

# 電子遊戲銷售

by 第四組

112029038 林資竣

112029023 游雲翊

112012067 秦晨恩

112029039 黃柏誠

112029051 黃柏堯

# 背景

組員中有人未來想開發電子遊戲，因此我們希望了解：一款遊戲是否能成為暢銷作品，究竟受到哪些因素影響？本研究透過分析全球電子遊戲銷售資料，觀察遊戲的平台、類型、上市年份、發行商及區域銷售等特徵，找出哪些因素最能影響遊戲「是否成為暢銷遊戲」。透過分析結果，我們希望能提供遊戲開發與行銷方向的參考，幫助未來設計出的遊戲更有機會成為熱門作品。

# 研究目的

建立「暢銷遊戲」定義 ( $\text{Global\_Sales} > 1.0$ )  
找出影響遊戲是否成為暢銷作品的關鍵特徵  
建立分類預測模型，預測遊戲是否可能成為暢銷遊戲  
根據分析與模型結果提出遊戲開發與行銷策略建議

# 資料描述

各種遊戲銷售紀錄 總共16598筆資料 11個蒐集項目

	A	B	C	D	E	F	G	H	I	J	K
1	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
2	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
3	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
4	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
5	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
6	5	Pokemon Red/Pokemon GB	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
7	6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
8	7	New Super Mario Bros DS	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
9	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
10	9	New Super Mario Bros Wii		2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
11	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
12	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
13	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
14	13	Pokemon Gold/Pokemon GB	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
15	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
16	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
17	16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studio	14.97	4.94	0.24	1.67	21.82
18	17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
19	18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81

研究所使用的資料集為公開的 Video Game Sales Dataset，其資料來源為 VGChartz 所彙整之全球遊戲銷售紀錄。資料內容涵蓋 1980 至 2020 年間發售的遊戲，包括遊戲基本資訊與各地區銷售量。

# 資料描述

	A	B	C	D	E	F	G	H	I	J	K
1	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
2	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
3	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
4	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
5	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
6	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
7	6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
8	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
9	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
10	9	New Super Mario Bros. 2	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
11	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
12	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
13	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
14	13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
15	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
16	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
17	16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
18	17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
19	18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81

蒐集項目含遊戲全球銷售**排名**、遊戲**名稱**、遊戲**平台**、發售**年份**、遊戲**類型**、**發行商**、**北美銷售量**、**歐洲銷售量**、**日本銷售量**、**其他地區銷售量**、**全球銷售總合**

# 欄位介紹

1	Rank	遊戲全球銷售排名	連續型資料
2	Name	遊戲名稱	類別型資料
3	Platform	遊戲平台	類別型資料
4	Year	發售年份	連續型資料
5	Genre	遊戲類型	類別型資料

# 欄位介紹

	Publisher	發行商	類別型資料
6	NA_Sales	北美銷售 (百萬套)	連續型資料
7	EU_Sales	歐洲銷售 (百萬套)	連續型資料
8	JP_Sales	日本銷售 (百萬套)	連續型資料
9	Other_Sales	其他地區銷售 (百萬套)	連續型資料
10	Global_Sales	全球銷售總和 (百萬套)	連續型資料
11			

# 資料型態

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Rank        16598 non-null   int64  
 1   Name         16598 non-null   object  
 2   Platform     16598 non-null   object  
 3   Year         16327 non-null   object  
 4   Genre        16598 non-null   object  
 5   Publisher    16540 non-null   object  
 6   NA_Sales     16598 non-null   float64 
 7   EU_Sales     16598 non-null   float64 
 8   JP_Sales     16598 non-null   float64 
 9   Other_Sales  16598 non-null   float64 
 10  Global_Sales 16596 non-null   float64 
dtypes: float64(5), int64(1), object(5)
memory usage: 1.4+ MB
None
```

連續型資料:

Rank  
NA\_Sales  
EU\_Sales  
JP\_Sales  
Other\_Sales  
Global\_Sales

類別型資料:

Year  
Name  
Platform  
Genre  
Publisher

# 資料前處理

# 特徵削除 (FEATURE REMOVAL)

Rank 為依照 `global_sales` 排序所得，與目標變數高度相關。

若納入模型會造成資料洩漏，因此在建模前予以移除Name。

純文字資料（非結構化），無法直接用於機器學習

字串種類太多 (16000+) → 會造成維度爆炸

與是否暢銷無明確因果關係  
因此刪除避免模型混亂

```
import pandas as pd

# 讀取資料 (csv 與 notebook 在同一資料夾)
df = pd.read_csv("vgsales.csv")

# 刪除 Rank 欄位
df = df.drop(columns=["Rank", "Name"])

# 查看前幾筆資料
df.head()
```

# 特徵削除

# (FEATURE REMOVAL)

	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37

# 檢查遺漏值

```
df.isnull().sum()
```

Platform	0
Year	271
Genre	0
Publisher	58
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	2
dtype:	int64

可以發現Year 欄位缺失 271 筆  
Publisher 缺失 58 筆  
Global\_Sales (目標變數) 缺失 2 筆  
其他欄位皆無缺失  
接著依照變數性質進行適當的缺失值處理

# 檢查遺漏值

```

import pandas as pd

# 讀取資料
df = pd.read_csv("vgsales.csv")

# 1. 刪除不需要的欄位 (Rank、Name)
df = df.drop(columns=["Rank", "Name"])

# 2. 缺失值處理
# Global_Sales : 重要欄位 → 刪除缺失值
df = df.dropna(subset=["Global_Sales"])

# Publisher : 類別欄位 → 補為 Unknown
df["Publisher"] = df["Publisher"].fillna("Unknown")

# Year : 年份不可補 → 刪除缺失
df = df.dropna(subset=["Year"])

# 3. 確認缺失值是否都處理完畢
print(df.isnull().sum())

# 查看處理後的筆數
print("處理後筆數:", len(df))

```

Platform	0
Year	0
Genre	0
Publisher	0
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0
dtype: int64	
處理後筆數：	16325

- ① 遺漏值檢查結果
- ② Year 缺失 271 筆 (刪除)
- ③ Publisher 缺失 58 筆 (補 Unknown)
- ④ Global\_Sales 缺失 2 筆 (刪除)
- ⑤ 其他欄位皆完整
- ⑥ 處理後無缺失值，剩餘 16,325 筆資料。

# 建立 HIT

為了將「遊戲銷售表現」轉換為模型可以判斷的分類問題，  
需要將連續的 Global\_Sales（全球銷售量）  
轉換為：

- 1 (Hit) = 暢銷遊戲
- 0 (Non-Hit) = 非暢銷遊戲

這樣模型才能進行 預測暢銷／不暢銷。

```
# 建立 Hit 欄位：1 = 暢銷、0 = 非暢銷  
df["Hit"] = (df["Global_Sales"] > 1.0).astype(int)  
  
