IMDB Movie Analysis

FINAL PROJECT-1

Project Description

The aim of this project is to establish a correlation between IMDB ratings and the influencing factors that impact these ratings. Conducting such an analysis holds significant v alue for producers, directors, and inv estors, as it prov ides v aluable insights to facilitate well-informed decisions concerning upcoming projects.

Factors chosen for analysing impact on rating are Genre, Duration, Language, Director, and Budget.

A mov ie with higher rating is deemed successful.

Approach

- Clean the raw data before analysing.
- Remove fields which are not necessary for analysis. Fields related to actor name, facebook likes, color, reviews, imdb link, and country are removed.
- Highlight cells with missing values and incorrect data type.
- Identify outliers for Duration, Budget, and Gross earning fields using Quartile function and calculating IQR, Upper range and Lower range.
- All Influencing factor fields and corresponding IMDB score field are copy pasted in different excel sheets for further cleaning.

Tech-Stack Used

- Mircrosoft Excel web version is used for analysis. Microsoft excel desktop app doesn’t allow saving the file in MacBook unless you have subscription.

- Link of excel file : IMDB movie analysis
- When opening this file in Google Sheets, certain modifications may occur. For a comprehensive view of all charts, please download the file and open it using the Excel desktop application.

Data cleaning summary

Convert

Convert the raw data into Table and remove 122 duplicate rows. Count of total rows after removing duplicate is 4921.

Readjust

Readjust columns with Movie Title being first column.

Count

Count of blank rows for each field after removing duplicate rows:

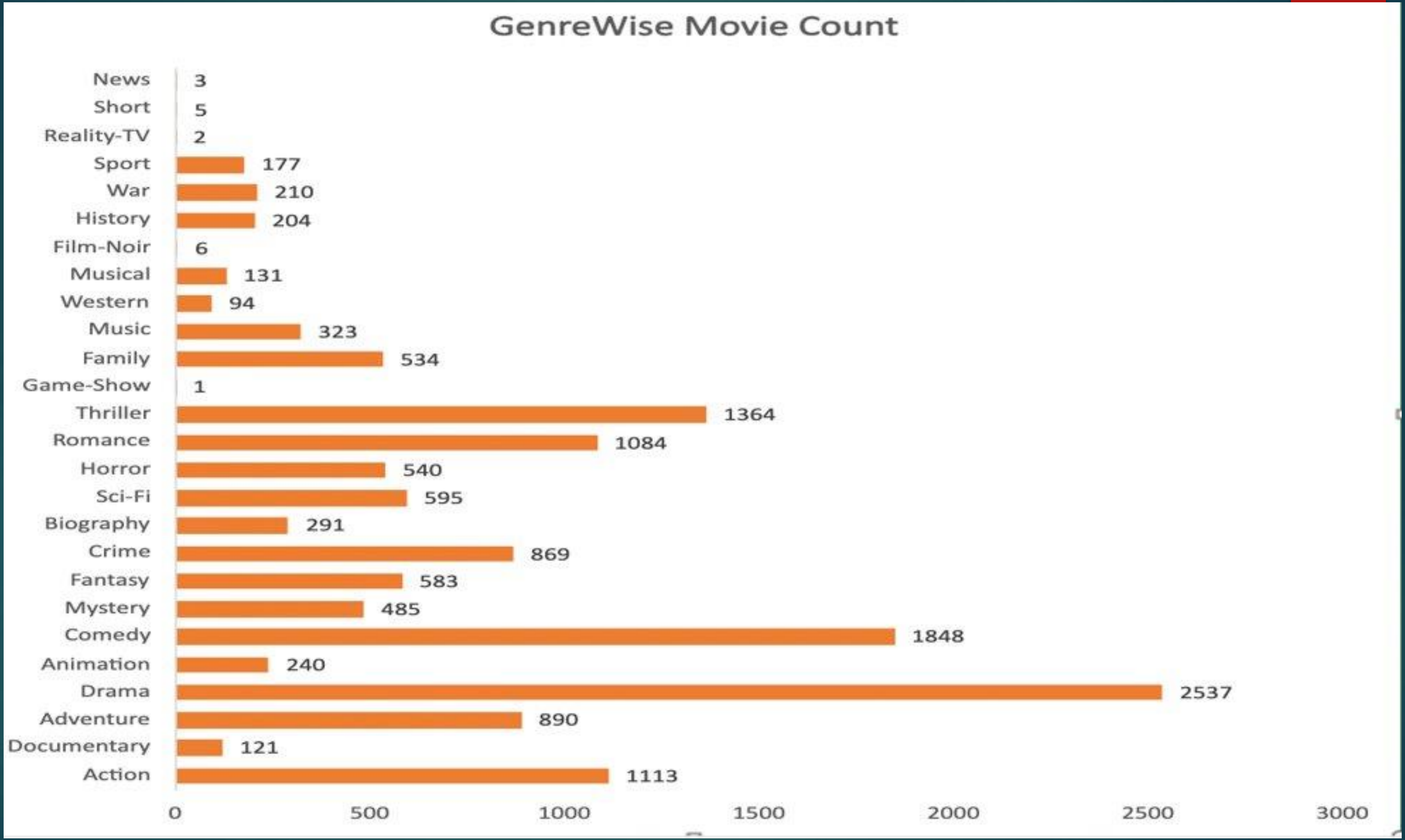
- Movie Title, Genres, imdb_score – 0
- Release Year – 106
- Director Name – 102
- Language – 12
- Duration – 15
- Gross earning - 865
- Budget - 485

