Proyecto SQL: Películas y directores de cine

Objetivos del análisis:

El objetivo del análisis es descubrir tendencias en las películas, como cuáles son los géneros predominantes en el cine, tanto en cantidad como en calidad (popularidad y calificaciones promedio), analizar los presupuestos y las ganancias para ver las rentabilidades, y por otro lado, analizar a los/las directores/as en busca de descubrir cuáles son los más exitosos en lo que hacen, su tendencia a realizar cierto género y la regularidad con la que realizan películas.

El conjunto de datos utilizado es de uso público:

https://www.kaggle.com/datasets/nayanack/movies-and-directors-dataset-for-film-analytics.

Previo al análisis realicé una limpieza y estandarización en Excel, así como la búsqueda de información faltante en algunos registros sobre el sexo de algunos/as directores, y también agregué una columna con el género principal de cada película a los efectos de ampliar en análisis.

Una vez cargados los datos en SQL Server procedí a su normalización, mediante la creación de claves primarias y foráneas, a modo de poder realizar consultas de tipo JOIN para un mejor análisis:

```sql
--Primero modifico la columna "ID_Director" de la tabla Directores para que no permita valores nulos,
--para luego transformarla en primary key:
ALTER TABLE Directores
ALTER COLUMN ID_Director INT NOT NULL;

-- Ahora transformo la columna "ID_Director" en primary key:
ALTER TABLE Directores
ADD CONSTRAINT PK_Director PRIMARY KEY (ID_Director);

--Ahora realizo lo mismo en la tabla "Películas" con la columna "ID_Peliculas" para transformarla en primary key,
--y también modifico "ID_Director" como clave foránea, para terminar de normalizar la base de datos.
ALTER TABLE Peliculas
ALTER COLUMN ID_Pelicula INT NOT NULL;

ALTER TABLE Peliculas
ADD CONSTRAINT PK_Pelicula PRIMARY KEY (ID_Pelicula);

ALTER TABLE Peliculas
ALTER COLUMN ID_Director INT;

-- Creo la Foreign Key:
ALTER TABLE Peliculas
ADD CONSTRAINT FK_Peliculas_Directores
FOREIGN KEY (ID_Director)
REFERENCES Directores(ID_Director);
```

Una vez normalizada la base de datos, procedí al análisis:

Al comenzar me di cuenta de una inconsistencia en los datos, y es que, al seleccionar todas las películas de un determinado género, no obtenía todos los resultados:

```
28  SELECT * FROM peliculas
29  WHERE Género = 'Ciencia ficción';
30
```

177 %

Results | Messages

	ID_Pelicula	Presupuesto	Puntuacion_Popularidad	Ingresos	Titulo	Género	Calificacion_Promedio	ID_Director	Año_Estreno	Mes_Estreno
1	46564	10500000,00	56	792910554,00	E.T. the Extra-Terrestrial	Ciencia ficción	73	4799	1982	Abril

Utilicé una sentencia con la cláusula “LIKE” para que me mostrara todas las variaciones de Ciencia ficción que podrían estar almacenadas en la base de datos, incluyendo aquellas con espacios adicionales o diferencias en las mayúsculas/minúsculas y obtuve lo siguiente:

```
27  SELECT Género, COUNT(*)
28  FROM peliculas
29  GROUP BY Género
30  HAVING Género LIKE '%Ciencia%';
```

161 %

Results | Messages

	Género	(No column name)
1	Ciencia ficción	1
2	Ciencia ficción	115

En conclusión, tenía valores con espacios adicionales, lo que provocaba problemas a la hora de realizar consultas, por lo que actualicé los valores de todos los géneros mediante las funciones “LTRIM” Y “RTRIM”, las cuales recortan los espacios en blanco a la izquierda y a la derecha de la cadena de caracteres:

```
29  UPDATE peliculas
30  SET Género = LTRIM(RTRIM(Género));
31
32
```

Messages

```
(1465 rows affected)

Completion time: 2024-10-21T14:46:52.2530872-03:00
```

Siguiendo con el análisis, cree una consulta que me mostrara cuántas películas había por cada género, seleccionando las primeras cinco con más cantidad:

```
38  SELECT TOP 5 COUNT(ID_Pelicula) AS Cantidad_Peliculas, Género
39  FROM Peliculas
40  GROUP BY Género
41  ORDER BY COUNT(ID_Pelicula) DESC;
42
```

46 %

Results | Messages

	Cantidad_Peliculas	Género
1	396	Comedia
2	329	Drama
3	141	Acción
4	116	Ciencia ficción
5	116	Suspenso

Podemos observar que las películas de comedia predominan en cantidad, seguidas por drama, acción, ciencia ficción y suspenso.

