



# Data Integration

urban-octo-spoon

Step3: Cleaning and Showcases

Isabella Lisica

Nureldien Gebril

Eyad Hamed

# Improve data quality, standarization

## Table movies

Before

movieId	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action Crime Thriller
7	Sabrina (1995)	Comedy Romance
8	Tom and Huck (1995)	Adventure Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action Adventure Thriller
11	American President, The (1995)	Comedy Drama Romance
12	Dracula: Dead and Loving It (1995)	Comedy Horror
13	Balto (1995)	Adventure Animation Children
14	Nixon (1995)	Drama
15	Cutthroat Island (1995)	Action Adventure Romance
16	Casino (1995)	Crime Drama
17	Sense and Sensibility (1995)	Drama Romance
18	Four Rooms (1995)	Comedy
19	Ace Ventura: When Nature Calls...	Comedy
20	Money Train (1995)	Action Comedy Crime Drama Thriller

After

movieId	title	genres	year
1	Toy Story	Adventure Animation Children Comedy Fantasy	1995
2	Jumanji	Adventure Children Fantasy	1995
3	Grumpier Old Men	Comedy Romance	1995
4	Waiting to Exhale	Comedy Drama Romance	1995
5	Father of the Bride Part II	Comedy	1995
6	Heat	Action Crime Thriller	1995
7	Sabrina	Comedy Romance	1995
8	Tom and Huck	Adventure Children	1995
9	Sudden Death	Action	1995
10	GoldenEye	Action Adventure Thriller	1995
11	American President	Comedy Drama Romance	1995
12	Dracula: Dead and Loving It	Comedy Horror	1995
13	Balto	Adventure Animation Children	1995
14	Nixon	Drama	1995
15	Cutthroat Island	Action Adventure Romance	1995
16	Casino	Crime Drama	1995
17	Sense and Sensibility	Drama Romance	1995
18	Four Rooms	Comedy	1995
19	Ace Ventura: When Nature Calls	Comedy	1995
20	Money Train	Action Comedy Crime Drama Thriller	1995
21	Get Shorty	Comedy Crime Thriller	1995

# Improve data quality, standarization

## Table revenues

Before

id	date	title	revenue	theaters	distributor
000020de-3a42-a942-5401-c68fcc662fa6	2007-11-18	Darfur Now	2531	22	Warner Independent Pictures (V
000047dc-1f39-51bb-b898-ab6a87cbce9	2015-01-09	Unbroken	2576345	3301	Universal Pictures
0001e185-4c4c-f41a-eb8f-2911020185c5	2015-10-31	Jurassic World	43465	164	Universal Pictures
00024499-c64c-1a2e-b57a-14acde4a0b2f	2009-09-01	G.I. Joe: The Rise of Cobra	595411	3467	Paramount Pictures
0002e659-a8e1-7bb1-e-4835-262379e9da25	2013-10-27	The Conjuring	6174	100	Warner Bros.
00036970-3d77-1f53-90ab-d8818b5e56d3	2014-01-23	Walking with Dinosaurs 3D	13466	454	Twentieth Century Fox
00046acb-5a08-8c83-20f3-2cbbdc4a09d8	2011-10-16	Don't Be Afraid of the Dark	8314	80	FilmDistrict
0005b94f-8d48-182f-a89a-186cd18dd400	2016-04-02	My Big Fat Greek Wedding 2	4889120	3179	Universal Pictures
0006742c-079a-a9bf-8556-326deb0d6b6d	2014-06-11	A Million Ways to Die in th...	914865	3160	Universal Pictures
0006ca00-6340-035b-045a-6ef2e19eb501	2002-05-27	Kissing Jessica Stein	24235	64	Fox Searchlight Pictures
0006f9a5-3273-d8b0-72b0-1f4321369b0c	2008-12-05	Religulous	10806	59	Lionsgate
000777e8-bfa9-c576-c2b5-217f9dbd0ee6	2013-06-25	Girl Rising	4870	16	GathrFilms
0008106b-3593-212c-2168-9136ffaeb9d	2019-12-21	Richard Jewell	1006211	2502	Warner Bros.
000c36c0-328b-bfc8-85b2-c4365f601d67	2008-04-11	Nim's Island	2378667	3518	Twentieth Century Fox
000c67e8-096e-8ae3-81e6-6ad825175ecd	2019-11-08	Jojo Rabbit	1143978	798	Fox Searchlight Pictures
000db900-d2d5-cde5-3a25-9545c47ebf42	2002-05-26	High Crimes	82135	322	Twentieth Century Fox
000dbf5f-223f-6b8c-a240-c13c6de40580	2018-11-07	Viper Club	3634	70	Roadside Attractions
000f4a28-c35f-0b8c-7108-eafcf875ded6e	2016-12-22	The Edge of Seventeen	9743	221	STX Entertainment
00101de4-288a-006b-02e5-200900eccee8	2004-06-30	Starsky & Hutch	4019	82	Warner Bros.
0010c149-d9c1-b03c-d3fb-b65dc08a9a68	2016-02-01	Trumbo	8700	78	Bleeker Street Media