Task –A : Movie Genre Analysis- Analyze the distribution of movie genres and their impact on the IMDB score.
Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

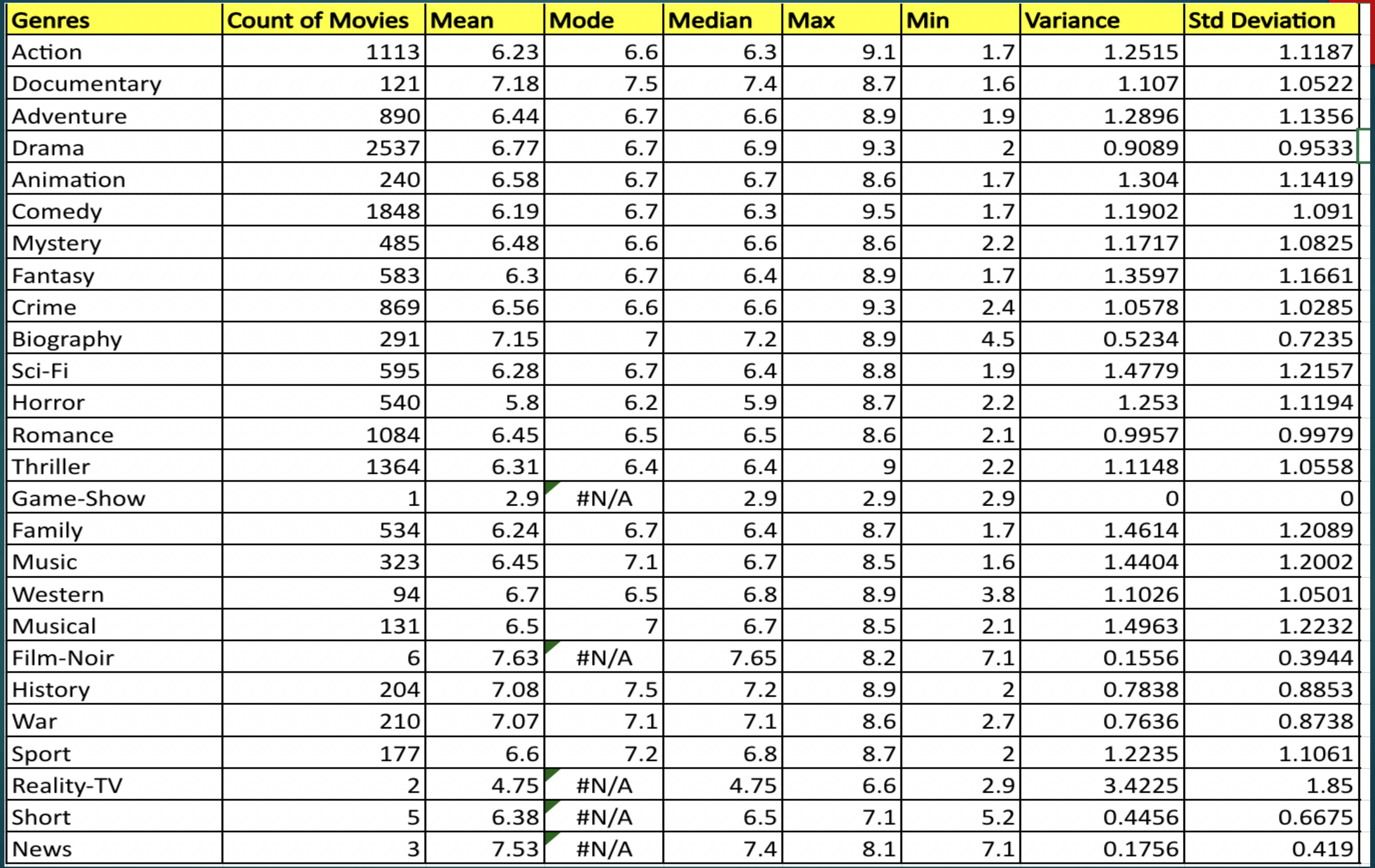
- Formula used to calculate Mode, Median, Variance with a condition is

  ```
  = MODE.SNGL(IF(ISNUMBER(SEARCH(D2, Table4[Genres])), Table4[imdb_score], ""))
  =MEDIAN(IF(ISNUMBER(SEARCH(D2, Table4[Genres])), Table4[imdb_score], ""))
  ```
- Use UNIQUE function with TEXTSPLIT to a list of Unique genres.