# 查看暢銷與非暢銷的筆數  
print(df["Hit"].value_counts())
```

```
Hit  
0    14293  
1     2032  
Name: count, dtype: int64
```



# 資料編碼

將「發行商（Publisher）進行編碼」，  
我們選擇計算每個 Publisher 成功遊戲數量  
排序 → 取前 10 名

代表的是：最容易出「暢銷作品」的發行商  
執行完後會得到：

✓ **top10\_publishers**

→ 10 個最會出暢銷遊戲的發行商（例如 Nintendo、EA、Activision 等）

✓ Top10Publisher 欄位 (0 / 1)

- 1 = 此發行商是前十名暢銷大品牌
- 0 = 不是大品牌

# 資料編碼

Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Hit	Top10Publisher
Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74	1	1
NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	1	1
Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82	1	1
Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33	1	1
GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37	1	1
GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26	1	1
DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01	1	1
Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02	1	1
Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62	1	1
NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31	1	1
DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76	1	1
DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42	1	1
GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1	1	1
Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72	1	1
Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22	1	1
X360	2010	Misc	Microsoft Game	14.97	4.94	0.24	1.67	21.82	1	0
PS3	2013	Action	Take-Two Interac	7.01	9.27	0.97	4.14	21.4	1	1
PS2	2004	Action	Take-Two Interac	9.43	0.4	0.41	10.57	20.81	1	1
SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61	1	1

# 資料編碼

由於 Platform 與 Genre 為類別變數，本研究使用 Label Encoding 將其轉換為數字代碼，以便模型能正確處理類別資訊。編碼僅代表類別代號，並無數字大小意義。

```
# 6. Label Encoding (平台 / 類型)
# =====
le_platform = LabelEncoder()
le_genre = LabelEncoder()

df["Platform_LE"] = le_platform.fit_transform(df["Platform"])
df["Genre_LE"] = le_genre.fit_transform(df["Genre"])

# 保留 Publisher (文字) 與 Top10Publisher (0/1)
# 不做 Publisher Label Encoding

# 刪除原本文字的 Platform、Genre
df = df.drop(columns=["Platform", "Genre"])

# =====
# 7. 匯出 Excel (推薦)
# =====
df.to_excel("vgsales_clean_labelencoded.xlsx", index=False)
print("✓ 已匯出：vgsales_clean_labelencoded.xlsx")

# =====
# 8. 查看結果
# =====
print(df.head())
print("欄位數量：" , df.shape[1])
```

# 資料編碼

A	B	C	D	E	F	G	H	I	J	K
Year	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Hit	Top10Publisher	Platform_1E	Genre_1E
2006	Nintendo	41.49	29.02	3.77	8.46	82.74	1	1	26	10
1985	Nintendo	29.08	3.58	6.81	0.77	40.24	1	1	11	4
2008	Nintendo	15.85	12.88	3.79	3.31	35.82	1	1	26	6
2009	Nintendo	15.75	11.01	3.28	2.96	33	1	1	26	10
1996	Nintendo	11.27	8.89	10.22	1	31.37	1	1	5	7
1989	Nintendo	23.2	2.26	4.22	0.58	30.26	1	1	5	5
2006	Nintendo	11.38	9.23	6.5	2.9	30.01	1	1	4	4
2006	Nintendo	14.03	9.2	2.93	2.85	29.02	1	1	26	3
2009	Nintendo	14.59	7.06	4.7	2.26	28.62	1	1	26	4
1984	Nintendo	26.93	0.63	0.28	0.47	28.31	1	1	11	8
2005	Nintendo	9.07	11	1.93	2.75	24.76	1	1	4	9
2005	Nintendo	9.81	7.57	4.13	1.92	23.42	1	1	4	6
1999	Nintendo	9	6.18	7.2	0.71	23.1	1	1	5	7
2007	Nintendo	8.94	8.03	3.6	2.15	22.72	1	1	26	10
2009	Nintendo	9.09	8.59	2.53	1.79	22	1	1	26	10
2010	Microsoft Game St	14.97	4.94	0.24	1.67	21.82	1	0	28	3
2013	Take-Two Interacti	7.01	9.27	0.97	4.14	21.4	1	1	17	0
2004	Take-Two Interacti	9.43	0.4	0.41	10.57	20.81	1	1	16	0
1990	Nintendo	12.78	3.75	3.54	0.55	20.61	1	1	23	4

# 資料編碼

為方便解讀資料，我們將 Platform 與 Genre 的原始類別與其編碼值輸出為對照表 (mapping)，並存於 Excel 的不同工作表中，提供後續模型解釋使用。

```
# -----
# 產生 Label Encoding 對照表
# -----



# Platform 對照表
platform_map = pd.DataFrame({
    "Platform": le_platform.classes_,
    "Platform_LE": range(len(le_platform.classes_))
})

# Genre 對照表
genre_map = pd.DataFrame({
    "Genre": le_genre.classes_,
    "Genre_LE": range(len(le_genre.classes_))
})

# 匯出 (可放在同一個 Excel 的不同 Sheet)
with pd.ExcelWriter("label_mapping.xlsx") as writer:
    platform_map.to_excel(writer, sheet_name="Platform_Map", index=False)
    genre_map.to_excel(writer, sheet_name="Genre_Map", index=False)

print("✓ 已匯出 Label Encoding 對照表：label_mapping.xlsx")
```

# 資料編碼

Platform	Platform_LE	Genre	Genre_LE
2600	0 PS	15 Action	0
3DO	1 PS2	16 Adventure	1
3DS	2 PS3	17 Fighting	2
DC	3 PS4	18 Misc	3
DS	4 PSP	19 Platform	4
GB	5 PSV	20 Puzzle	5
GBA	6 SAT	21 Racing	6
GC	7 SCD	22 Role-Playing	7
GEN	8 SNES	23 Shooter	8
GG	9 TG16	24 Simulation	9
N64	10 WS	25 Sports	10
NES	11 Wii	26 Strategy	11
NG	12 WiiU		
PC	13 X360		
PCFX	14 XB		
	15 XOne		

# 敘述統計

全球銷售 **Global\_Sales** 平均為 0.54 百萬套，中位數為 0.17 百萬套

→ 大部分遊戲銷量偏低

北美 (NA) 銷售普遍高於其他區域

→ 均值 0.26，高於 EU (0.15) 與 JP (0.08)

標準差顯示銷量分布差異大

→ 市場由少數大作拉高平均值

暢銷遊戲 (Hit=1) 僅約 12%

→ 產生**資料不平衡現象**

