IMDb Dataset

資料庫管理 期末報告

22 May, 2025

生技三 4111030338 林竣睿

Outlines

- **資料集簡介**
- **資料前處理**
- **ER model**
- **SQL**
- **資料視覺化**

Dataset Introduction
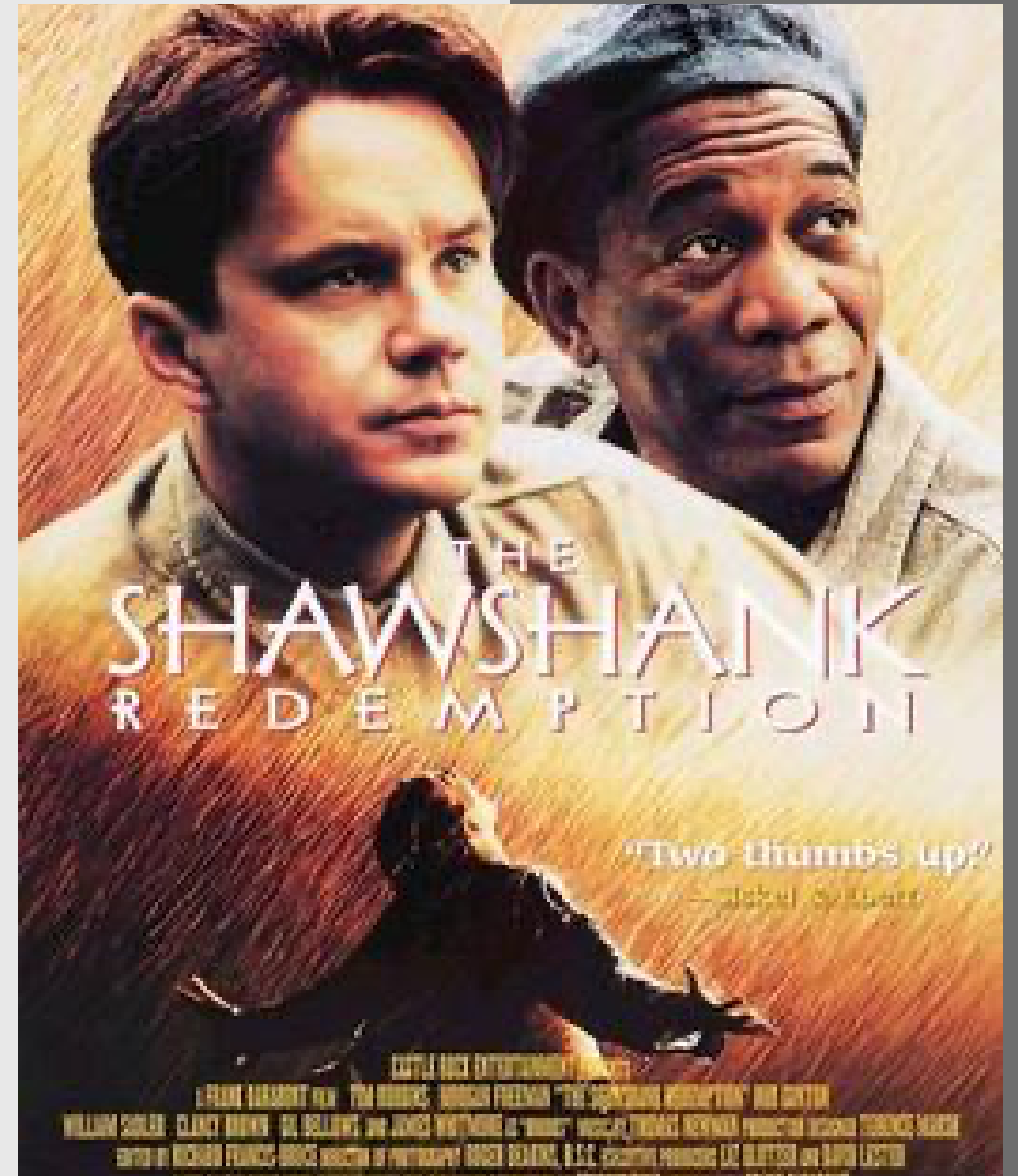
來源：
- 本資料集自Kaggle下載
  (https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset)
- 紀錄2021 年底時 IMDb評分前250的電影

