美國地區
電影首週票房預測

Prediction for Opening Week Box Office of Movies in US

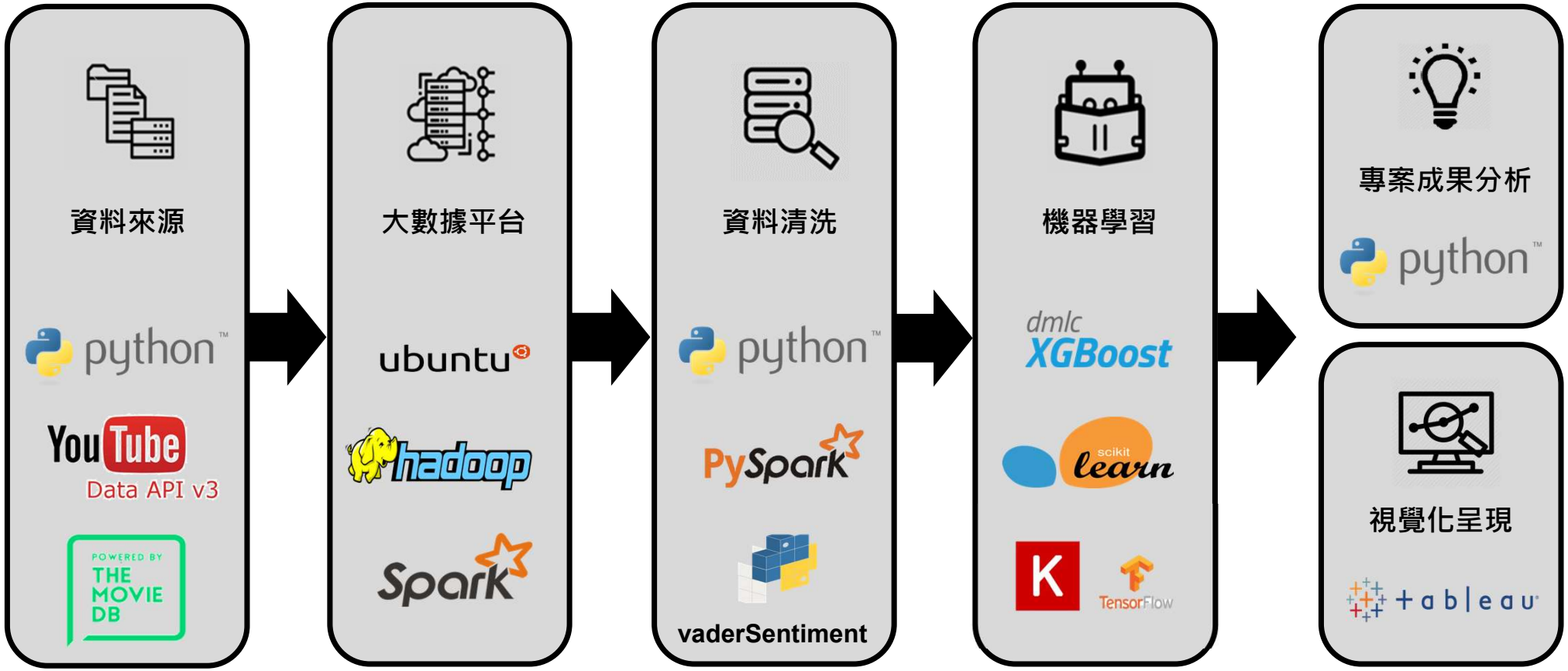
製作人：楊雅筑

專案範圍

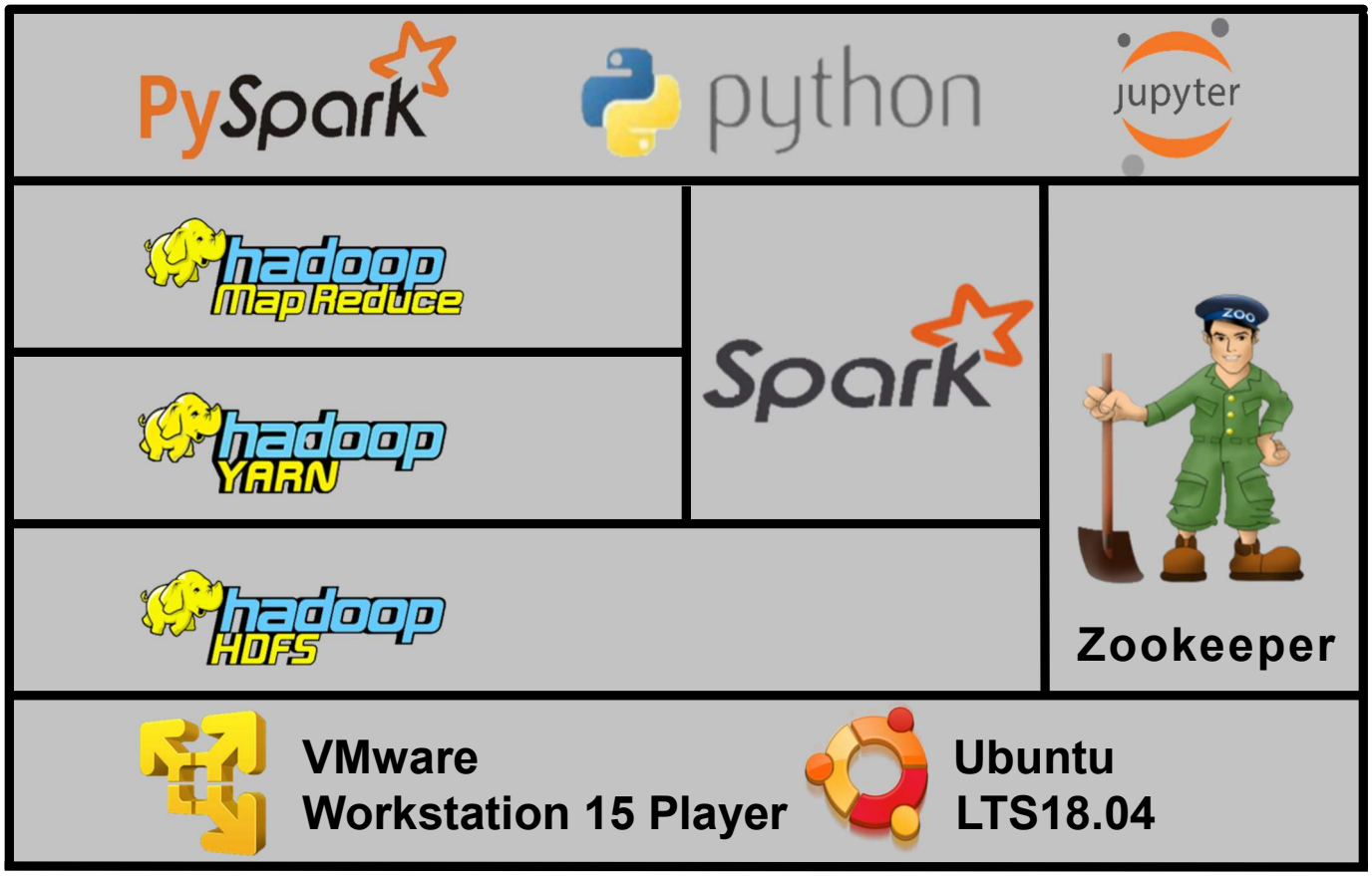
- 鑒於資料完整性，選擇美國市場在2009年至2019年9月近十年的資料，利用演員卡司、檔期、類型等電影資訊，以及觀眾在 Youtube 官方預告片上的留言聲量，進行預測。