```
# 數值欄位描述統計  
df.describe()
```

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Hit
count	16325.000000	16325.000000	16325.000000	16325.000000	16325.000000	16325.000000
mean	0.265447	0.147572	0.078663	0.048331	0.540290	0.124472
std	0.821636	0.508794	0.311576	0.189896	1.565819	0.330129
min	0.000000	0.000000	0.000000	0.000000	0.010000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.060000	0.000000
50%	0.080000	0.020000	0.000000	0.010000	0.170000	0.000000
75%	0.240000	0.110000	0.040000	0.040000	0.480000	0.000000
max	41.490000	29.020000	10.220000	10.570000	82.740000	1.000000

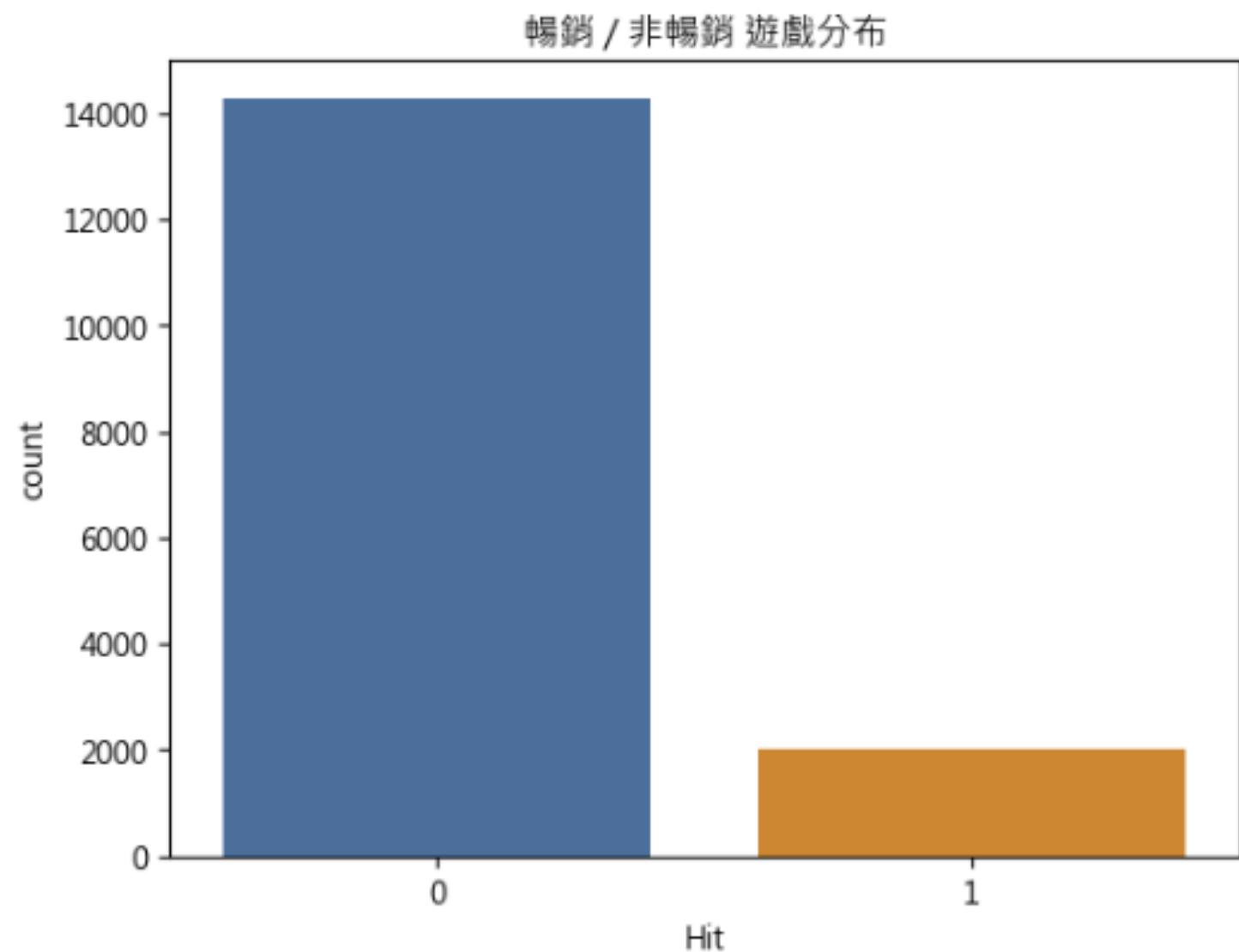
# 視覺化分析

## 平台 HIT 分布圖 (暢銷 VS 非暢銷)

本圖顯示暢銷與非暢銷遊戲的分布。

可以看到**非暢銷遊戲**佔大多數，資料呈現明顯**不平衡**。後續模型需要考量此特性，並找出哪些因素能有效預測暢銷。

```
sns.countplot(x="Hit", data=df)  
plt.title("暢銷 / 非暢銷 遊戲分布")  
plt.show()
```





## 不同平台的暢銷率 (HIT RATE)