Seleccioné los géneros y la calificación promedio por cada uno, para ver si coinciden los géneros más realizados con las mejores calificaciones y ver si puede ser un factor por el cual se eligen producir los mismos:

```
46  SELECT ROUND(AVG(Calificacion_Promedio),2) AS Calificacion_promedio, Género
47  FROM Peliculas
48  GROUP BY Género
49  ORDER BY AVG(Calificacion_Promedio) DESC;
```

161 %

Results | Messages

	Calificacion_promedio	Género
1	71,83	Crimen
2	69,46	Bélico
3	69	Romance
4	68,64	Documental
5	67,84	Drama
6	67,33	Histórico
7	66,73	Biografía
8	66,5	Western
9	64,95	Thriller
10	64,51	Suspenso
11	63,86	Ciencia ficción
12	63,62	Musical
13	63,07	Superhéroes
14	62,4	Acción
15	62,38	Fantasía
16	61,82	Aventura
17	61,26	Terror
18	61,08	Animación
19	60,35	Comedia

En principio se puede ver que no son los géneros más realizados los que tienen mejor calificación promedio.

Por último, para terminar con el análisis de los géneros, cree una consulta que me permitió ver el presupuesto y las ganancias promedio, para luego analizar el beneficio y descubrir cuáles son los cinco géneros con más retorno:

```
52 SELECT TOP 5 Género, ROUND(AVG(Presupuesto),2) AS Presupuesto_promedio,
53 ROUND(AVG(Ingresos),2) AS Ingresos_promedio,
54 ROUND((AVG(Ingresos)-AVG(Presupuesto)),2) AS Beneficio_Promedio
55 FROM Peliculas
56 GROUP BY Género
57 ORDER BY (AVG(Ingresos)-AVG(Presupuesto)) DESC;
```

150 %

Results | Messages

	Género	Presupuesto_promedio	Ingresos_promedio	Beneficio_Promedio
1	Superhéroes	161259259,26	490545113,30	329285854,04
2	Fantasía	97376923,08	317786890,97	220409967,89
3	Ciencia ficción	87685793,13	297349184,23	209663391,10
4	Histórico	120888888,89	284712793,11	163823904,22
5	Aventura	93934660,71	249176402,20	155241741,48

En este caso vemos un resultado que coincide con lo que sucede en la actualidad: la tendencia de las películas de superhéroes, que tienen mucho éxito en taquilla.

Pasando de lo macro a lo micro, ahora toca analizar un poco las películas:

Para comenzar realicé una consulta simple para visualizar las diez películas que más beneficio obtuvieron:

```
72 SELECT TOP 10 Titulo, Género, ROUND(Ingresos-Presupuesto,2) AS Beneficios
73 FROM Peliculas
74 ORDER BY Beneficios DESC;
```

165 %

Results | Messages

	Titulo	Género	Beneficios
1	Avatar	Ciencia ficción	2550965087,00
2	Titanic	Drama	1645034188,00
3	Furious 7	Acción	1316249360,00
4	The Lord of the Rings: The Return of the King	Fantasía	1024888979,00
5	Transformers: Dark of the Moon	Ciencia ficción	928746996,00
6	Skyfall	Acción	908561013,00
7	Transformers: Age of Extinction	Ciencia ficción	881405097,00
8	Pirates of the Caribbean: Dead Man's Chest	Aventura	865659812,00
9	Jurassic Park	Ciencia ficción	857100000,00
10	Harry Potter and the Philosopher's Stone	Fantasía	851475550,00

También cree una expresión de tabla común en la que utilicé la función “ROW_NUMBER” para asignar un numero de fila a cada género, ordenando las películas por la calificación promedio de mayor a menor, lo que le da a cada película dentro de los géneros un ranking basado en su calificación. Luego basé la consulta principal en esta tabla, seleccionando solo las películas con el ranking 1, es decir, la película con la calificación más alta de cada género:

```
62  WITH RankingPeliculas AS (
63  SELECT Titulo, Género, Calificacion_Promedio,
64  ROW_NUMBER() OVER (PARTITION BY Género ORDER BY Calificacion_Promedio DESC) AS Ranking
65  FROM Peliculas)
66  SELECT Titulo, Género, Calificacion_Promedio
67  FROM RankingPeliculas
68  WHERE Ranking = 1
69  ORDER BY Calificacion_Promedio DESC;
```

