Trainity Assignment – 5

Cleaning the data:: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

1)Your task: Clean the data

Solution:

1)When it comes to data analysis the most crucial step is the data cleaning process
2)So we will first drop the unwanted columns from the dataset and follow by removing the null values
The coloumns to be removed are

- Color
- director_facebook_likes
- actor_3_facebook_likes
- actor_1_facebook_likes
- cast_total_facebook_likes
- facenumber_in_poster
- plot_keywords
- movie_imdb_link
- actor_2_facebook_likes
- aspect_ratio
- movie_facebook_likes

These columns are dropped because they have no impact on our data analysis tasks

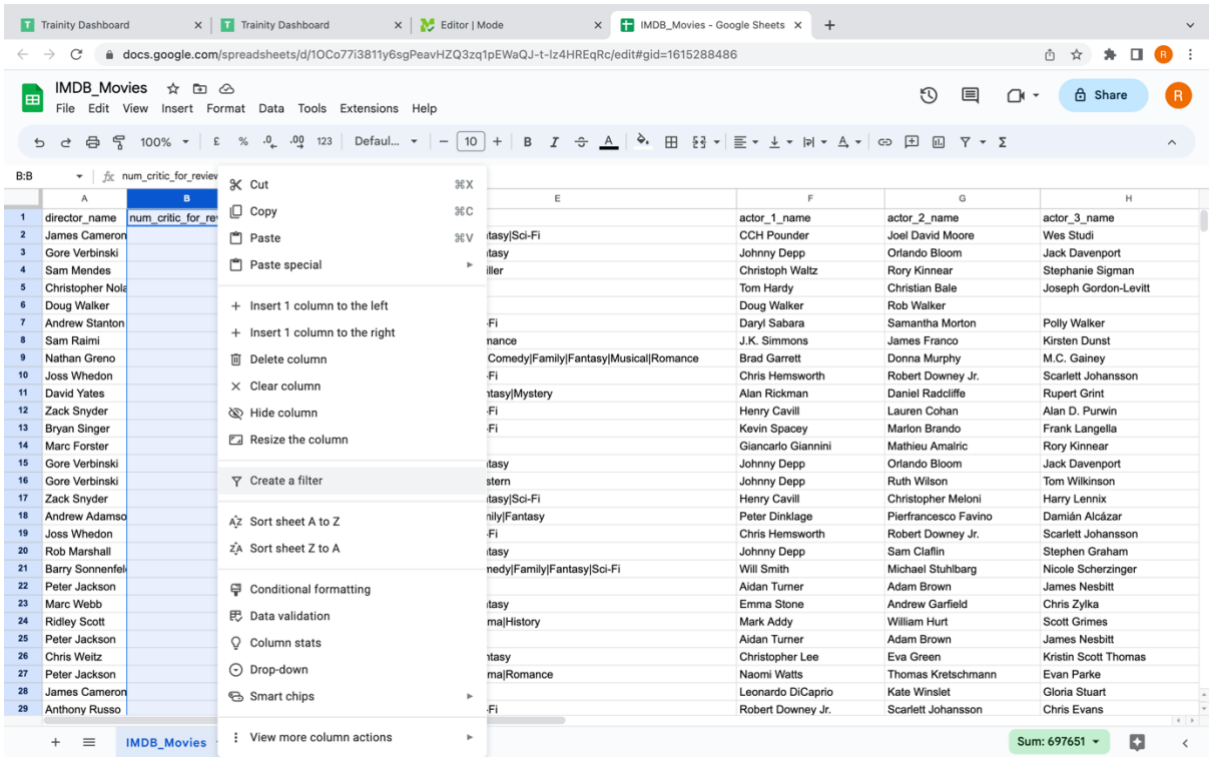
3)After removing these coloumns we see there are blank rows these blank rows can give our data analysis output a wrong result so we must remove these blank rows

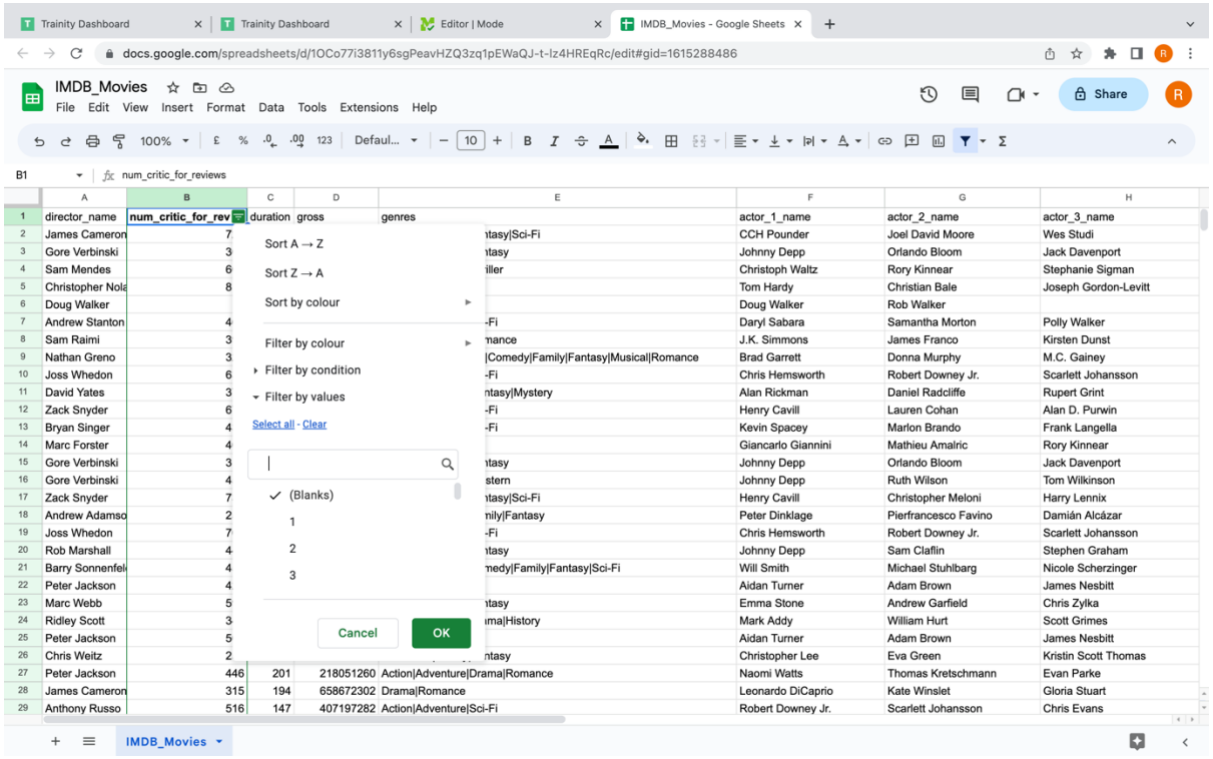
We can do this by creating a filter and filtering out the blank values for each coloumn and then removing that particular row

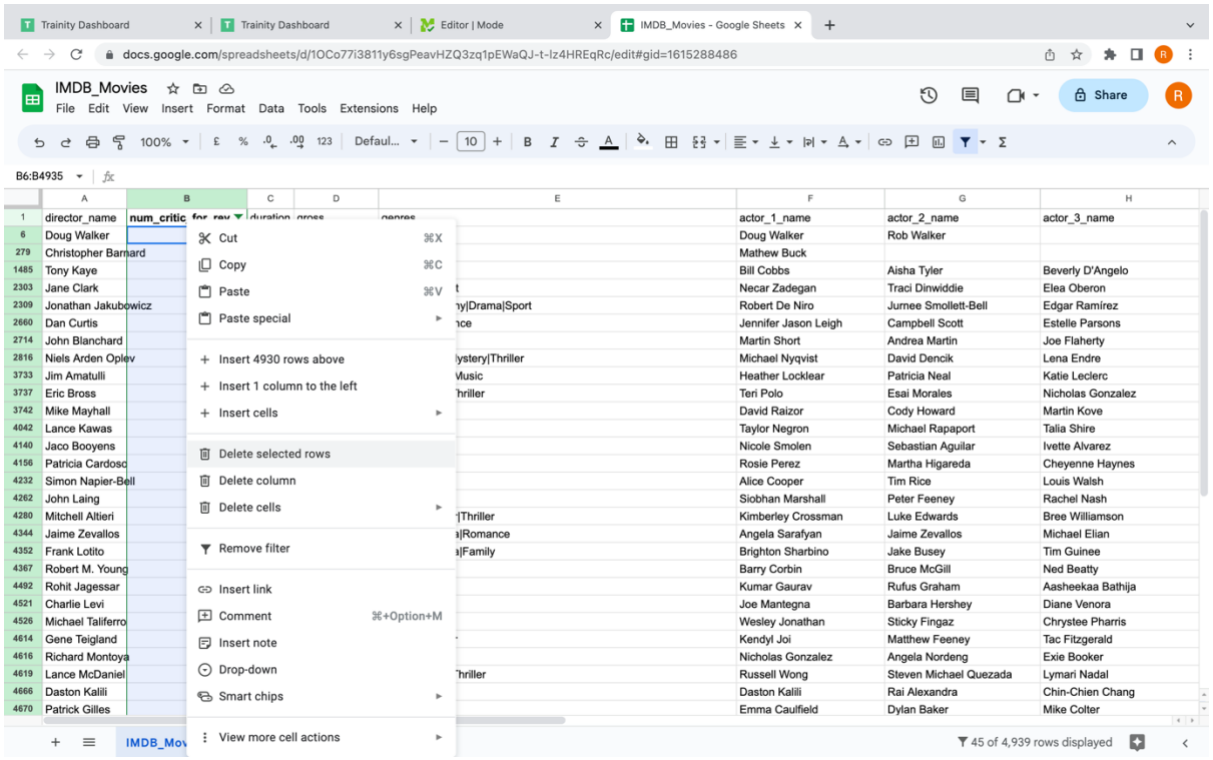
4) Finally we check for duplicates if any and remove it

