TRAINITY DATA ANALYTICS PROJECT - 5

IMDB

MOVIE ANASYSIS

BY

ASHWIN K

CONTENTS

1 PROJECT DESCRIPTION

2 APPROACH

3 TECH-STACK USED

4 INSIGHTS

5 SOLUTIONS

6 RESULT

PROJECT DESCRIPTION

- THE MAIN FOCUS OF THE PROJECT IS TO MAKE USE OF A DATASET THAT CONTAINS COMPREHENSIVE INFORMATION ABOUT NUMEROUS MOVIES FROM THE WELL-KNOWN ONLINE DATABASE IMDB. IMDB SERVES AS A HIGHLY REGARDED AND ALL-ENCOMPASSING REPOSITORY FOR A WIDE RANGE OF MEDIA.

- BY ADDRESSING SPECIFIC INQUIRIES, THE PRIMARY OBJECTIVE IS TO IDENTIFY A PROBLEM OR CHALLENGE THAT NECESSITATES A SOLUTION AND DETERMINE THE MOST SUITABLE DATA TO EFFECTIVELY TACKLE IT.

- DURING THIS PROCESS, DATA CLEANSING AND EMPLOYING DATA ANALYSIS TECHNIQUES MAY BE REQUIRED TO THOROUGHLY EXAMINE THE DATASET, UNVEIL VALUABLE INSIGHTS, AND REACH WELL-FOUNDED CONCLUSIONS.

- BY UNDERTAKING THESE TASKS, THE PROJECT AIMS TO LEVERAGE THE ABUNDANT INFORMATION WITHIN THE IMDB DATASET, UTILIZING DATA ANALYSIS SKILLS TO DERIVE INSIGHTS, ENHANCE DECISION-MAKING PROCESSES, AND GAIN VALUABLE KNOWLEDGE WITHIN THE REALM OF MOVIES.

APPROACH

1

THE DATASET UNDERGOES A DATA CLEANING PROCESS, INVOLVING THE REMOVAL OF UNNECESSARY COLUMNS AND ELIMINATION OF BLANK SPACES TO ENSURE DATA INTEGRITY AND CONSISTENCY.

2

THE CLEANED DATASET IS THOROUGHLY EXAMINED TO IDENTIFY ANY ISSUES OR PROBLEMS THAT MAY REQUIRE RESOLUTION.

3

IN ORDER TO ACCOMPLISH THE GIVEN TASKS, A VARIETY OF ADVANCED EXCEL FORMULAS ARE EMPLOYED.

4

THE FORMULAS FACILITATE THE EXTRACTION OF DESIRED RESULTS, LEVERAGING THE CAPABILITIES OF EXCEL TO ANALYZE AND MANIPULATE THE DATA EFFECTIVELY.

5

BY APPLYING THESE TECHNIQUES, THE PROJECT AIMS TO ACHIEVE ACCURATE AND RELIABLE OUTCOMES BASED ON THE PROVIDED TASKS AND OBJECTIVES.

TECH-STACK USED

MICROSOFT
EXCEL

INSIGHTS

- BY SUCCESSFULLY COMPLETING THIS PROJECT, I HAVE ACQUIRED VALUABLE EXPERTISE IN THE ANALYSIS OF DATASETS AND THE COMPREHENSION OF INTERRELATIONSHIPS AMONG VARIOUS COLUMNS WITHIN A DATABASE.

- I HAVE DEVELOPED A KEEN ABILITY TO IDENTIFY AND RESOLVE ISSUES OR ANOMALIES THAT MAY BE PRESENT IN THE DATASET, ALLOWING ME TO DELVE DEEPER INTO THEIR UNDERLYING CAUSES. FURTHERMORE, I HAVE CULTIVATED PROFICIENCY IN UTILIZING AN ARRAY OF ADVANCED EXCEL FEATURES, GREATLY ENHANCING MY DATA ANALYSIS CAPABILITIES.

- THESE FEATURES HAVE EMPOWERED ME TO PERFORM INTRICATE CALCULATIONS, MANIPULATE DATA WITH PRECISION, AND EXTRACT PROFOUND INSIGHTS FROM THE DATASET AT HAND. THROUGH THE REFINEMENT OF THESE SKILLS, I AM NOW WELL-EQUIPPED TO APPROACH FUTURE DATA ANALYSIS TASKS WITH INCREASED EFFECTIVENESS AND EXTRACT VALUABLE INFORMATION FROM DATASETS IN A MORE COMPREHENSIVE MANNER.

ROOT CAUSE ANALYSIS

PROBLEM STATEMENT

THE IMDB ANALYSIS INDICATES A SIGNIFICANT DECREASE IN THE IMDB SCORE FOR A PARTICULAR MOVIE. LET'S PERFORM A ROOT CAUSE ANALYSIS USING THE "DIRECTOR_NAME," "NUM_CRITIC_FOR_REVIEWS," "DURATION," "GROSS," AND "GENRES" COLUMNS TO IDENTIFY THE UNDERLYING REASONS.

FIVE WHYS ROOT CAUSE ANALYSIS

Q:WHY DID THE IMDB SCORE OF THE MOVIE DECREASE?

A:BECAUSE THE MOVIE RECEIVED A LOWER RATING FROM BOTH CRITICS AND USERS.

Q:WHY DID THE MOVIE RECEIVE A LOWER RATING FROM CRITICS AND USERS?

A:BECAUSE THE MOVIE'S DURATION WAS TOO LONG, LEADING TO AUDIENCE FATIGUE AND DECREASED ENGAGEMENT.

Q:WHY WAS THE MOVIE'S DURATION TOO LONG?

A:BECAUSE THE DIRECTOR, IDENTIFIED IN THE "DIRECTOR_NAME" COLUMN, FAILED TO EDIT THE MOVIE EFFECTIVELY AND REMOVE UNNECESSARY SCENES.

FIVE WHYS ROOT CAUSE ANALYSIS

Q:WHY DID THE DIRECTOR FAIL TO EDIT THE MOVIE EFFECTIVELY?

A:BECAUSE THE DIRECTOR PRIORITIZED MAINTAINING THE ARTISTIC INTEGRITY OF THE FILM OVER PACING AND AUDIENCE SATISFACTION.

Q:WHY DID THE DIRECTOR PRIORITIZE ARTISTIC INTEGRITY OVER PACING AND AUDIENCE SATISFACTION?

A:BECAUSE THE DIRECTOR HAD A STRONG ARTISTIC VISION FOR THE MOVIE AND BELIEVED IT WAS NECESSARY TO CONVEY THE INTENDED MESSAGE, REGARDLESS OF AUDIENCE PREFERENCES.