```
platform_hit = df.groupby("Platform_LE")["Hit"].mean().sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=platform_hit.index, y=platform_hit.values)
plt.title("不同平台的暢銷率")
plt.xlabel("Platform LE")
plt.ylabel("Hit Rate")
plt.show()
```

# 不同平台的暢銷率 (HIT RATE)

平台之間的暢銷率**差異顯著**

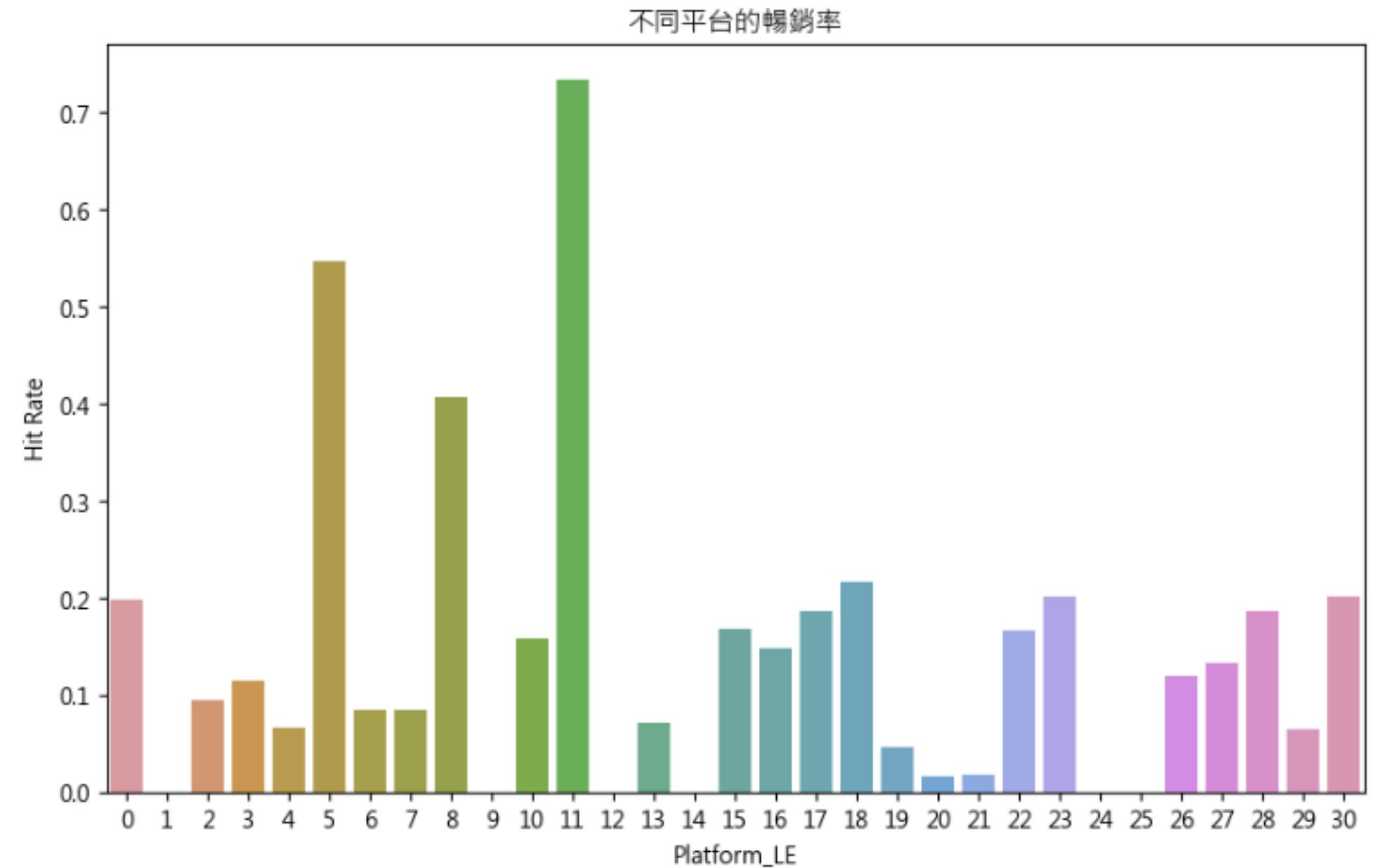
不同平台的 Hit Rate 從接近 0% 到超  
過 70% 不等，**差距非常大**。

少數平台擁有明顯**較高的暢銷率**，表  
示該平台較具市場競爭力。

多數平台 Hit Rate 偏低，意味著遊戲  
雖然多，但較少能成為暢銷作品。

編碼的數字僅為平台代號，並非排  
序。

**平台是影響遊戲銷售的重要因素之**



# 不同 GENRE (類型) 的暢銷率 (HIT RATE)

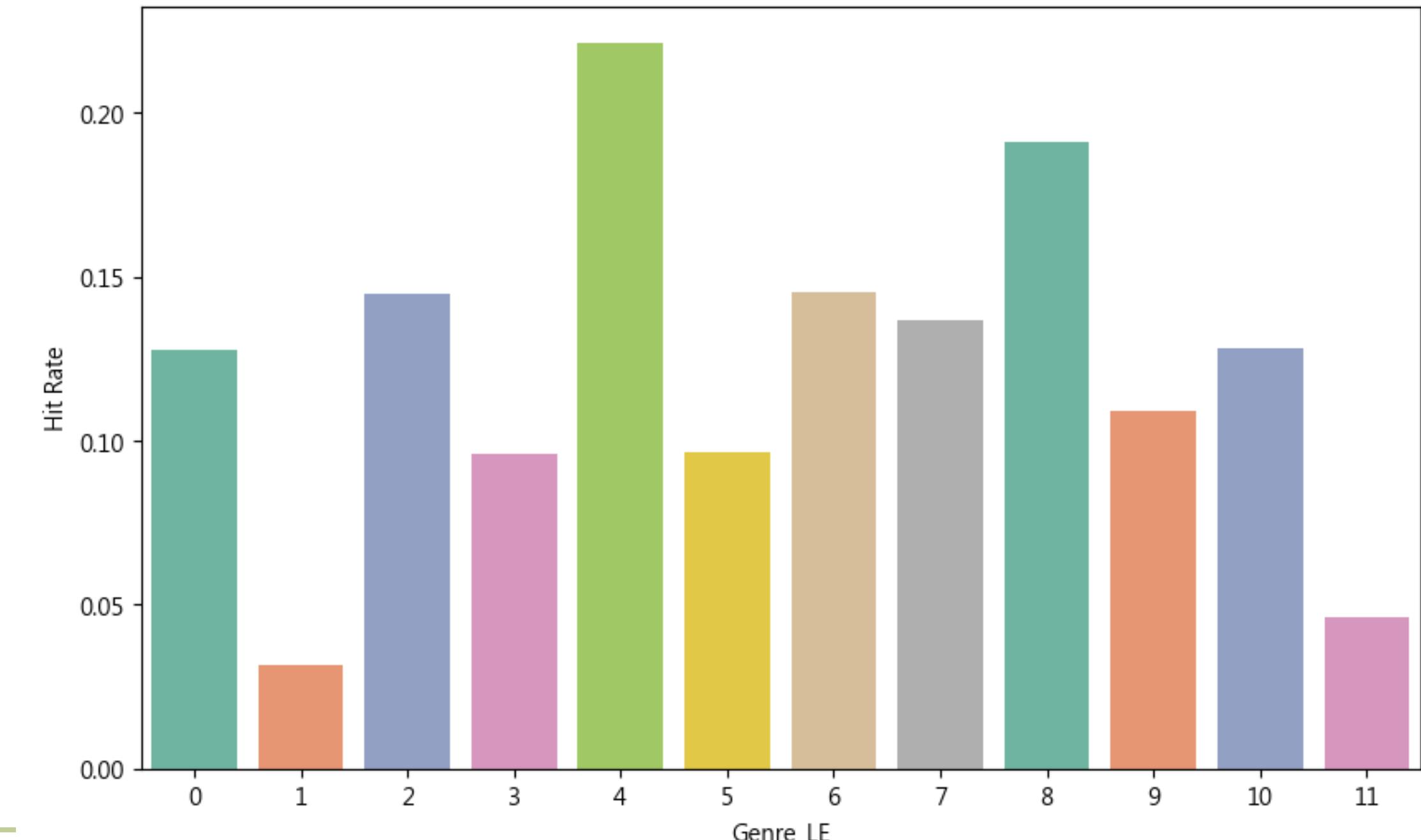
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

plt.rcParams['font.family'] = 'Microsoft JhengHei'
plt.rcParams['axes.unicode_minus'] = False

plt.figure(figsize=(10,6))
sns.barplot(x=genre_hit.index, y=genre_hit.values, palette="Set2")
plt.title("不同遊戲類型的暢銷率")
plt.xlabel("Genre LE")
plt.ylabel("Hit Rate")
plt.show()
```