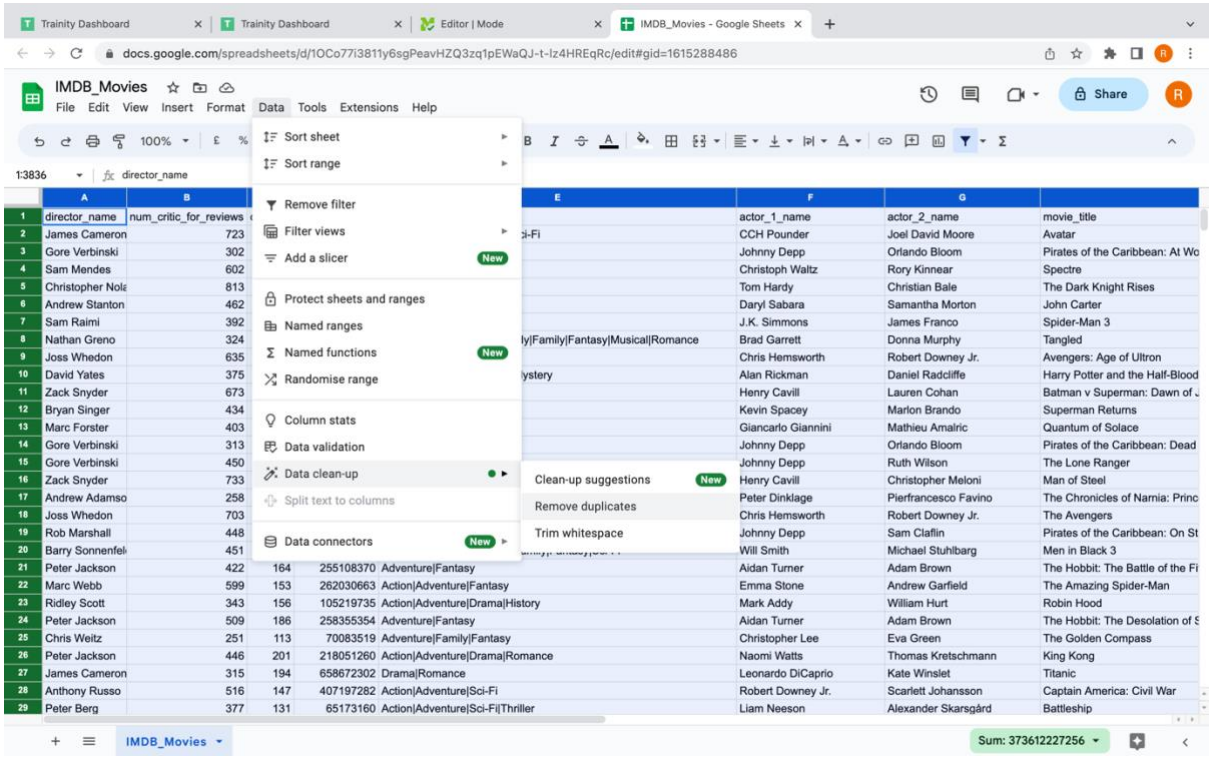
Final Dataset after cleaning

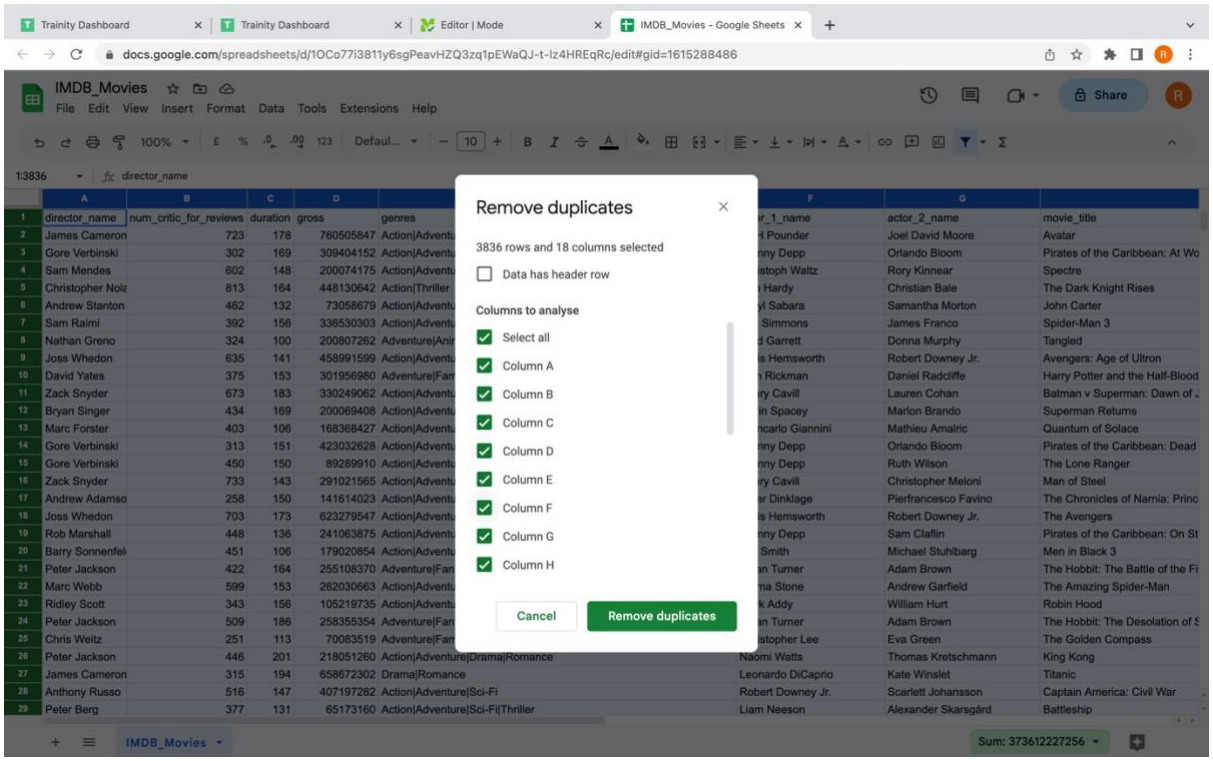
https://docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit?usp=sharing

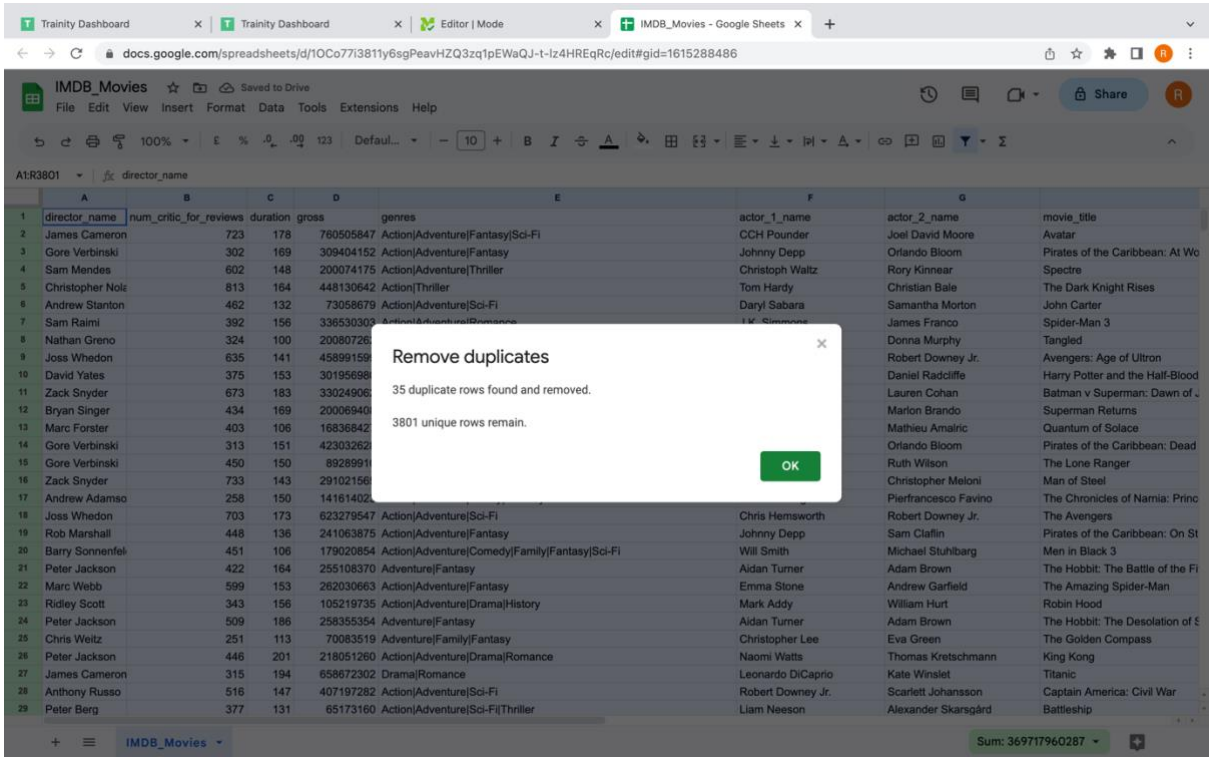

Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
File Edit View Insert Format Data Tools Extensions Help
Share
B:B
num_critic_for_reviev
Cut ⌘X
Copy ⌘C
Paste ⌘V
Paste special
Insert 1 column to the left
Insert 1 column to the right
Delete column
Clear column
Hide column
Resize the column
Create a filter
Sort sheet A to Z
Sort sheet Z to A
Conditional formatting
Data validation
Column stats
Drop-down
Smart chips
View more column actions
director_name
num_critic_for_re
actor_1_name
actor_2_name
actor_3_name
James Cameron
CCH Pounder
Joel David Moore
Wes Studi
Gore Verbinski
Johnny Depp
Orlando Bloom
Jack Davenport
Sam Mendes
Christoph Waltz
Rory Kinnear
Stephanie Sigman
Christopher Nola
Tom Hardy
Christian Bale
Joseph Gordon-Levitt
Doug Walker
Doug Walker
Rob Walker
Andrew Stanton
Daryl Sabara
Samantha Morton
Polly Walker
Sam Raimi
J.K. Simmons
James Franco
Kirsten Dunst
Nathan Greno
Brad Garrett
Donna Murphy
M.C. Gainey
Joss Whedon
Chris Hemsworth
Robert Downey Jr.
Scarlett Johansson
David Yates
Alan Rickman
Daniel Radcliffe
Rupert Grint
Zack Snyder
Henry Cavill
Lauren Cohan
Alan D. Purwin
Bryan Singer
Kevin Spacey
Marlon Brando
Frank Langella
Marc Forster
Giancarlo Giannini
Mathieu Amalric
Rory Kinnear
Gore Verbinski
Johnny Depp
Orlando Bloom
Jack Davenport
Gore Verbinski
Johnny Depp
Ruth Wilson
Tom Wilkinson
Zack Snyder
Henry Cavill
Christopher Meloni
Harry Lennix
Andrew Adamso
Peter Dinklage
Pierfrancesco Favino
Damián Alcázar
Joss Whedon
Chris Hemsworth
Robert Downey Jr.
Scarlett Johansson
Rob Marshall
Johnny Depp
Sam Claflin
Stephen Graham
Barry Sonnenfel
Will Smith
Michael Stuhlbarg
Nicole Scherzinger
Peter Jackson
Aidan Turner
Adam Brown
James Nesbitt
Marc Webb
Emma Stone
Andrew Garfield
Chris Zylka
Ridley Scott
Mark Addy
William Hurt
Scott Grimes
Peter Jackson
Aidan Turner
Adam Brown
James Nesbitt
Chris Weitz
Christopher Lee
Eva Green
Kristin Scott Thomas
Peter Jackson
Naomi Watts
Thomas Kretschmann
Evan Parke
James Cameron
Leonardo DiCaprio
Kate Winslet
Gloria Stuart
Anthony Russo
Robert Downey Jr.
Scarlett Johansson
Chris Evans
IMDB_Movies
Sum: 697651


Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
File Edit View Insert Format Data Tools Extensions Help
Share
B1
num_critic_for_reviews
director_name
num_critic_for_rev
duration
gross
genres
actor_1_name
actor_2_name
actor_3_name
Sort A → Z
Sort Z → A
Sort by colour
Filter by colour
Filter by condition
Filter by values
Select all - Clear
(Blanks)
1
2
3
Cancel
OK
James Cameron 7
Gore Verbinski 3
Sam Mendes 6
Christopher Nola 8
Doug Walker
Andrew Stanton 4
Sam Raimi 3
Nathan Greno 3
Joss Whedon 6
David Yates 3
Zack Snyder 6
Bryan Singer 4
Marc Forster 4
Gore Verbinski 3
Gore Verbinski 4
Zack Snyder 7
Andrew Adamso 2
Joss Whedon 7
Rob Marshall 4
Barry Sonnenfel 4
Peter Jackson 4
Marc Webb 5
Ridley Scott 3
Peter Jackson 5
Chris Weitz 2
Peter Jackson 446 201 218051260 Action|Adventure|Drama|Romance
James Cameron 315 194 658672302 Drama|Romance
Anthony Russo 516 147 407197282 Action|Adventure|Sci-Fi
IMDB_Movies


Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
File Edit View Insert Format Data Tools Extensions Help
Share
B6:B4935
Cut ⌘X
Copy ⌘C
Paste ⌘V
Paste special
Insert 4930 rows above
Insert 1 column to the left
Insert cells
Delete selected rows
Delete column
Delete cells
Remove filter
Insert link
Comment ⌘+Option+M
Insert note
Drop-down
Smart chips
View more cell actions
45 of 4,939 rows displayed