After

revenueId	title	revenue	distributor	year
54	Fairy Tail: Dragon Cry	172249.00000	FUNimation Entertainment	2017
97	Swimfan	4932608.00000	Twentieth Century Fox	2002
100	City by the Sea	3738191.00000	Warner Bros.	2002
105	Spider-Man/Men in Black IIDouble Bill	936685.00000	Sony Pictures Entertainment (SPE)	2002
147	Rent	4751680.00000	Revolution Studios	2005
184	Quattro Noza	717.00000	Lionsgate	2005
210	Ice Age: Continental Drift	16728341.00000	Twentieth Century Fox	2012
215	Savages	2740400.00000	Universal Pictures	2012
217	Katy Perry: Part of Me	1313660.00000	Paramount Pictures	2012
236	Speak Bachchan	61256.00000	FIP	2012
240	The Obama Effect	52922.00000	Arc Entertainment	2012
345	Blade II	12274000.00000	New Line Cinema	2002
346	E.T. the Extra-Terrestrial20th Anni...	5728335.00000	Universal Pictures	2002
352	Sorority Boys	1687977.00000	Walt Disney Studios Motion Pictures	2002
469	Tomb Raider	5934618.00000	Warner Bros.	2018
480	7 Days in Entebbe	409790.00000	Focus Features	2018
483	Vertigo2018 Re-release	252880.00000	Fathom Events	2018
519	Intent to Destroy: Death, Denial & ...	1118.00000	GathrFilms	2018
545	Atlas Shrugged: Who Is John Galt?	35553.00000	Atlas Distribution Company	2014
626	Tyler Perry's Boo! A Madea Halloween	7281034.00000	Lionsgate	2016
627	Jack Reacher: Never Go Back	5223311.00000	Paramount Pictures	2016

# Improve data quality, standarization

## Table ratings

Before

	userId	movieId	rating	timestamp
►	1	296	5.0	1147880044
	1	306	3.5	1147868817
	1	307	5.0	1147868828
	1	665	5.0	1147878820
	1	899	3.5	1147868510
	1	1088	4.0	1147868495
	1	1175	3.5	1147868826
	1	1217	3.5	1147878326
	1	1237	5.0	1147868839
	1	1250	4.0	1147868414
	1	1260	3.5	1147877857
	1	1653	4.0	1147868097
	1	2011	2.5	1147868079
	1	2012	2.5	1147868068
	1	2068	2.5	1147869044
	1	2161	3.5	1147868609
	1	2351	4.5	1147877957
	1	2573	4.0	1147878923
	1	2632	5.0	1147878248
	1	2692	5.0	1147869100