ROOT CAUSE ANALYSIS CONCLUSION

THE DECREASE IN IMDB SCORE CAN BE ATTRIBUTED TO THE DIRECTOR'S DECISION TO PRIORITIZE ARTISTIC INTEGRITY OVER PACING AND AUDIENCE SATISFACTION. THE MOVIE'S DURATION BECAME TOO LONG, LEADING TO AUDIENCE FATIGUE AND DECREASED ENGAGEMENT, RESULTING IN LOWER RATINGS FROM BOTH CRITICS AND USERS.

SOLUTIONS

A) CLEANING THE DATA

NUMBER OF NULL VALUES	1677
NUMBER OF DUPLICATES REMOVED	47
NUMBER OF DROPPED COLUMNS	14

COLUMNS DROPPED

COLOR, DIRECTOR_FACEBOOK_LIKES, ACTOR_3_FACEBOOK_LIKES, ACTOR_2_NAME, ACTOR_1_FACEBOOK_LIKES, CAST_TOTAL_FACEBOOK_LIKES, ACTOR_3_NAME, FACENUMBER_IN_POSTS, PLOT_KEYWORDS, MOVIE_IMDB_LINK, CONTENT_RATING, ACTOR_2_FACEBOOK_LIKES, ASPECT_RATIO, MOVIE_FACEBOOK_LIKES

DRIVE LINK

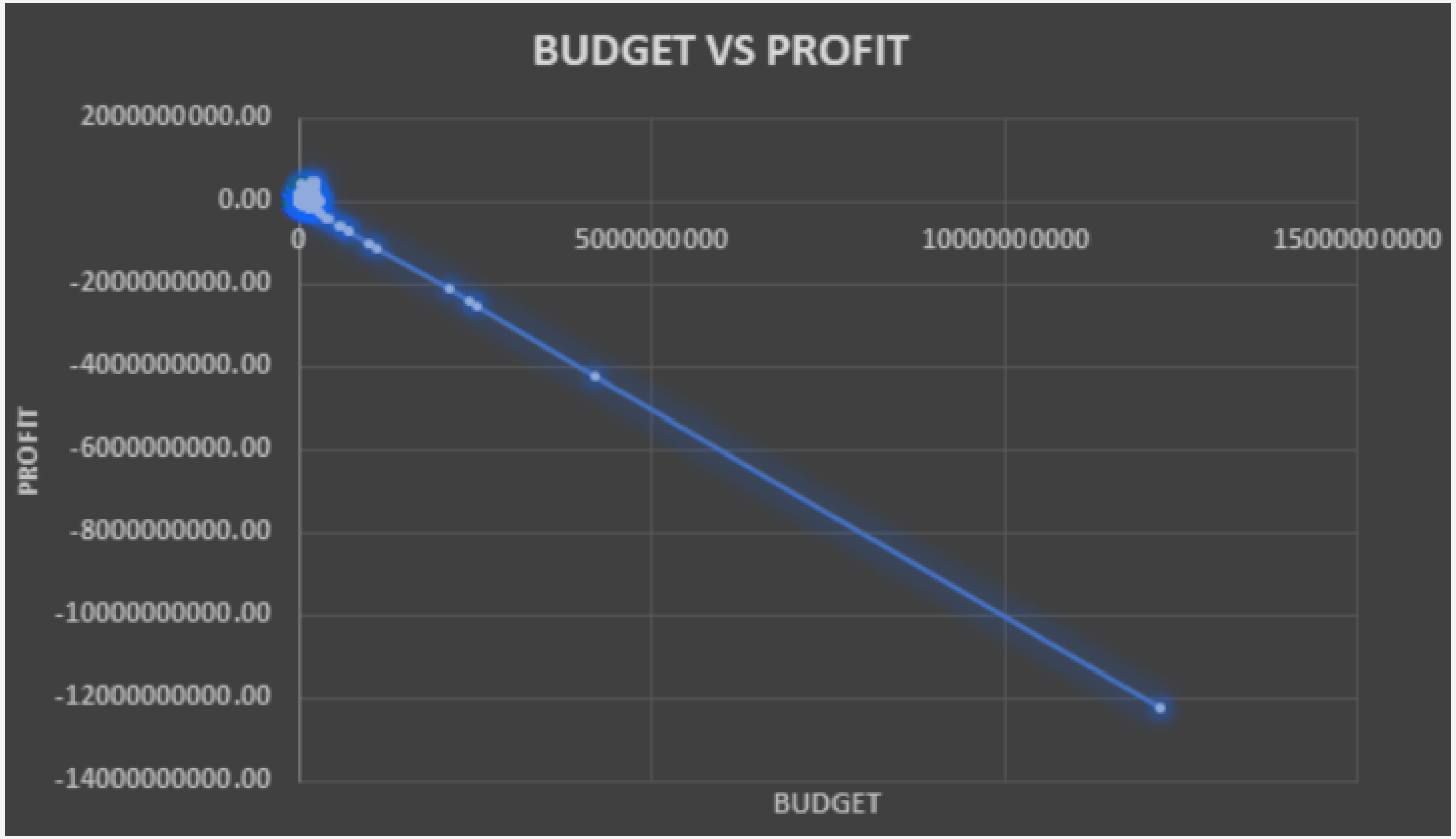
https://drive.google.com/file/d/13OMTfheIJMEzT8TniRBiMqAYwO3LMaVY/view?usp=share_link