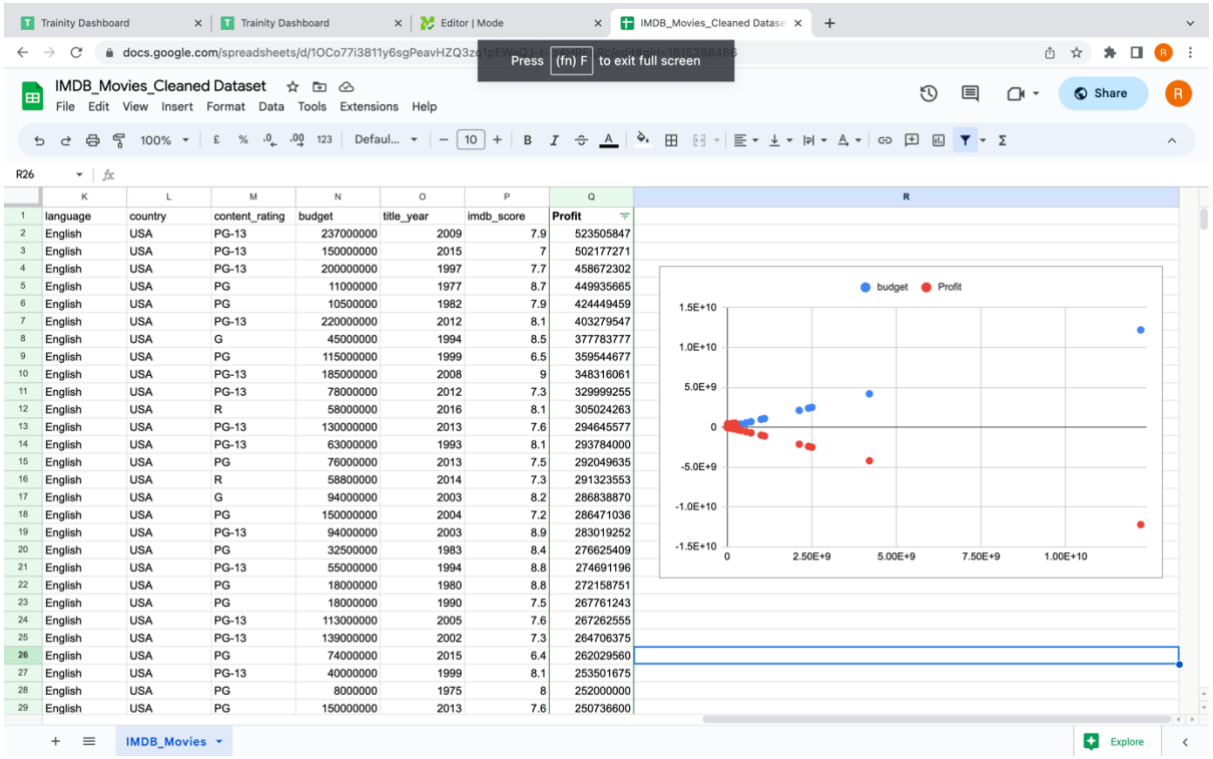

Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
File Edit View Insert Format Data Tools Extensions Help
Share
1:3836
director_name
Sort sheet
Sort range
Remove filter
Filter views
Add a slicer New
Protect sheets and ranges
Named ranges
Named functions New
Randomise range
Column stats
Data validation
Data clean-up
Split text to columns
Data connectors New
Clean-up suggestions New
Remove duplicates
Trim whitespace
director_name num_critic_for_reviews
James Cameron 723
Gore Verbinski 302
Sam Mendes 602
Christopher Nol 813
Andrew Stanton 462
Sam Raimi 392
Nathan Greno 324
Joss Whedon 635
David Yates 375
Zack Snyder 673
Bryan Singer 434
Marc Forster 403
Gore Verbinski 313
Gore Verbinski 450
Zack Snyder 733
Andrew Adamso 258
Joss Whedon 703
Rob Marshall 448
Barry Sonnenfel 451
Peter Jackson 422 164 255108370 Adventure|Fantasy
Marc Webb 599 153 262030663 Action|Adventure|Fantasy
Ridley Scott 343 156 105219735 Action|Adventure|Drama|History
Peter Jackson 509 186 258355354 Adventure|Fantasy
Chris Weitz 251 113 70083519 Adventure|Family|Fantasy
Peter Jackson 446 201 218051260 Action|Adventure|Drama|Romance
James Cameron 315 194 658672302 Drama|Romance
Anthony Russo 516 147 407197282 Action|Adventure|Sci-Fi
Peter Berg 377 131 65173160 Action|Adventure|Sci-Fi|Thriller
actor_1_name actor_2_name movie_title
CCH Pounder Joel David Moore Avatar
Johnny Depp Orlando Bloom Pirates of the Caribbean: At Wo
Christoph Waltz Rory Kinnear Spectre
Tom Hardy Christian Bale The Dark Knight Rises
Daryl Sabara Samantha Morton John Carter
J.K. Simmons James Franco Spider-Man 3
Brad Garrett Donna Murphy Tangled
Chris Hemsworth Robert Downey Jr. Avengers: Age of Ultron
Alan Rickman Daniel Radcliffe Harry Potter and the Half-Blood
Henry Cavill Lauren Cohan Batman v Superman: Dawn of
Kevin Spacey Marlon Brando Superman Returns
Giancarlo Giannini Mathieu Amalric Quantum of Solace
Johnny Depp Orlando Bloom Pirates of the Caribbean: Dead
Johnny Depp Ruth Wilson The Lone Ranger
Henry Cavill Christopher Meloni Man of Steel
Peter Dinklage Pierfrancesco Favino The Chronicles of Narnia: Princ
Chris Hemsworth Robert Downey Jr. The Avengers
Johnny Depp Sam Claflin Pirates of the Caribbean: On St
Will Smith Michael Stuhlbarg Men in Black 3
Aidan Turner Adam Brown The Hobbit: The Battle of the Fi
Emma Stone Andrew Garfield The Amazing Spider-Man
Mark Addy William Hurt Robin Hood
Aidan Turner Adam Brown The Hobbit: The Desolation of S
Christopher Lee Eva Green The Golden Compass
Naomi Watts Thomas Kretschmann King Kong
Leonardo DiCaprio Kate Winslet Titanic
Robert Downey Jr. Scarlett Johansson Captain America: Civil War
Liam Neeson Alexander Skarsgård Battleship
IMDB_Movies
Sum: 373612227256


Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
File Edit View Insert Format Data Tools Extensions Help
Share
Remove duplicates
3836 rows and 18 columns selected
Data has header row
Columns to analyse
Select all
Column A
Column B
Column C
Column D
Column E
Column F
Column G
Column H
Cancel
Remove duplicates
IMDB_Movies
Sum: 373612227256

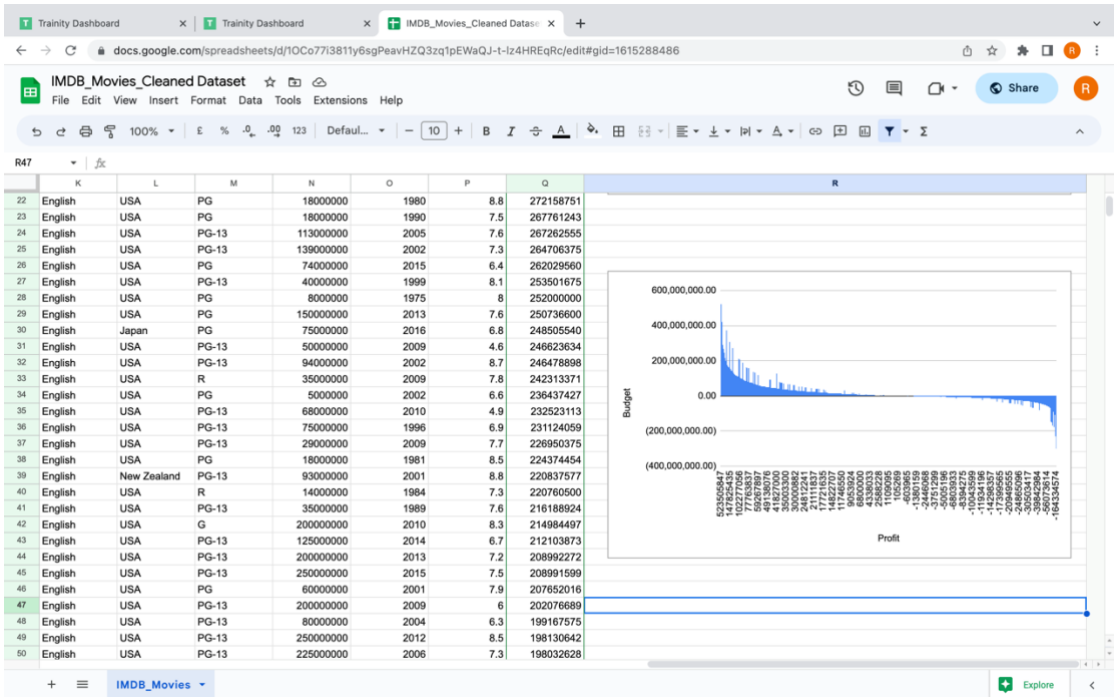

Trainity Dashboard
Trainity Dashboard
Editor | Mode
IMDB_Movies - Google Sheets
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help
Share
Remove duplicates
35 duplicate rows found and removed.
3801 unique rows remain.
OK
IMDB_Movies
Sum: 369717960287

2) **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

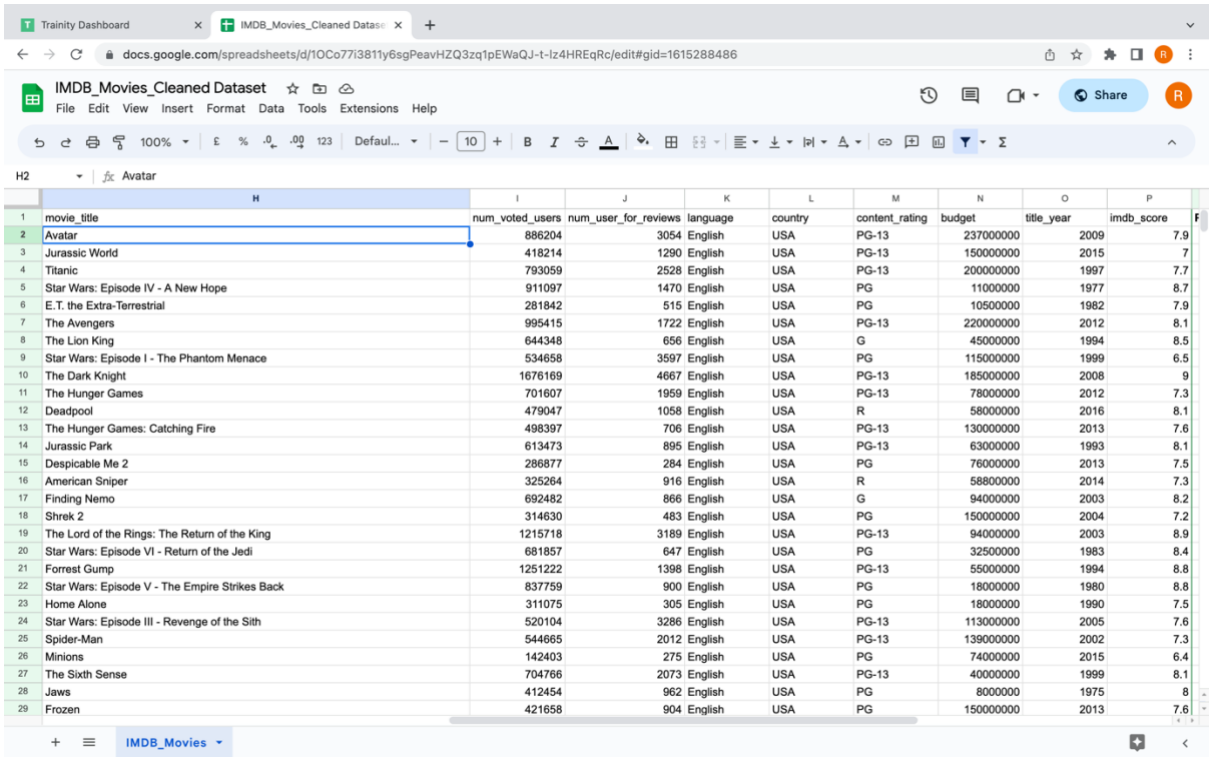
Your task: Find the movies with the highest profit?



We use scatterplot for observing outliers



Since the data is sorted in descending order of profit the movies showing at top are movies with highest profit



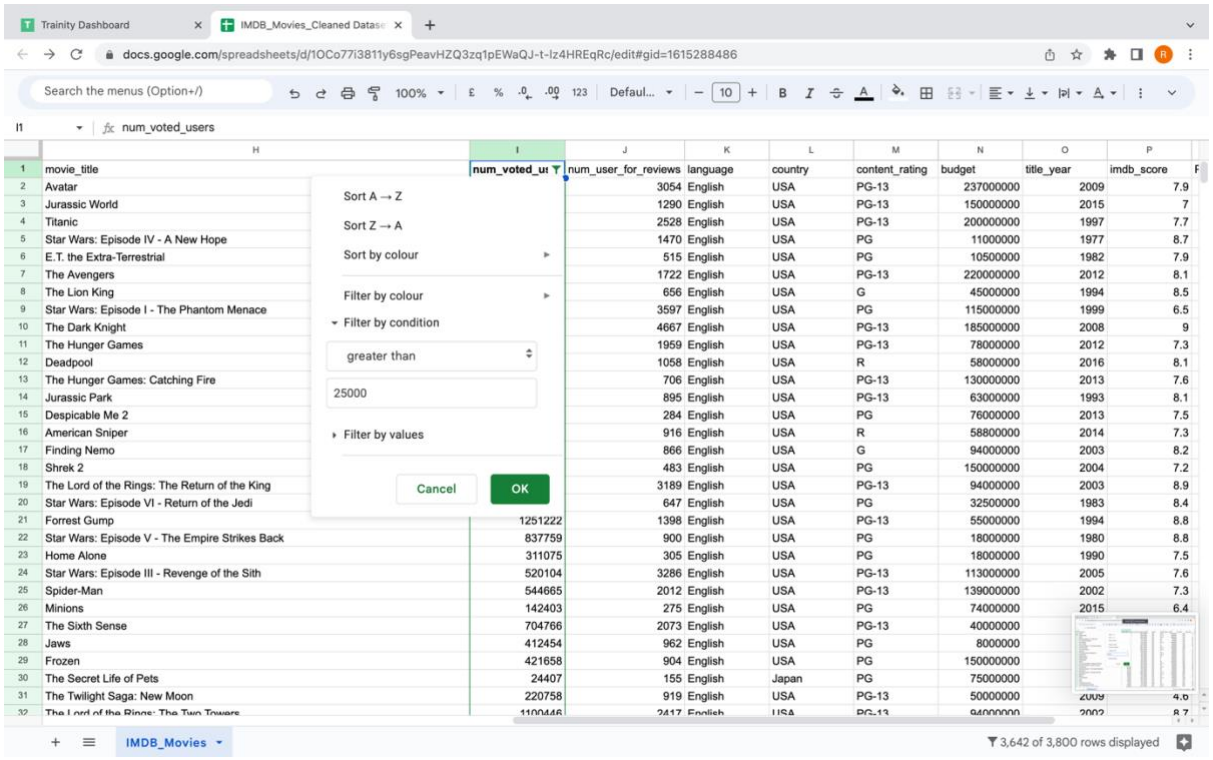
movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
Avatar	886204	3054	English	USA	PG-13	237000000	2009	7.9
Jurassic World	418214	1290	English	USA	PG-13	150000000	2015	7
Titanic	793059	2528	English	USA	PG-13	200000000	1997	7.7
Star Wars: Episode IV - A New Hope	911097	1470	English	USA	PG	11000000	1977	8.7
E.T. the Extra-Terrestrial	281842	515	English	USA	PG	10500000	1982	7.9
The Avengers	995415	1722	English	USA	PG-13	220000000	2012	8.1
The Lion King	644348	656	English	USA	G	45000000	1994	8.5
Star Wars: Episode I - The Phantom Menace	534658	3597	English	USA	PG	115000000	1999	6.5
The Dark Knight	1676169	4667	English	USA	PG-13	185000000	2008	9
The Hunger Games	701607	1959	English	USA	PG-13	78000000	2012	7.3
Deadpool	479047	1058	English	USA	R	58000000	2016	8.1
The Hunger Games: Catching Fire	498397	706	English	USA	PG-13	130000000	2013	7.6
Jurassic Park	613473	895	English	USA	PG-13	63000000	1993	8.1
Despicable Me 2	286877	284	English	USA	PG	76000000	2013	7.5
American Sniper	325264	916	English	USA	R	58800000	2014	7.3
Finding Nemo	692482	866	English	USA	G	94000000	2003	8.2
Shrek 2	314630	483	English	USA	PG	150000000	2004	7.2
The Lord of the Rings: The Return of the King	1215718	3189	English	USA	PG-13	94000000	2003	8.9
Star Wars: Episode VI - Return of the Jedi	681857	647	English	USA	PG	32500000	1983	8.4
Forrest Gump	1251222	1398	English	USA	PG-13	55000000	1994	8.8
Star Wars: Episode V - The Empire Strikes Back	837759	900	English	USA	PG	18000000	1980	8.8
Home Alone	311075	305	English	USA	PG	18000000	1990	7.5
Star Wars: Episode III - Revenge of the Sith	520104	3286	English	USA	PG-13	113000000	2005	7.6
Spider-Man	544665	2012	English	USA	PG-13	139000000	2002	7.3
Minions	142403	275	English	USA	PG	74000000	2015	6.4
The Sixth Sense	704766	2073	English	USA	PG-13	40000000	1999	8.1
Jaws	412454	962	English	USA	PG	8000000	1975	8
Frozen	421658	904	English	USA	PG	150000000	2013	7.6