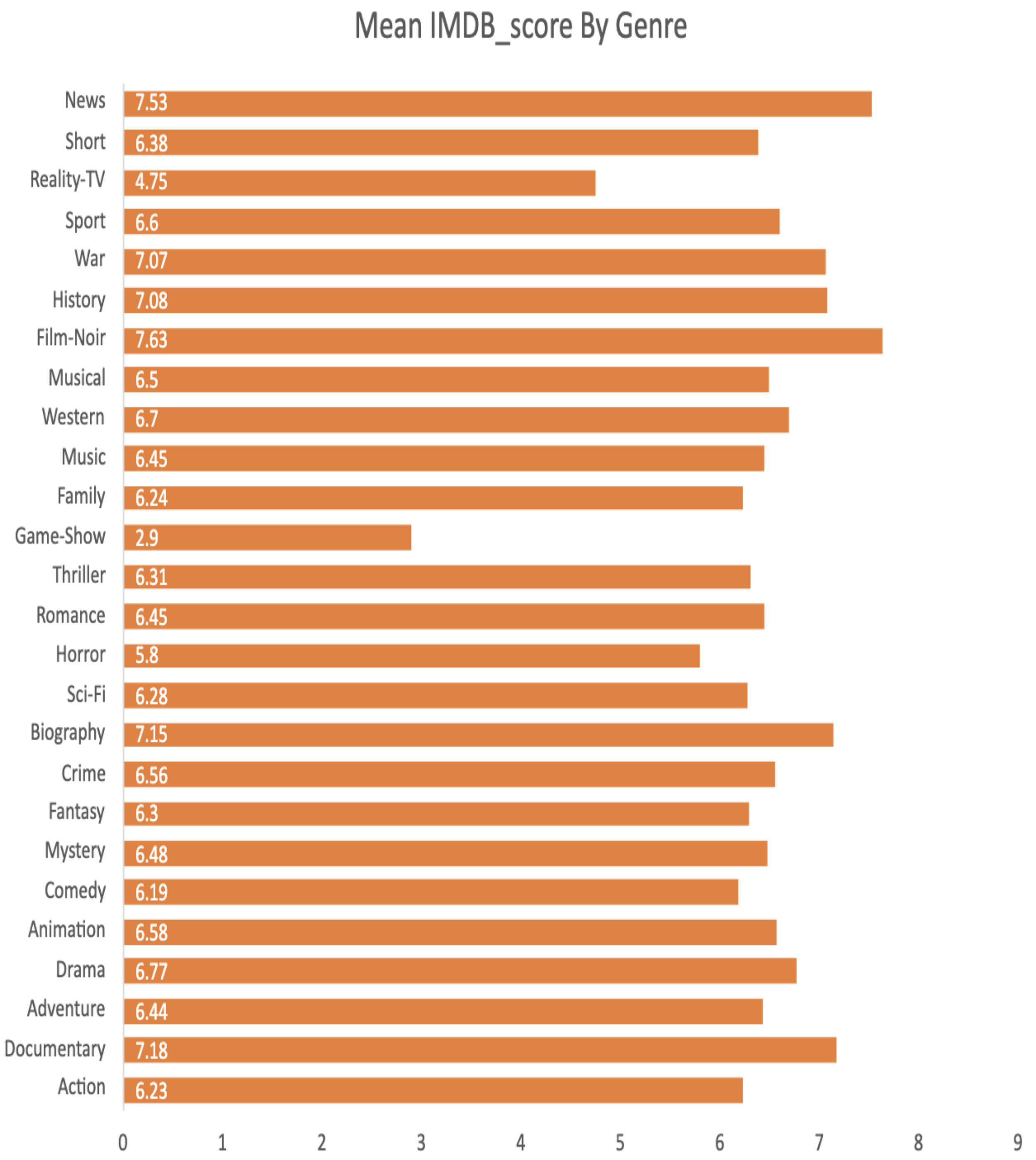
  ```
  = UNIQUE(TEXTSPLIT(UNIQUE(Table4[Genres]),"|"," "))
  ```
- The most common Genres for movies are :
  - Drama - 2537 movies
  - Comedy - 1848 movies
  - Thriller – 1364 movies
  - Action - 1113 movies
  - Romance – 1084 movies


GenreWise Movie Count
News 3
Short 5
Reality-TV 2
Sport 177
War 210
History 204
Film-Noir 6
Musical 131
Western 94
Music 323
Family 534
Game-Show 1
Thriller 1364
Romance 1084
Horror 540
Sci-Fi 595
Biography 291
Crime 869
Fantasy 583
Mystery 485
Comedy 1848
Animation 240
Drama 2537
Adventure 890
Documentary 121
Action 1113
0
500
1000
1500
2000
2500
3000

