統計

透過基本描述性統計來觀察資料長相，其中較值得注意的是，電影 FB 粉絲數最小值為 0，且第一四分位數也為 0，代表有將近 25%的電影在 FB 沒有任何粉絲。

	評論	票房	IMDB用戶投票數量	預算	上線日期	IMDB評分	電影FB粉絲數
count	3887.000000	3.887000e+03	3.887000e+03	3.887000e+03	3887.000000	3887.000000	3887.000000
mean	163.348083	5.110642e+07	1.026879e+05	3.755932e+07	2003.100077	6.012092	9147.428351
std	124.033304	6.981997e+07	1.507585e+05	9.413565e+07	9.916819	1.094885	21311.758372
min	1.000000	1.620000e+02	2.200000e+01	-2.147484e+09	1927.000000	1.000000	0.000000
25%	73.000000	6.852056e+06	1.733050e+04	1.000000e+07	1999.000000	5.000000	0.000000
50%	134.000000	2.803125e+07	5.065300e+04	2.400000e+07	2005.000000	6.000000	210.000000
75%	222.000000	6.547616e+07	1.242320e+05	5.000000e+07	2010.000000	7.000000	11000.000000
max	813.000000	7.605058e+08	1.689764e+06	2.127520e+09	2016.000000	9.000000	349000.000000

圖二、IMDB 資料集之描述性統計值

進一步查看電影 FB 粉絲數的資料分佈，針對「電影 FB 粉絲數」、「評論」、「IMDB 用戶投票數量」三個欄位，可以看到有類似的趨勢，可進一步透過相關性分析驗證這幾個欄位的關聯性。



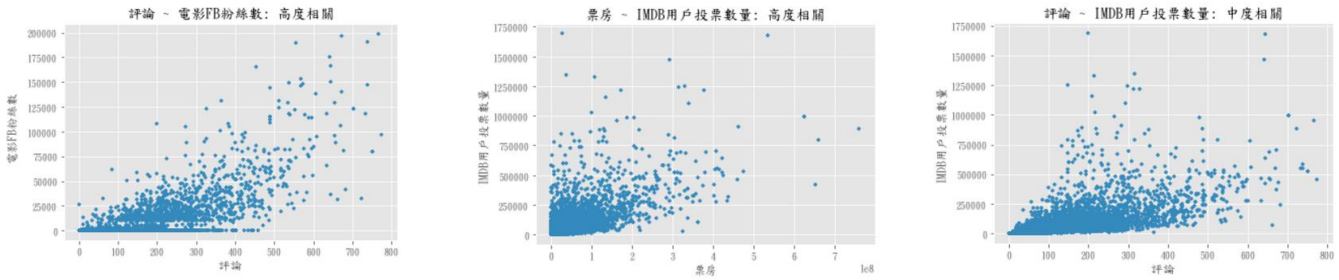
圖三、電影 FB 粉絲數之資料分佈圖

三、 相關性分析

從相關性矩陣可以看到，「評論」與「電影 FB 粉絲數」的相關係數為 0.70，屬於高度相關；「評論」與「IMDB 用戶投票數量」的相關係數為 0.59，屬於中度相關；從矩陣中也另外發現，「IMDB 用戶投票數量」與「票房」的相關係數為 0.62，同樣屬於高度相關。

	評論	票房	IMDB用戶投票數量	預算	上線日期	IMDB評分	電影FB粉絲數
評論	1.000000	0.471881	0.594285	0.190543	0.393583	0.331470	0.708199
票房	0.471881	1.000000	0.627553	0.297022	0.044074	0.200088	0.373321
IMDB用戶投票數量	0.594285	0.627553	1.000000	0.171225	0.012632	0.458345	0.512752
預算	0.190543	0.297022	0.171225	1.000000	0.096000	-0.005873	0.137128
上線日期	0.393583	0.044074	0.012632	0.096000	1.000000	-0.123856	0.303564
IMDB評分	0.331470	0.200088	0.458345	-0.005873	-0.123856	1.000000	0.272780
電影FB粉絲數	0.708199	0.373321	0.512752	0.137128	0.303564	0.272780	1.000000