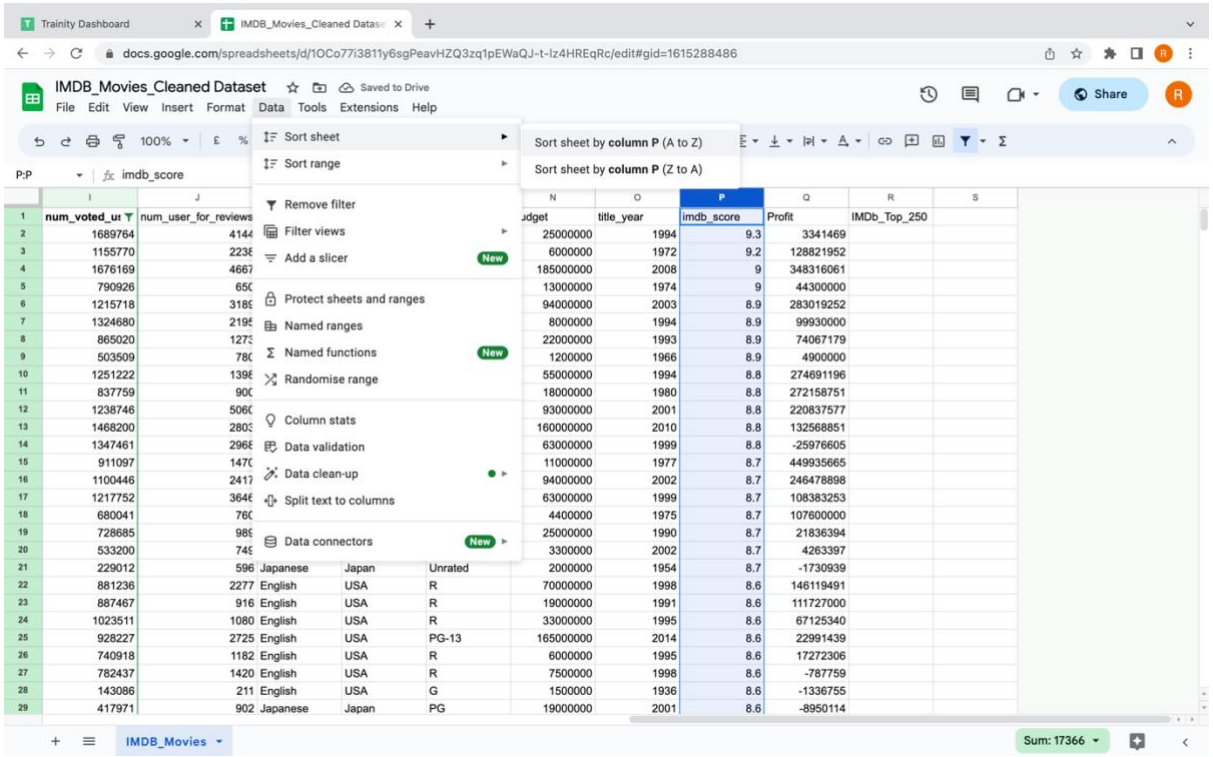
3)Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

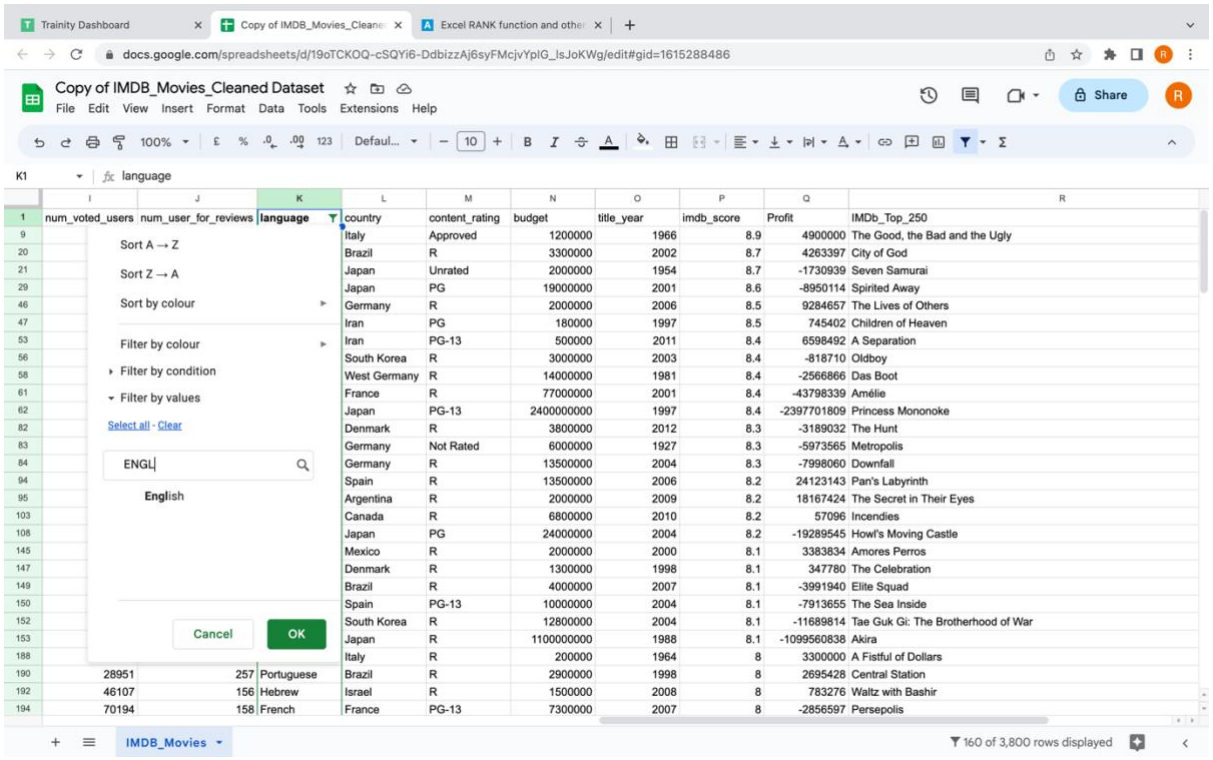
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

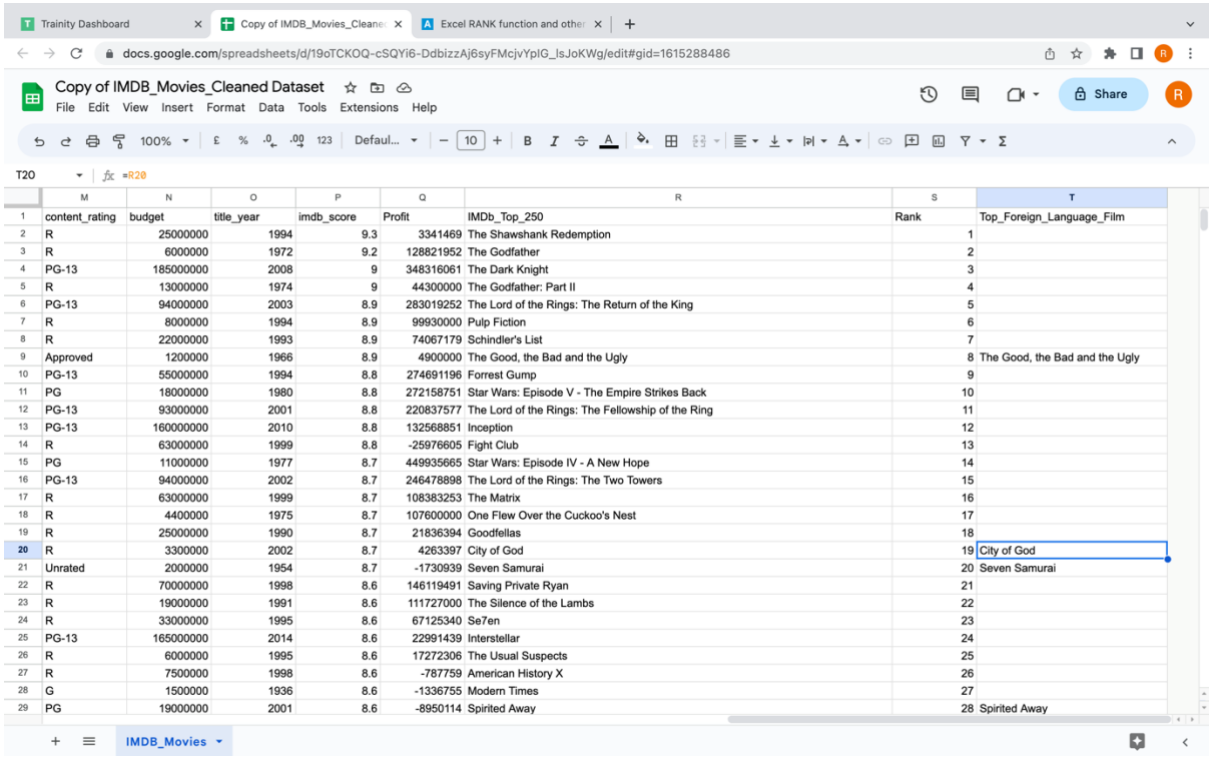
- First create a new coloumn names IMDb_Top_250
- Now we will have to filter out the rows whose num_voted_user>25000 using sort and filter tab
- Now arranging the dataset on the basis of imdb_score in descending order
- Since we need only top 250 we are selecting only top 250 rows
- Now we will create a new column using the rank() function to assign rank to the rows
- Now we use filter option and filter the other language and store it in the foreign language coloumn


Trainity Dashboard
IMDB_Movies_Cleaned Dataset
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
Search the menus (Option+/)
num_voted_users
Sort A → Z
Sort Z → A
Sort by colour
Filter by colour
Filter by condition
greater than
25000
Filter by values
Cancel
OK
IMDB_Movies
3,642 of 3,800 rows displayed


Trainity Dashboard
IMDB_Movies_Cleaned Dataset
docs.google.com/spreadsheets/d/1OCo77i3811y6sgPeavHZQ3zq1pEWaQJ-t-lz4HREqRc/edit#gid=1615288486
IMDB_Movies_Cleaned Dataset
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help
Share
Sort sheet
Sort range
Sort sheet by column P (A to Z)
Sort sheet by column P (Z to A)
Remove filter
Filter views
Add a slicer
New
Protect sheets and ranges
Named ranges
Named functions
New
Randomise range
Column stats
Data validation
Data clean-up
Split text to columns
Data connectors
New
P:P
imdb_score
IMDB_Movies
Sum: 17366



	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score	Profit	IMDb_Top_250
9				Italy	Approved	1200000	1966	8.9	4900000	The Good, the Bad and the Ugly
20				Brazil	R	3300000	2002	8.7	4263397	City of God
21				Japan	Unrated	2000000	1954	8.7	-1730939	Seven Samurai
29				Japan	PG	19000000	2001	8.6	-8950114	Spirited Away
46				Germany	R	2000000	2006	8.5	9284657	The Lives of Others
47				Iran	PG	180000	1997	8.5	745402	Children of Heaven
53				Iran	PG-13	500000	2011	8.4	6598492	A Separation
56				South Korea	R	3000000	2003	8.4	-818710	Oldboy
58				West Germany	R	14000000	1981	8.4	-2566866	Das Boot
61				France	R	77000000	2001	8.4	-43798339	Amélie
62				Japan	PG-13	2400000000	1997	8.4	-2397701809	Princess Mononoke
82				Denmark	R	3800000	2012	8.3	-3189032	The Hunt
83				Germany	Not Rated	6000000	1927	8.3	-5973565	Metropolis
84				Germany	R	13500000	2004	8.3	-7998060	Downfall
94				Spain	R	13500000	2006	8.2	24123143	Pan's Labyrinth
95				Argentina	R	2000000	2009	8.2	18167424	The Secret in Their Eyes
103				Canada	R	6800000	2010	8.2	57096	Incendies
108				Japan	PG	24000000	2004	8.2	-19289545	Howl's Moving Castle
145				Mexico	R	2000000	2000	8.1	3383834	Amores Perros
147				Denmark	R	1300000	1998	8.1	347780	The Celebration
149				Brazil	R	4000000	2007	8.1	-3991940	Elite Squad
150				Spain	PG-13	10000000	2004	8.1	-7913655	The Sea Inside
152				South Korea	R	12800000	2004	8.1	-11689814	Tae Guk Gi: The Brotherhood of War
153				Japan	R	1100000000	1988	8.1	-1099560838	Akira
188				Italy	R	200000	1964	8	3300000	A Fistful of Dollars
190	28951	257	Portuguese	Brazil	R	2900000	1998	8	2695428	Central Station
192	46107	156	Hebrew	Israel	R	1500000	2008	8	783276	Waltz with Bashir
194	70194	158	French	France	PG-13	7300000	2007	8	-2856597	Persepolis