After

	ratingId	userId	movieId	rating
►	1	1	296	5.00
	2	1	306	3.50
	3	1	307	5.00
	4	1	665	5.00
	5	1	899	3.50
	6	1	1088	4.00
	7	1	1175	3.50
	8	1	1217	3.50
	9	1	1237	5.00
	10	1	1250	4.00
	11	1	1260	3.50
	12	1	1653	4.00
	13	1	2011	2.50
	14	1	2012	2.50
	15	1	2068	2.50
	16	1	2161	3.50
	17	1	2351	4.50
	18	1	2573	4.00
	19	1	2632	5.00
	20	1	2692	5.00
	21	1	2843	4.50

## Movies table

Remove duplicates based on title, year	216 records
Remove empty title records	52 records
Remove (different languages) in the title	4744 records
Remove (, The)	1437 records
Year column ->	Int

## Before

movieId	title	genres
28	Persuasion (1995)	Drama Romance
29	City of Lost Children, The (Cité des enfants perdus, La) (1995)	Adventure Drama Fantasy Mystery Sci-Fi
30	Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)	Crime Drama
31	Dangerous Minds (1995)	Drama

## Revenues table

Remove duplicates (title)	183
revenue column ->	Decimal(20.5)

## Ratings table

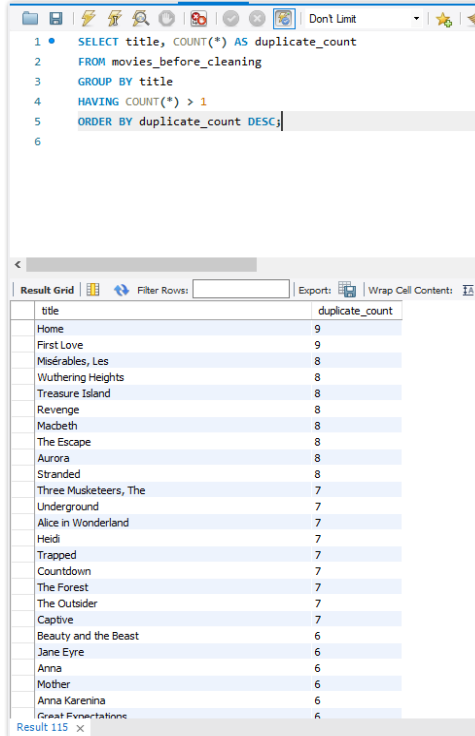
Rating column ->	Decimal(5,2)
------------------	--------------

## After

movieId	title	genres	year
28	Persuasion	Drama Romance	1995
29	City of Lost Children	Adventure Drama Fantasy Mystery Sci-Fi	1995
30	Shanghai Triad	Crime Drama	1995
31	Dangerous Minds	Drama	1995

# Comparison cleaned vs. Dirty data

## Movies table

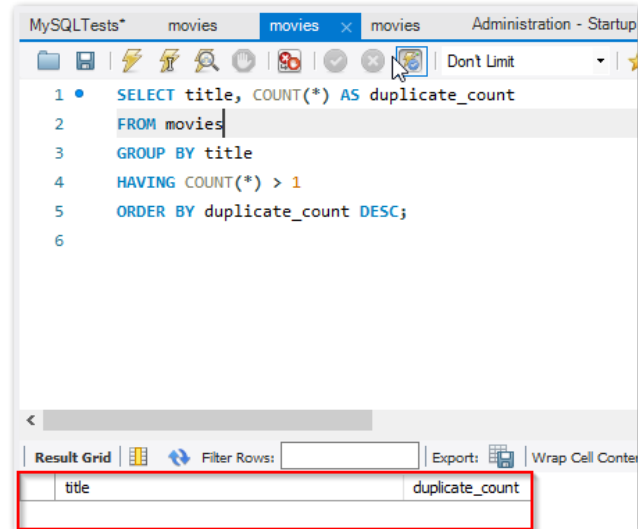


SQL Query:

```
1 • SELECT title, COUNT(*) AS duplicate_count
2 FROM movies_before_cleaning
3 GROUP BY title
4 HAVING COUNT(*) > 1
5 ORDER BY duplicate_count DESC;
6
```

