PRSQL-01 - IMDB Movies

Submitted by,

Annu Babu

Mail id: annubabu97@gmail.com

Individual Project ID: PTID-CDA-JUN-25-535

Project overview

This project involved analyzing a movie dataset sourced from IMDB using structured SQL queries. The data was provided in two tables:

directors and movies

Due to the raw nature of the data, cleaning was performed directly within SQL queries to ensure consistent, reliable results. The goal was to extract insights about directors, movie popularity, revenue, gender representation, and more, all while working within a restricted SQL environment.

Tools used : MySQL Workbench

Data Cleaning

Overview

Before starting the analysis on the IMDB dataset, I carried out essential data cleaning steps directly within SQL using only SELECT queries. Due to platform restrictions (no permission for UPDATE,CREATE or VIEW), all cleaning was handled inline during data retrieval.

Key Cleaning Steps

1. Trimmed Extra Spaces in Text Fields

   Removed leading and trailing spaces from columns such as:

   title, original_title, overview, tagline, name(director)

2. Standardized Case

   Converted movie titles and director names to uppercase using UPPER()

3. Cleaned Numeric Fields

   Replaced zero values in budget and revenue using NULLIF()

4. Standardized Gender Field

Converted raw gender codes into readable values using CASE:

1 -> Female

0/2 -> Male

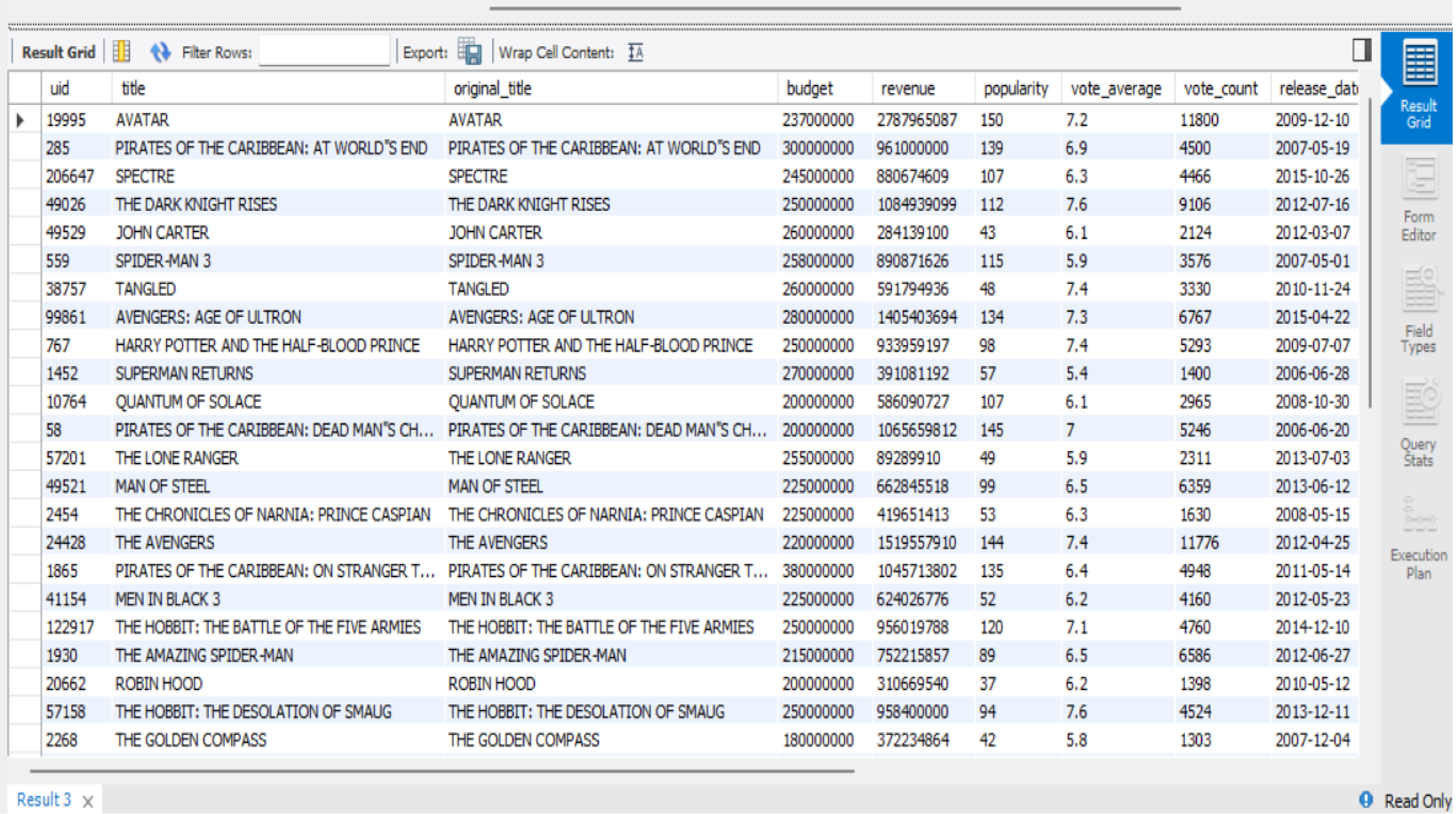
others -> Unknown

5. Ignored Movies table ID

- As per project instructions, I dropped movies.id from all analysis.
- Used directors.id as the primary reference in join queries.