B) MOVIES WITH HIGHEST PROFIT

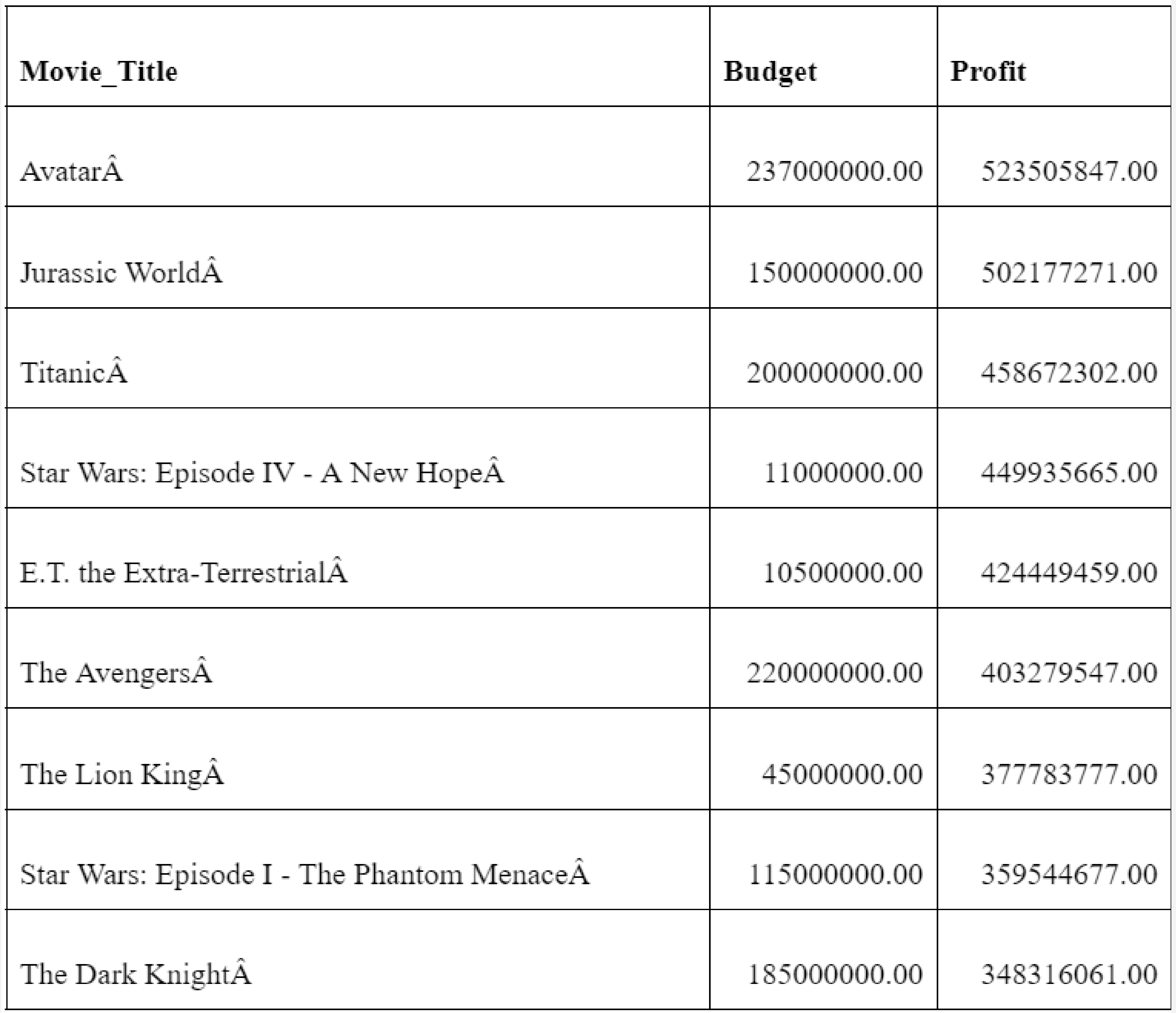
SCATTER CHART FOR FINDING OUTLIERS

OUTLIERS

(2127519898,-2127109510)
(2400000000,-2397701809)
(2500000000,-2499804112)
(4200000000,-4199788333)
(12215500000,-12213298588)



TOP 10 HIGHEST PROFITABLE MOVIES

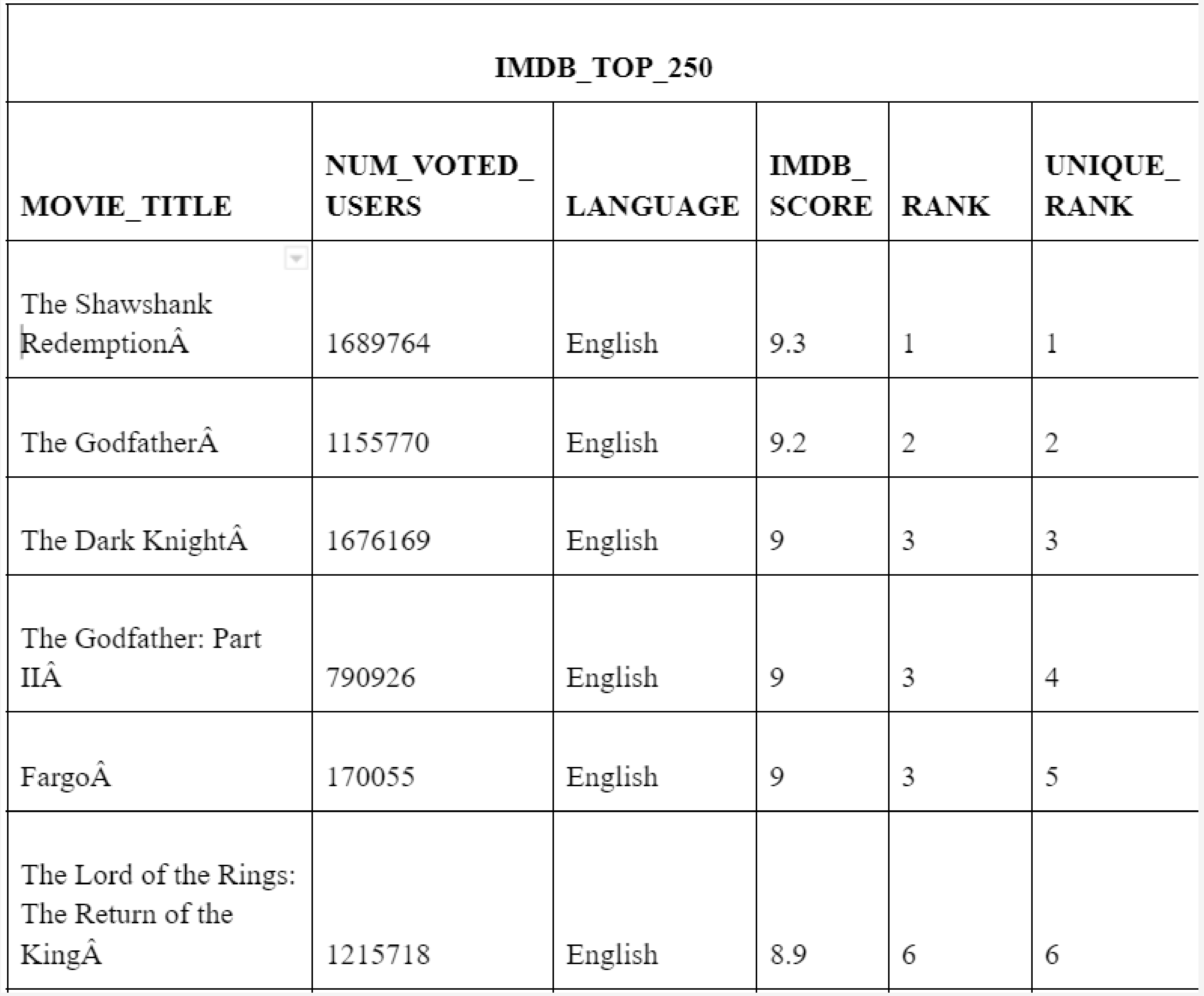


Movie_Title	Budget	Profit
AvatarÂ	237000000.00	523505847.00
Jurassic WorldÂ	150000000.00	502177271.00
TitanicÂ	200000000.00	458672302.00
Star Wars: Episode IV - A New HopeÂ	11000000.00	449935665.00
E.T. the Extra-TerrestrialÂ	10500000.00	424449459.00
The AvengersÂ	220000000.00	403279547.00
The Lion KingÂ	45000000.00	377783777.00
Star Wars: Episode I - The Phantom MenaceÂ	115000000.00	359544677.00
The Dark KnightÂ	185000000.00	348316061.00