Result Grid:

title	duplicate_count
Home	9
First Love	9
Misérables, Les	8
Wuthering Heights	8
Treasure Island	8
Revenge	8
Macbeth	8
The Escape	8
Aurora	8
Stranded	8
Three Musketeers, The	7
Underground	7
Alice in Wonderland	7
Heidi	7
Trapped	7
Countdown	7
The Forest	7
The Outsider	7
Captive	7
Beauty and the Beast	6
Jane Eyre	6
Anna	6
Mother	6
Anna Karenina	6
Great Expectations	6



SQL Query:

```
1 • SELECT title, COUNT(*) AS duplicate_count
2 FROM movies
3 GROUP BY title
4 HAVING COUNT(*) > 1
5 ORDER BY duplicate_count DESC;
6
```

Result Grid:

title	duplicate_count
-------	-----------------

# Comparison cleaned vs. Dirty data

## Revenues table

id	date	title	revenue	theaters	distributor
000020de-3a42-a942-5401-c68fcc662fa6	2007-11-18	Darfur Now	2531	22	Warner Independent Pictures (WIP)
000047dc-1f39-51bb-b898-ab6a87cbcee9	2015-01-09	Unbroken	2576345	3301	Universal Pictures
0001e185-4c4c-f41a-eb8f-2911020185c5	2015-10-31	Jurassic World	43465	164	Universal Pictures
00024499-c64c-1a2e-b57a-14acde4a0b2f	2009-09-01	G.I. Joe: The Rise of C...	595411	3467	Paramount Pictures
0002e659-a8e1-7bb1-6485-262379e9da25	2013-10-27	The Conjuring	6174	100	Warner Bros.
00036970-3d77-1f53-90ab-d8818b5e56d3	2014-01-23	Walking with Dinosaurs...	13466	454	Twentieth Century Fox
00046acb-5a08-8c83-20f3-2cbbdc4a09d8	2011-10-16	Don't Be Afraid of the ...	8314	80	FilmDistrict
0005b94f-8d48-182f-a89a-186cd18dd400	2016-04-02	My Big Fat Greek Wedd...	4889120	3179	Universal Pictures
0006742c-079a-a9bf-8556-326deb0d6b6d	2014-06-11	A Million Ways to Die in...	914865	3160	Universal Pictures
0006ca00-6340-035b-045a-6ef2e19eb501	2002-05-27	Kissing Jessica Stein	24235	64	Fox Searchlight Pictures

revenueId	title	revenue	distributor	year
54	Fairy Tail: Dragon Cry	172249.00000	FUNimation Entertainment	2017
97	Swimfan	4932608.00000	Twentieth Century Fox	2002
100	City by the Sea	3738191.00000	Warner Bros.	2002
105	Spider-Man/Men in Black IIDouble Bill	936685.00000	Sony Pictures Entertainment (SPE)	2002
147	Rent	4751680.00000	Revolution Studios	2005
184	Quattro Noza	717.00000	Lionsgate	2005
210	Ice Age: Continental Drift	16728341.00000	Twentieth Century Fox	2012
215	Savages	2740400.00000	Universal Pictures	2012
217	Katy Perry: Part of Me	1313660.00000	Paramount Pictures	2012
236	Speak Bachchan	61256.00000	FIP	2012

# Comparison cleaned vs. Dirty data

## Ratings table

Dirty

userId	movieId	rating	timestamp
1	296	5.0	1147880044
1	306	3.5	1147868817
1	307	5.0	1147868828
1	665	5.0	1147878820
1	899	3.5	1147868510
1	1088	4.0	1147868495
1	1175	3.5	1147868826
1	1217	3.5	1147878326
1	1237	5.0	1147868839
1	1250	4.0	1147868414
1	1260	3.5	1147877857
1	1653	4.0	1147868097
1	2011	2.5	1147868079