- 近十年平均首週票房約佔總票房40%，且首週票房對於電影上映時間長短有關鍵性影響，故針對首週票房建立預測模型。

專案流程與使用工具


資料來源
python
YouTube
Data API v3
POWERED BY THE MOVIE DB
大數據平台
ubuntu
hadoop
Spark
資料清洗
python
PySpark
vaderSentiment
機器學習
dmlc XGBoost
scikit learn
K
TensorFlow
專案成果分析
python
視覺化呈現
tableau

Hadoop平台架構


PySpark
python
jupyter
hadoop MapReduce
hadoop YARN
hadoop HDFS
Spark
Zookeeper
VMware Workstation 15 Player
Ubuntu LTS18.04

叢集架構

硬體設備介紹

實體主機：6台	虛擬主機：12台
CPU：Intel® Core™ i7-4790 @3.60 GHz 4核8緒	CPU：每台均配置2顆cpu
RAM：DDR3 8GB*4 (1600MHZ)	RAM：每台均配置12GB
HDD：1TB (7200RPM)	HDD：500GB

資料來源

Web API

電影資訊庫網站	Youtube電影預告片
電影類型 演員卡司 年份 上映日期	上架日期 留言內容 留言數量

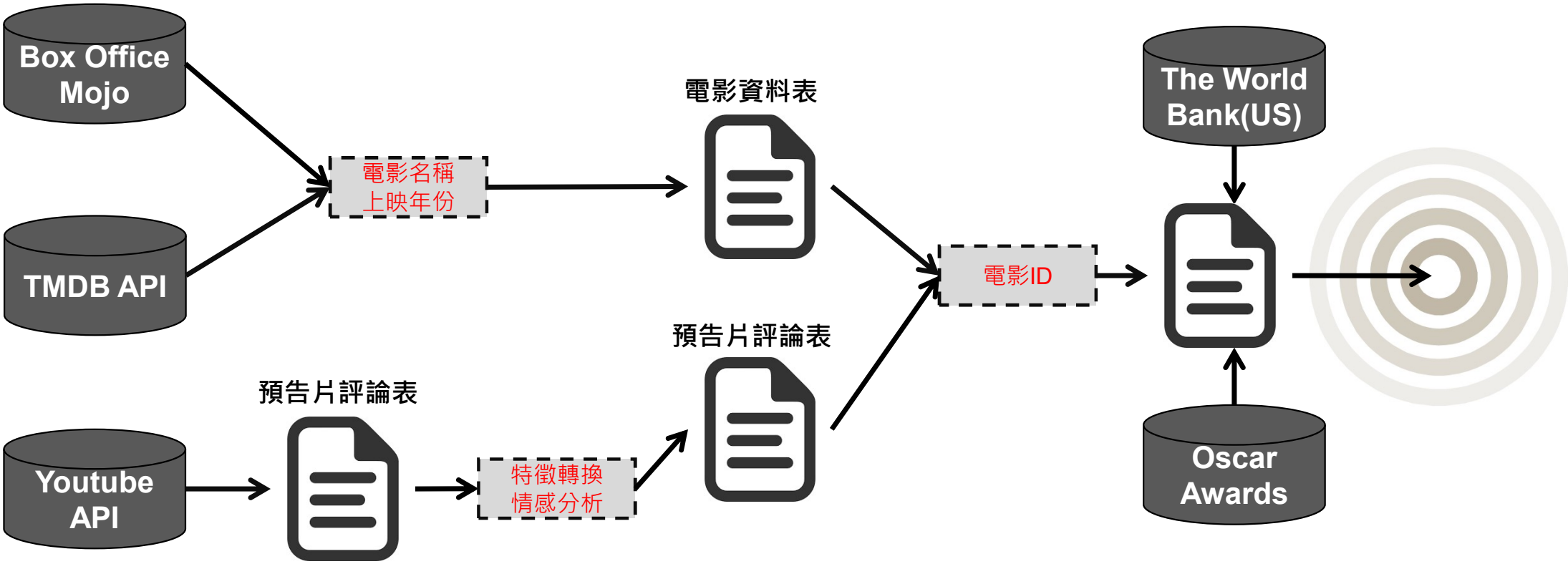
網路爬蟲

電影票房網站
電影發行商 首週票房 總票房

數據集下載

世界銀行網站	奧斯卡金像獎
美國GDP 通膨指數 失業率	得獎演員 受提名演員

資料處理流程


Box Office Mojo
TMDB API
電影名稱
上映年份
電影資料表
The World Bank(US)
電影ID
預告片評論表
Youtube API
預告片評論表
特徵轉換
情感分析
預告片評論表
Oscar Awards