# 不同 GENRE (類型) 的暢銷率 (HIT RATE)

不同遊戲類型的暢銷率



## 不同 GENRE (類型) 的暢銷率 (HIT RATE)

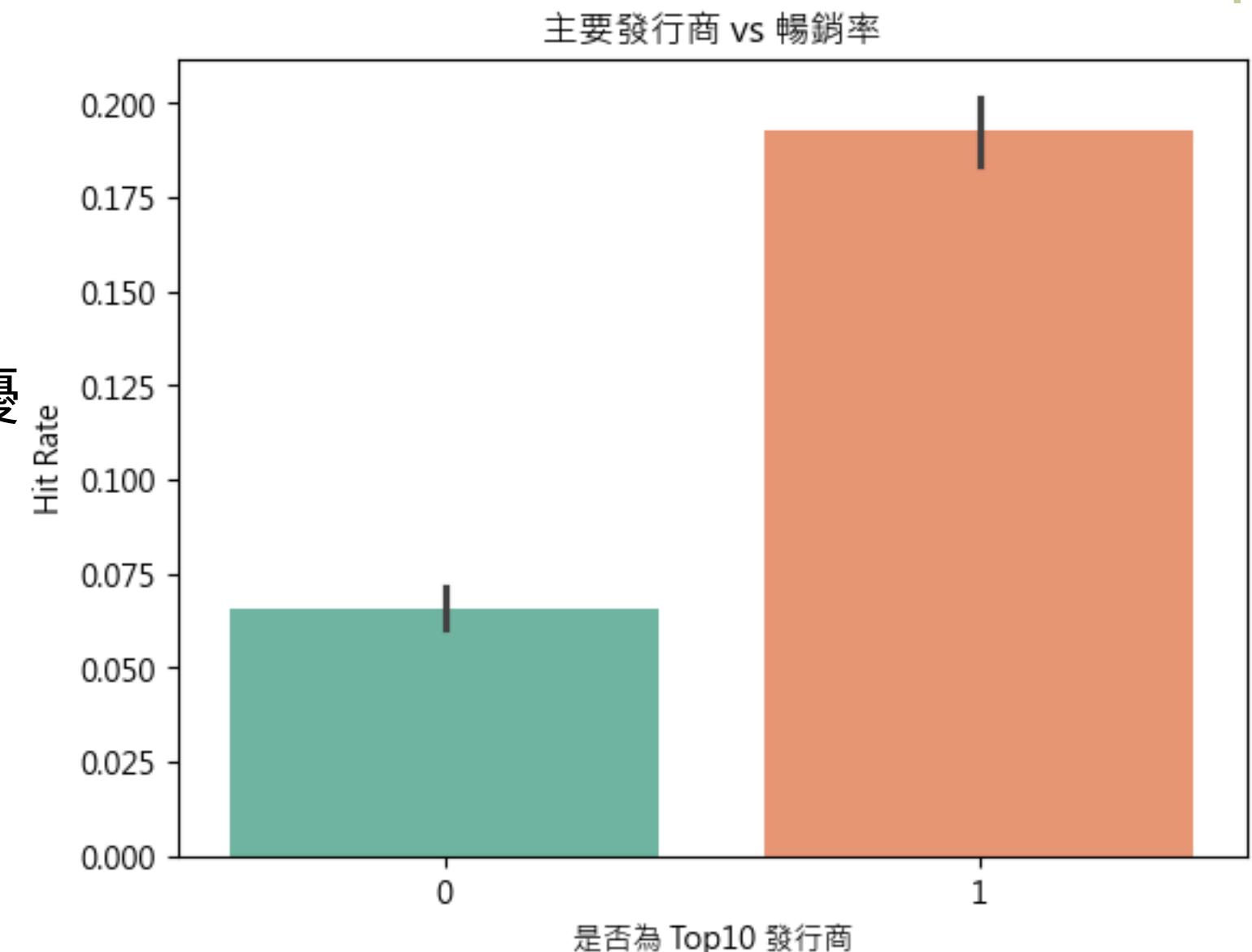
- ◆ 不同遊戲類型的暢銷率存在**明顯差異**。  
部分類型 (如  $\text{Genre\_LE} = 4, 8$ ) 的平均暢銷率**較高**，表示這些類型的遊戲**更容易達到高銷售量**。
  - ◆ 某些類型 (如  $\text{Genre\_LE} = 1, 11$ ) 的暢銷率相對**較低**，顯示該類型的遊戲**較不容易成為暢銷作品**。
  - ◆ 整體呈現類別間的不均衡性，表示「**遊戲類型**」確實對銷售表現有**一定影響**，但影響程度仍屬中度，並非決定性因素。
- 👉 可結論：遊戲類型會影響暢銷機率，但**非主要決定因素**，仍需與其他變數一起分析。

# TOP10PUBLISHER 與暢銷的關係

```
sns.barplot(x="Top10Publisher", y="Hit", data=df, palette="Set2")
plt.title("主要發行商 vs 暢銷率")
plt.xlabel("是否為 Top10 發行商")
plt.ylabel("Hit Rate")
plt.show()
```

# TOP10 PUBLISHER 與暢銷的關係

- ◆ 由 Top10 遊戲發行商所推出的遊戲，暢銷概率明顯較高。  
(Hit Rate 顯著高於非 Top10 的發行商)
- ◆ 非 Top10 發行商的平均暢銷率偏低，代表品牌知名度、行銷資源、通路能力等因素在銷售量上具有優勢效果。
- ◆ 主要發行商的影響力明顯大於遊戲類型與平台，顯示「發行商」是影響銷售的重要關鍵。
- 👉 可結論：Top10 發行商具有更高成功率，品牌與行銷能力對遊戲是否暢銷有顯著作用。



# 熱力圖分析

```
import seaborn as sns
import matplotlib.pyplot as plt

# 設定中文字體
plt.rcParams['font.family'] = 'Microsoft JhengHei'
plt.rcParams['axes.unicode_minus'] = False

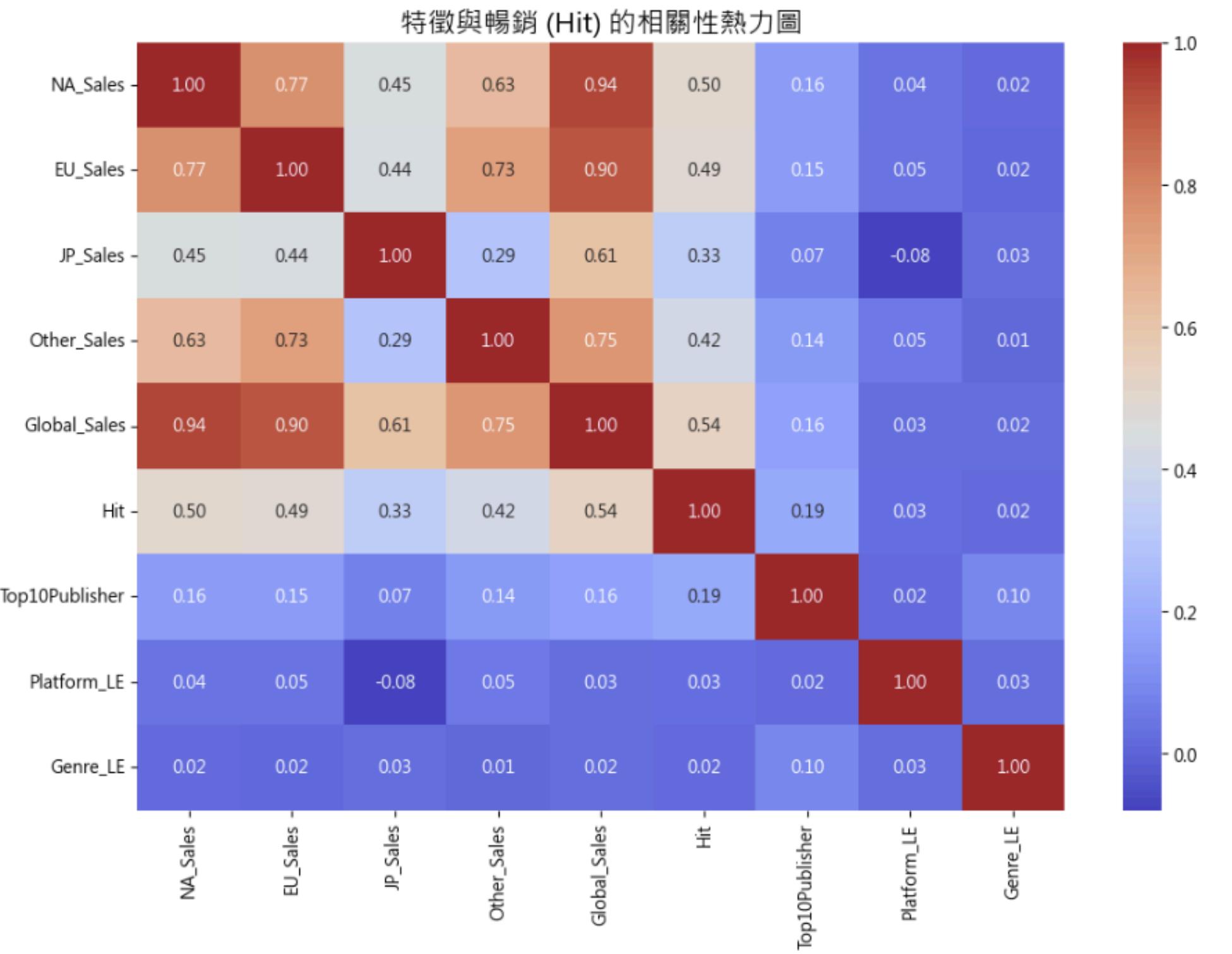
# 選擇要放進熱力圖的數值欄位
corr_cols = ["Year", "NA_Sales", "EU_Sales", "JP_Sales", "Other_Sales",
             "Global_Sales", "Hit", "Top10Publisher", "Platform_LE", "Genre_LE"]

plt.figure(figsize=(12, 8))
sns.heatmap(df[corr_cols].corr(numeric_only=True), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("特徵與暢銷 (Hit) 的相關性熱力圖", fontsize=16)
plt.show()
```