Clean

ratingId	userId	movieId	rating
1	1	296	5.00
2	1	306	3.50
3	1	307	5.00
4	1	665	5.00
5	1	899	3.50
6	1	1088	4.00
7	1	1175	3.50
8	1	1217	3.50
9	1	1237	5.00
10	1	1250	4.00
11	1	1260	3.50
12	1	1653	4.00
13	1	2011	2.50



# Showcases

## Example1: show top 10 revenue movies in 2000

MySQLTests\* movies movies Administration - Startup / Shutdo...

```
1 • SELECT m.title, r.revenue
2 FROM movies_before_cleaning m
3 INNER JOIN revenues r ON m.title = r.title
4 WHERE r.year = 2000
5 ORDER BY r.revenue DESC
6 LIMIT 10;
7
```

Result Grid

title	revenue
X-Men	20785331.00000
Mission: Impossible II	16487335.00000
Charlie's Angels	13660578.00000
Charlie's Angels	13660578.00000
Gladiator	13615219.00000
Gladiator	13615219.00000
Nutty Professor II: The Klumps	12352020.00000
Unbreakable	12300000.00000
Meet the Parents	11972120.00000
What Women Want	11329000.00000

MySQLTests\* movies movies Administration - Startup / Shutdo...

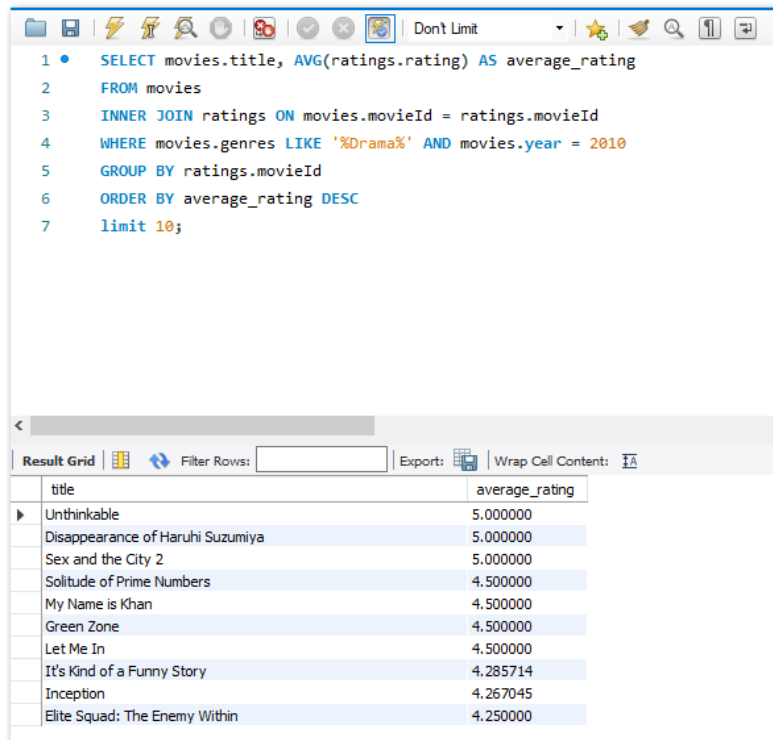
```
1 • SELECT m.title, r.revenue
2 FROM movies m
3 INNER JOIN revenues r ON m.title = r.title
4 WHERE r.year = 2000
5 ORDER BY r.revenue DESC
6 LIMIT 10;
7
```

Result Grid

title	revenue
X-Men	20785331.00000
How the Grinch Stole Christmas	19744225.00000
Mission: Impossible II	16487335.00000
Charlie's Angels	13660578.00000
Gladiator	13615219.00000
Nutty Professor II: The Klumps	12352020.00000
Unbreakable	12300000.00000
Meet the Parents	11972120.00000
What Women Want	11329000.00000
Cast Away	11120000.00000

# Showcases

## Example2: show top 10 rated drama movies in 2010



The screenshot shows a SQL query editor window with a toolbar at the top. The query is as follows:

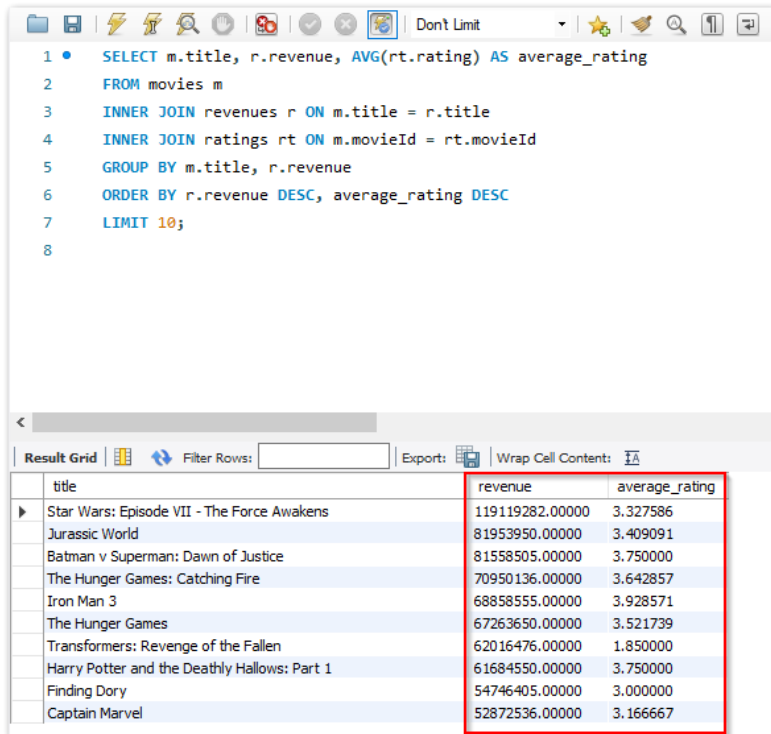
```
1 • SELECT movies.title, AVG(ratings.rating) AS average_rating
2 FROM movies
3 INNER JOIN ratings ON movies.movieId = ratings.movieId
4 WHERE movies.genres LIKE '%Drama%' AND movies.year = 2010
5 GROUP BY ratings.movieId
6 ORDER BY average_rating DESC
7 limit 10;
```

Below the query editor, the results are displayed in a table with two columns: 'title' and 'average\_rating'. The table shows the top 10 rated drama movies from 2010.

title	average_rating
Unthinkable	5.000000
Disappearance of Haruhi Suzumiya	5.000000
Sex and the City 2	5.000000
Solitude of Prime Numbers	4.500000
My Name is Khan	4.500000
Green Zone	4.500000
Let Me In	4.500000
It's Kind of a Funny Story	4.285714
Inception	4.267045
Elite Squad: The Enemy Within	4.250000