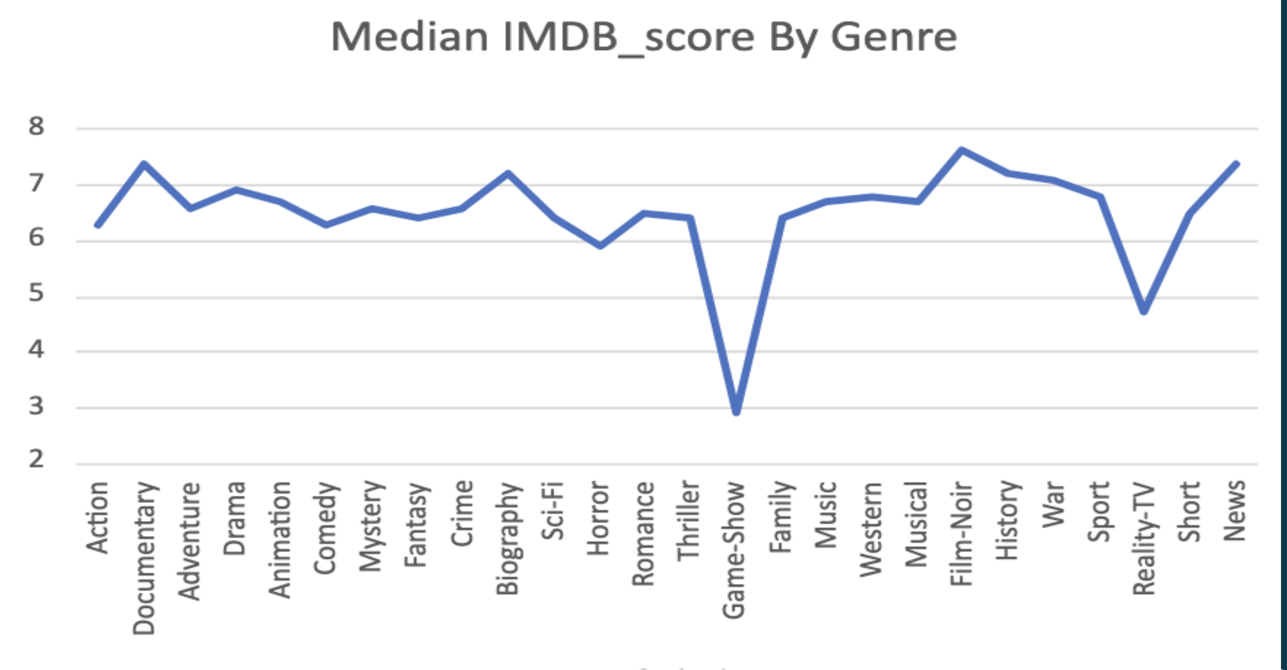
Descriptive Statistics for each genres are :