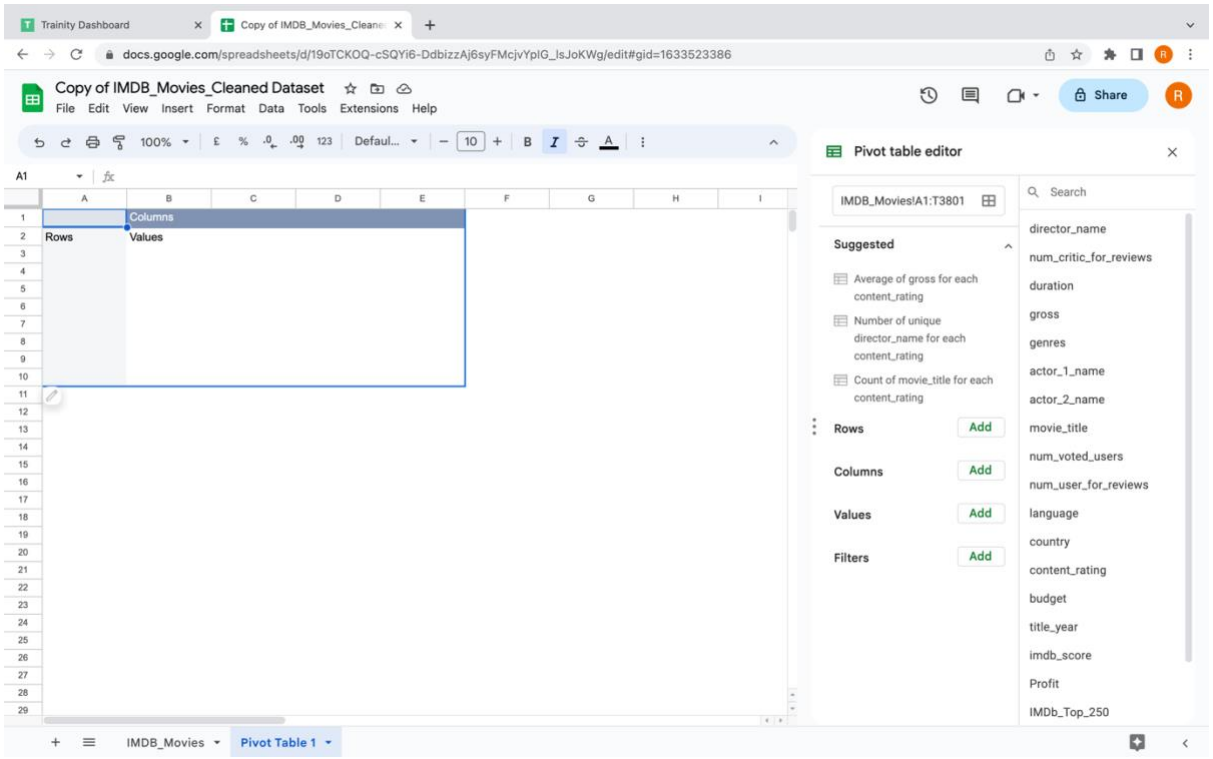
	content_rating	budget	title_year	imdb_score	Profit	IMDb_Top_250	Rank	Top_Foreign_Language_Film
2	R	25000000	1994	9.3	3341469	The Shawshank Redemption	1	
3	R	6000000	1972	9.2	128821952	The Godfather	2	
4	PG-13	185000000	2008	9	348316061	The Dark Knight	3	
5	R	13000000	1974	9	44300000	The Godfather: Part II	4	
6	PG-13	94000000	2003	8.9	283019252	The Lord of the Rings: The Return of the King	5	
7	R	8000000	1994	8.9	99930000	Pulp Fiction	6	
8	R	22000000	1993	8.9	74067179	Schindler's List	7	
9	Approved	1200000	1966	8.9	4900000	The Good, the Bad and the Ugly	8	The Good, the Bad and the Ugly
10	PG-13	55000000	1994	8.8	274691196	Forrest Gump	9	
11	PG	18000000	1980	8.8	272158751	Star Wars: Episode V - The Empire Strikes Back	10	
12	PG-13	93000000	2001	8.8	220837577	The Lord of the Rings: The Fellowship of the Ring	11	
13	PG-13	160000000	2010	8.8	132568851	Inception	12	
14	R	63000000	1999	8.8	-25976605	Fight Club	13	
15	PG	11000000	1977	8.7	449935665	Star Wars: Episode IV - A New Hope	14	
16	PG-13	94000000	2002	8.7	246478898	The Lord of the Rings: The Two Towers	15	
17	R	63000000	1999	8.7	108383253	The Matrix	16	
18	R	4400000	1975	8.7	107600000	One Flew Over the Cuckoo's Nest	17	
19	R	25000000	1990	8.7	21836394	Goodfellas	18	
20	R	3300000	2002	8.7	4263397	City of God	19	City of God
21	Unrated	2000000	1954	8.7	-1730939	Seven Samurai	20	Seven Samurai
22	R	70000000	1998	8.6	146119491	Saving Private Ryan	21	
23	R	19000000	1991	8.6	111727000	The Silence of the Lambs	22	
24	R	33000000	1995	8.6	67125340	Se7en	23	
25	PG-13	165000000	2014	8.6	22991439	Interstellar	24	
26	R	6000000	1995	8.6	17272306	The Usual Suspects	25	
27	R	7500000	1998	8.6	-787759	American History X	26	
28	G	1500000	1936	8.6	-1336755	Modern Times	27	
29	PG	19000000	2001	8.6	-8950114	Spirited Away	28	Spirited Away

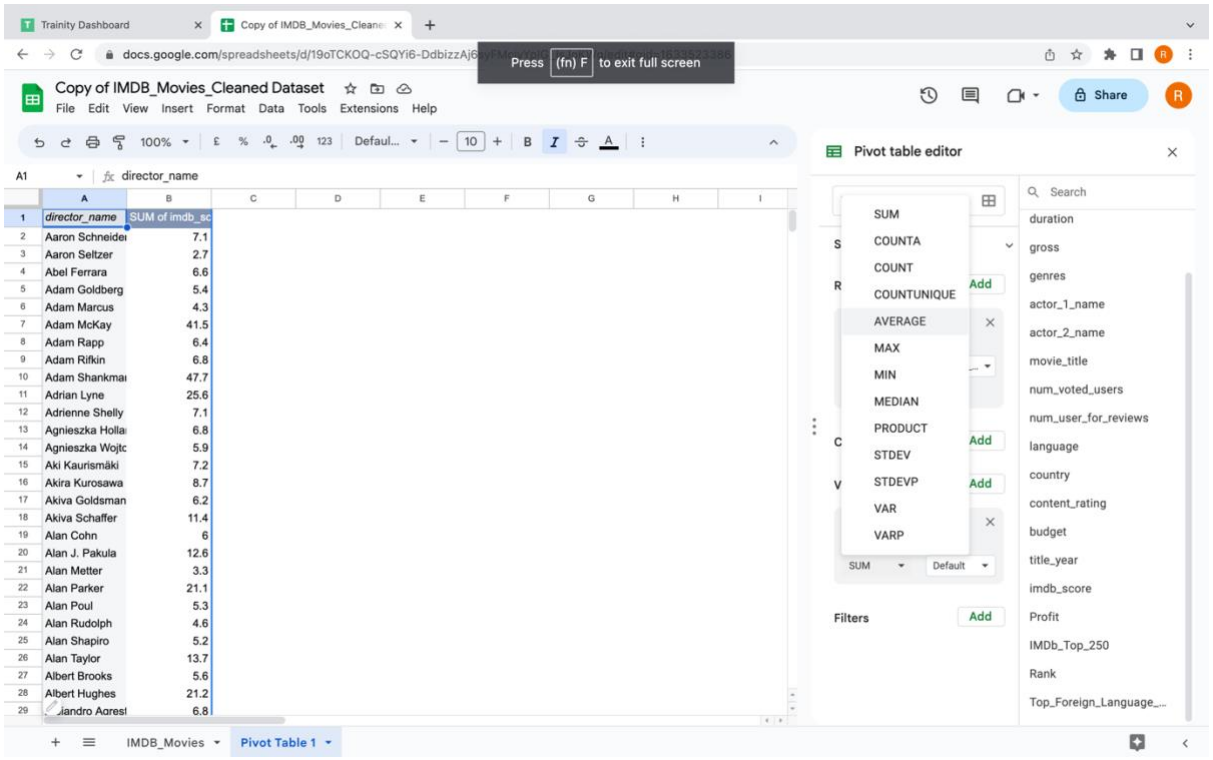
4) **Best Directors:** TGroup the column using the director_name column.

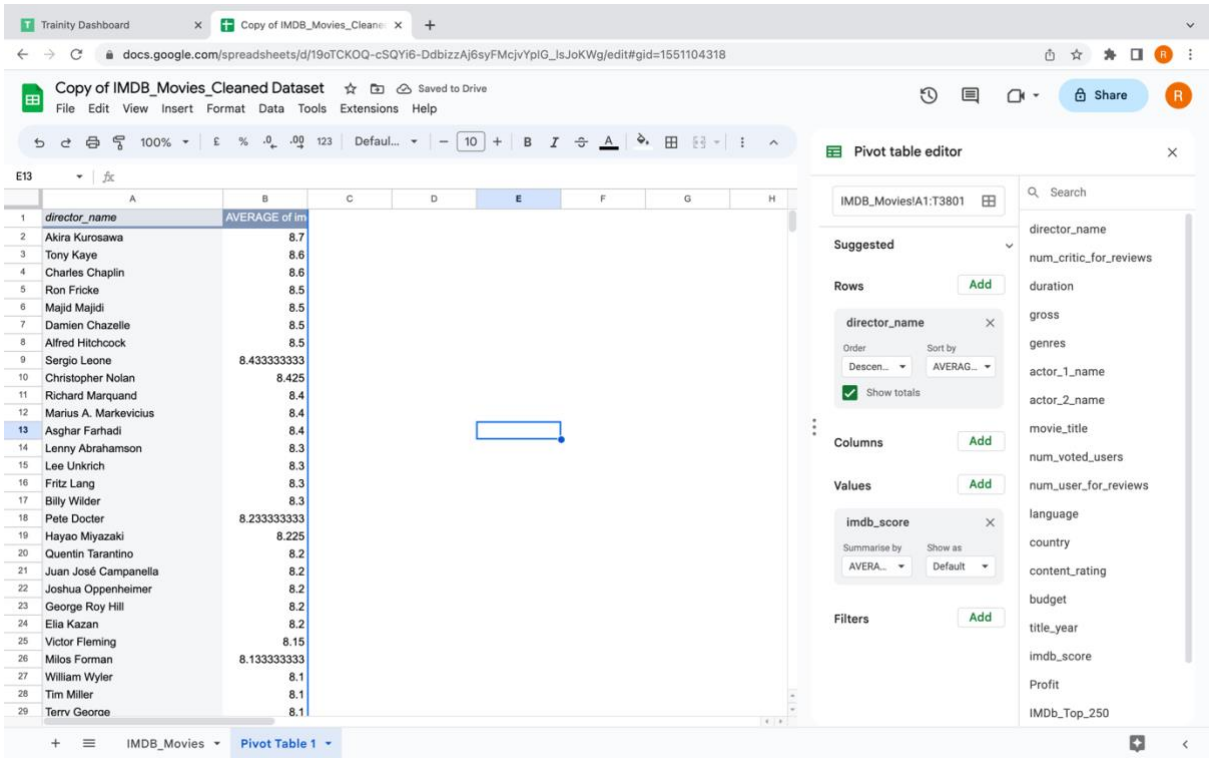
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

Your task: Find the best directors

- We will select the imdb_score column from the dataset.
- Now by using Pivot tables we will add director_name into series section of pivot table
- As per the question we must calculate mean (ie.average) of the imdb score
- Now Finally Sort the data on the basis of imdb average in descending order