# 熱力圖分析

銷量決定暢銷：  
遊戲是否暢銷 (Hit) 主要取決於全球  
總銷售額 (相關係數  $r=0.54$ )。

平台與類型影響低：  
編碼後的遊戲平台和類型  
(Platform\_LE / Genre\_LE) 與銷售表現的線性相關性極低，表明這些特徵的編碼數值無法有效預測遊戲銷量。



# 實驗設計



## 實驗設計

本研究將「是否為暢銷遊戲」視為二元分類問題：

**目標變數** (Y)：HIT，1=暢銷 ( $\text{GLOBAL\_SALES} > 1.0$ )，0=非暢銷

**解釋變數** (X) 包含：

遊戲平台 (PLATFORM)

遊戲類型 (GENRE)

發行商相關特徵 (例如 TOP10PUBLISHER)

發售年份 (YEAR)

各區域銷售量 (NA / EU / JP / OTHER\_SALES)

**研究目標**：建立分類模型，預測遊戲是否可能成為暢銷作品，並分析各特徵的重要性。

## 資料切分

為了建立能泛化至未來新遊戲的預測模型，本研究將資料分為：

- 訓練集 (TRAINING SET): 80%
- 供模型學習特徵與銷售表現之間的關係。
- 測試集 (TESTING SET): 20%
- 用於評估模型在未見資料上的預測能力。

由於暢銷遊戲 ( $HIT=1$ ) 比例僅佔約 12%，資料呈現不平衡，因此採用 STRATIFIED SAMPLING (分層抽樣) 的方式，以確保訓練集與測試集中  $HIT=1$  的比例都與原始資料一致，避免模型因類別不平衡而偏向預測非暢銷遊戲。

# 資料切分

特徵欄位 (X) 包含：

- 年份 (YEAR)
- 各區域銷售 (NA / EU / JP / OTHER)
- 是否為前十大發行商 (TOP10PUBLISHER)
- 平台編碼 (PLATFORM\_LE)
- 類型編碼 (GENRE\_LE)

目標欄位 (Y)：

- HIT (0/1) 是否為暢銷遊戲

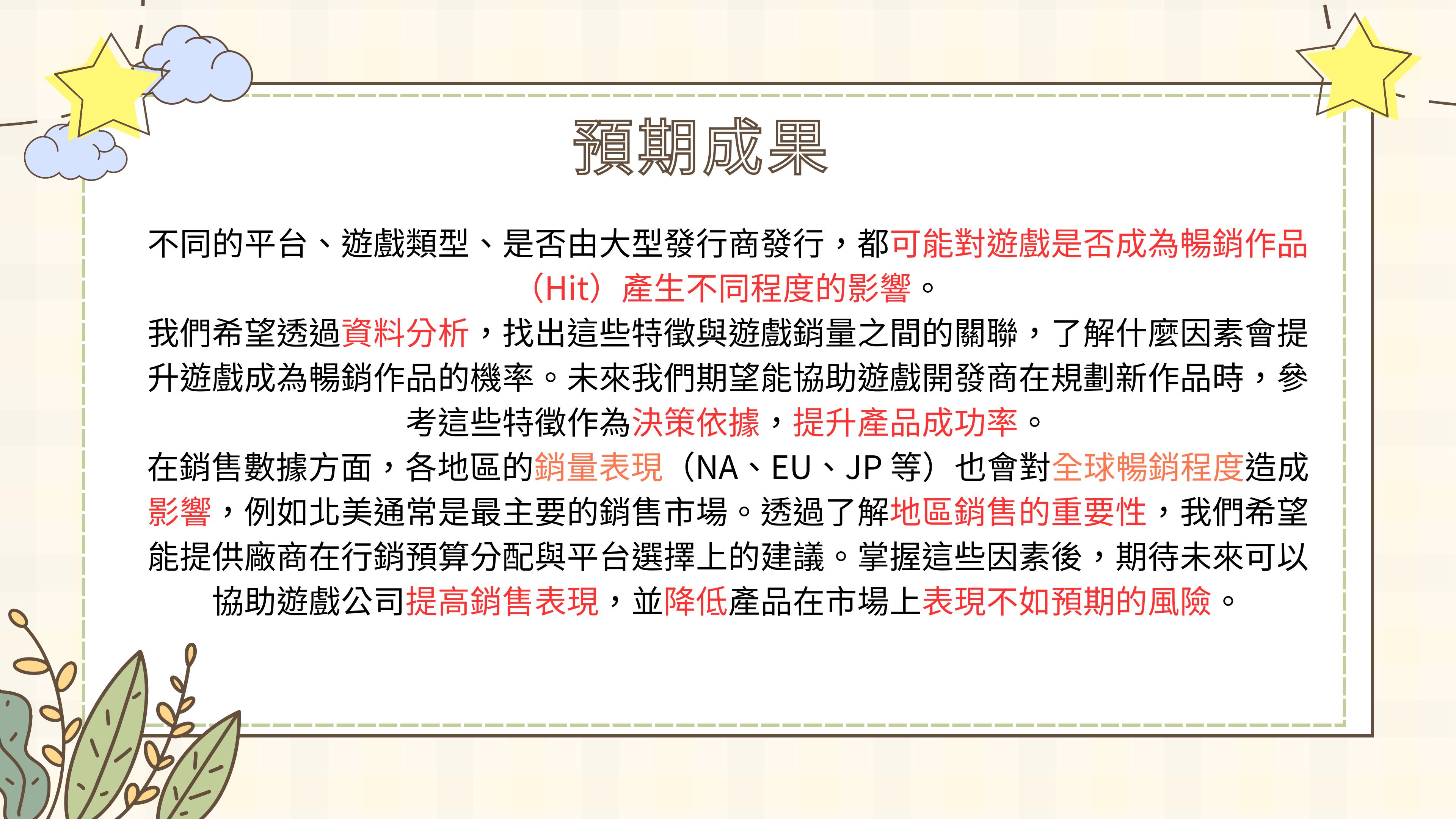
```
# 特徵欄位：都是已經編碼好或是數值型
feature_cols = [
    "Year",
    "NA_Sales", "EU_Sales", "JP_Sales", "Other_Sales",
    "Top10Publisher",      # 是否前十大發行商 (0/1)
    "Platform_LE",        # 平台編碼
    "Genre_LE"             # 類型編碼
]

X = df[feature_cols]
y = df["Hit"]
from sklearn.model_selection import train_test_split

# 分成 80% 訓練 + 20% 測試
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,          # 測試資料 20%
    random_state=42,
    stratify=y              # 保留 Hit=1 的比例
)

print("訓練資料筆數：" , len(X_train))
print("測試資料筆數：" , len(X_test))
```