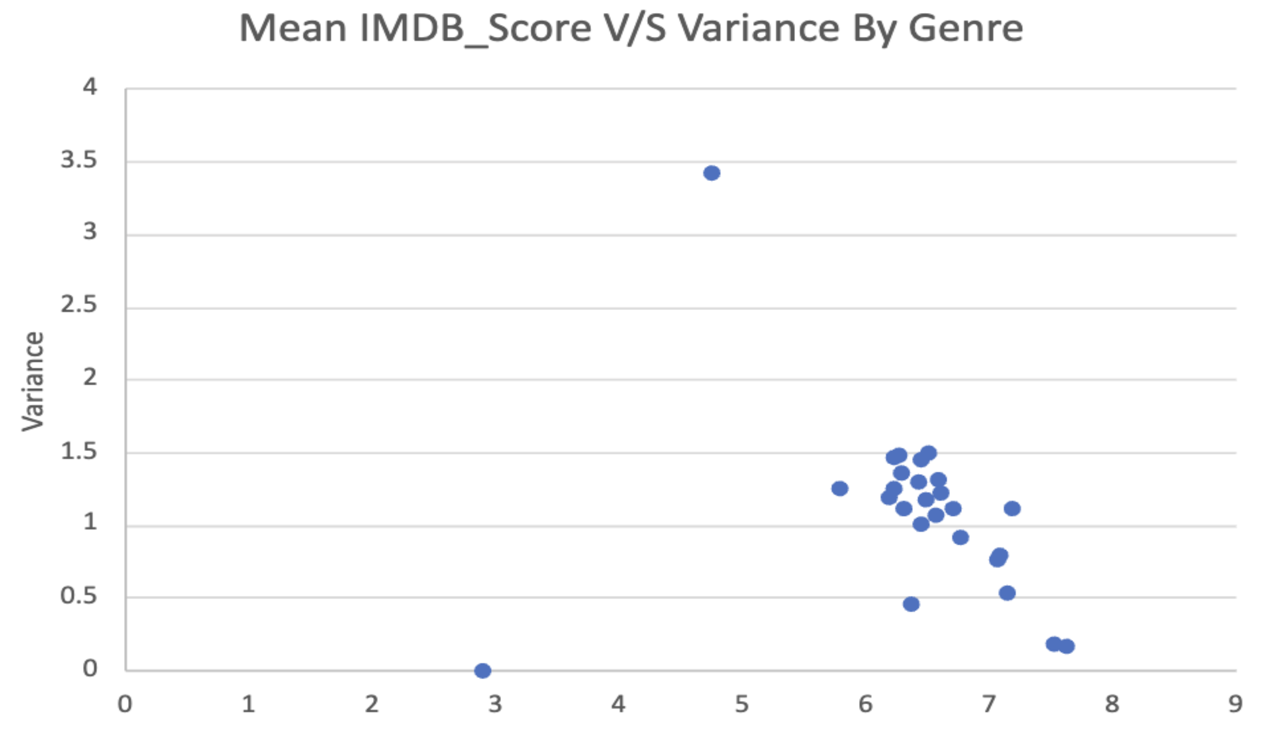
Genres	Count of Movies	Mean	Mode	Median	Max	Min	Variance	Std Deviation
Action	1113	6.23	6.6	6.3	9.1	1.7	1.2515	1.1187
Documentary	121	7.18	7.5	7.4	8.7	1.6	1.107	1.0522
Adventure	890	6.44	6.7	6.6	8.9	1.9	1.2896	1.1356
Drama	2537	6.77	6.7	6.9	9.3	2	0.9089	0.9533
Animation	240	6.58	6.7	6.7	8.6	1.7	1.304	1.1419
Comedy	1848	6.19	6.7	6.3	9.5	1.7	1.1902	1.091
Mystery	485	6.48	6.6	6.6	8.6	2.2	1.1717	1.0825
Fantasy	583	6.3	6.7	6.4	8.9	1.7	1.3597	1.1661
Crime	869	6.56	6.6	6.6	9.3	2.4	1.0578	1.0285
Biography	291	7.15	7	7.2	8.9	4.5	0.5234	0.7235
Sci-Fi	595	6.28	6.7	6.4	8.8	1.9	1.4779	1.2157
Horror	540	5.8	6.2	5.9	8.7	2.2	1.253	1.1194
Romance	1084	6.45	6.5	6.5	8.6	2.1	0.9957	0.9979
Thriller	1364	6.31	6.4	6.4	9	2.2	1.1148	1.0558
Game-Show	1	2.9	#N/A	2.9	2.9	2.9	0	0
Family	534	6.24	6.7	6.4	8.7	1.7	1.4614	1.2089
Music	323	6.45	7.1	6.7	8.5	1.6	1.4404	1.2002
Western	94	6.7	6.5	6.8	8.9	3.8	1.1026	1.0501
Musical	131	6.5	7	6.7	8.5	2.1	1.4963	1.2232
Film-Noir	6	7.63	#N/A	7.65	8.2	7.1	0.1556	0.3944
History	204	7.08	7.5	7.2	8.9	2	0.7838	0.8853
War	210	7.07	7.1	7.1	8.6	2.7	0.7636	0.8738
Sport	177	6.6	7.2	6.8	8.7	2	1.2235	1.1061
Reality-TV	2	4.75	#N/A	4.75	6.6	2.9	3.4225	1.85
Short	5	6.38	#N/A	6.5	7.1	5.2	0.4456	0.6675
News	3	7.53	#N/A	7.4	8.1	7.1	0.1756	0.419



Mean IMDB_score By Genre



Median IMDB_score By Genre



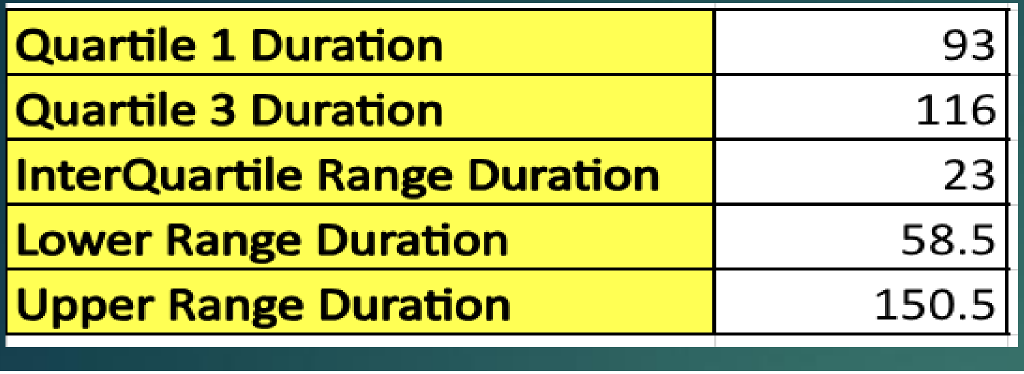
Mean IMDB_Score V/S Variance By Genre

Summary of above chats :