THE MOVIE **AVATARÂ**
HAS THE **HIGHEST PROFIT**

C) TOP 250 MOVIES LIST

SAMPLE OUTPUT FOR TOP 250 MOVIES



IMDB_TOP_250

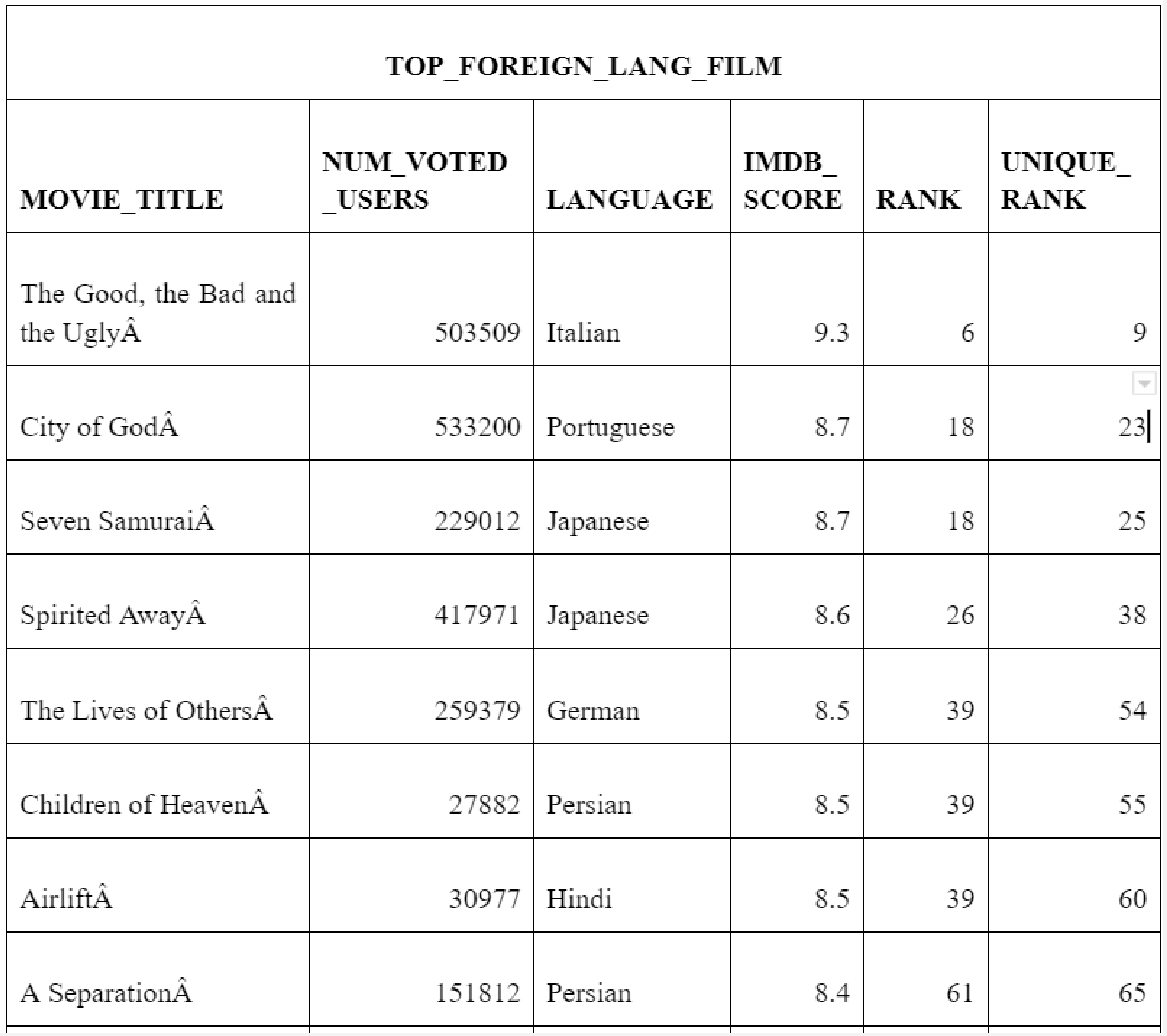
MOVIE_TITLE	NUM_VOTED_ USERS	LANGUAGE	IMDB_ SCORE	RANK	UNIQUE_ RANK
The Shawshank RedemptionÂ	1689764	English	9.3	1	1
The GodfatherÂ	1155770	English	9.2	2	2
The Dark KnightÂ	1676169	English	9	3	3
The Godfather: Part IIÂ	790926	English	9	3	4
FargoÂ	170055	English	9	3	5
The Lord of the Rings: The Return of the KingÂ	1215718	English	8.9	6	6

DRIVE LINK FOR ACCESSING FULL OUTPUT

https://drive.google.com/file/d/1tjJFNQ1mRcWSpOUNBBKo7VDi232IGHO9/view?usp=share_link