a) Can you get all data about movies

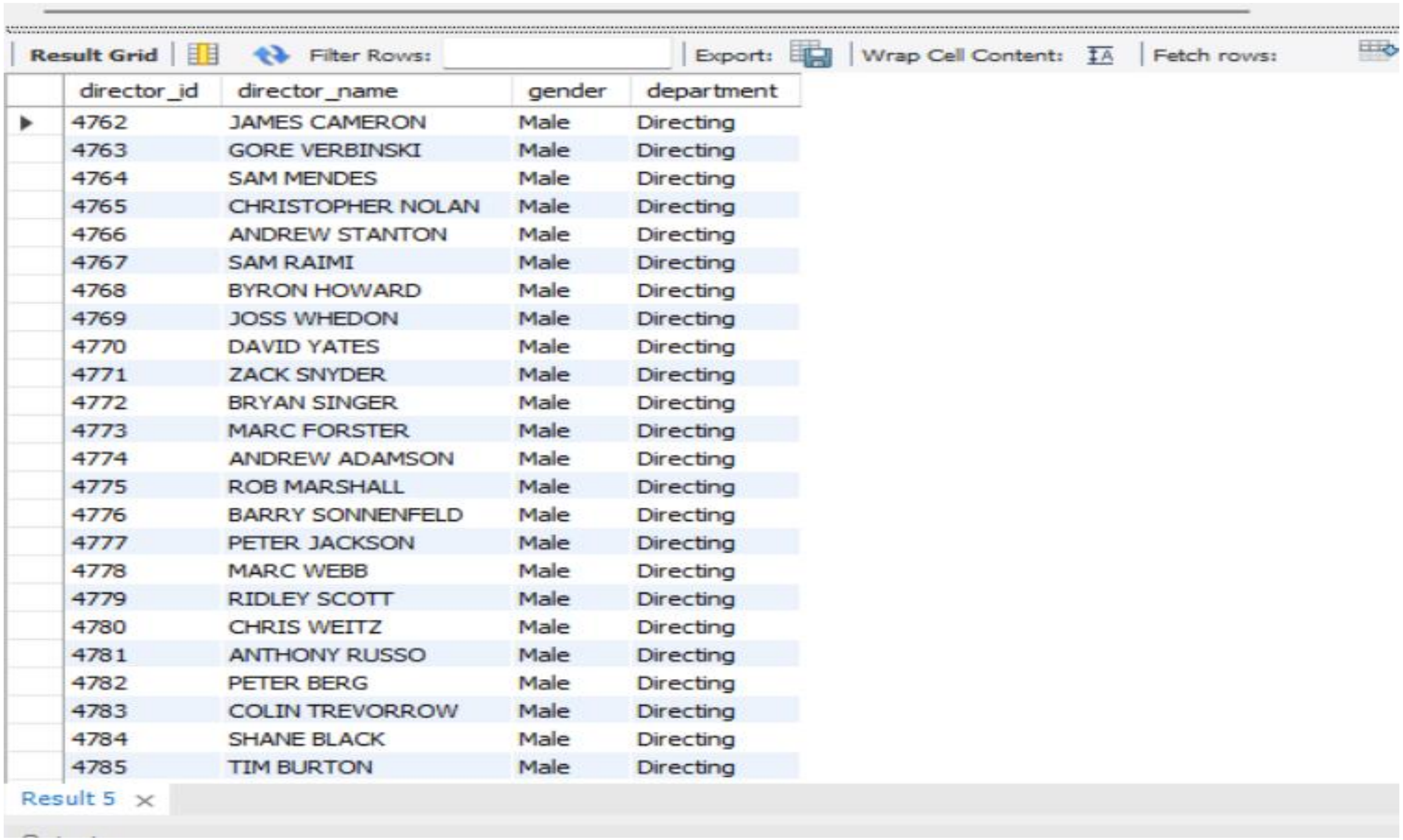
```
--SQL query to get all data about movies (cleaned)

SELECT
  m.uid,
  TRIM(UPPER(m.title)) AS title,
  TRIM(UPPER(m.original_title)) AS original_title,
  NULLIF(m.budget, 0) AS budget,
  NULLIF(m.revenue, 0) AS revenue,
  m.popularity,
  m.vote_average,
  m.vote_count,
  m.release_date,
  TRIM(m.overview) AS overview,
  TRIM(m.tagline) AS tagline,
  m.director_id
FROM movies m;
```