欄位：
- 包含：rank, year, rating, genre等

注意事項:
- 部分資料有缺失值
- 時間欄位需要進行轉換

Dataset Introduction

	id	rank	name	year	rating	genre	run_time	budget	box_office
0	e9e7	1	The Shawshank Redemption	1994	9.3	Drama	142	25000000.0	2.888450e+07
1	7219	2	The Godfather	1972	9.2	Crime,Drama	175	6000000.0	2.503418e+08
2	df0d	3	The Dark Knight	2008	9.0	Action,Crime,Drama	152	185000000.0	1.006234e+09
3	a443	4	The Godfather Part II	1974	9.0	Crime,Drama	202	13000000.0	4.796192e+07
4	8dae	5	12 Angry Men	1957	9.0	Crime,Drama	96	350000.0	9.550000e+02

certificate	tagline	casts	directors	writers
R	Fear can hold you prisoner. Hope can set you f...	Tim Robbins,Morgan Freeman,Bob Gunton,William ...	Frank Darabont	Stephen King,Frank Darabont
R	An offer you can't refuse.	Marlon Brando,Al Pacino,James Caan,Diane Keato...	Francis Ford Coppola	Mario Puzo,Francis Ford Coppola
PG-13	Why So Serious?	Christian Bale,Heath Ledger,Aaron Eckhart,Mich...	Christopher Nolan	Jonathan Nolan,Christopher Nolan,David S. Goyer
R	All the power on earth can't change destiny.	Al Pacino,Robert De Niro,Robert Duvall,Diane K...	Francis Ford Coppola	Francis Ford Coppola,Mario Puzo
Approved	Life Is In Their Hands -- Death Is On Their Mi...	Henry Fonda,Lee J. Cobb,Martin Balsam,John Fie...	Sidney Lumet	Reginald Rose

| #	Column	Non-Null Count	Dtype
0	id	250 non-null	object
1	rank	250 non-null	int64
2	name	250 non-null	object
3	year	250 non-null	int64
4	rating	250 non-null	float64
5	genre	250 non-null	object
6	run_time	250 non-null	int64
7	budget	231 non-null	float64
8	box_office	238 non-null	float64
9	certificate	250 non-null	object
10	tagline	250 non-null	object
11	casts	250 non-null	object
12	directors	250 non-null	object
13	writers	250 non-null	object