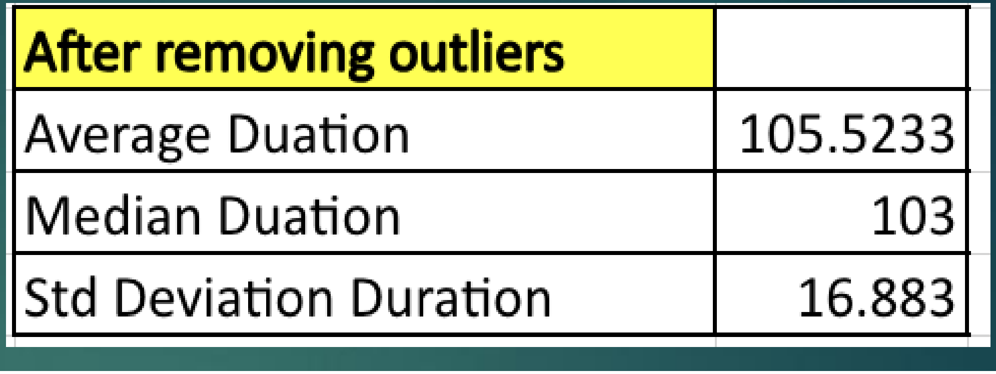
- In descriptive statistics table, MODE values are not applicable for Game-show, Reality-TV, Short & News because of unique imdb scores of movies under these genres. Max scores(>9)are given to a movie with genres like Action, Drama, Crime, and comedy.
- Mean IMDB score v/s Genre chart shows that a movie is more likely to have a score above 7 if it has genres like News, War, Film-Noir, History, Documentry, and Biography. Although, the number of movies are significantly less for these genres.
- Median IMDB score v/s genre chart conveys the same information as mean chart.
- Mean IMDB score v/s Variance scatter plot shows that a movie with higher average score is more likely to have lesser variance.

Task-B : Movie Duration Analysis- Analyze the distribution of movie durations and its impact on the IMDB score.
Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

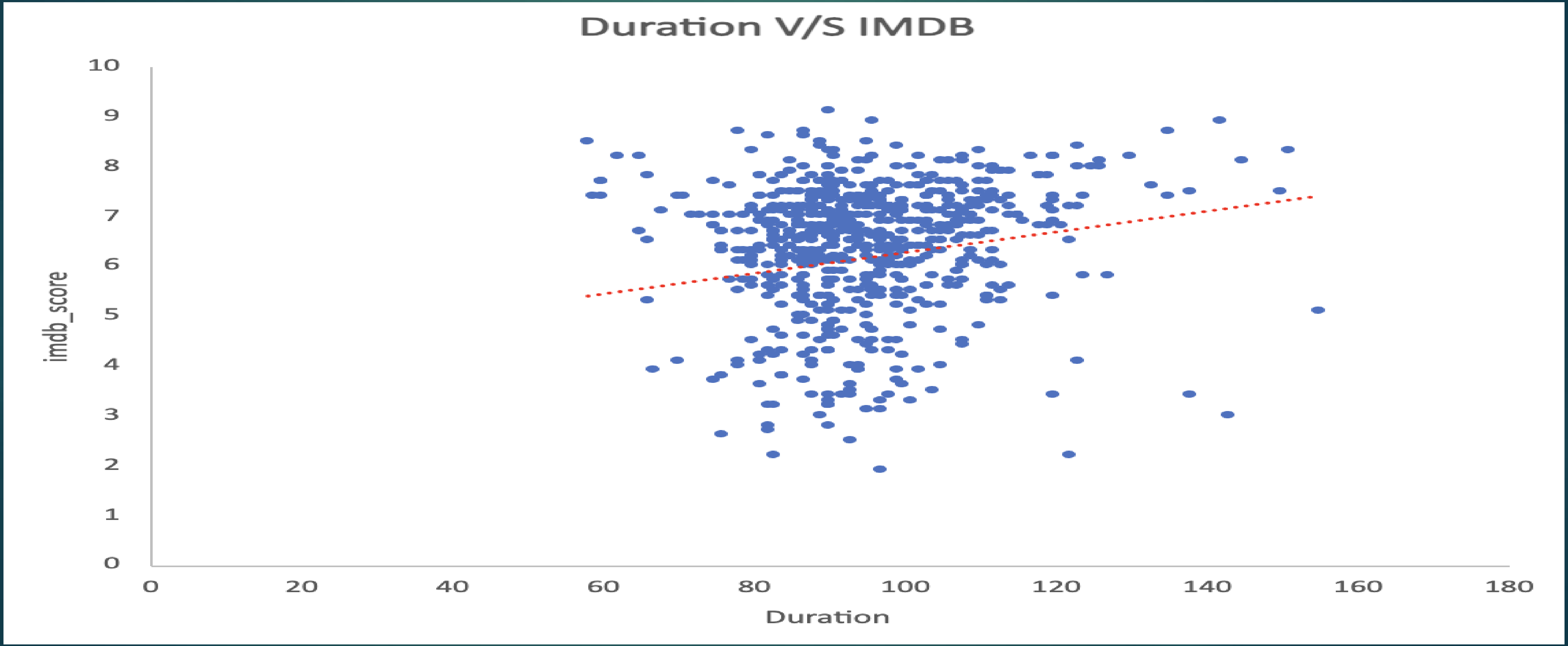
- Copy-paste Duration and IMDB score fields in a separate excel sheet.
- Count of blank cells for the Duration field = 15. Remove all the rows containing blank cells
- Calculate Quartiles using quartile function. to identify the outliers. Count of outliers = 237 cells. Remove the rows containing outliers. Count of Rows after removing outliers = 4670
- Data is ready for analysis now.