特徵工程

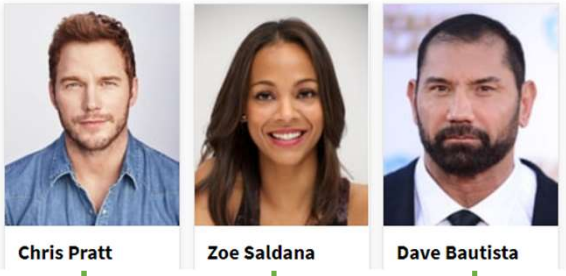
- 一部電影含多種電影類型，一部電影含多位演員
  → 視為元素組合與單一元素兩種特徵

電影類型

- Action (動作)
- Adventure (冒險)
- Comedy (喜劇)
- Science Fiction (科幻)

動作類型：**True** / False
冒險類型：**True** / False
喜劇類型：**True** / False
科幻類型：**True** / False
驚悚類型：True / **False**
…

電影演員

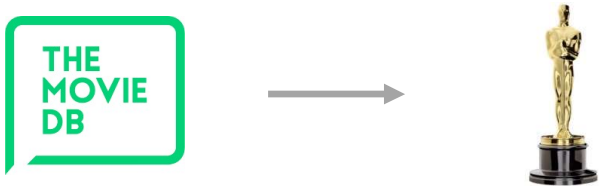


第一演員　第二演員　第三演員

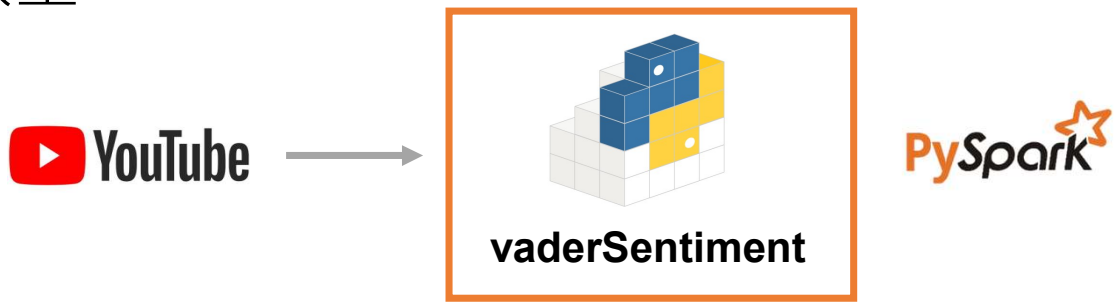
Source：https://www.themoviedb.org

特徵工程

- 卡司：比對奧斯卡金像獎資料，計算演員以及導演得獎、提名數量



- 評論：透過情感分析套件(vaderSentiment)，計算預告片正負面評論數量

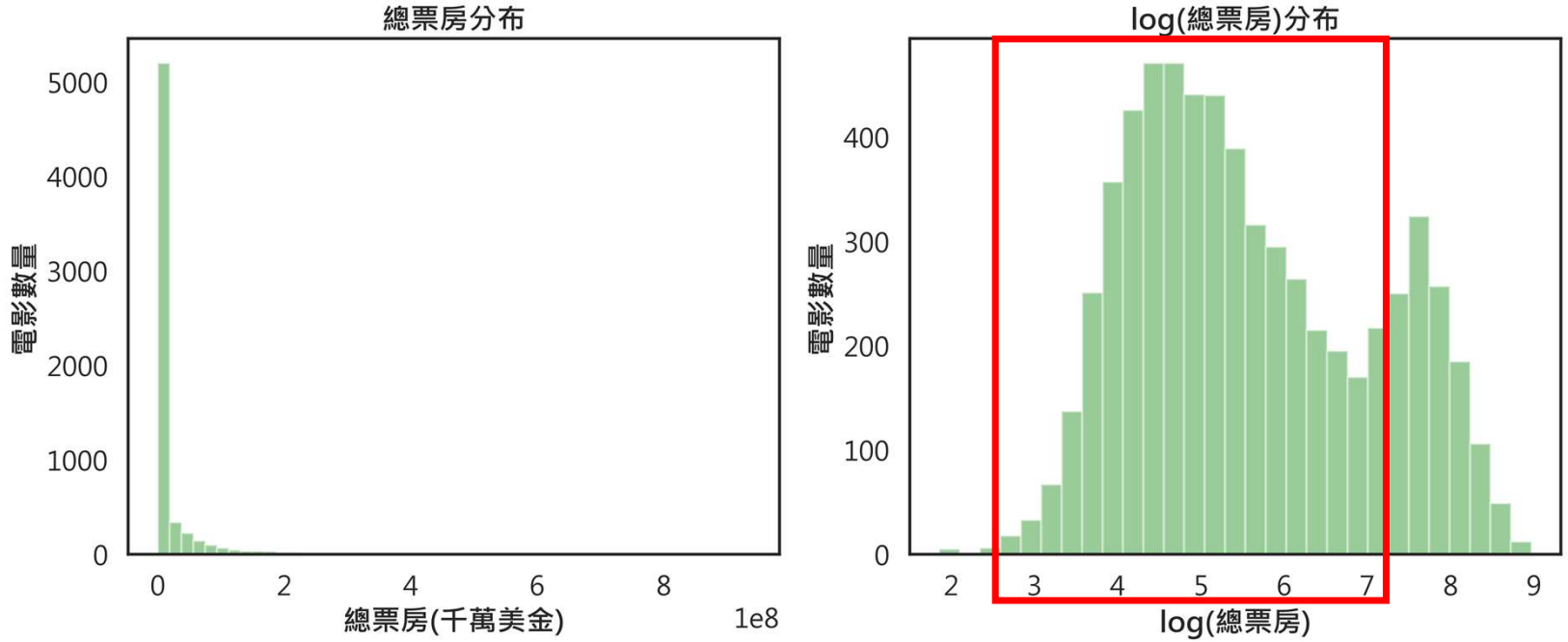


- 最終以**54種欄位特徵**建立機器學習模型

電影產業市場概況分析

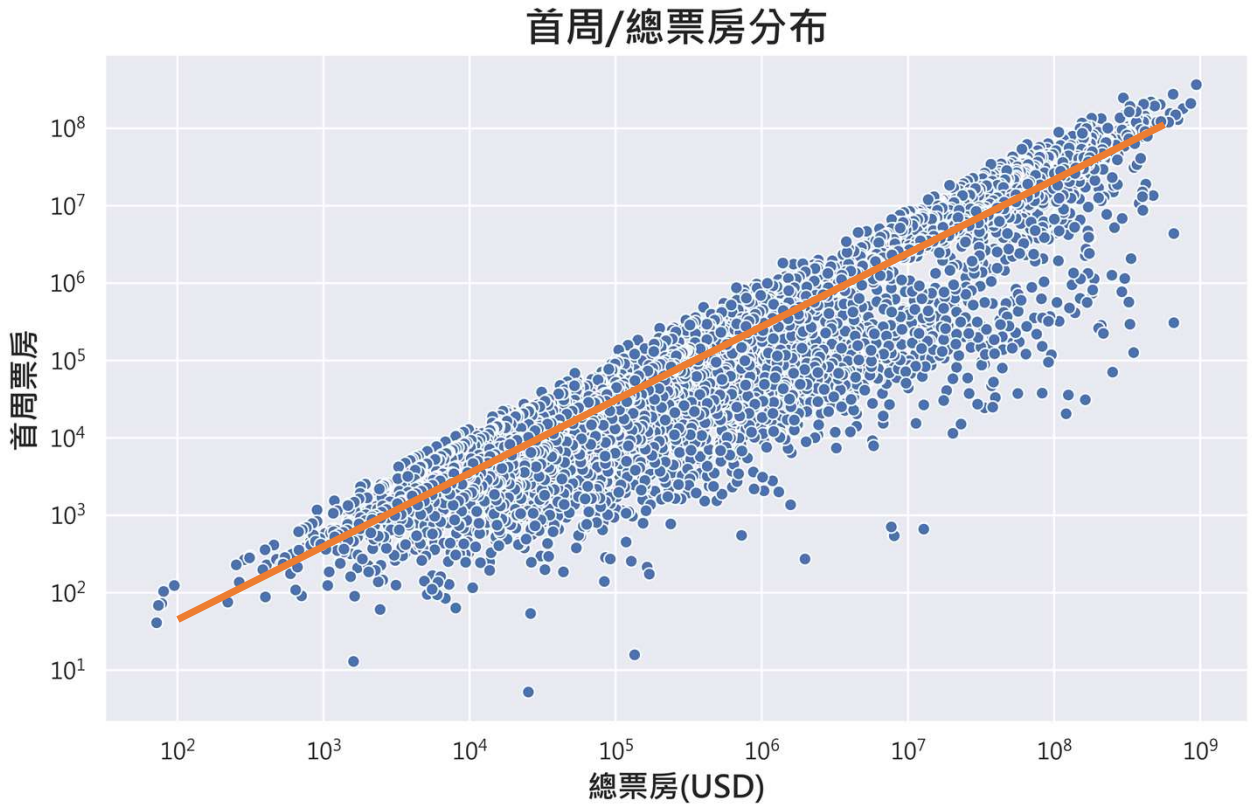
- 電影總票房分布

- 票房分布極端