Data Preprocessing

資料清理
- 移除不符合欄位定義的值
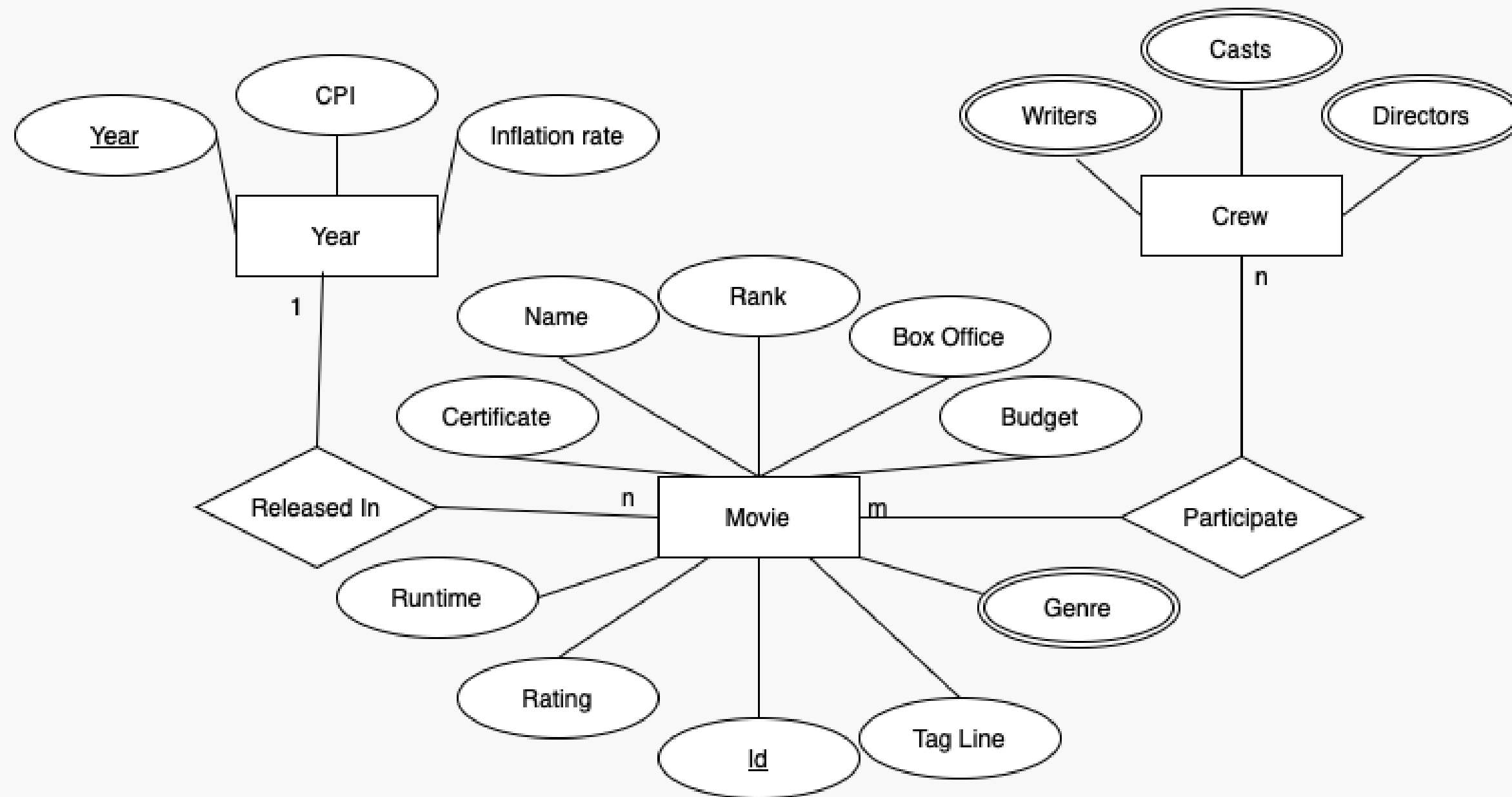- 驗證欄位資料是否合理
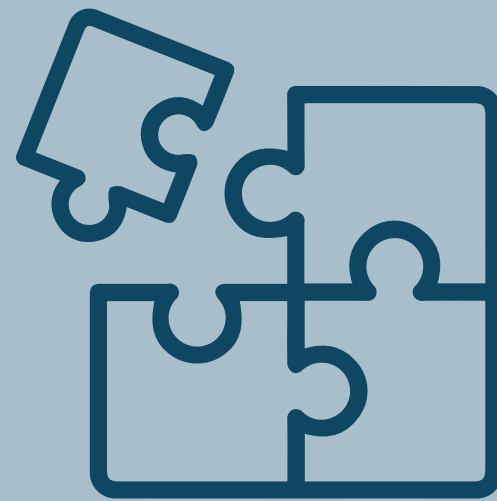
資料合併
- 引入CPI 資料
- 將更新資料引入至資料集

資料轉換
- 將文字資料編碼使資料容易分析、利用
- 轉換時間格式,並採用Bining建立相關性

ER Model

- **Movie**
  - 為主要的Instance，有一個電影的主要屬性。
- **Crew**
  - 為一部電影的團隊，包含演員、導演、編劇等。
  - 透過Participate 關係與Movie 聯繫。
- **Year**
  - 電影的上映年份。
  - 部分電影年代較久遠，因此引入CPI 來提供比較依據。