C) TOP 250 MOVIES LIST

SAMPLE OUTPUT FOR TOP FOREIGN LANGUAGE MOVIES



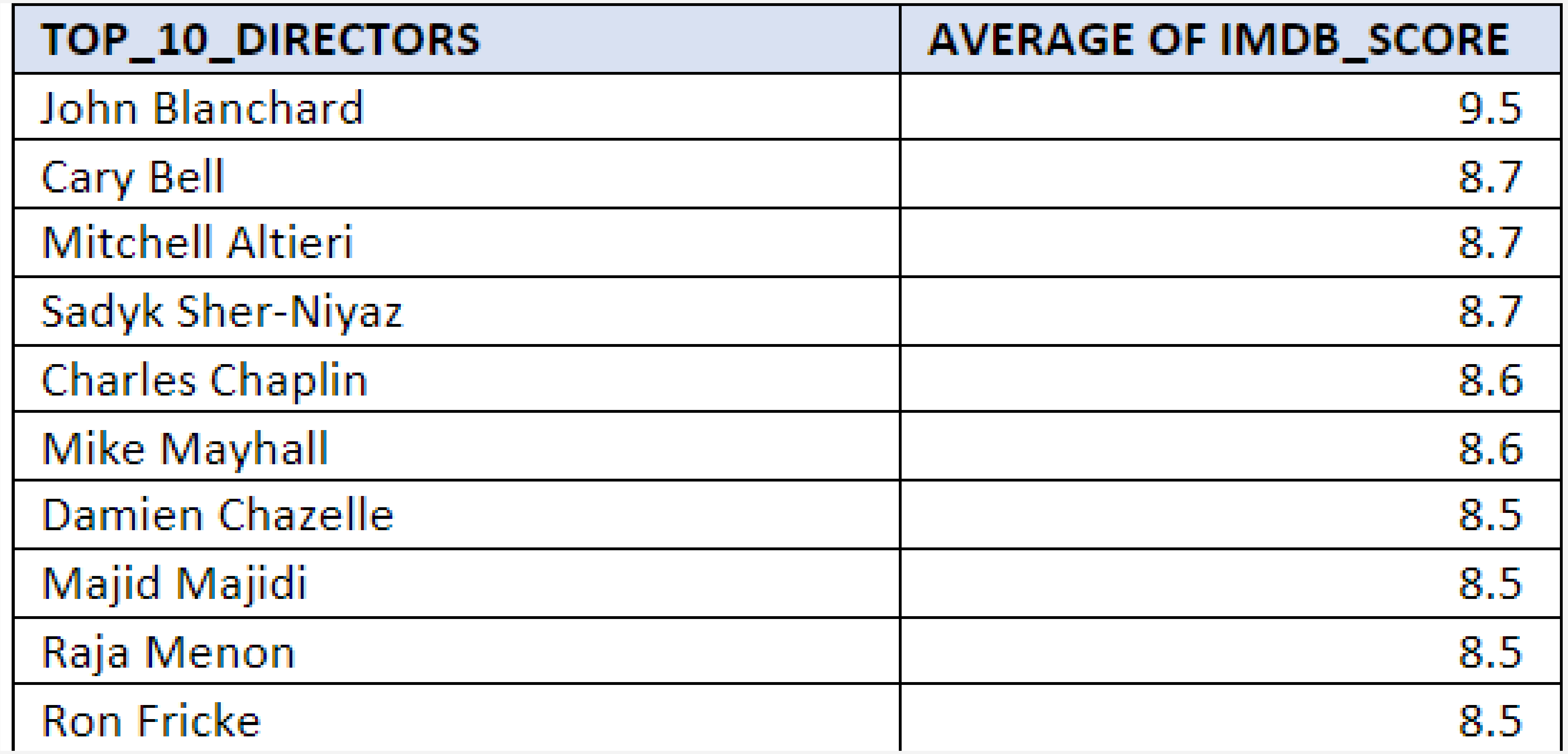
TOP_FOREIGN_LANG_FILM

MOVIE_TITLE	NUM_VOTED_USERS	LANGUAGE	IMDB_SCORE	RANK	UNIQUE_RANK
The Good, the Bad and the UglyÂ	503509	Italian	9.3	6	9
City of GodÂ	533200	Portuguese	8.7	18	23
Seven SamuraiÂ	229012	Japanese	8.7	18	25
Spirited AwayÂ	417971	Japanese	8.6	26	38
The Lives of OthersÂ	259379	German	8.5	39	54
Children of HeavenÂ	27882	Persian	8.5	39	55
AirliftÂ	30977	Hindi	8.5	39	60
A SeparationÂ	151812	Persian	8.4	61	65

DRIVE LINK FOR ACCESSING FULL OUTPUT

https://drive.google.com/file/d/1L4p7oKIV87gVvH-FyWM2YLYePnb637yR/view?usp=share_link

D) BEST DIRECTORS



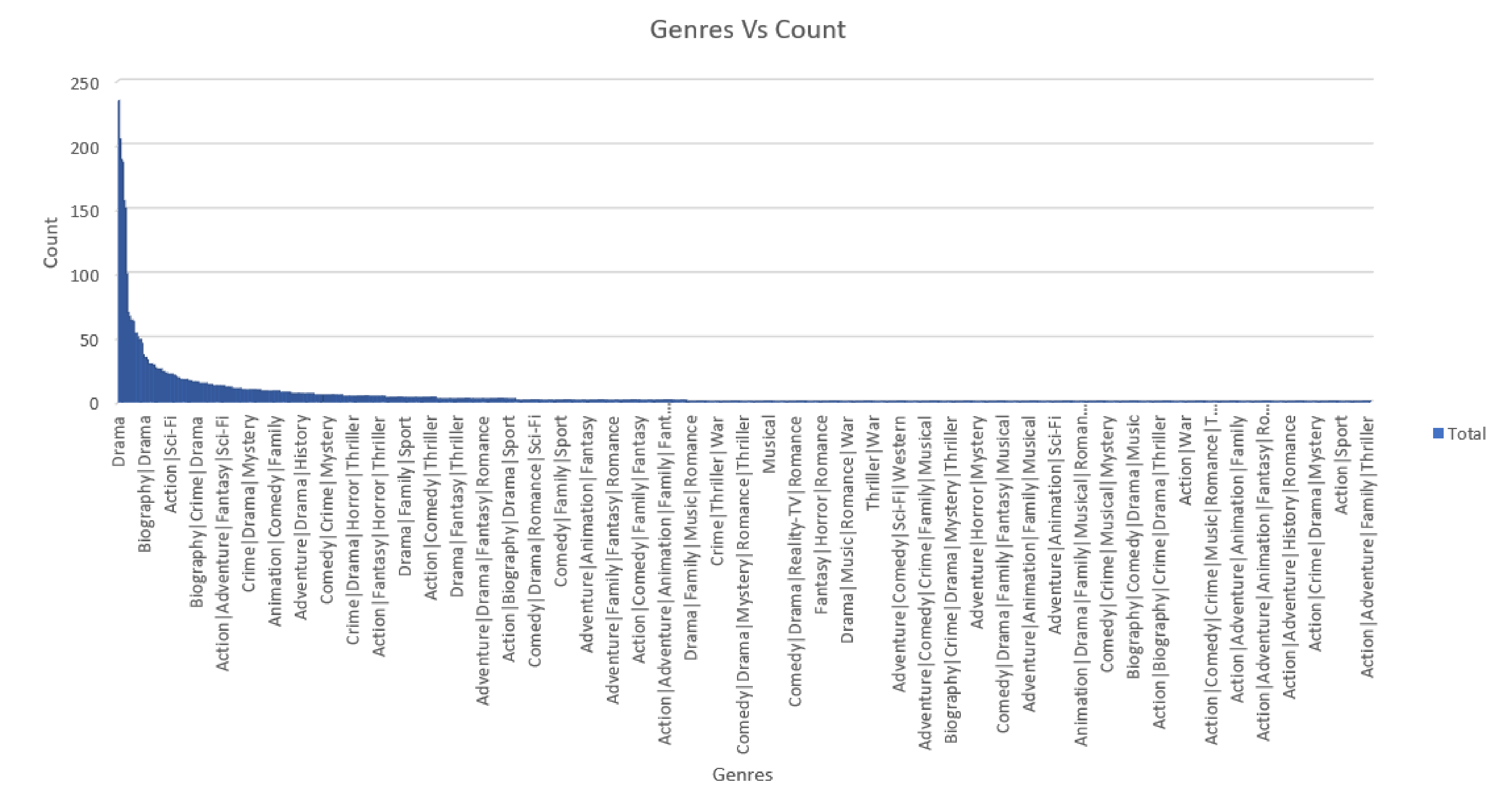
TOP_10_DIRECTORS	AVERAGE OF IMDB_SCORE
John Blanchard	9.5
Cary Bell	8.7
Mitchell Altieri	8.7
Sadyk Sher-Niyaz	8.7
Charles Chaplin	8.6
Mike Mayhall	8.6
Damien Chazelle	8.5
Majid Majidi	8.5
Raja Menon	8.5
Ron Fricke	8.5