- 總票房百萬以下電影佔70%



電影產業市場概況分析

- 首週票房與總票房

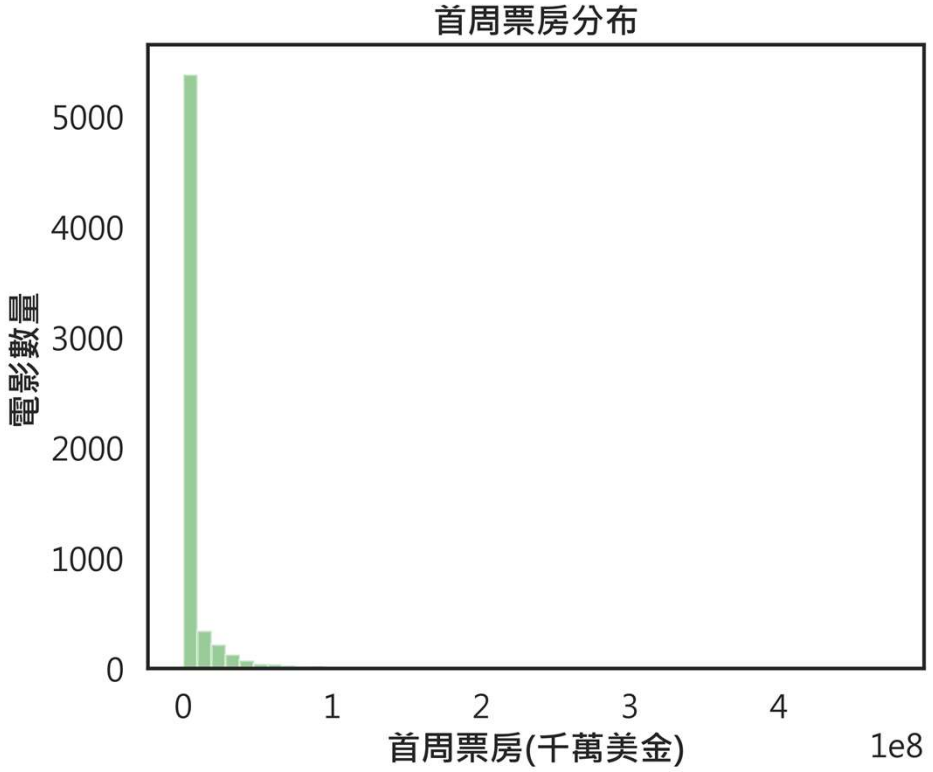


樣本數：6368部電影

特徵工程

- 資料分佈離散度高：高票房的電影資料易被當作離群值

總票房金額分佈(萬美元)		首週票房金額分佈(萬美元)	
mean	1827.750191	mean	683.981240
std	5506.506645	std	2296.342253
min	0.007200	min	0.001100
25%	2.937175	25%	0.817475
50%	21.248950	50%	4.367250
75%	577.910700	75%	88.860600
max	93666.222500	max	39085.605400



→ **依據首週票房金額分為4類**

1) Level 1：大於4千萬美元以上
2) Level 2：1千萬美元~4千萬美元
3) Level 3：1百萬美元~1千萬美元
4) Level 4：小於1百萬美元

模型建置：多元分類

隨機森林(Random Forest)

- 決策樹(Decision Tree)為基礎的集成學習法(ensemble method)
- Bagging 隨機抽樣樣本與特徵，建立多個決策樹，對於各決策樹分類結果以平均機率值估計取得最終分類結果

樹的數量
樹深
抽樣的特徵數量
分類權重
Out-of-bag