SQL

- **查詢特定年份間，票房表現最佳的電影:**

  SELECT name, year, box_office
  FROM Movies
  WHERE year >= 2000 **AND** year <= 2020
  ORDER BY box_office **DESC**
  LIMIT 5;

	name	year	box_office
1	Avengers: Endgame	2019	2799439100.0
2	Avengers: Infinity War	2018	2052415039.0
3	Harry Potter and the Deathly Hallows: Part 2	2011	1342359942.0
4	The Lord of the Rings: The Return of the King	2003	1146457748.0
5	The Dark Knight Rises	2012	1081169825.0

- **查詢特定導演的作品:**

  SELECT m.name, m.year, m.rating, cr.directors
  FROM Movies m
  JOIN Crews cr **ON** m.id = cr.id
  WHERE cr.directors **LIKE** '%Christopher Nolan%'
  AND m.rating >= 8.0
  ORDER BY m.rating **DESC**;

	name	year	rating	directors
1	The Dark Knight	2008	9.0	Christopher Nolan
2	Inception	2010	8.8	Christopher Nolan
3	Interstellar	2014	8.6	Christopher Nolan
4	The Prestige	2006	8.5	Christopher Nolan
5	Memento	2000	8.4	Christopher Nolan
6	The Dark Knight Rises	2012	8.4	Christopher Nolan
7	Batman Begins	2005	8.2	Christopher Nolan

SQL

- **查詢特定類型電影的平均評分與上映數量:**

SELECT genre,
      COUNT(*) AS movie_count,
      ROUND(AVG(rating), 2) AS average_rating
FROM Movies
WHERE genre LIKE '%Action%'
GROUP BY genre
HAVING movie_count >= 2
ORDER BY average_rating DESC
LIMIT 5;

	genre	movie_count	average_rating
1	Action,Adventure,Drama	5	8.7
2	Action,Drama,Mystery	2	8.5
3	Action,Sci-Fi	3	8.47
4	Action,Crime,Drama	5	8.44
5	Action,Drama	3	8.43

- **查詢高投資回報率的電影:**

SELECT name, year, budget, box_office,
      ROUND((box_office / budget), 2) AS
return_on_investment
FROM Movies
WHERE budget > 0 AND box_office IS NOT NULL
ORDER BY return_on_investment DESC
LIMIT 10;

	name	year	budget	box_office	return_on_investment
1	Rocky	1976	960000.0	117250402.0	122.14
2	Gone with the Wind	1939	3977000.0	402382193.0	101.18
3	Star Wars: Episode IV - A New Hope	1977	11000000.0	775398007.0	70.49
4	Jaws	1975	7000000.0	476512065.0	68.07
5	A Separation	2011	500000.0	22926076.0	45.85
6	The Intouchables	2011	9500000.0	426588510.0	44.9
7	The Godfather	1972	6000000.0	250341816.0	41.72
8	The Exorcist	1973	11000000.0	441306145.0	40.12
9	Psycho	1960	806947.0	32052925.0	39.72
10	The Lives of Others	2006	2000000.0	77356942.0	38.68