LINK TO ACCESS

https://docs.google.com/spreadsheets/d/1U-LMyi42SJzPdtpMRONNrNf9FpAodWhy/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

THE BEST DIRECTOR IS **JOHN BLANCHARD.**

E) POPULAR GENRES



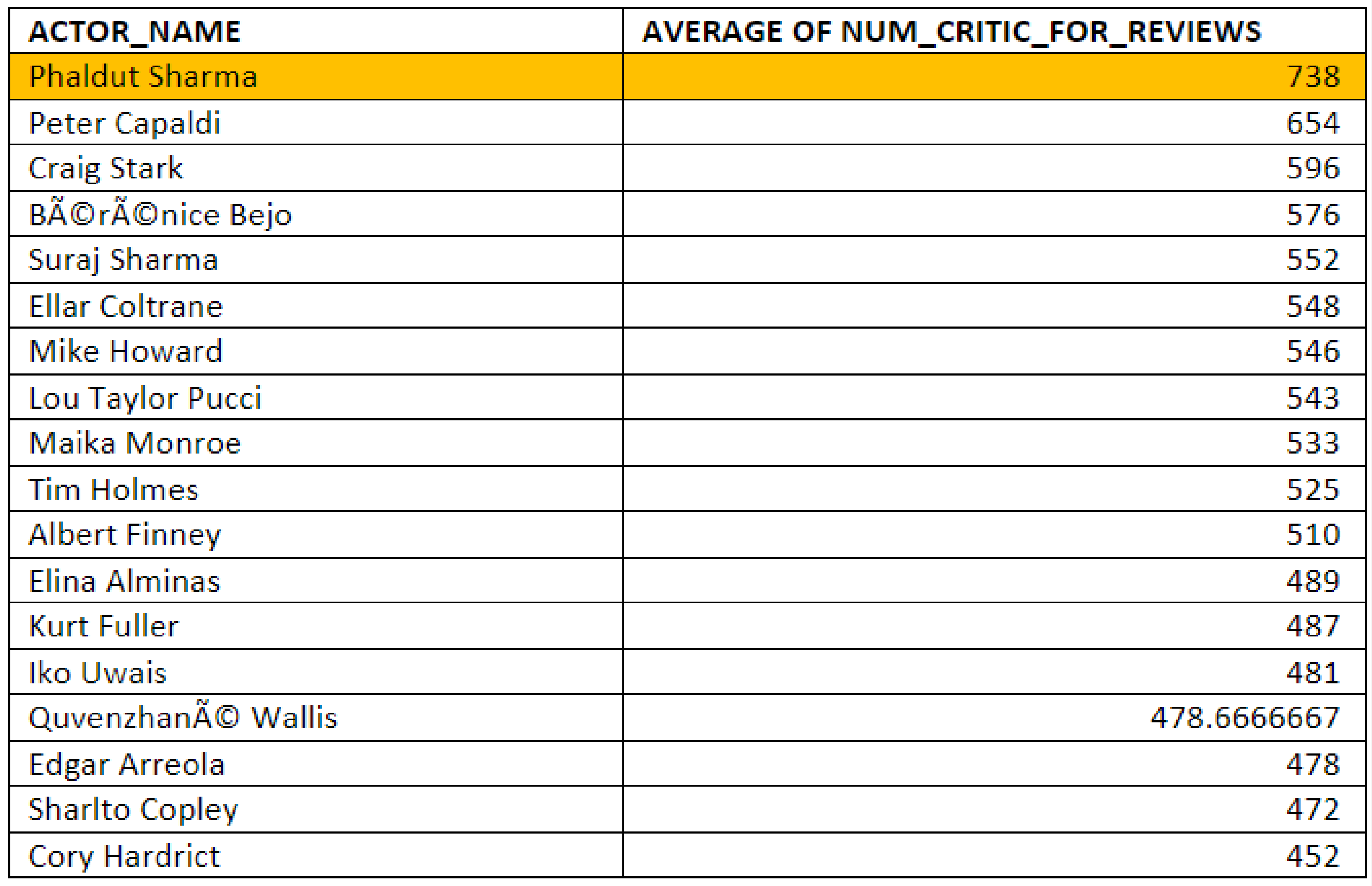
LINK TO ACCESS

https://docs.google.com/spreadsheets/d/1189PXgdpQWI_PCNjOXxZhcE1qhh1nE5p/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

THE MOST POPULAR GENRES IS **DRAMA.**

F) CHARTS

ACTORS WITH AVERAGE OF NUMBER OF CRITIC FOR REVIEWS



ACTOR_NAME	AVERAGE OF NUM_CRITIC_FOR_REVIEWS
Phaldut Sharma	738
Peter Capaldi	654
Craig Stark	596
BÃ©rÃ©nice Bejo	576
Suraj Sharma	552
Ellar Coltrane	548
Mike Howard	546
Lou Taylor Pucci	543
Maika Monroe	533
Tim Holmes	525
Albert Finney	510
Elina Alminas	489
Kurt Fuller	487
Iko Uwais	481
QuvenzhanÃ© Wallis	478.6666667
Edgar Arreola	478
Sharlto Copley	472
Cory Hardrict	452