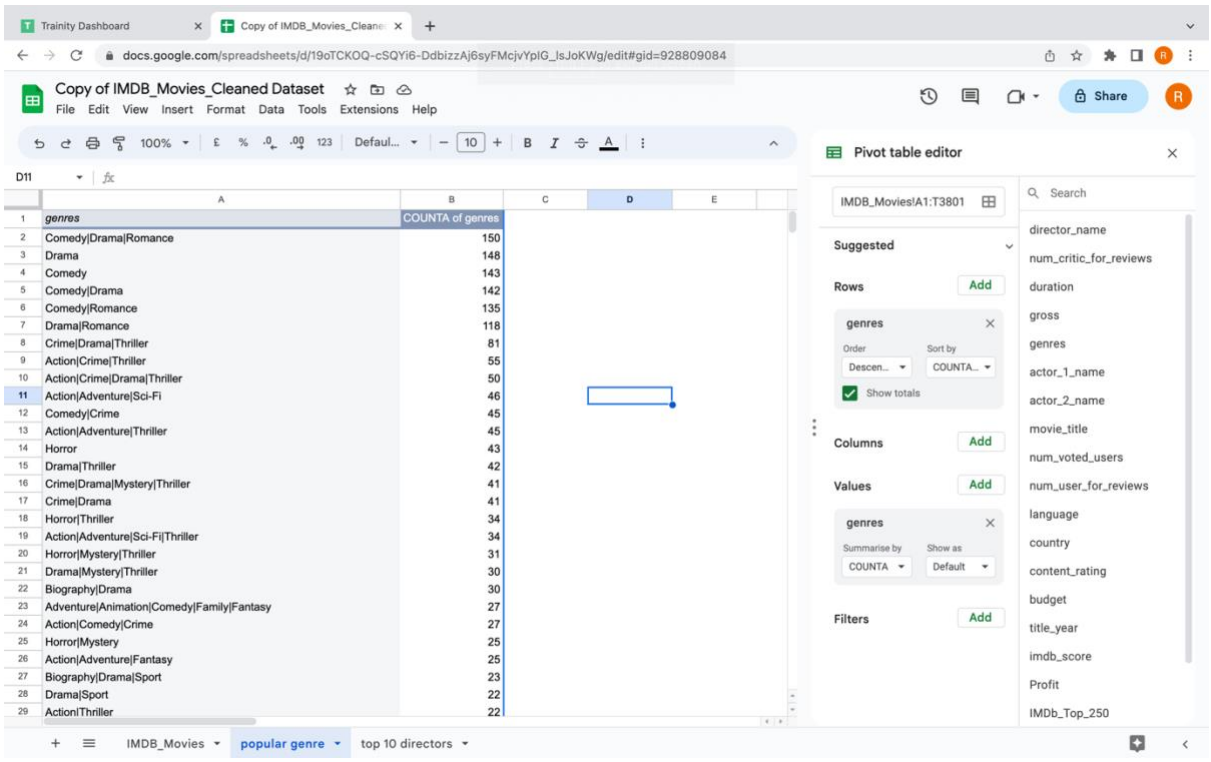

Trainity Dashboard
Copy of IMDB_Movies_Cleaned Dataset
Press (fn) F to exit full screen
Share
Pivot table editor
director_name
SUM of imdb_sc
Aaron Schneider 7.1
Aaron Seltzer 2.7
Abel Ferrara 6.6
Adam Goldberg 5.4
Adam Marcus 4.3
Adam McKay 41.5
Adam Rapp 6.4
Adam Rifkin 6.8
Adam Shankman 47.7
Adrian Lyne 25.6
Adrienne Shelly 7.1
Agnieszka Holland 6.8
Agnieszka Wojtc 5.9
Aki Kaurismäki 7.2
Akira Kurosawa 8.7
Akiva Goldsman 6.2
Akiva Schaffer 11.4
Alan Cohn 6
Alan J. Pakula 12.6
Alan Metter 3.3
Alan Parker 21.1
Alan Poul 5.3
Alan Rudolph 4.6
Alan Shapiro 5.2
Alan Taylor 13.7
Albert Brooks 5.6
Albert Hughes 21.2
Alejandro Agresti 6.8
SUM
COUNTA
COUNT
COUNTUNIQUE
AVERAGE
MAX
MIN
MEDIAN
PRODUCT
STDEV
STDEVP
VAR
VARP
Filters
Add
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
title_year
imdb_score
Profit
IMDb_Top_250
Rank
IMDB_Movies
Pivot Table 1


Trainity Dashboard
Copy of IMDB_Movies_Cleaned Dataset
Saved to Drive
Share
Pivot table editor
IMDB_Movies!A1:T3801
director_name
AVERAGE of im
Akira Kurosawa 8.7
Tony Kaye 8.6
Charles Chaplin 8.6
Ron Fricke 8.5
Majid Majidi 8.5
Damien Chazelle 8.5
Alfred Hitchcock 8.5
Sergio Leone 8.433333333
Christopher Nolan 8.425
Richard Marquand 8.4
Marius A. Markevicius 8.4
Asghar Farhadi 8.4
Lenny Abrahamson 8.3
Lee Unkrich 8.3
Fritz Lang 8.3
Billy Wilder 8.3
Pete Docter 8.233333333
Hayao Miyazaki 8.225
Quentin Tarantino 8.2
Juan José Campanella 8.2
Joshua Oppenheimer 8.2
George Roy Hill 8.2
Elia Kazan 8.2
Victor Fleming 8.15
Milos Forman 8.133333333
William Wyler 8.1
Tim Miller 8.1
Terry George 8.1
Suggested
Rows
Add
director_name
Order
Descen...
Sort by
AVERAG...
Show totals
Columns
Values
imdb_score
Summarise by
AVERA...
Show as
Default
Filters
num_critic_for_reviews
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
title_year
imdb_score
Profit
IMDb_Top_250
IMDB_Movies
Pivot Table 1

5) **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres

- Firstly we will use pivot table to solve this problem
- We will add genre column in rows and in values and use count function to count the occurrences
- Now arrange it in descending order and we have solved the problem



genres	COUNTA of genres
Comedy|Drama|Romance	150
Drama	148
Comedy	143
Comedy|Drama	142
Comedy|Romance	135
Drama|Romance	118
Crime|Drama|Thriller	81
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46
Comedy|Crime	45
Action|Adventure|Thriller	45
Horror	43
Drama|Thriller	42
Crime|Drama|Mystery|Thriller	41
Crime|Drama	41
Horror|Thriller	34
Action|Adventure|Sci-Fi|Thriller	34
Horror|Mystery|Thriller	31
Drama|Mystery|Thriller	30
Biography|Drama	30
Adventure|Animation|Comedy|Family|Fantasy	27
Action|Comedy|Crime	27
Horror|Mystery	25
Action|Adventure|Fantasy	25
Biography|Drama|Sport	23
Drama|Sport	22
Action|Thriller	22

6) **Charts:** Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

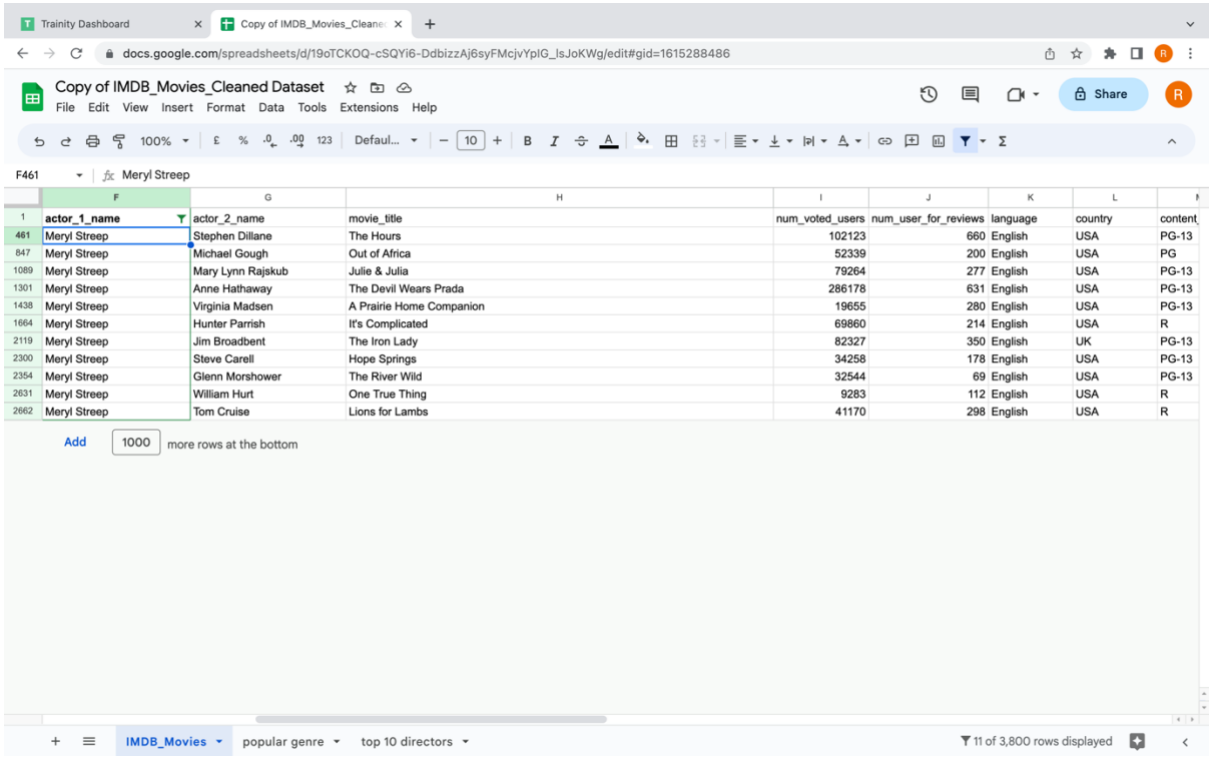
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

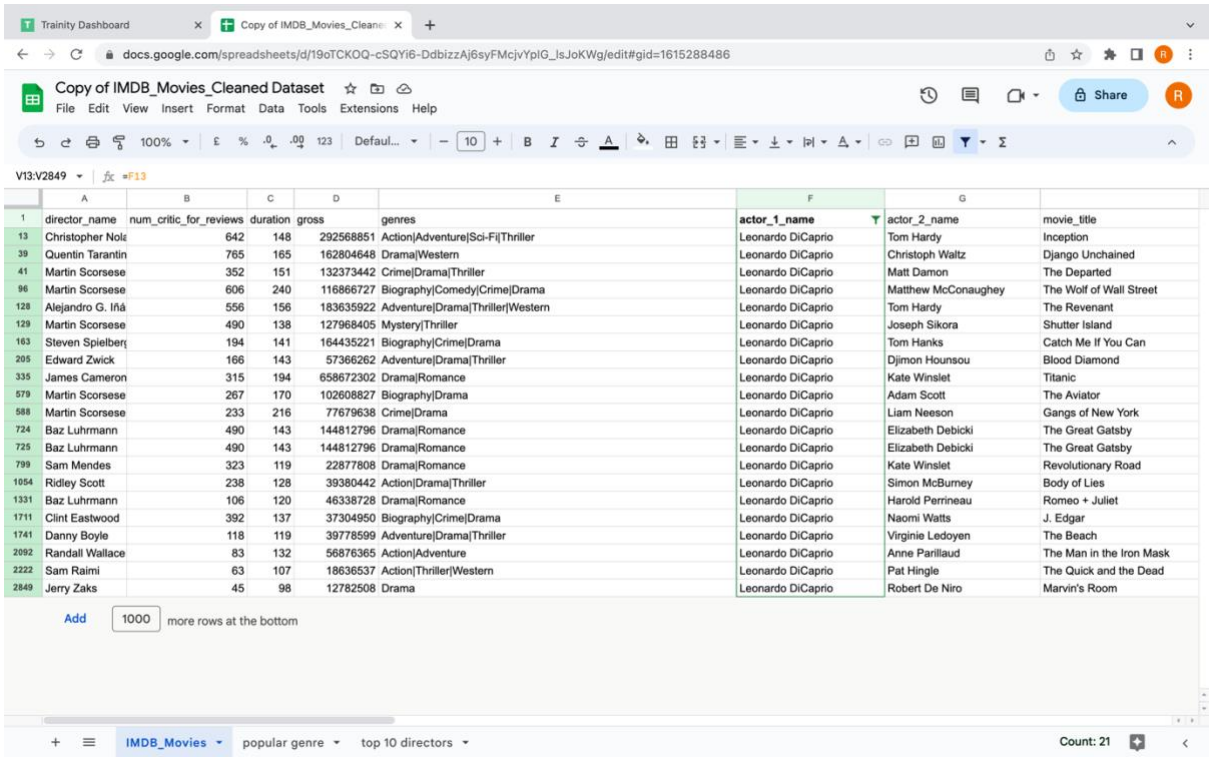
Observe the change in number of voted users over decades using a bar chart. Create a

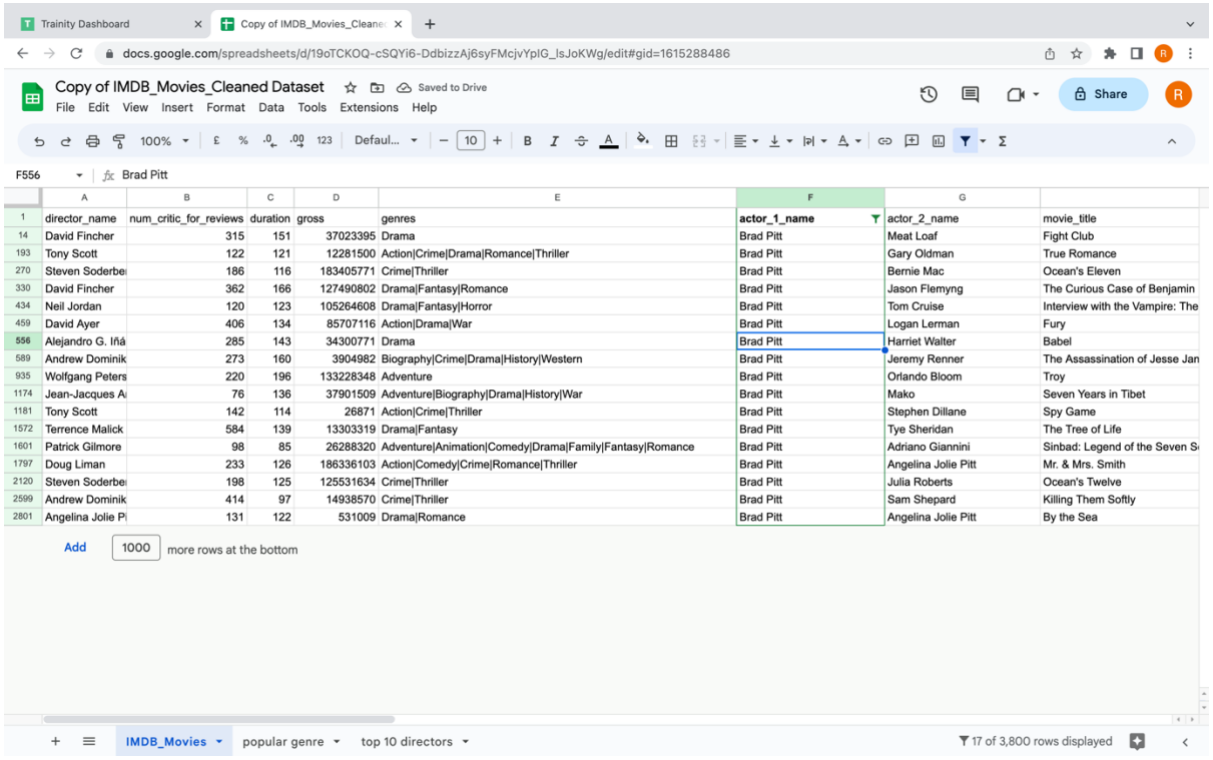
column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