SQL
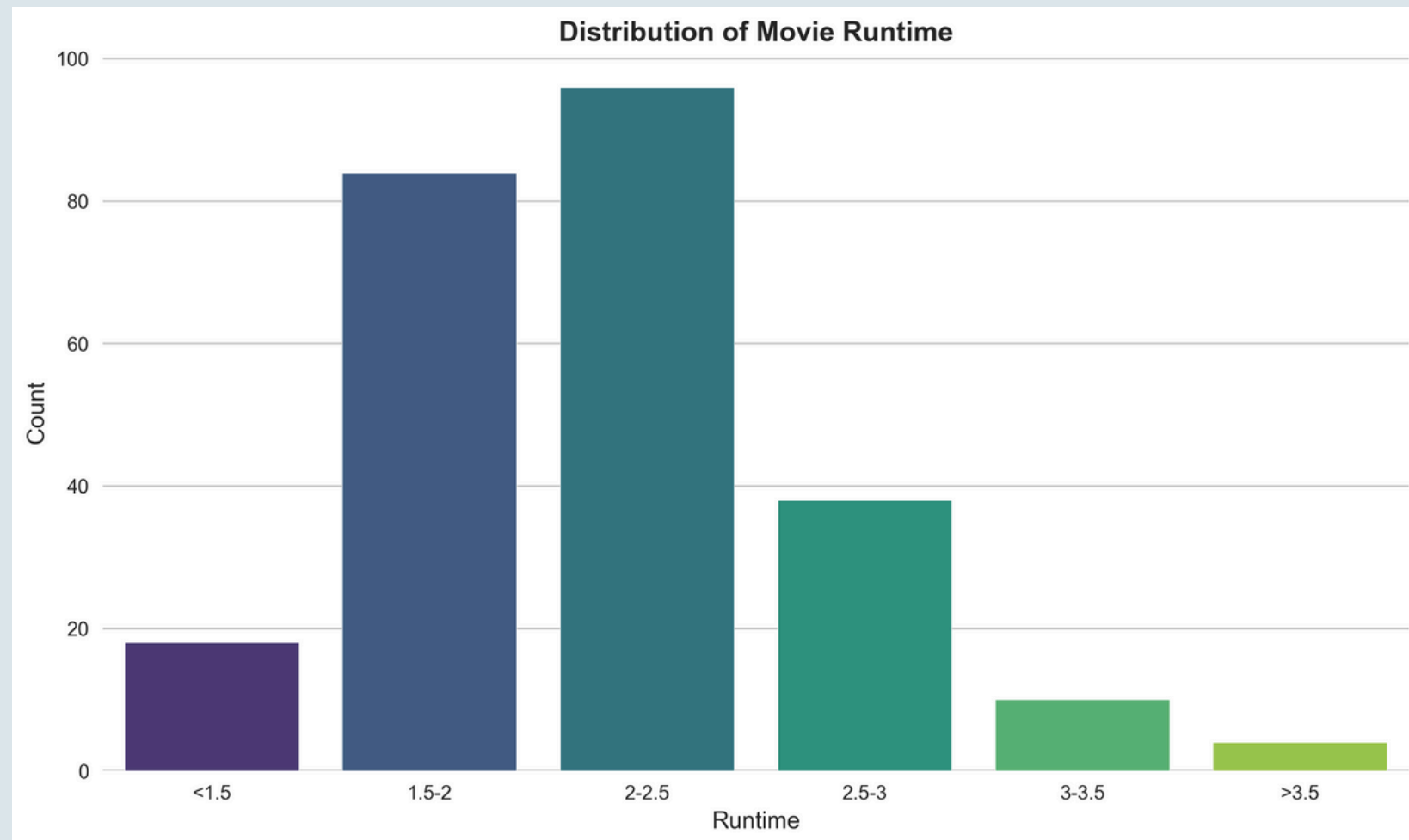
- **計算電影的實際票房收入（考慮通貨膨脹）：**
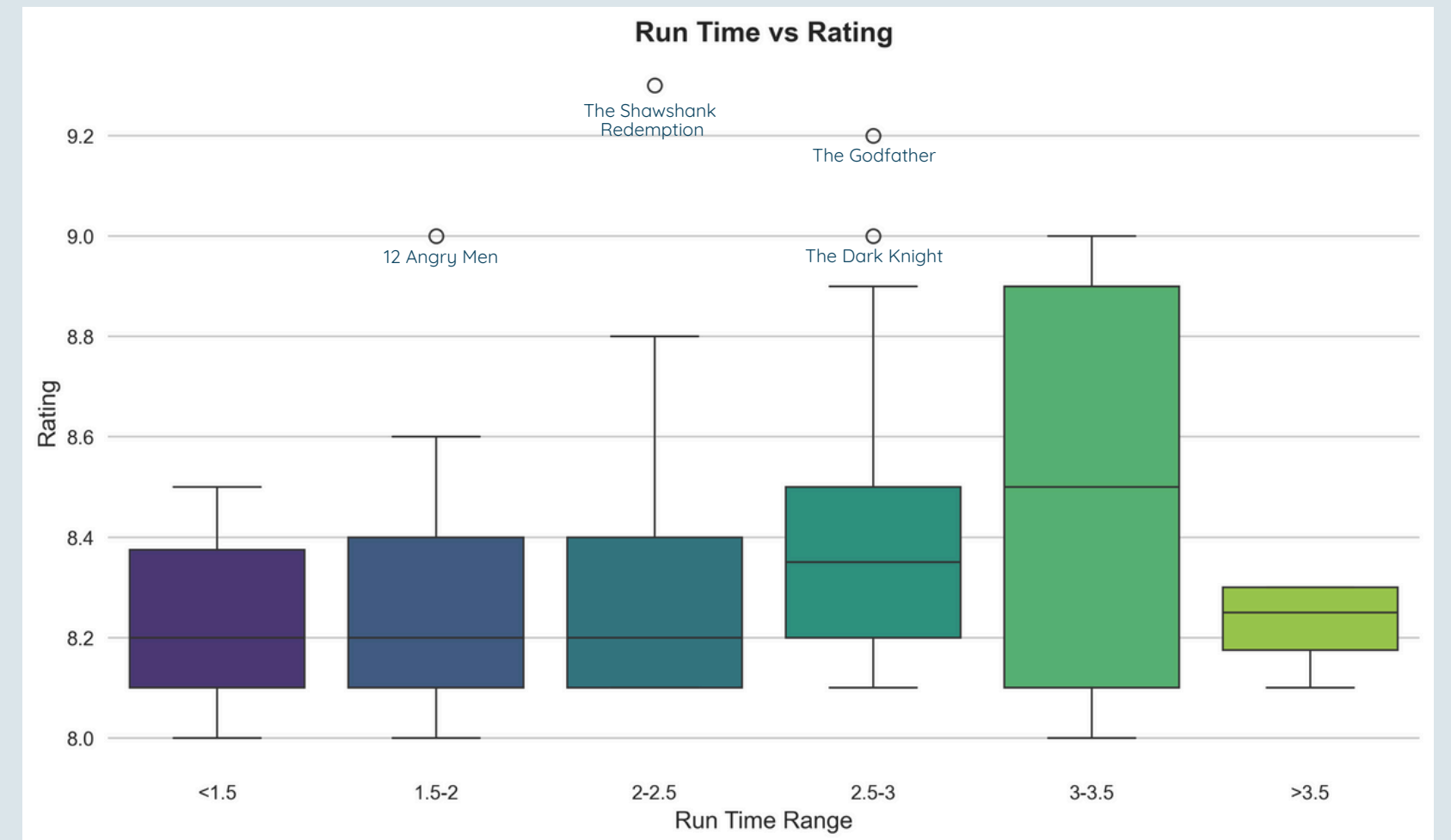
```
SELECT m.name, m.year, m.box_office,
        (m.box_office * (SELECT c2."Annual Average CPI(-U)"
                        FROM Cpi c2
                        WHERE c2.year = 2024) / c."Annual Average CPI(-U)") AS adjusted_box_office
FROM Movies m
JOIN Cpi c ON m.year = c.year
WHERE m.box_office IS NOT NULL
ORDER BY adjusted_box_office DESC
LIMIT 10;
```