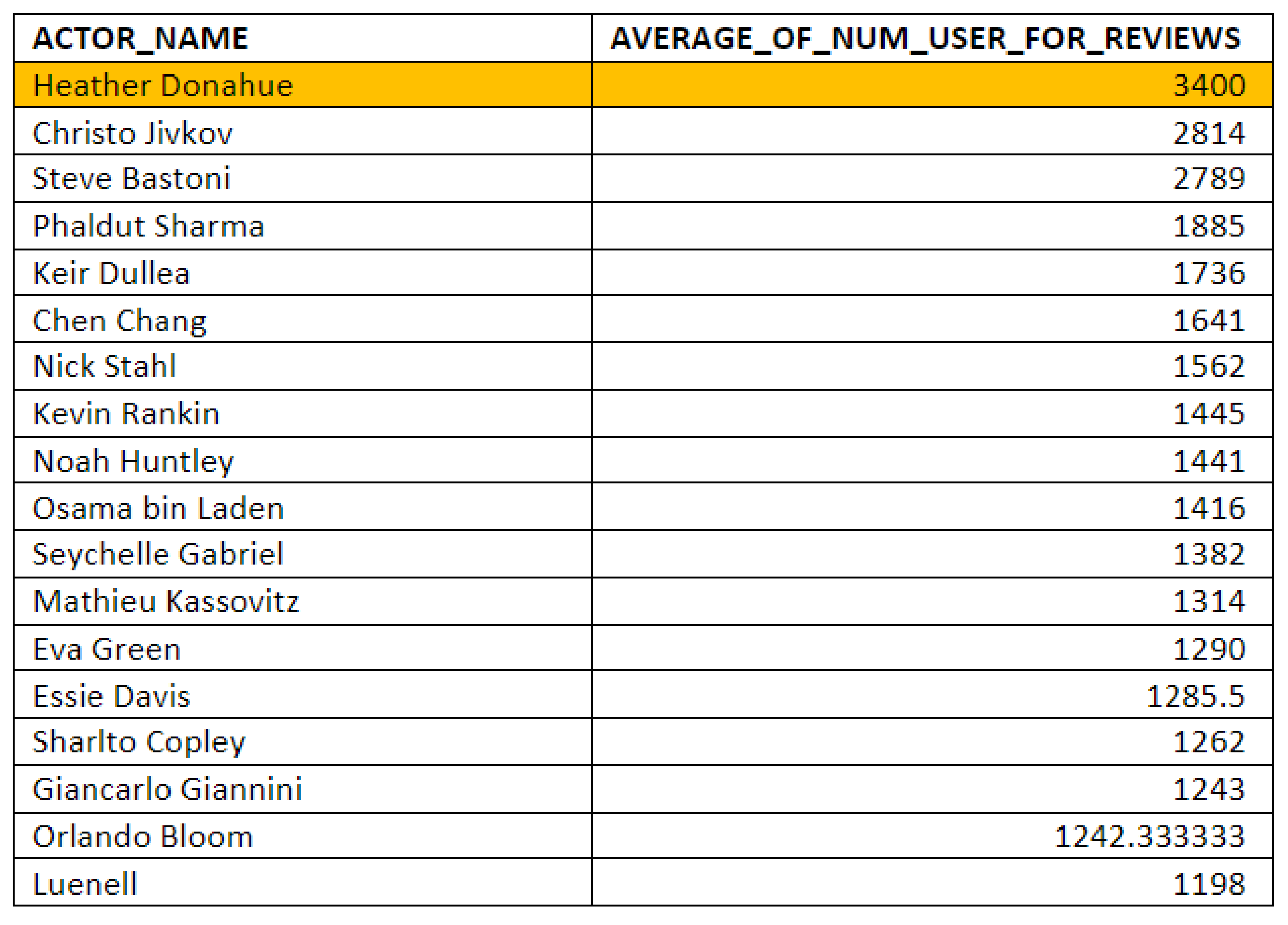
THE ACTOR **PHALDUT SHARMA** IS THE CRITIC FAVORITE.

LINK TO ACCESS

https://docs.google.com/spreadsheets/d/1rTrhrWt04hcuNKf3VflKxFiKbrFDFgmz/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

F) CHARTS

ACTORS WITH AVERAGE OF NUMBER OF USER FOR REVIEWS



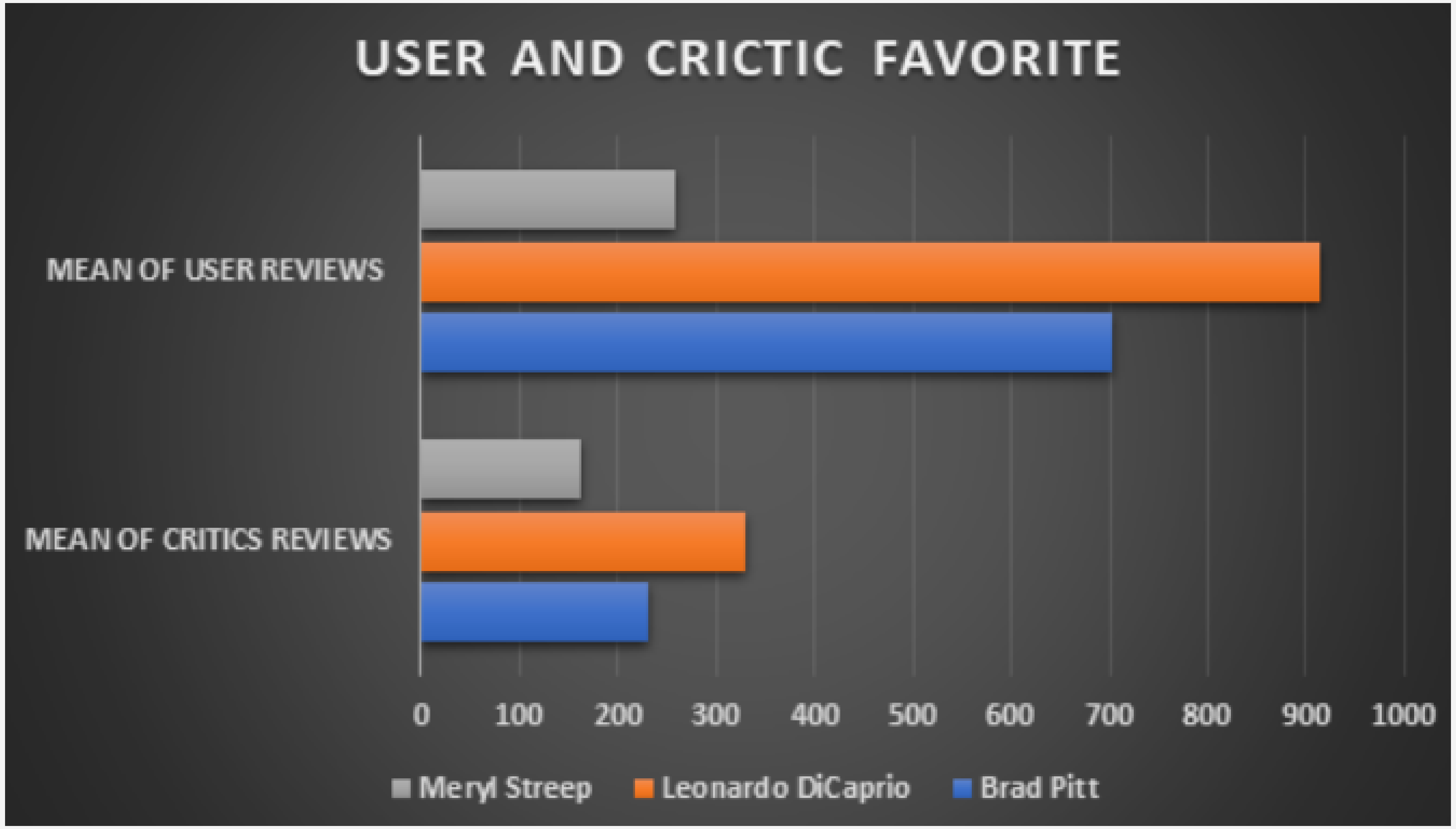
ACTOR_NAME	AVERAGE_OF_NUM_USER_FOR_REVIEWS
Heather Donahue	3400
Christo Jivkov	2814
Steve Bastoni	2789
Phaldut Sharma	1885
Keir Dullea	1736
Chen Chang	1641
Nick Stahl	1562
Kevin Rankin	1445
Noah Huntley	1441
Osama bin Laden	1416
Seychelle Gabriel	1382
Mathieu Kassovitz	1314
Eva Green	1290
Essie Davis	1285.5
Sharlto Copley	1262
Giancarlo Giannini	1243
Orlando Bloom	1242.333333
Luenell	1198