	precision	recall	f1-score
1	0.64	0.64	0.64
2	0.51	0.50	0.51
3	0.31	0.13	0.19
4	0.89	0.95	0.92
accuracy			0.82
macro avg	0.59	0.56	0.56
weighted avg	0.79	0.82	0.80

模型建置：多元分類

dmlc XGBoost 梯度提升決策樹(Gradient Boosting Decision Tree)

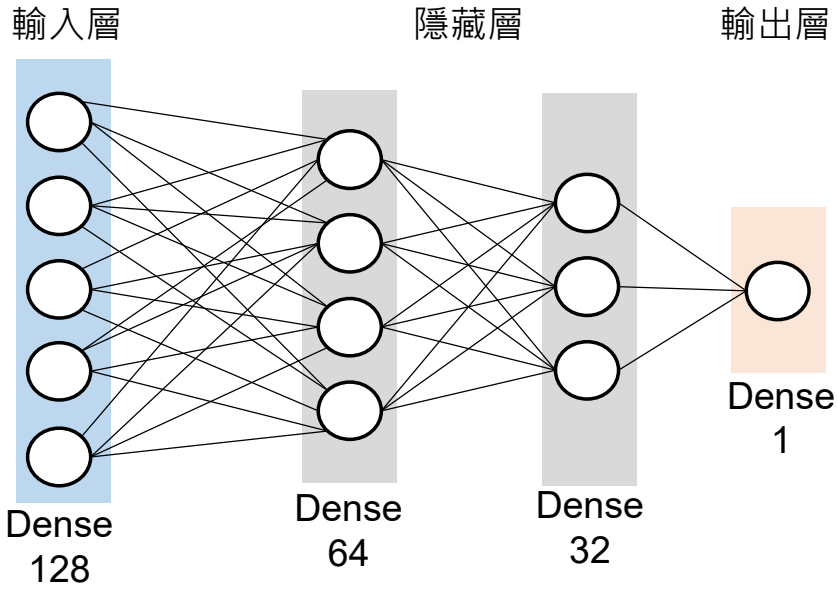
- 決策樹(Decision Tree)為基礎的集成學習法(ensemble method)
- Boosting 根據前一個決策樹模型的錯誤率學習，迭代多顆決策樹擬合為一顆決策樹，取得最終分類結果

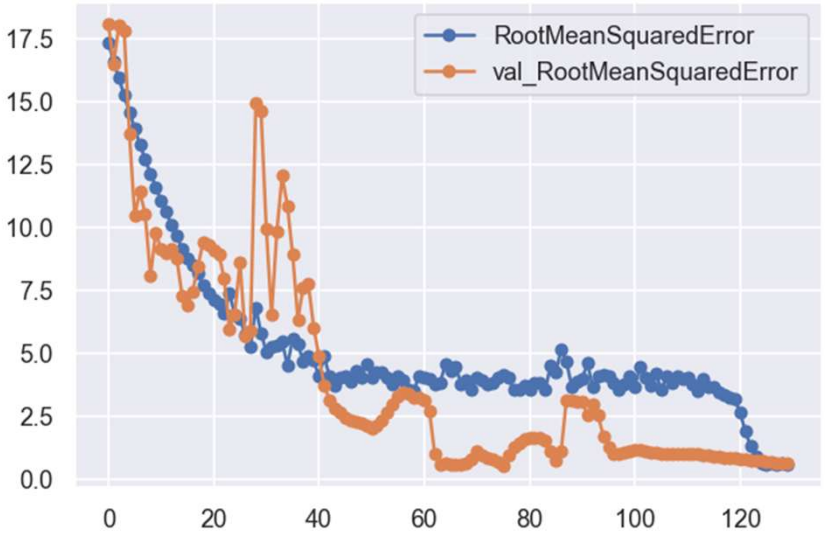
樹深
葉節點數量
學習率
L1, L2正則項

	precision	recall	f1-score
0	0.70	0.64	0.67
1	0.58	0.64	0.61
2	0.41	0.16	0.23
3	0.92	0.98	0.95
accuracy			0.85
macro avg	0.65	0.60	0.61
weighted avg	0.83	0.85	0.83

模型建置：預測Level 1首週票房金額

類神經網路(Artificial Neural Network)