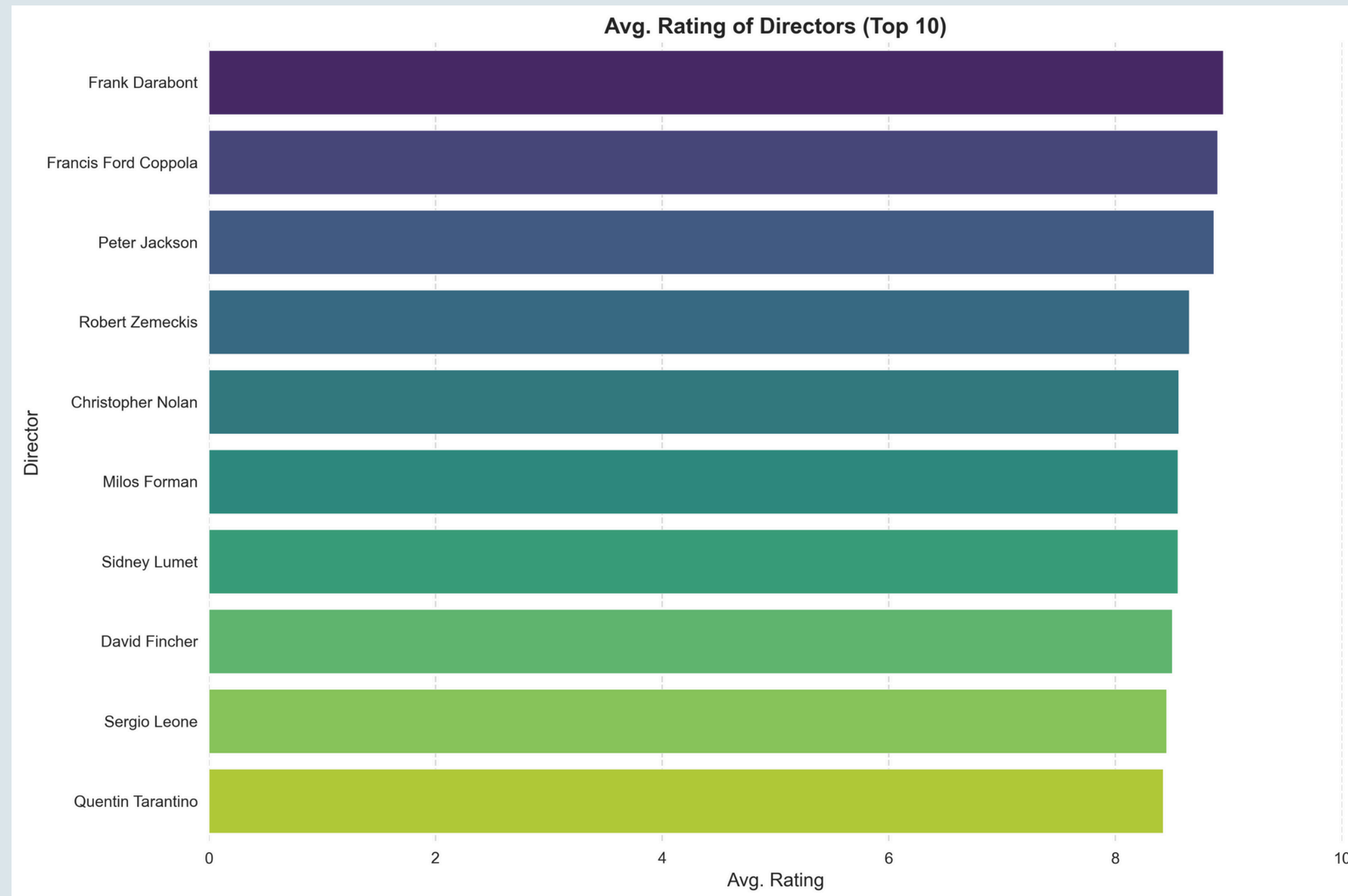|	name	year	box_office	adjusted_box_office
1	Gone with the Wind	1939	402382193.0	9101364135.19424
2	Star Wars: Episode IV - A New Hope	1977	775398007.0	4022856986.81188
3	Avengers: Endgame	2019	2799439100.0	3442094849.58936
4	The Exorcist	1973	441306145.0	3124924594.32432
5	Jaws	1975	476512065.0	2784672736.72862
6	Avengers: Infinity War	2018	2052415039.0	2569809989.09438
7	Jurassic Park	1993	1109802321.0	2414684081.12388
8	Spider-Man: No Way Home	2021	1921847111.0	2229626316.23026
9	The Lion King	1994	968511805.0	2054656622.75304
10	Star Wars: Episode V - The Empire Strikes Back	1980	538375067.0	2054188362.43689