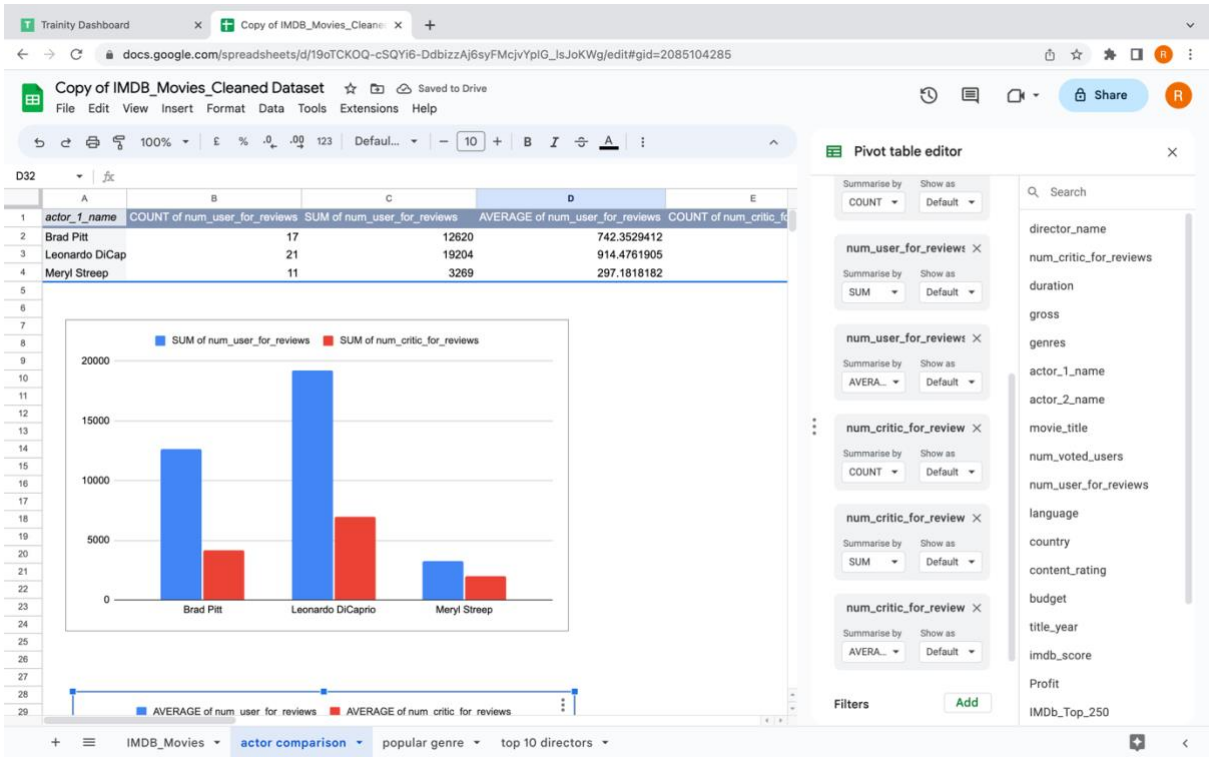
Your task: Find the critic-favorite and audience-favorite actors

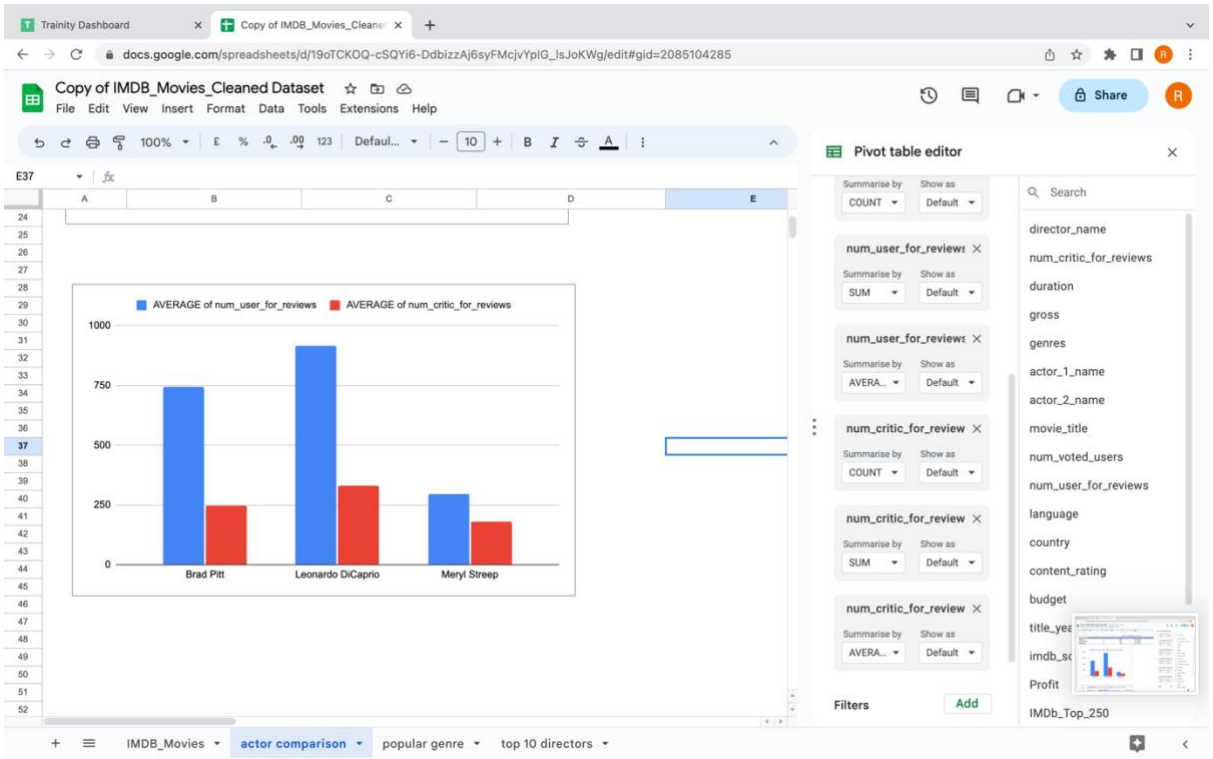


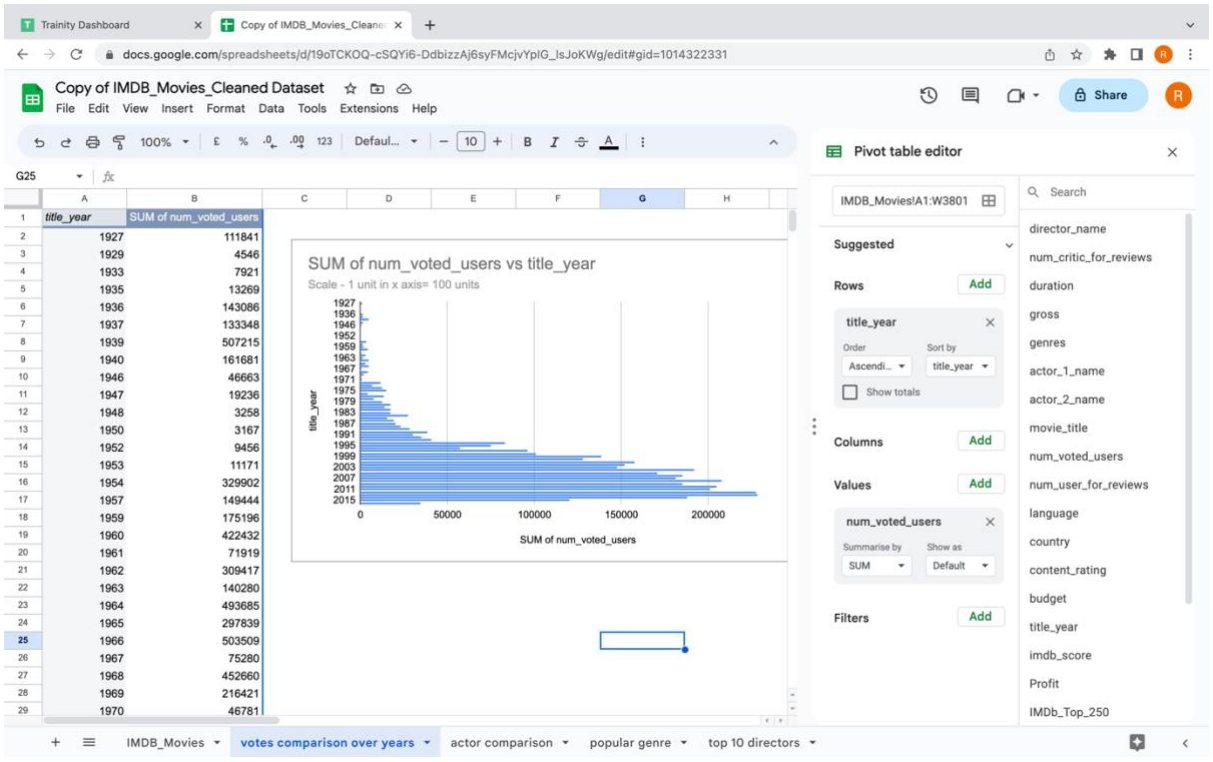
	actor_1_name	actor_2_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content
461	Meryl Streep	Stephen Dillane	The Hours	102123	660	English	USA	PG-13
847	Meryl Streep	Michael Gough	Out of Africa	52339	200	English	USA	PG
1089	Meryl Streep	Mary Lynn Rajskub	Julie & Julia	79264	277	English	USA	PG-13
1301	Meryl Streep	Anne Hathaway	The Devil Wears Prada	286178	631	English	USA	PG-13
1438	Meryl Streep	Virginia Madsen	A Prairie Home Companion	19655	280	English	USA	PG-13
1664	Meryl Streep	Hunter Parrish	It's Complicated	69860	214	English	USA	R
2119	Meryl Streep	Jim Broadbent	The Iron Lady	82327	350	English	UK	PG-13
2300	Meryl Streep	Steve Carell	Hope Springs	34258	178	English	USA	PG-13
2354	Meryl Streep	Glenn Morshower	The River Wild	32544	69	English	USA	PG-13
2631	Meryl Streep	William Hurt	One True Thing	9283	112	English	USA	R
2662	Meryl Streep	Tom Cruise	Lions for Lambs	41170	298	English	USA	R