訓練資料筆數： 13060  
測試資料筆數： 3265



## 預期成果

不同的平台、遊戲類型、是否由大型發行商發行，都可能對遊戲是否成為暢銷作品 (Hit) 產生不同程度的影響。

我們希望透過資料分析，找出這些特徵與遊戲銷量之間的關聯，了解什麼因素會提升遊戲成為暢銷作品的機率。未來我們期望能協助遊戲開發商在規劃新作品時，參考這些特徵作為決策依據，提升產品成功率。

在銷售數據方面，各地區的銷量表現 (NA、EU、JP 等) 也會對全球暢銷程度造成影響，例如北美通常是最主要的銷售市場。透過了解地區銷售的重要性，我們希望能提供廠商在行銷預算分配與平台選擇上的建議。掌握這些因素後，期待未來可以協助遊戲公司提高銷售表現，並降低產品在市場上表現不如預期的風險。

## Q&A

### 1. 連續型資料的平均數和標準差

	Rank	830.060	479.185
NA_Sales		2.647	8.167
EU_Sales		1.467	5.054
JP_Sales		778	3.093
Other_Sales		481	1.886
Global_Sales		5.375	15.551

## Q&A

### 2. 類別型資料的分布

類別型欄位：['NAME', 'PLATFORM', 'YEAR', 'GENRE', 'PUBLISHER']

Name	
Need for Speed: Most Wanted	12
LEGO Marvel Super Heroes	9
Ratatouille	9
Madden NFL 07	9
FIFA 14	9
..	
Farming 2017 - The Simulation	1
Kinect Adventures!	1
Wii Fit Plus	1
Wii Fit	1
Nintendogs	1

## Q&A

### 2. 類別型資料的分布

類別型欄位: **['NAME', 'PLATFORM', 'YEAR', 'GENRE', 'PUBLISHER']**

Platform					
DS	2163	PSV	413	GEN	27
PS2	2161	PS4	336	NG	12
PS3	1328	N64	319	SCD	6
Wii	1325	SNES	239	WS	6
X360	1265	XOne	213	3DO	3
PSP	1212	SAT	173	TG16	2
PS	1196	WiiU	143	2007	1
PC	960	2600	133	GG	1
XB	824	GB	98	2010	1
GBA	822	NES	98	PCFX	1
GC	556	DC	52		
3DS	509				

## Q&A

### 2. 類別型資料的分布

類別型欄位：['NAME', 'PLATFORM', 'YEAR', 'GENRE', 'PUBLISHER']

==== Year 類別分布 ===

Year	Count	Year	Count	Year	Count	Year	Count	Year	Count
2009	1431	2004	763	1994	121	1993	60	2020	1
2008	1428	2012	657	1981	46	1992	43	41	41
2010	1258	2015	614	1991	41	1982	36	21	21
2007	1201	2014	582	1986	21	1983	17	17	17
2011	1139	2013	546	1989	17	1990	16	16	16
2006	1008	2001	482	1987	16	1988	15	14	14
2005	941	1998	379	1984	14	1985	14	9	9
2002	829	2000	349	1980	2	2017	3	3	3
2003	775	2016	344	2020	1	41	41	Adventure	2
		1999	338						
		1997	289						
		1996	263						
		1995	219						

## Q&A

### 2. 類別型資料的分布

類別型欄位：['NAME', 'PLATFORM', 'YEAR', 'GENRE', 'PUBLISHER']

==== Genre 類別分布 ===

Genre	
Action	3316
Sports	2346
Misc	1739
Role-Playing	1488
Shooter	1310
Adventure	1284
Racing	1249
Platform	886
Simulation	867
Fighting	848
Strategy	681
Puzzle	582
Sony Computer Entertainment	1
Idea Factory	1



## Q&A

### 2. 類別型資料的分布

類別型欄位：['NAME', 'PLATFORM', 'YEAR', 'GENRE', 'PUBLISHER']

==== Publisher 類別分布 ===

Publisher	
Electronic Arts	1351
Activision	975
Namco Bandai Games	932
Ubisoft	921
Konami Digital Entertainment	832
...	
Media Entertainment	1
New World Computing	1
Interchannel-Holon	1
Rain Games	1
UIG Entertainment	1