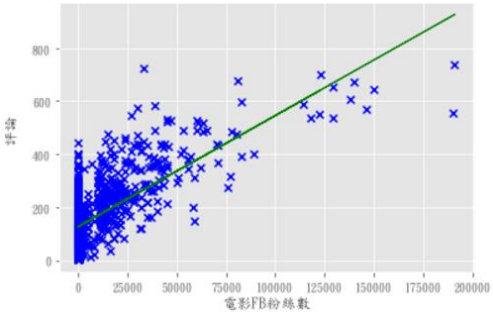
圖四、相關性分析



圖五、相關性散佈圖

四、 線性迴歸分析

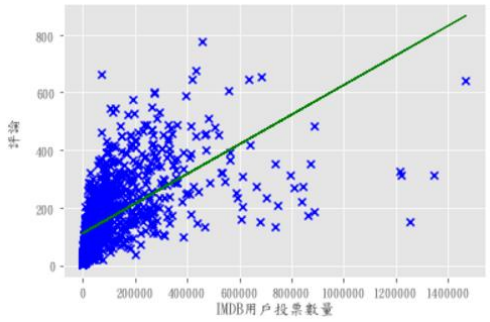
延續相關性分析的結果，由於「評論」分別與「電影 FB 粉絲數」及「IMDB 用戶投票數量」皆有高度相關及中度相關，因此將「評論」作為因變數，「電影 FB 粉絲數」及「IMDB 用戶投票數量」作為自變數，透過線性迴歸方法來建立「評論」的預測模型。皆依 7:3 的比例，將資料集切分為訓練集與測試集。



圖六、模型一

模型一：「電影 FB 粉絲數」預測「評論」

得出模型參數為 0.0042，截距為 124.8699，均方誤差為 8324.32，即得到模型為$y = 0.0042x + 124.8699$。模型在測試集上的得分為 0.47。



圖七、模型二

模型二：「IMDB 用戶投票數量」預測「評論」

得出模型參數為 0.0005，截距為 110.4097，均方誤差為 10676.50，即得到模型為$y = 0.0005x + 110.4097$。模型在測試集上的得分為 0.31。

模型三：「電影 FB 粉絲數」與「IMDB 用戶投票數量」預測「評論」

得出「IMDB 用戶投票數量(x_1)」的模型參數為 37.5613，「電影 FB 粉絲數(x_2)」的模型參數為 70.4413，截距為 164.1975，均方誤差為 6081.97，即得到模型為$y = 37.5613x_1 + 70.4413x_2 + 164.1975$。模型在測試集上的得分為 0.56。

分析洞察結果

就統計結果而言，根據模型得分，透過「電影 FB 粉絲數」及「IMDB 用戶投票數量」來預測「評論」的得分為 0.56，比模型一和模型二的得分還高，表示使用兩個自變數去預測較精準，但也還有許多解釋空間，本次分析是選出高度相關之變數，之後可以再嘗試將其他屬於中度相關之變數也一同加入，也許會得到更高的得分。

就實務意涵而言，電影 FB 粉絲數越多，評論數就越多，且 IMDB 用戶投票數量也越高，這可能代表著，願意在 FB 公開表示喜愛該電影的觀眾，同樣也會在 IMDB 上給予支持票，並且在 IMDB 上給予評論。

本次分析是針對評論數進行預測，IMDB 評分也是很有趣的預測議題。從相關性分析看到，IMDB 評分和這三種支持電影的方式沒有任一項達到高度相關的標準，只有 IMDB 用戶投票數有達到中度相關，表示喜愛電影的觀眾不一定會給予較高的 IMDB 評分。由於本資料集有篩選掉部份欄位，原資料集有提供演員姓名和演員的粉絲數等詳細資料，也許 IMDB 評分高低與演員的受歡迎程度更有關係，之後也可以嘗試分析看看。