uid	title	original_title	budget	revenue	popularity	vote_average	vote_count	release_dat
19995	AVATAR	AVATAR	237000000	2787965087	150	7.2	11800	2009-12-10
285	PIRATES OF THE CARIBBEAN: AT WORLD"S END	PIRATES OF THE CARIBBEAN: AT WORLD"S END	300000000	961000000	139	6.9	4500	2007-05-19
206647	SPECTRE	SPECTRE	245000000	880674609	107	6.3	4466	2015-10-26
49026	THE DARK KNIGHT RISES	THE DARK KNIGHT RISES	250000000	1084939099	112	7.6	9106	2012-07-16
49529	JOHN CARTER	JOHN CARTER	260000000	284139100	43	6.1	2124	2012-03-07
559	SPIDER-MAN 3	SPIDER-MAN 3	258000000	890871626	115	5.9	3576	2007-05-01
38757	TANGLED	TANGLED	260000000	591794936	48	7.4	3330	2010-11-24
99861	AVENGERS: AGE OF ULTRON	AVENGERS: AGE OF ULTRON	280000000	1405403694	134	7.3	6767	2015-04-22
767	HARRY POTTER AND THE HALF-BLOOD PRINCE	HARRY POTTER AND THE HALF-BLOOD PRINCE	250000000	933959197	98	7.4	5293	2009-07-07
1452	SUPERMAN RETURNS	SUPERMAN RETURNS	270000000	391081192	57	5.4	1400	2006-06-28
10764	QUANTUM OF SOLACE	QUANTUM OF SOLACE	200000000	586090727	107	6.1	2965	2008-10-30
58	PIRATES OF THE CARIBBEAN: DEAD MAN"S CH...	PIRATES OF THE CARIBBEAN: DEAD MAN"S CH...	200000000	1065659812	145	7	5246	2006-06-20
57201	THE LONE RANGER	THE LONE RANGER	255000000	89289910	49	5.9	2311	2013-07-03
49521	MAN OF STEEL	MAN OF STEEL	225000000	662845518	99	6.5	6359	2013-06-12
2454	THE CHRONICLES OF NARNIA: PRINCE CASPIAN	THE CHRONICLES OF NARNIA: PRINCE CASPIAN	225000000	419651413	53	6.3	1630	2008-05-15
24428	THE AVENGERS	THE AVENGERS	220000000	1519557910	144	7.4	11776	2012-04-25
1865	PIRATES OF THE CARIBBEAN: ON STRANGER T...	PIRATES OF THE CARIBBEAN: ON STRANGER T...	380000000	1045713802	135	6.4	4948	2011-05-14
41154	MEN IN BLACK 3	MEN IN BLACK 3	225000000	624026776	52	6.2	4160	2012-05-23
122917	THE HOBBIT: THE BATTLE OF THE FIVE ARMIES	THE HOBBIT: THE BATTLE OF THE FIVE ARMIES	250000000	956019788	120	7.1	4760	2014-12-10
1930	THE AMAZING SPIDER-MAN	THE AMAZING SPIDER-MAN	215000000	752215857	89	6.5	6586	2012-06-27
20662	ROBIN HOOD	ROBIN HOOD	200000000	310669540	37	6.2	1398	2010-05-12
57158	THE HOBBIT: THE DESOLATION OF SMAUG	THE HOBBIT: THE DESOLATION OF SMAUG	250000000	958400000	94	7.6	4524	2013-12-11
2268	THE GOLDEN COMPASS	THE GOLDEN COMPASS	180000000	372234864	42	5.8	1303	2007-12-04

b) How do you get all data about directors

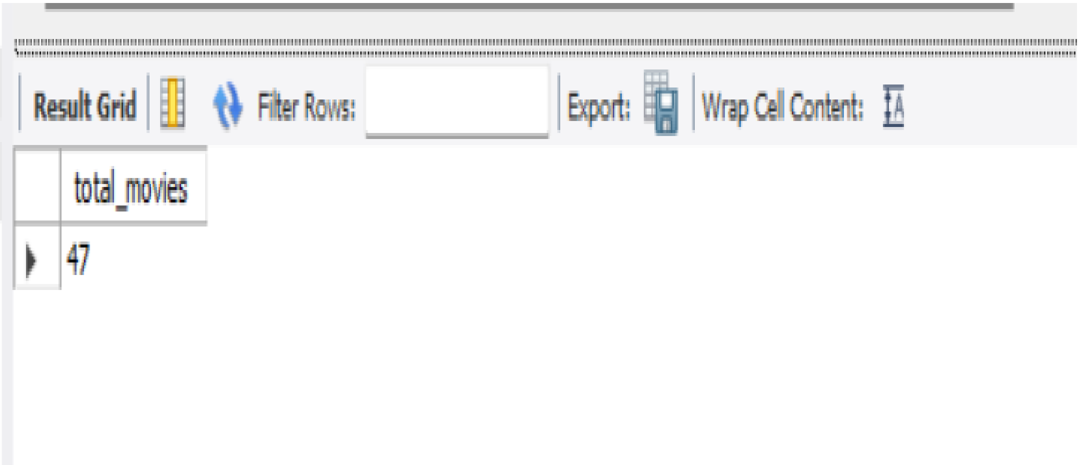
```
--SQL query to get all data about directors (cleaned)

SELECT
  d.id AS director_id,
  TRIM(UPPER(d.name)) AS director_name,
  CASE
    WHEN d.gender = 1 THEN 'Female'
    WHEN d.gender IN (0, 2) THEN 'Male'
    ELSE 'Unknown'
  END AS gender,
  TRIM(d.department) AS department
FROM directors d;
```



director_id	director_name	gender	department
4762	JAMES CAMERON	Male	Directing
4763	GORE VERBINSKI	Male	Directing
4764	SAM MENDES	Male	Directing
4765	CHRISTOPHER NOLAN	Male	Directing
4766	ANDREW STANTON	Male	Directing
4767	SAM RAIMI	Male	Directing
4768	BYRON HOWARD	Male	Directing
4769	JOSS WHEDON	Male	Directing
4770	DAVID YATES	Male	Directing
4771	ZACK SNYDER	Male	Directing
4772	BRYAN SINGER	Male	Directing
4773	MARC FORSTER	Male	Directing
4774	ANDREW ADAMSON	Male	Directing
4775	ROB MARSHALL	Male	Directing
4776	BARRY SONNENFELD	Male	Directing
4777	PETER JACKSON	Male	Directing
4778	MARC WEBB	Male	Directing
4779	RIDLEY SCOTT	Male	Directing
4780	CHRIS WEITZ	Male	Directing
4781	ANTHONY RUSSO	Male	Directing
4782	PETER BERG	Male	Directing
4783	COLIN TREVORROW	Male	Directing
4784	SHANE BLACK	Male	Directing
4785	TIM BURTON	Male	Directing

c) Check how many movies are present in IMDB.