Quartile 1 Duration	93
Quartile 3 Duration	116
InterQuartile Range Duration	23
Lower Range Duration	58.5
Upper Range Duration	150.5



After removing outliers	
Average Duation	105.5233
Median Duation	103
Std Deviation Duration	16.883



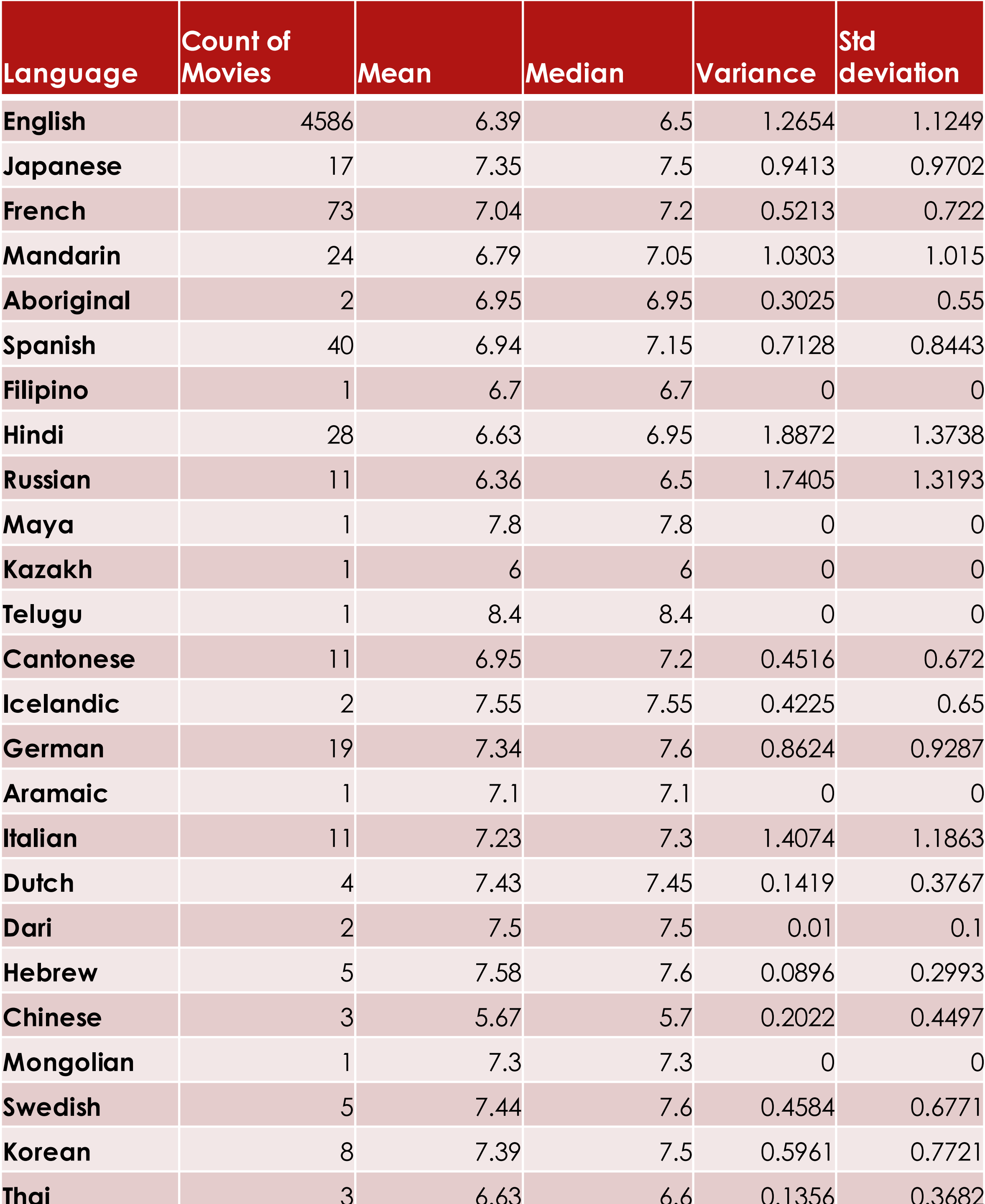
Summary of the above chart :