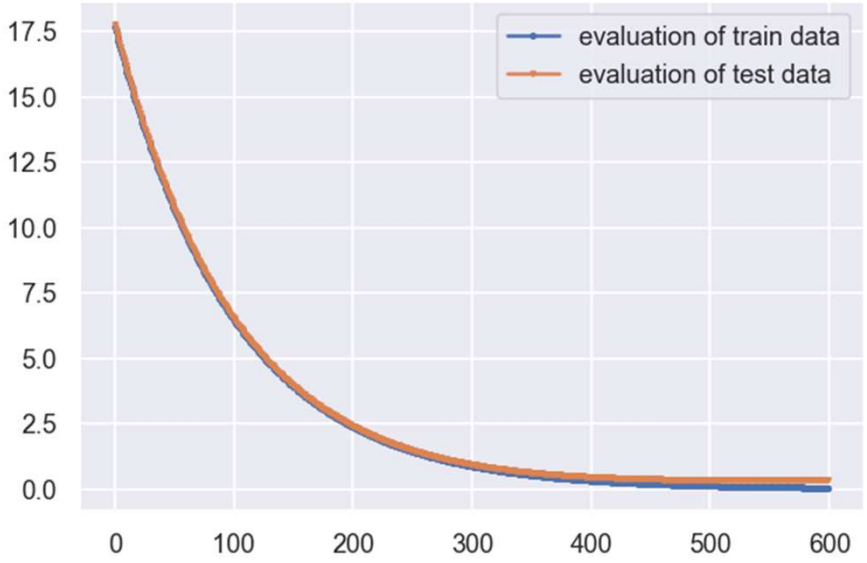



```
Epoch 130/130
169/169 [==============================]
6 - val_root_mean_squared_error: 0.6455
```

激發函數：ReLU
優化器：Adam演算法
MSE平方誤差值學習，RMSE均方根誤差評估

模型建置：預測Level 1首週票房金額

dmlc XGBoost 梯度提升決策樹(GBDT+gblinear)




```
In [176]: 1  model.eval(data_test)
Out[176]: '[0]\teval-rmse:0.325719'
```