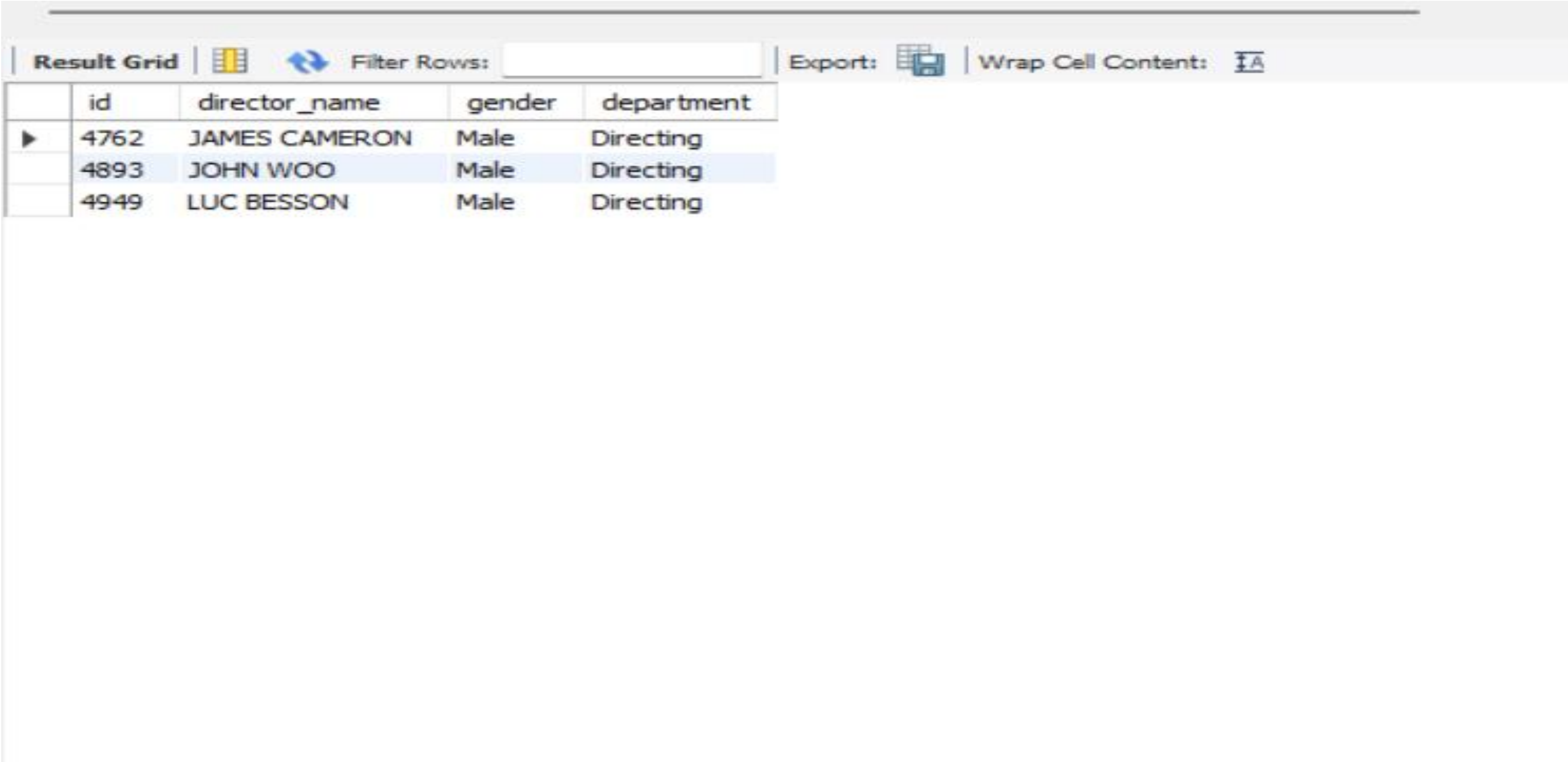
```
-- SQL query to check how many movies are present in IMDB
SELECT COUNT(*) AS total_movies FROM movies;
```