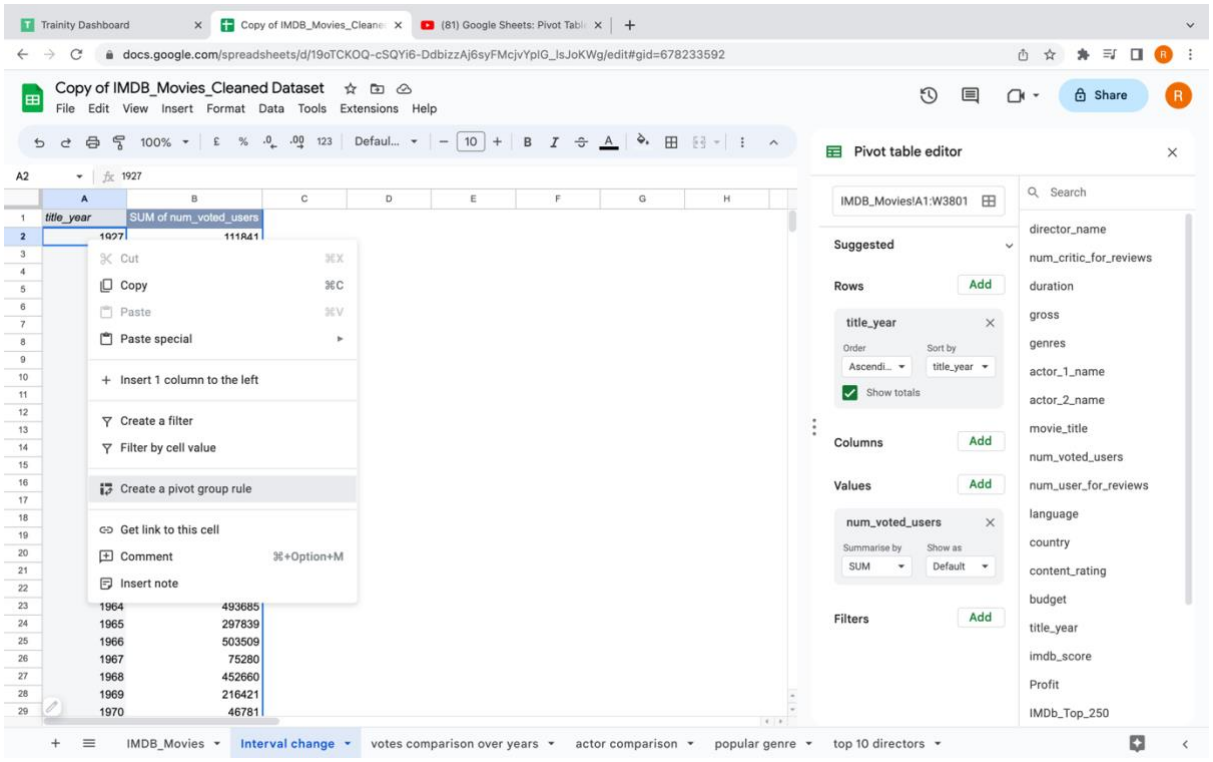

Trainity Dashboard
Copy of IMDB_Movies_Cleane
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=1615288486
Copy of IMDB_Movies_Cleaned Dataset
File Edit View Insert Format Data Tools Extensions Help
Share
V13:V2849
=F13
director_name num_critic_for_reviews duration gross genres actor_1_name actor_2_name movie_title
Christopher Nola 642 148 292568851 Action|Adventure|Sci-Fi|Thriller Leonardo DiCaprio Tom Hardy Inception
Quentin Tarantin 765 165 162804648 Drama|Western Leonardo DiCaprio Christoph Waltz Django Unchained
Martin Scorsese 352 151 132373442 Crime|Drama|Thriller Leonardo DiCaprio Matt Damon The Departed
Martin Scorsese 606 240 116866727 Biography|Comedy|Crime|Drama Leonardo DiCaprio Matthew McConaughey The Wolf of Wall Street
Alejandro G. Iñá 556 156 183635922 Adventure|Drama|Thriller|Western Leonardo DiCaprio Tom Hardy The Revenant
Martin Scorsese 490 138 127968405 Mystery|Thriller Leonardo DiCaprio Joseph Sikora Shutter Island
Steven Spielber 194 141 164435221 Biography|Crime|Drama Leonardo DiCaprio Tom Hanks Catch Me If You Can
Edward Zwick 166 143 57366262 Adventure|Drama|Thriller Leonardo DiCaprio Djimon Hounsou Blood Diamond
James Cameron 315 194 658672302 Drama|Romance Leonardo DiCaprio Kate Winslet Titanic
Martin Scorsese 267 170 102608827 Biography|Drama Leonardo DiCaprio Adam Scott The Aviator
Martin Scorsese 233 216 77679638 Crime|Drama Leonardo DiCaprio Liam Neeson Gangs of New York
Baz Luhrmann 490 143 144812796 Drama|Romance Leonardo DiCaprio Elizabeth Debicki The Great Gatsby
Baz Luhrmann 490 143 144812796 Drama|Romance Leonardo DiCaprio Elizabeth Debicki The Great Gatsby
Sam Mendes 323 119 22877808 Drama|Romance Leonardo DiCaprio Kate Winslet Revolutionary Road
Ridley Scott 238 128 39380442 Action|Drama|Thriller Leonardo DiCaprio Simon McBurney Body of Lies
Baz Luhrmann 106 120 46338728 Drama|Romance Leonardo DiCaprio Harold Perrineau Romeo + Juliet
Clint Eastwood 392 137 37304950 Biography|Crime|Drama Leonardo DiCaprio Naomi Watts J. Edgar
Danny Boyle 118 119 39778599 Adventure|Drama|Thriller Leonardo DiCaprio Virginie Ledoyen The Beach
Randall Wallace 83 132 56876365 Action|Adventure Leonardo DiCaprio Anne Parillaud The Man in the Iron Mask
Sam Raimi 63 107 18636537 Action|Thriller|Western Leonardo DiCaprio Pat Hingle The Quick and the Dead
Jerry Zaks 45 98 12782508 Drama Leonardo DiCaprio Robert De Niro Marvin's Room
Add 1000 more rows at the bottom
IMDB_Movies popular genre top 10 directors
Count: 21

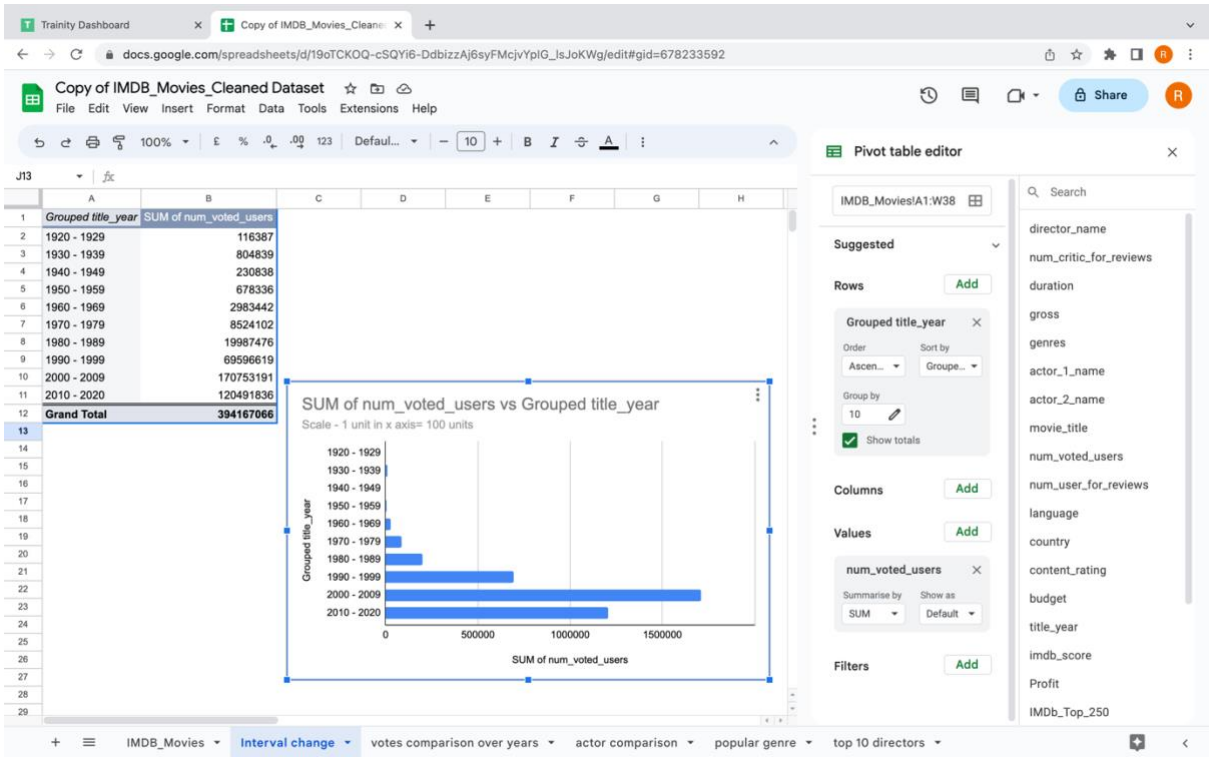

Trainity Dashboard
Copy of IMDB_Movies_Cleane
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=1615288486
Copy of IMDB_Movies_Cleaned Dataset
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help
Share
F556
Brad Pitt
director_name num_critic_for_reviews duration gross genres actor_1_name actor_2_name movie_title
David Fincher 315 151 37023395 Drama Brad Pitt Meat Loaf Fight Club
Tony Scott 122 121 12281500 Action|Crime|Drama|Romance|Thriller Brad Pitt Gary Oldman True Romance
Steven Soderbe 186 116 183405771 Crime|Thriller Brad Pitt Bernie Mac Ocean's Eleven
David Fincher 362 166 127490802 Drama|Fantasy|Romance Brad Pitt Jason Flemyng The Curious Case of Benjamin
Neil Jordan 120 123 105264608 Drama|Fantasy|Horror Brad Pitt Tom Cruise Interview with the Vampire: The
David Ayer 406 134 85707116 Action|Drama|War Brad Pitt Logan Lerman Fury
Alejandro G. Iñá 285 143 34300771 Drama Brad Pitt Harriet Walter Babel
Andrew Dominik 273 160 3904982 Biography|Crime|Drama|History|Western Brad Pitt Jeremy Renner The Assassination of Jesse Jan
Wolfgang Peters 220 196 133228348 Adventure Brad Pitt Orlando Bloom Troy
Jean-Jacques A 76 136 37901509 Adventure|Biography|Drama|History|War Brad Pitt Mako Seven Years in Tibet
Tony Scott 142 114 26871 Action|Crime|Thriller Brad Pitt Stephen Dillane Spy Game
Terrence Malick 584 139 13303319 Drama|Fantasy Brad Pitt Tye Sheridan The Tree of Life
Patrick Gilmore 98 85 26288320 Adventure|Animation|Comedy|Drama|Family|Fantasy|Romance Brad Pitt Adriano Giannini Sinbad: Legend of the Seven S
Doug Liman 233 126 186336103 Action|Comedy|Crime|Romance|Thriller Brad Pitt Angelina Jolie Pitt Mr. & Mrs. Smith
Steven Soderbe 198 125 125531634 Crime|Thriller Brad Pitt Julia Roberts Ocean's Twelve
Andrew Dominik 414 97 14938570 Crime|Thriller Brad Pitt Sam Shepard Killing Them Softly
Angelina Jolie P 131 122 531009 Drama|Romance Brad Pitt Angelina Jolie Pitt By the Sea
Add 1000 more rows at the bottom
IMDB_Movies popular genre top 10 directors
17 of 3,800 rows displayed


Trainity Dashboard
Copy of IMDB_Movies_Cleaned Dataset
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=2085104285
Copy of IMDB_Movies_Cleaned Dataset
File Edit View Insert Format Data Tools Extensions Help
Saved to Drive
Share
D32
actor_1_name | COUNT of num_user_for_reviews | SUM of num_user_for_reviews | AVERAGE of num_user_for_reviews | COUNT of num_critic_fo
Brad Pitt | 17 | 12620 | 742.3529412
Leonardo DiCap | 21 | 19204 | 914.4761905
Meryl Streep | 11 | 3269 | 297.1818182
SUM of num_user_for_reviews
SUM of num_critic_for_reviews
20000
15000
10000
5000
0
Brad Pitt
Leonardo DiCaprio
Meryl Streep
AVERAGE of num_user_for_reviews
AVERAGE of num_critic_for_reviews
Pivot table editor
Summarise by
Show as
COUNT
Default
num_user_for_reviews
SUM
AVERA...
num_critic_for_review
Filters
Add
Search
director_name
num_critic_for_reviews
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
title_year
imdb_score
Profit
IMDb_Top_250
IMDB_Movies
actor comparison
popular genre
top 10 directors


Trainity Dashboard
Copy of IMDB_Movies_Cleaned Dataset
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=2085104285
Copy of IMDB_Movies_Cleaned Dataset
File Edit View Insert Format Data Tools Extensions Help
Share
E37
AVERAGE of num_user_for_reviews
AVERAGE of num_critic_for_reviews
1000
750
500
250
0
Brad Pitt
Leonardo DiCaprio
Meryl Streep
Pivot table editor
Summarise by
Show as
COUNT
Default
num_user_for_reviews
SUM
AVERA...
num_critic_for_review
Filters
Add
Search
director_name
num_critic_for_reviews
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
Profit
IMDb_Top_250
IMDB_Movies
actor comparison
popular genre
top 10 directors


Trainity Dashboard
Copy of IMDB_Movies_Cleaned Dataset
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=1014322331
Copy of IMDB_Movies_Cleaned Dataset
File Edit View Insert Format Data Tools Extensions Help
Share
Pivot table editor
IMDB_Movies!A1:W3801
Search
Suggested
Rows
Add
title_year
Order
Ascendi...
Sort by
title_year
Show totals
Columns
Values
num_voted_users
Summarise by
SUM
Show as
Default
Filters
director_name
num_critic_for_reviews
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
title_year
imdb_score
Profit
IMDb_Top_250
SUM of num_voted_users vs title_year
Scale - 1 unit in x axis= 100 units
SUM of num_voted_users
IMDB_Movies
votes comparison over years
actor comparison
popular genre
top 10 directors


Trainity Dashboard
(81) Google Sheets: Pivot Tabl
docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit#gid=678233592
Copy of IMDB_Movies_Cleaned Dataset
File Edit View Insert Format Data Tools Extensions Help
Share
Pivot table editor
IMDB_Movies!A1:W3801
Search
Suggested
Rows
Add
title_year
Order
Ascendi...
Sort by
title_year
Show totals
Columns
Values
num_voted_users
Summarise by
SUM
Show as
Default
Filters
director_name
num_critic_for_reviews
duration
gross
genres
actor_1_name
actor_2_name
movie_title
num_voted_users
num_user_for_reviews
language
country
content_rating
budget
title_year
imdb_score
Profit
IMDb_Top_250
Cut
Copy
Paste
Paste special
Insert 1 column to the left
Create a filter
Filter by cell value
Create a pivot group rule
Get link to this cell
Comment
Insert note
IMDB_Movies
Interval change
votes comparison over years
actor comparison
popular genre
top 10 directors



Grouped title_year	SUM of num_voted_users
1920 - 1929	116387
1930 - 1939	804839
1940 - 1949	230838
1950 - 1959	678336
1960 - 1969	2983442
1970 - 1979	8524102
1980 - 1989	19987476
1990 - 1999	69596619
2000 - 2009	170753191
2010 - 2020	120491836
Grand Total	**394167066**

Link to Excel File:

https://docs.google.com/spreadsheets/d/19oTCKOQ-cSQYi6-DdbizzAj6syFMcjvYpIG_lsJoKWg/edit?usp=sharing