$137,728,787


```
In [185]: 1  int(np.exp(y_pred))
Out[185]: 115968432
```

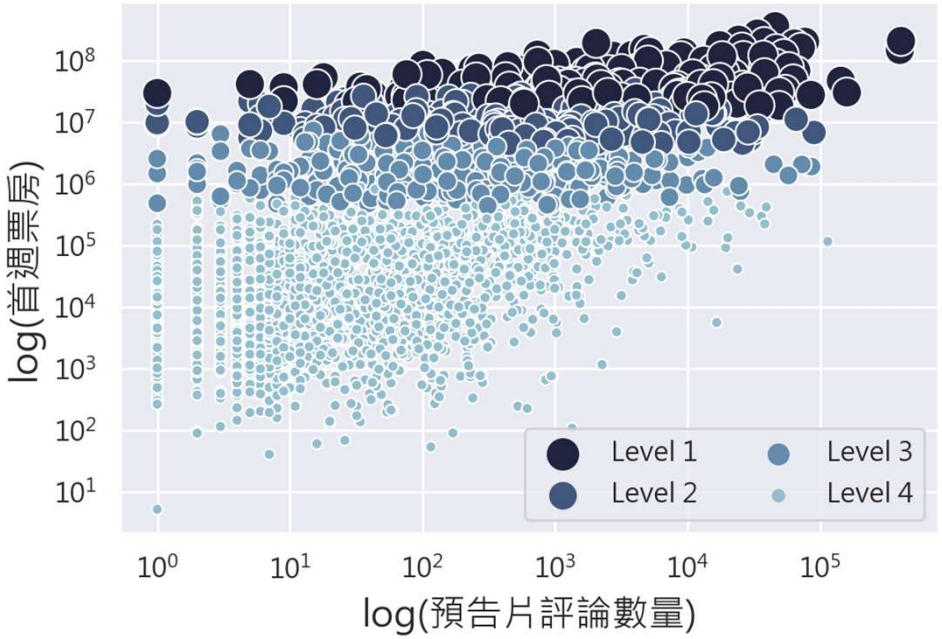
Source：https://www.themoviedb.org

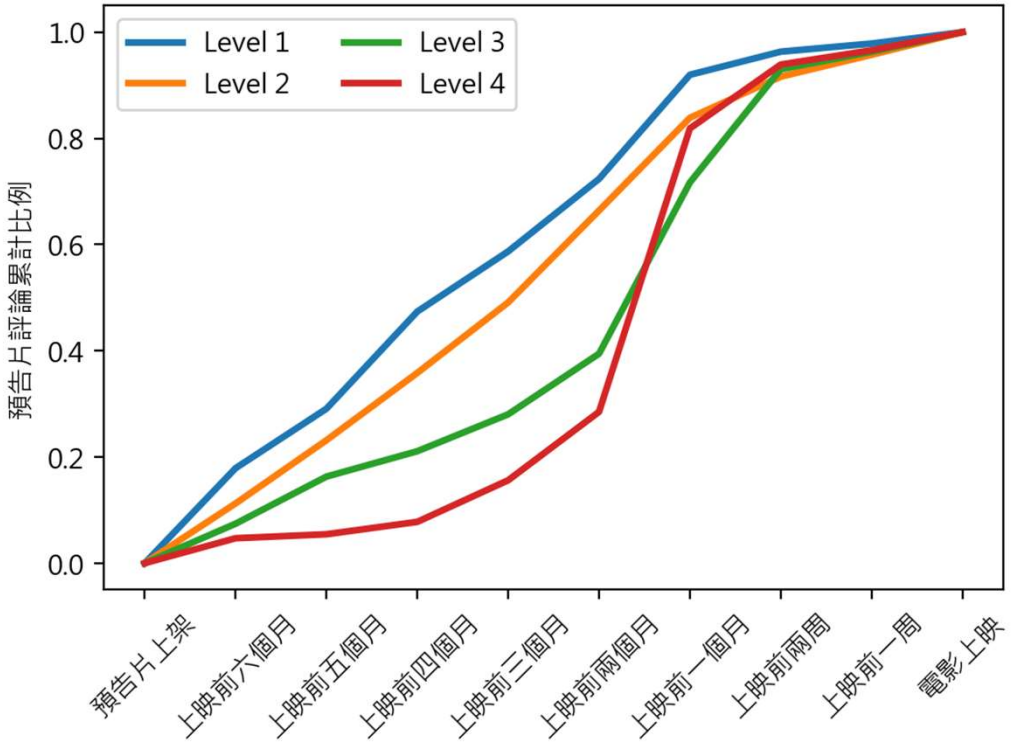
專案成果與商業應用

1. 提供投資者與電影製片業者票房價格落點
2. 由分類模型中，可看出**各價格區段**的重要區分特徵：
   - 預告片評論
   - 演員

成果分析：預告片評論

- 預告片評論與首週票房成正比
- 高票房的預告片**初期**評論比例較高





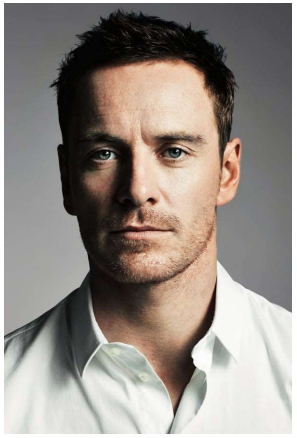
成果分析：演員

- 第二演員欄位重要性高於第一演員
  - 高票房的演員與低票房演員的重複性：第一演員(79%) > 第二演員(69%)
- 票房毒藥？
  - 高票房電影的第一要角演員名單中，其主演電影中低票房數量比例高，且主演作品數量大於平均值者

Nicolas Cage

Meryl Streep

Michael Fassbender



Source：https://www.themoviedb.org

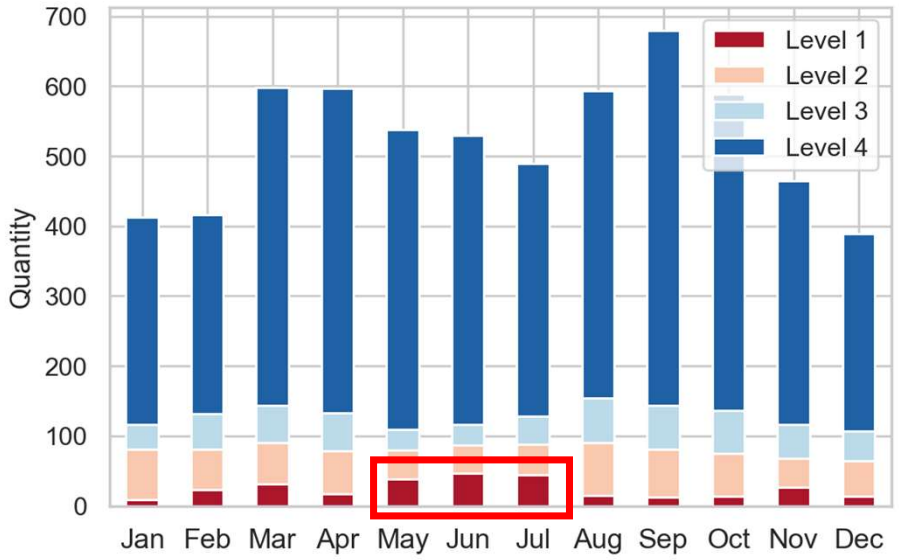
專案成果與商業應用

1. 預測票房價格
2. 由分類模型中，可看出各價格區段的重要區分特徵：
   - 預告片評論
   - 演員
3. 由高票房電影價格預測模型中，看出**票房表現**重要區分特徵：
   - **上映檔期** ⬆
   - **電影類型** ⬆
   - 預告片評論⬇
   - 演員⬇