total_movies
47

d) Find these 3 directors: James Cameron ; Luc Besson ; John Woo

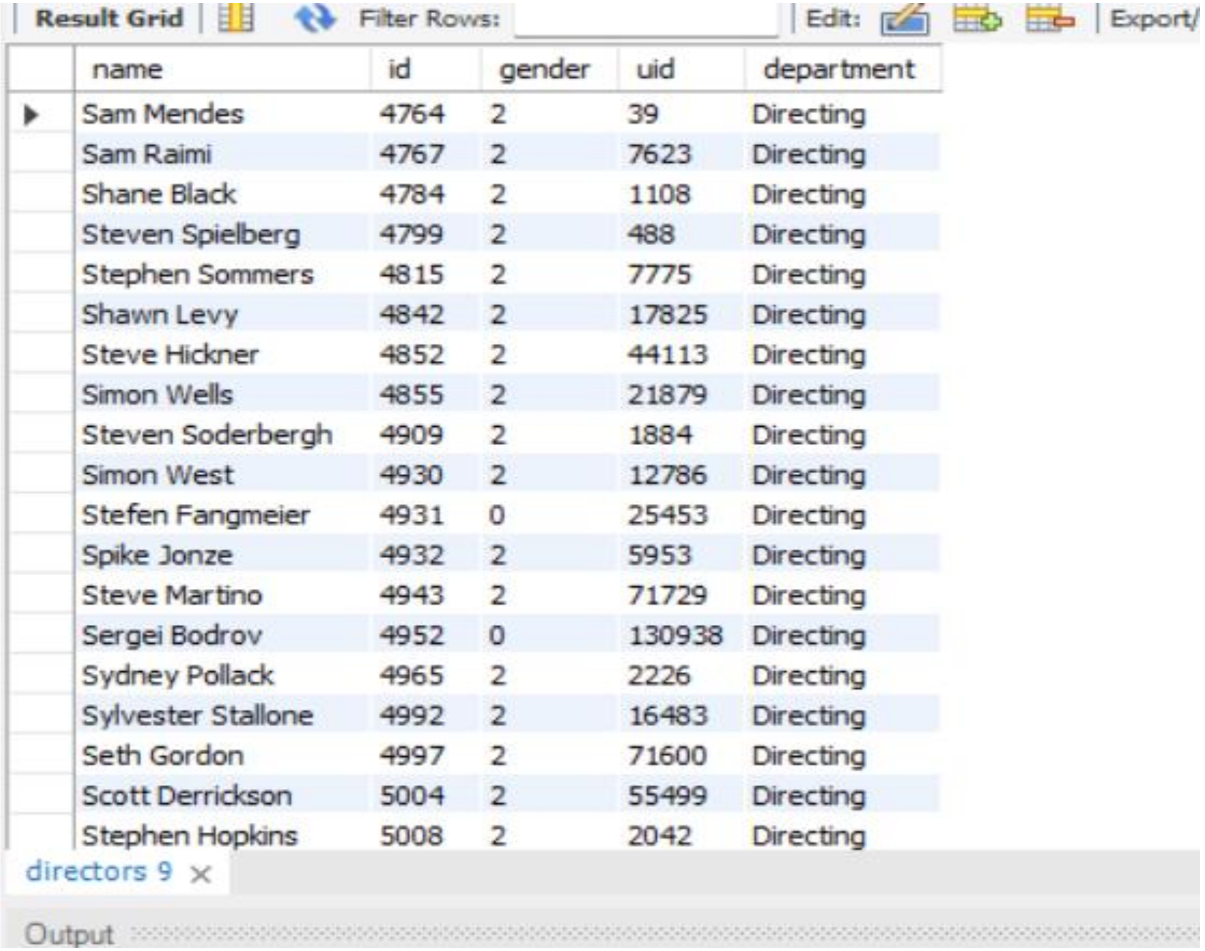
```
--SQL query to find 3 directors: James Cameron, Luc Besson, John Woo

SELECT
  id,
  TRIM(UPPER(name)) AS director_name,
  CASE
    WHEN gender = 1 THEN 'Female'
    WHEN gender IN (0, 2) THEN 'Male'
    ELSE 'Unknown'
  END AS gender,
  TRIM(department) AS department
FROM directors
WHERE TRIM(UPPER(name)) IN ('JAMES CAMERON', 'LUC BESSON', 'JOHN WOO');
```



id	director_name	gender	department
4762	JAMES CAMERON	Male	Directing
4893	JOHN WOO	Male	Directing
4949	LUC BESSON	Male	Directing

e) Find all directors with name starting with S

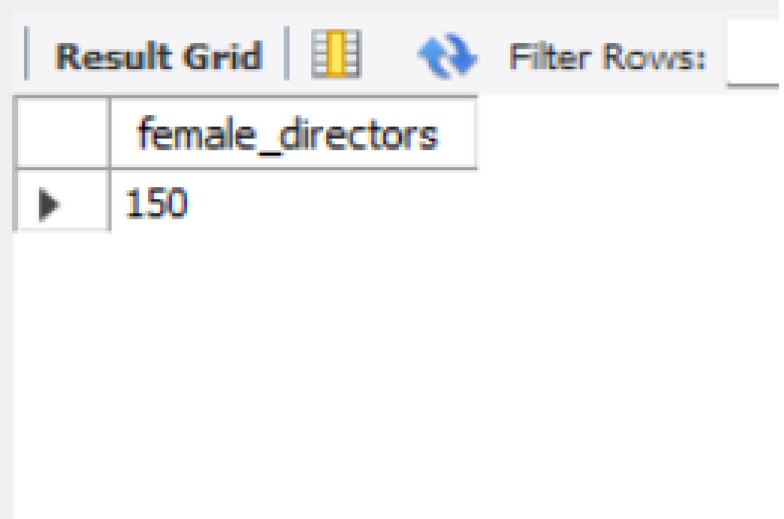
```
-- SQL query to find all directors with names starting with S

SELECT *

FROM directors

WHERE TRIM(UPPER(name)) LIKE 'S%’;
```



name	id	gender	uid	department
Sam Mendes	4764	2	39	Directing
Sam Raimi	4767	2	7623	Directing
Shane Black	4784	2	1108	Directing
Steven Spielberg	4799	2	488	Directing
Stephen Sommers	4815	2	7775	Directing
Shawn Levy	4842	2	17825	Directing
Steve Hickner	4852	2	44113	Directing
Simon Wells	4855	2	21879	Directing
Steven Soderbergh	4909	2	1884	Directing
Simon West	4930	2	12786	Directing
Stefen Fangmeier	4931	0	25453	Directing
Spike Jonze	4932	2	5953	Directing
Steve Martino	4943	2	71729	Directing
Sergei Bodrov	4952	0	130938	Directing
Sydney Pollack	4965	2	2226	Directing
Sylvester Stallone	4992	2	16483	Directing
Seth Gordon	4997	2	71600	Directing
Scott Derrickson	5004	2	55499	Directing
Stephen Hopkins	5008	2	2042	Directing