150 %

Results | Messages

	Titulo	Género	Calificacion_Promedio
1	Pulp Fiction	Crimen	83
2	Fight Club	Drama	83
3	The Dark Knight	Superhéroes	82
4	The Silence of the Lambs	Suspenso	81
5	The Lord of the Rings: The Return of the King	Fantasía	81
6	Interstellar	Ciencia ficción	81
7	Memento	Thriller	81
8	The Grand Budapest Hotel	Comedia	80
9	Apocalypse Now	Bélico	80
10	The Last Waltz	Documental	79
11	Gladiator	Histórico	79
12	The Thing	Terror	78
13	Django Unchained	Western	78
14	Raiders of the Lost Ark	Aventura	77
15	Kill Bill: Vol. 1	Acción	77
16	Fantastic Mr. Fox	Animación	75
17	Shine	Biografía	73
18	The Bridges of Madison County	Romance	73
19	The Phantom of the Opera	Musical	70

Siguiendo con la última parte del análisis, los directores:

Comencé por investigar cuáles son los directores que más películas realizaron en cada género, para eso volví a utilizar una expresión de tabla común, con un inner join dentro a modo de unir la tabla de películas con la de directores:

```
78  WITH RankingDirectores AS(
79  SELECT
80  COUNT(Peliculas.ID_Pelicula) AS Cantidad_peliculas,
81  Peliculas.Género,
82  Directores.Nombre,
83  ROW_NUMBER() OVER (PARTITION BY Peliculas.Género ORDER BY COUNT(Peliculas.ID_Pelicula) DESC) AS Ranking
84  FROM Peliculas
85  INNER JOIN Directores
86  ON Peliculas.ID_Director = Directores.ID_Director
87  GROUP BY Peliculas.Género, Directores.Nombre
88  )
89  SELECT Nombre, Género, Cantidad_peliculas
90  FROM RankingDirectores
91  WHERE Ranking = 1
92  ORDER BY Cantidad_peliculas DESC;
```

136 %

Results | Messages

	Nombre	Género	Cantidad_peliculas
1	Woody Allen	Comedia	15
2	Martin Scorsese	Drama	12
3	Wes Craven	Terror	8
4	Steven Spielberg	Ciencia ficción	7
5	Tony Scott	Acción	7
6	Michael Moore	Documental	6
7	David Fincher	Suspenso	6
8	Peter Jackson	Fantasía	6
9	Bryan Singer	Superhéroes	5
10	Robert Rodriguez	Aventura	5
11	Clint Eastwood	Bélico	3
12	George Miller	Animación	3
13	Quentin Tarantino	Crimen	3
14	Brian De Palma	Thriller	3
15	Ridley Scott	Histórico	3
16	Kenny Ortega	Musical	3
17	Lawrence Kasd...	Western	2
18	Tim Burton	Biografía	2
19	Lasse Hallström	Romance	1

Mediante estos resultados podemos observar las preferencias de los directores a la hora de realizar ciertos géneros.

Siguiendo con los directores cree una subconsulta que me devuelve la calificación promedio de las películas, y la utilicé para filtrar y obtener aquellos directores que tienen una calificación promedio por encima de la media.

En la imagen muestro los primeros 10 resultados para no utilizar toda la hoja:

```sql
SELECT Directores.Nombre, ROUND(AVG(Peliculas.Calificacion_Promedio),2) AS Calificacion_Promedio
FROM Directores
INNER JOIN Peliculas
ON Directores.ID_Director = Peliculas.ID_Director
GROUP BY Directores.Nombre
HAVING AVG(Peliculas.Calificacion_promedio) > (SELECT AVG(Calificacion_Promedio)
FROM Peliculas)
ORDER BY Calificacion_Promedio DESC;
```

136 %

Results | Messages

	Nombre	Calificacion_Promedio
1	Christopher Nolan	78
2	Quentin Tarantino	77,75
3	David Lynch	74,4
4	Wes Anderson	74,14
5	David Fincher	73,4
6	Peter Jackson	73,33
7	James Cameron	73,29
8	Martin Scorsese	73
9	Alejandro González Iñárritu	72,33
10	Paul Thomas Anderson	72,17

Para finalizar con el análisis cree una consulta que me devuelve a los 10 directores que más beneficios totales generaron con sus películas:

```sql
SELECT TOP 10 Directores.Nombre, ROUND(SUM(Peliculas.Ingresos) - SUM(Peliculas.Presupuesto),2) AS Beneficios_totales
FROM Directores
INNER JOIN Peliculas
ON Directores.ID_Director = Peliculas.ID_Director
GROUP BY Directores.Nombre
ORDER BY Beneficios_totales DESC;
```

136 %

Results | Messages

	Nombre	Beneficios_totales
1	Steven Spielberg	7016239164,00
2	Peter Jackson	5205642820,00
3	James Cameron	5136669439,00
4	Michael Bay	4422524638,00
5	Christopher Nolan	3222483234,00
6	Chris Columbus	3098631503,00
7	Robert Zemeckis	2600622002,00
8	Francis Lawrence	2349457182,00
9	Tim Burton	2340418241,00
10	Sam Raimi	2208951462,00