Data Visualisation

Distribution of Movie Runtime

- 透過分群，可以發現大多數電影時長介於 1.5 ~ 2.5 小時

Run Time vs Rating

The Shawshank Redemption
The Godfather
12 Angry Men
The Dark Knight

- 片長評分箱型圖。可以得到評分特別高的電影。

Data Visualisation

Avg. Rating of Directors (Top 10)

- 透過平均分數，計算排名前10的導演

Data Visualisation

Top 10 movies' Budget and Box office(CPI Adjusted)
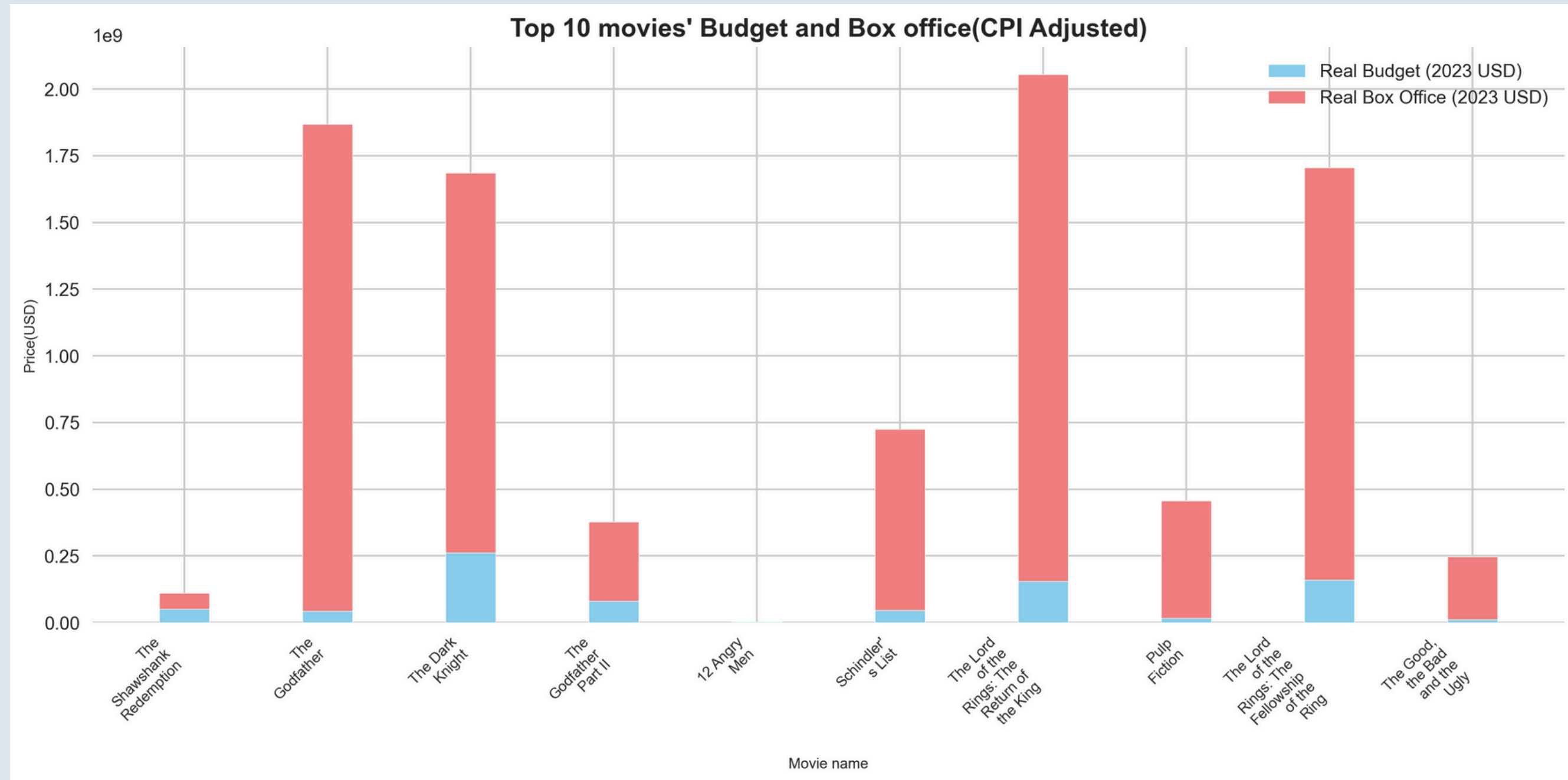
- 排名前10電影，其預算與票房關係

Thank you