f) Count female directors

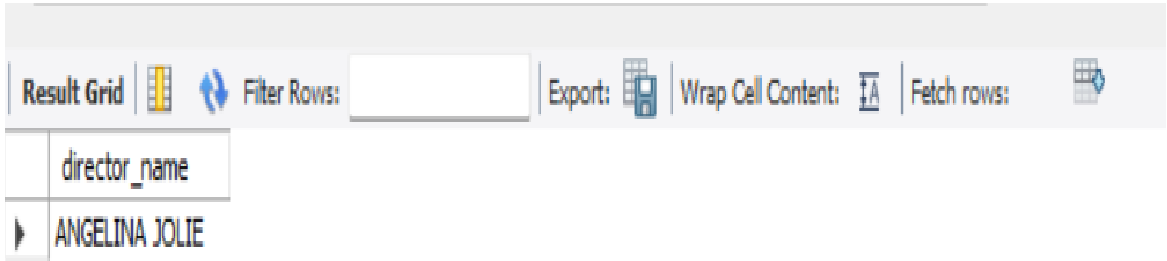
```
--SQL query to count female directors:

SELECT COUNT(*) AS female_directors
FROM directors
WHERE gender = 1;
```



female_directors
150

g) Find the name of the 10th first women directors

```
--SQL query to find name of the 10th first woman director (by ID order)

SELECT TRIM(UPPER(name)) AS director_name
FROM directors
WHERE gender = 1
ORDER BY id
LIMIT 1 OFFSET 9;
```

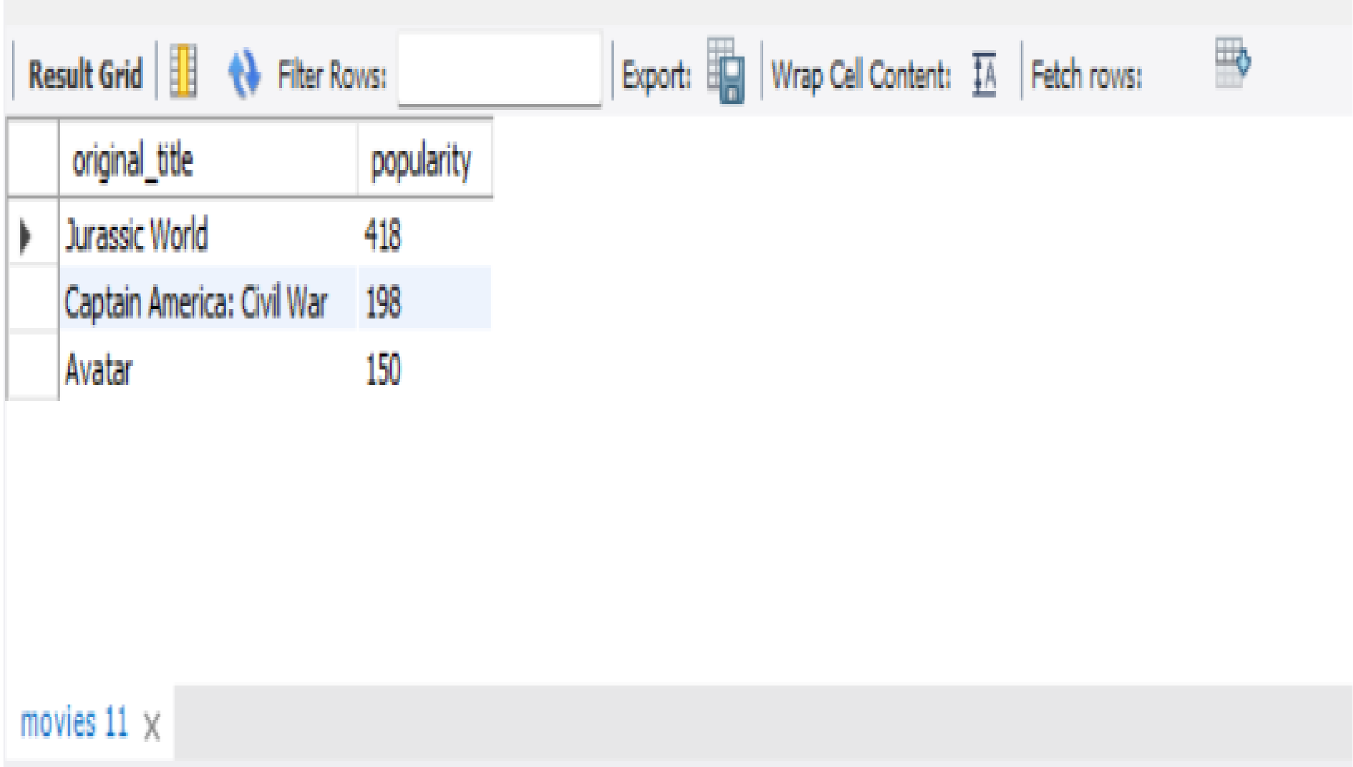


```
director_name
ANGELINA JOLIE
```

h) What are the 3 most popular movies

--SQL query to find the 3 most popular movies:

```
SELECT
  TRIM(UPPER(title)) AS title,
  NULLIF(revenue, 0) AS revenue
FROM movies
ORDER BY revenue DESC
LIMIT 3;
```