## Q&A

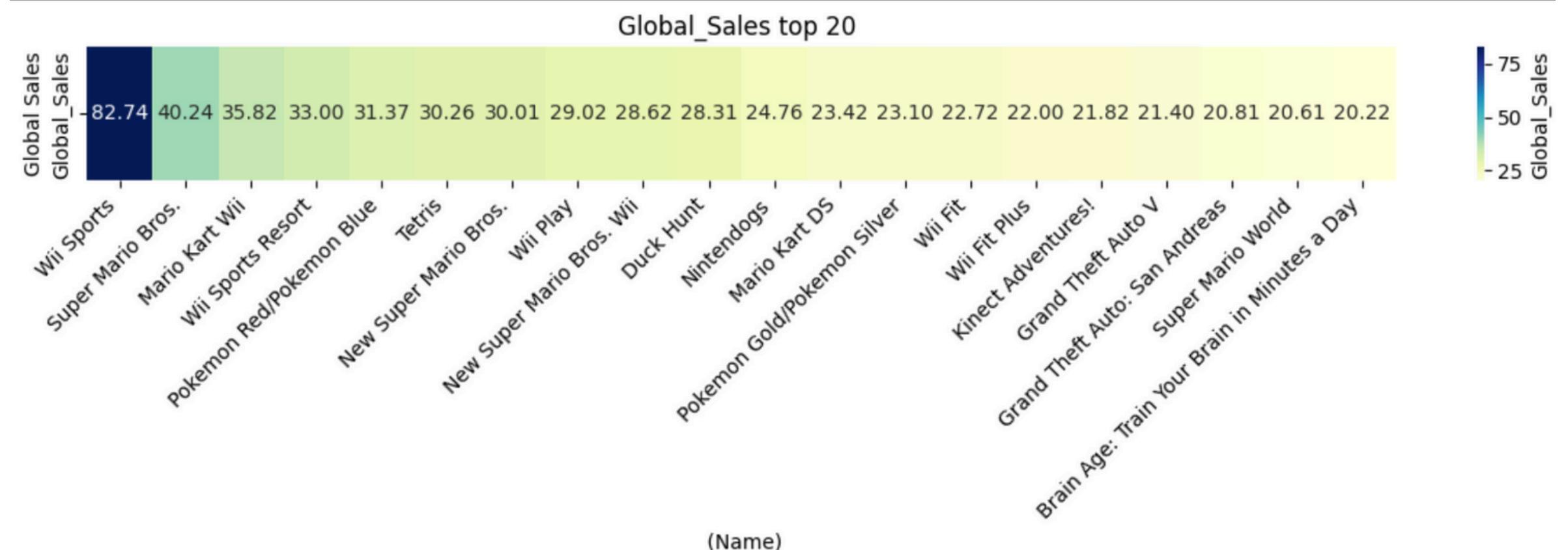
### 3. 各類型遊戲全球銷售總量加總

各類型遊戲銷售量加總：

Genre	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Action	861.77	516.48	158.65	184.92	1722.84
Sports	670.09	371.34	134.76	132.65	1309.24
Shooter	575.16	310.45	38.18	101.90	1026.20
Role-Playing	326.50	187.57	350.29	59.38	923.83
Platform	445.99	200.65	130.65	51.51	829.13
Misc	396.92	211.77	106.67	73.92	789.87
Racing	356.93	236.31	56.61	76.68	726.76
Fighting	220.74	100.00	87.15	36.19	444.05
Simulation	181.78	113.02	63.54	31.36	389.98
Puzzle	122.01	50.52	56.68	12.47	242.21
Adventure	101.93	63.74	51.87	16.70	234.47
Strategy	67.83	44.84	49.10	11.23	173.27

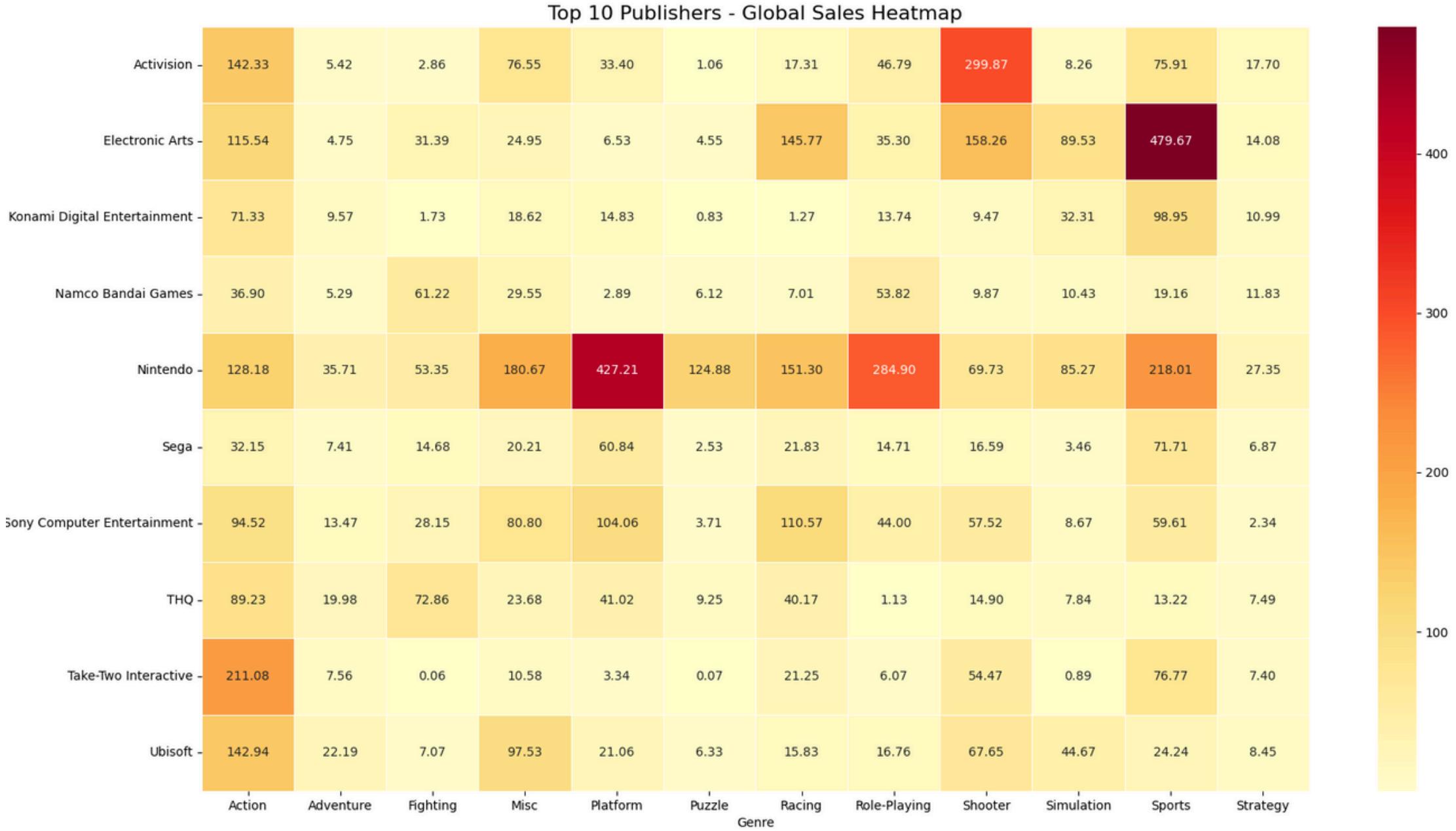
# Q&A

## 4.不同遊戲的全球銷售量(TOP20)



# Q&A

## 5.不同發行商在不同遊戲類型上的全球銷售表現差異(取前10名)



顏色越深，代表該發行商在該類型的銷售量越高  
顏色越淡則代表銷售量較低或該類型不常推出

# 參考文獻

<https://medium.com/python-資料視覺化/python-商業數據分析之可視化繪圖-第19講-熱力圖-seaborn-heatmap-cf1b17d7964e>

<https://medium.com/python-資料視覺化/python-商業數據分析之可視化繪圖-第8講-散點圖-scatter-chart-d693c18a40a6>

<https://ithelp.ithome.com.tw/m/articles/10322270>

# 工作分配

112029038 林資竣	做簡報 資料介紹 研究目的 資料描述 資料型態 資料前處理 資料視覺化 實驗設計 整體報告修改
112029039 黃柏誠	做簡報 預期結果 Q&A
112029023 游雲翊	做簡報 問題思考 問題研究製作 Q&A
112029051 黃柏堯	做簡報 背景 研究目的 資料描述 資料型態 資料前處理 資料 視覺化 實驗設計 整體報告修改
112012067 秦晨恩	做簡報 問題思考 報告

THANK YOU