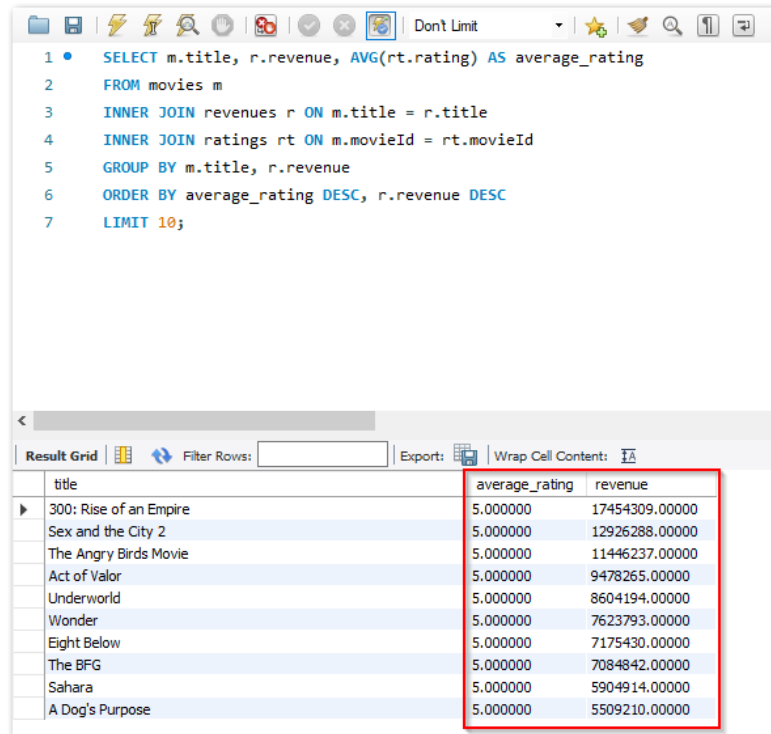
# Showcases

## Example3: find out if there is a relation between a movie rating and movie revenues.



```
1 • SELECT m.title, r.revenue, AVG(rt.rating) AS average_rating
2 FROM movies m
3 INNER JOIN revenues r ON m.title = r.title
4 INNER JOIN ratings rt ON m.movieId = rt.movieId
5 GROUP BY m.title, r.revenue
6 ORDER BY r.revenue DESC, average_rating DESC
7 LIMIT 10;
8
```

title	revenue	average_rating
Star Wars: Episode VII - The Force Awakens	119119282.00000	3.327586
Jurassic World	81953950.00000	3.409091
Batman v Superman: Dawn of Justice	81558505.00000	3.750000
The Hunger Games: Catching Fire	70950136.00000	3.642857
Iron Man 3	68858555.00000	3.928571
The Hunger Games	67263650.00000	3.521739
Transformers: Revenge of the Fallen	62016476.00000	1.850000
Harry Potter and the Deathly Hallows: Part 1	61684550.00000	3.750000
Finding Dory	54746405.00000	3.000000
Captain Marvel	52872536.00000	3.166667



```
1 • SELECT m.title, r.revenue, AVG(rt.rating) AS average_rating
2 FROM movies m
3 INNER JOIN revenues r ON m.title = r.title
4 INNER JOIN ratings rt ON m.movieId = rt.movieId
5 GROUP BY m.title, r.revenue
6 ORDER BY average_rating DESC, r.revenue DESC
7 LIMIT 10;
```

title	average_rating	revenue
300: Rise of an Empire	5.000000	17454309.00000
Sex and the City 2	5.000000	12926288.00000
The Angry Birds Movie	5.000000	11446237.00000
Act of Valor	5.000000	9478265.00000
Underworld	5.000000	8604194.00000
Wonder	5.000000	7623793.00000
Eight Below	5.000000	7175430.00000
The BFG	5.000000	7084842.00000
Sahara	5.000000	5904914.00000
A Dog's Purpose	5.000000	5509210.00000

# Improvement ideas

## Movies table

195243	काशी - In Search of Ganga	Drama Thriller	2018
150669	أهواك	(no genres listed)	2015
185889	ארבינקא	Comedy Crime Romance	1967
185313	Я хуюеу	Children Comedy	2018
172575	Юленька	Drama Horror Thriller	2009
193211	Чы в лесу шишки?	(no genres listed)	1965
206907	Чудесный колокольчик	Animation	1949
188053	Цветик-семицветик	Animation	1948
186775	Упырь	Action Horror	1997
196469	Ужас, который всегда с тобой	(no genres listed)	2007
193721	Терем-теремок	Animation	1971
193709	Стрекоза и муравей	(no genres listed)	1961
176659	Средь бела дня...	Drama	1982
188129	Спадок чарівника Бахрама	(no genres listed)	1975
188323	Сопілочка і глечик	(no genres listed)	1950
172499	Совершенно серьезно	(no genres listed)	1961
199155	Собака Павлова	Comedy Drama Romance	2005



**INTEGRATION**

Thank you!