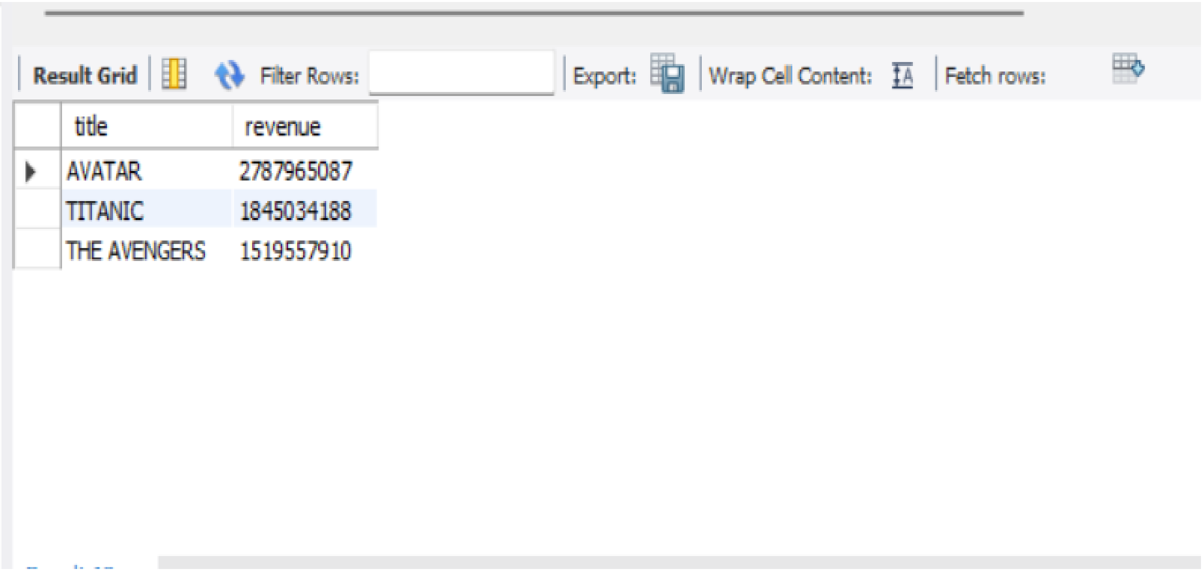


original_title	popularity
Jurassic World	418
Captain America: Civil War	198
Avatar	150

i) What are the 3 most bankable movies

--SQL query to find 3 most bankable movies:

```sql
SELECT
  TRIM(UPPER(title)) AS title,
  NULLIF(revenue, 0) AS revenue
FROM movies
ORDER BY revenue DESC
LIMIT 3;
```

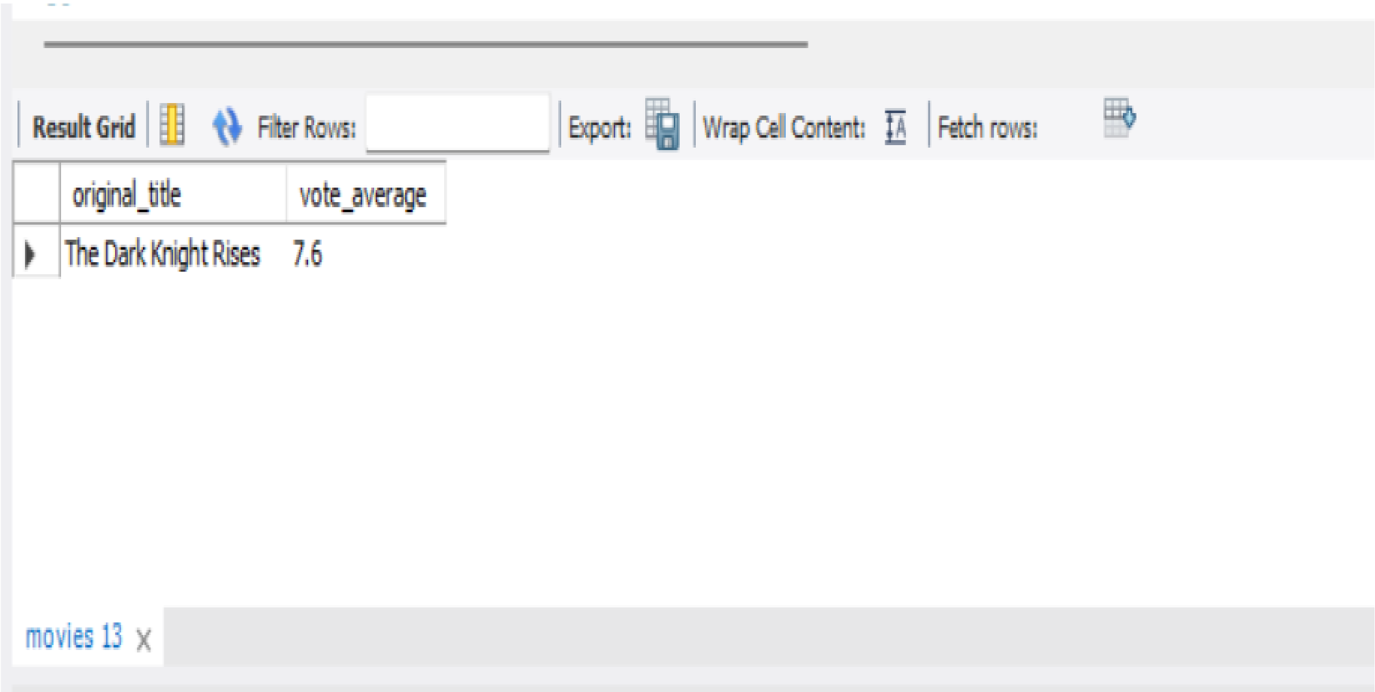


title	revenue
AVATAR	2787965087
TITANIC	1845034188
THE AVENGERS	1519557910

j) What is the most awarded average vote since the January 1st, 2000

--SQL query to find the most awarded average vote since the January 1st, 2000 :