THE ACTOR **HEATHER DONAHUE** IS THE USER FAVORITE.

LINK TO ACCESS

https://docs.google.com/spreadsheets/d/1nVp2zZyFDce3TnSgLWCgo_oAmKo6dWVg/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

F) CHARTS



WHILE CONSIDERING MERYL STREEP, LEONARDO DI CAPIRO AND BRAD PITT.

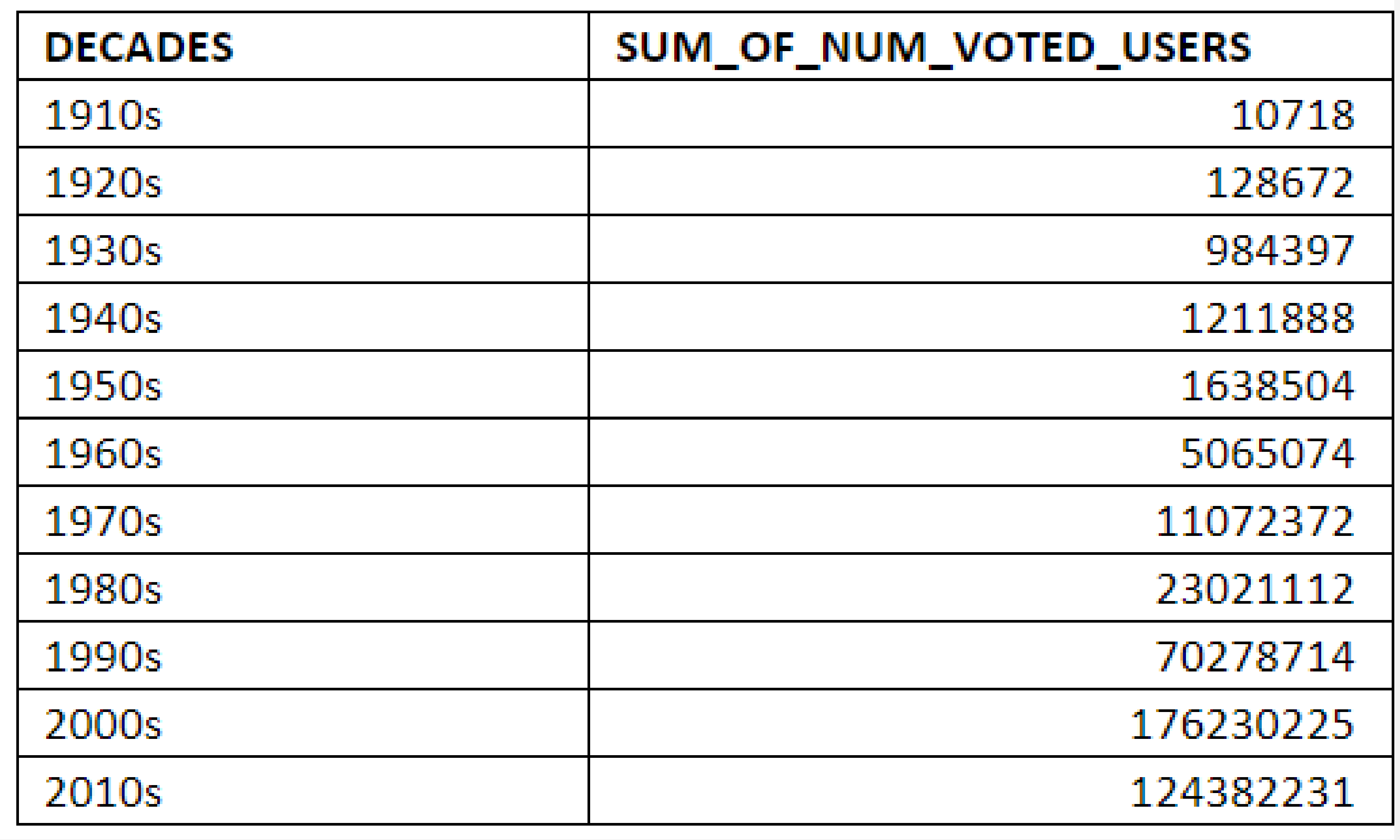
LEONARDO DI CAPIRO IS THE USER AND CRITIC FAVORITE.

LINK TO ACCESS

https://docs.google.com/spreadsheets/d/1nVp2zZyFDce3TnSgLWCgo_oAmKo6dWVg/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

F) CHARTS

DECADES VS TOTAL NUMBER OF VOTED USERS



DECADES	SUM_OF_NUM_VOTED_USERS
1910s	10718
1920s	128672
1930s	984397
1940s	1211888
1950s	1638504
1960s	5065074
1970s	11072372
1980s	23021112
1990s	70278714
2000s	176230225
2010s	124382231

THE DECADE **2000S** HAS THE **MORE** NUMBER OF VOTED USERS.

RESULT

THIS PROJECT ENHANCED MY DATASET ANALYSIS SKILLS AND UNDERSTANDING OF COLUMN INTERRELATIONSHIPS. I CAN NOW IDENTIFY AND ADDRESS DATASET ISSUES AND ANOMALIES, UNCOVERING THEIR ROOT CAUSES. PROFICIENCY IN ADVANCED EXCEL FEATURES IMPROVED MY DATA ANALYSIS CAPABILITIES, ENABLING COMPLEX CALCULATIONS AND EFFECTIVE DATA MANIPULATION. THESE SKILLS EQUIP ME TO TACKLE FUTURE ANALYSIS TASKS EFFICIENTLY AND DERIVE VALUABLE INSIGHTS FROM DATASETS.

LINK TO ACCESS COMPLETE EXCEL SHEET

https://docs.google.com/spreadsheets/d/1wwyspns-xnu8_TyxJ_spOmawhS0morkj/edit?usp=sharing&ouid=111617584332759410517&rtpof=true&sd=true

THANK YOU