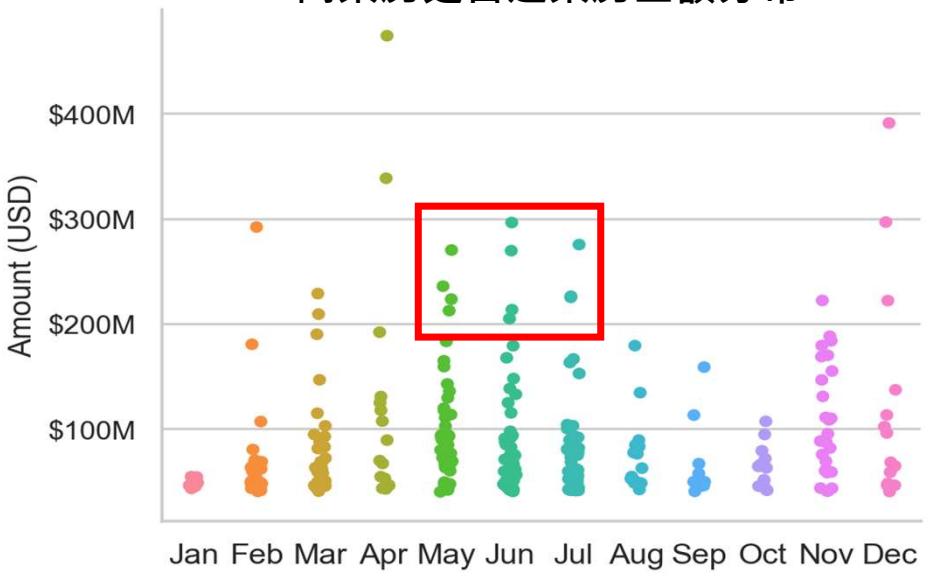
成果分析：上映檔期

- **暑假期間（5~7月份）**上映，電影間競爭高，但票房金額分布普遍表現較好

近10年各月份電影上映累計數量

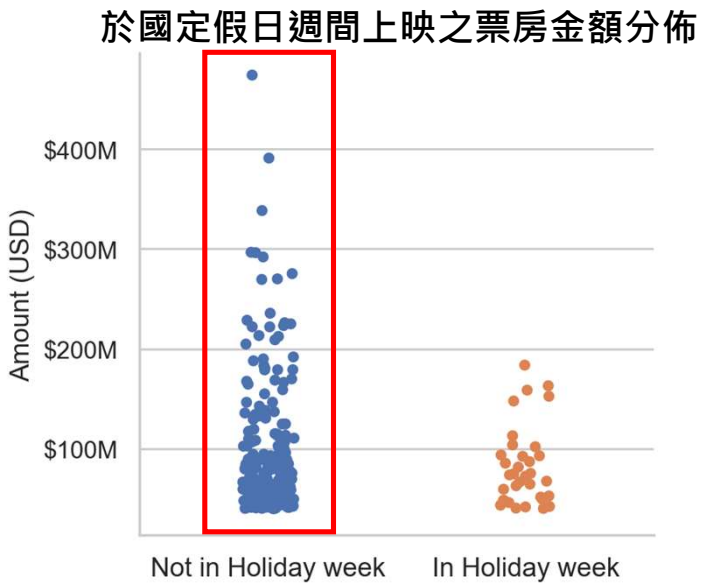


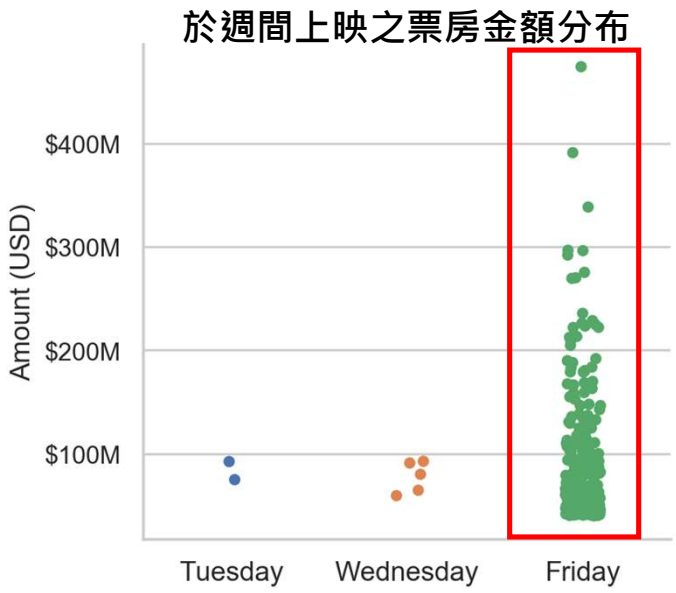
高票房之首週票房金額分布



成果分析：上映檔期

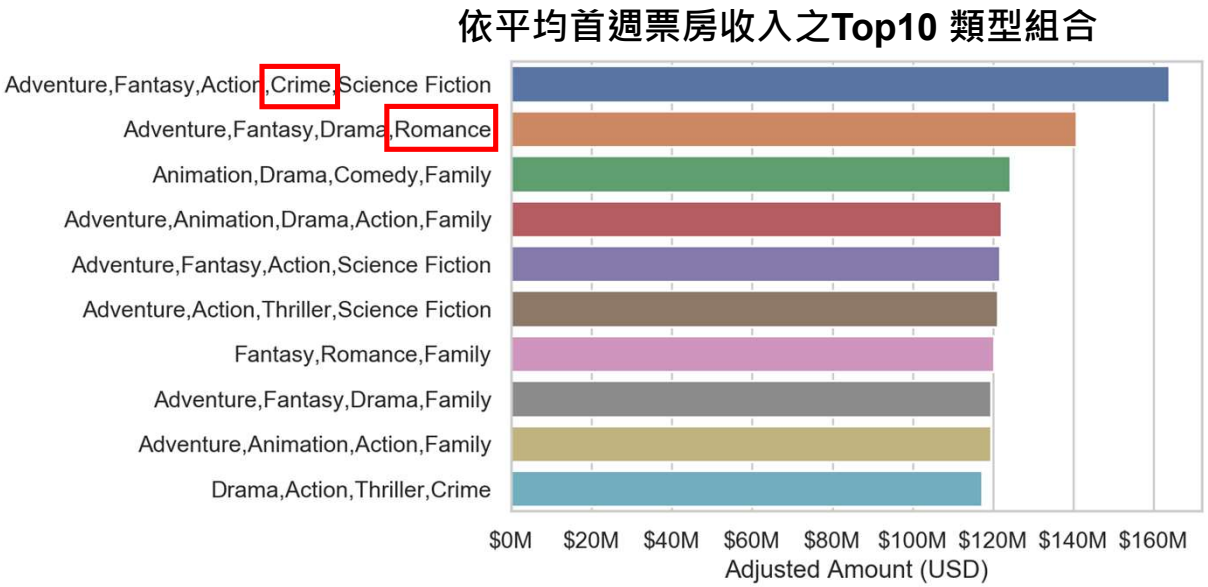
- **暑假期間（5~7月份）**上映，電影間競爭高，但票房金額分布普遍表現較好
- **不在國定假日週**上映對票房表現較佳
- 於**星期五**上映是不變的定律

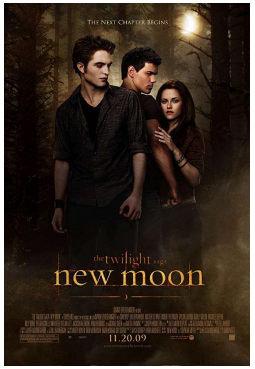




成果分析：電影類型

- **歷史、紀錄片**類型完全不受觀眾青睞
- 累積票房收入最高的電影類型組合：**冒險+動作+科幻**
- 在類型組合中多加上**犯罪、愛情**元素會增加票房表現





Source：https://www.themoviedb.org

感謝聆聽