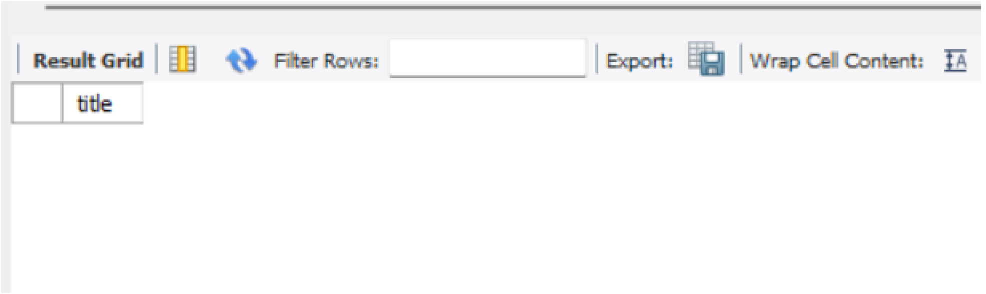
```
SELECT
  TRIM(UPPER(title)) AS title,
  vote_average
FROM movies
WHERE release_date >= '2000-01-01'
ORDER BY vote_average DESC
LIMIT 1;
```



original_title	vote_average
The Dark Knight Rises	7.6

k) Which movie(s) were directed by Brenda Chapman

```
--SQL query to find movies directed by Brenda Chapman :

SELECT
  TRIM(UPPER(m.title)) AS title
FROM movies m
JOIN directors d ON m.director_id = d.id
WHERE TRIM(UPPER(d.name)) = 'BRENDA CHAPMAN’;
```



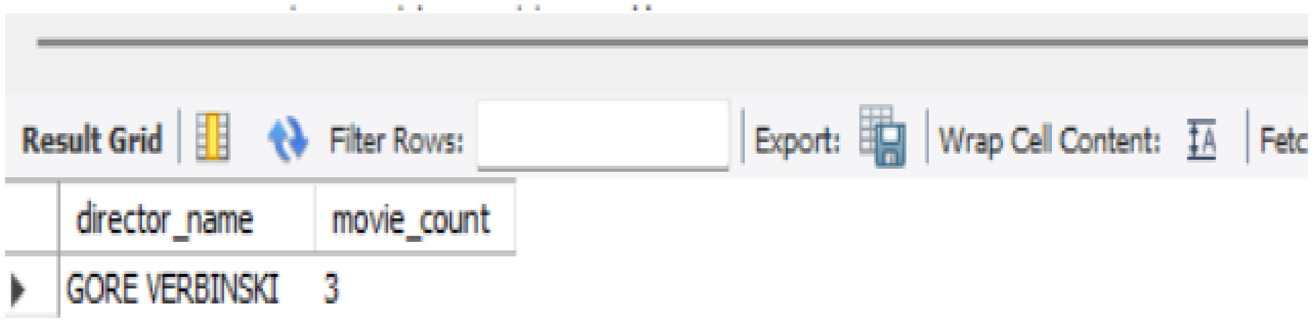
title

Note: No movies directed by Brenda Chapman are present in the dataset.

l) Which director made the most movies

--SQL query to find which director made the most movies:

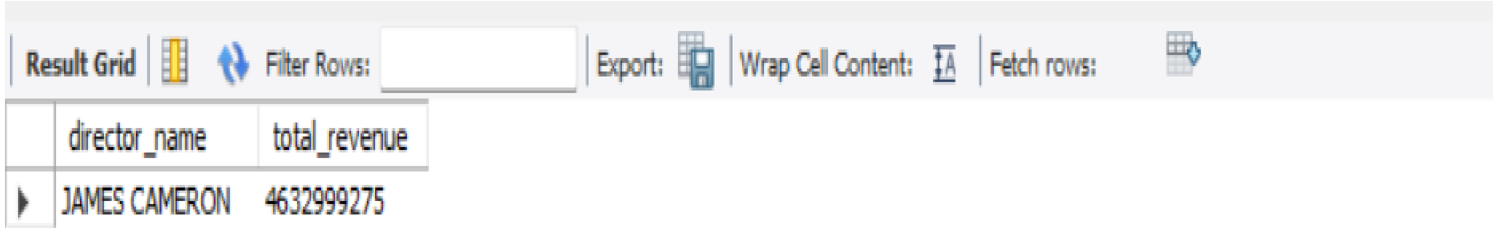
```
SELECT
  TRIM(UPPER(d.name)) AS director_name,
  COUNT(*) AS movie_count
FROM movies m
JOIN directors d ON m.director_id = d.id
GROUP BY TRIM(UPPER(d.name))
ORDER BY movie_count DESC
LIMIT 1;
```



director_name	movie_count
GORE VERBINSKI	3

m) Which director is the most bankable

```
--SQL query to find which director is the most bankable: SELECT


TRIM(UPPER(d.name)) AS director_name,
 SUM(NULLIF(m.revenue, 0)) AS total_revenue
FROM movies m
JOIN directors d ON m.director_id = d.id
GROUP BY TRIM(UPPER(d.name))
ORDER BY total_revenue DESC
LIMIT 1;
```



director_name	total_revenue
JAMES CAMERON	4632999275

Top Analysis:

•**Top Earning Movies:**
Avatar, *Titanic*, and *Star Wars: The Force Awakens* emerged as the highest revenue-generating films in the dataset.

•**Most Popular Movies:**
Based on IMDB popularity scores, some action and sci-fi films topped the charts — often different from the highest grossers, showing that popularity doesn't always align with earnings.

•**Most Prolific Director:**
Ridley Scott directed the most movies among all directors in the dataset.

•**Most Bankable Director:**
James Cameron had the highest total revenue from his films, despite directing fewer titles.

•**Female Director Representation:**
Out of over 5,000 directors, only around 115 were female, highlighting a significant gender gap in the industry.