The above chart is a scatter plot, generated after having removed outliers with Duration > 150.5 and Duration < 58.5

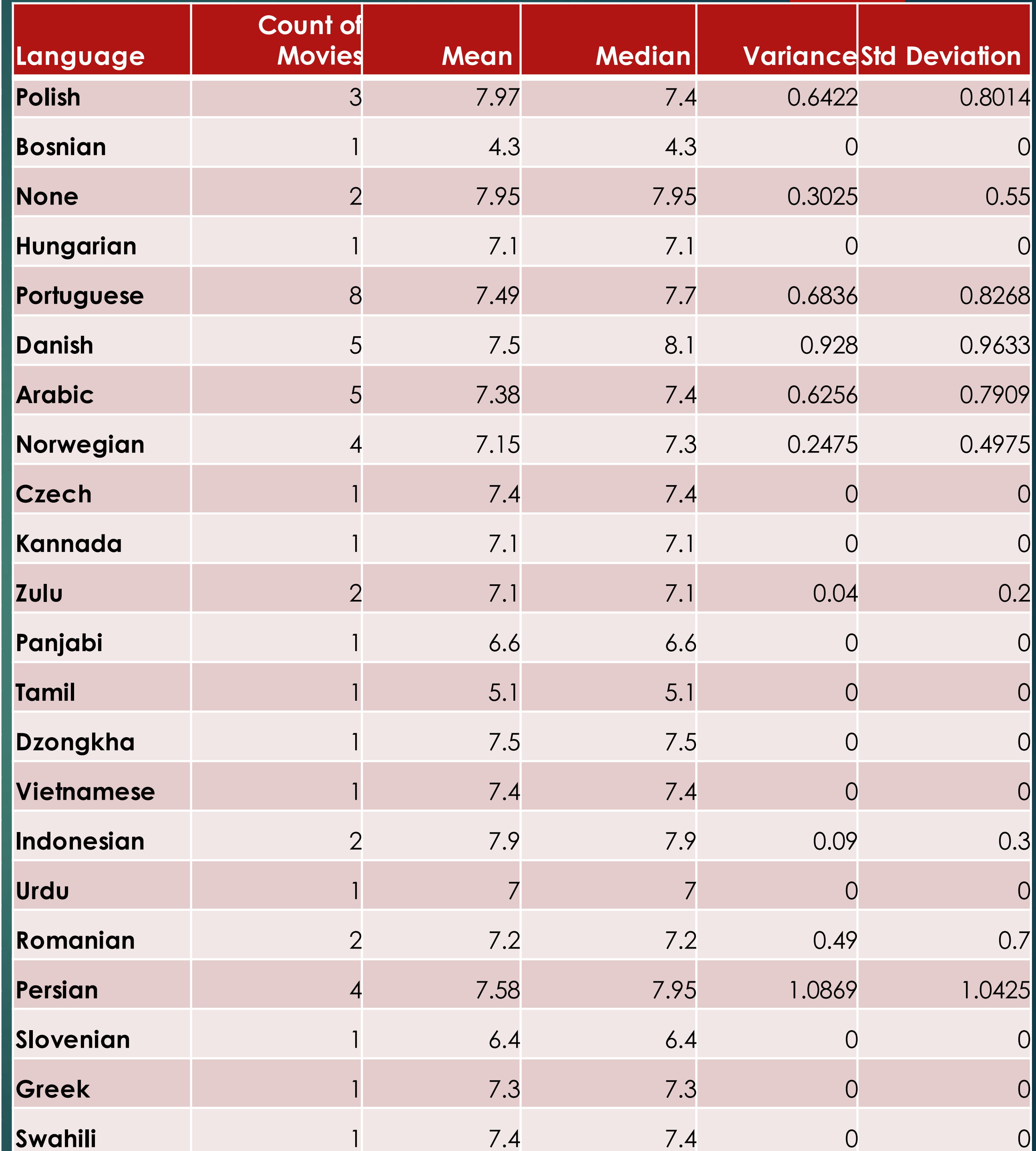
The Trend line in the scatter plot shows that a movie of longer duration is more likely to have a higher imdb score.

Task – C : Language Analysis - Situation: Examine the distribution of movies based on their language.
Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

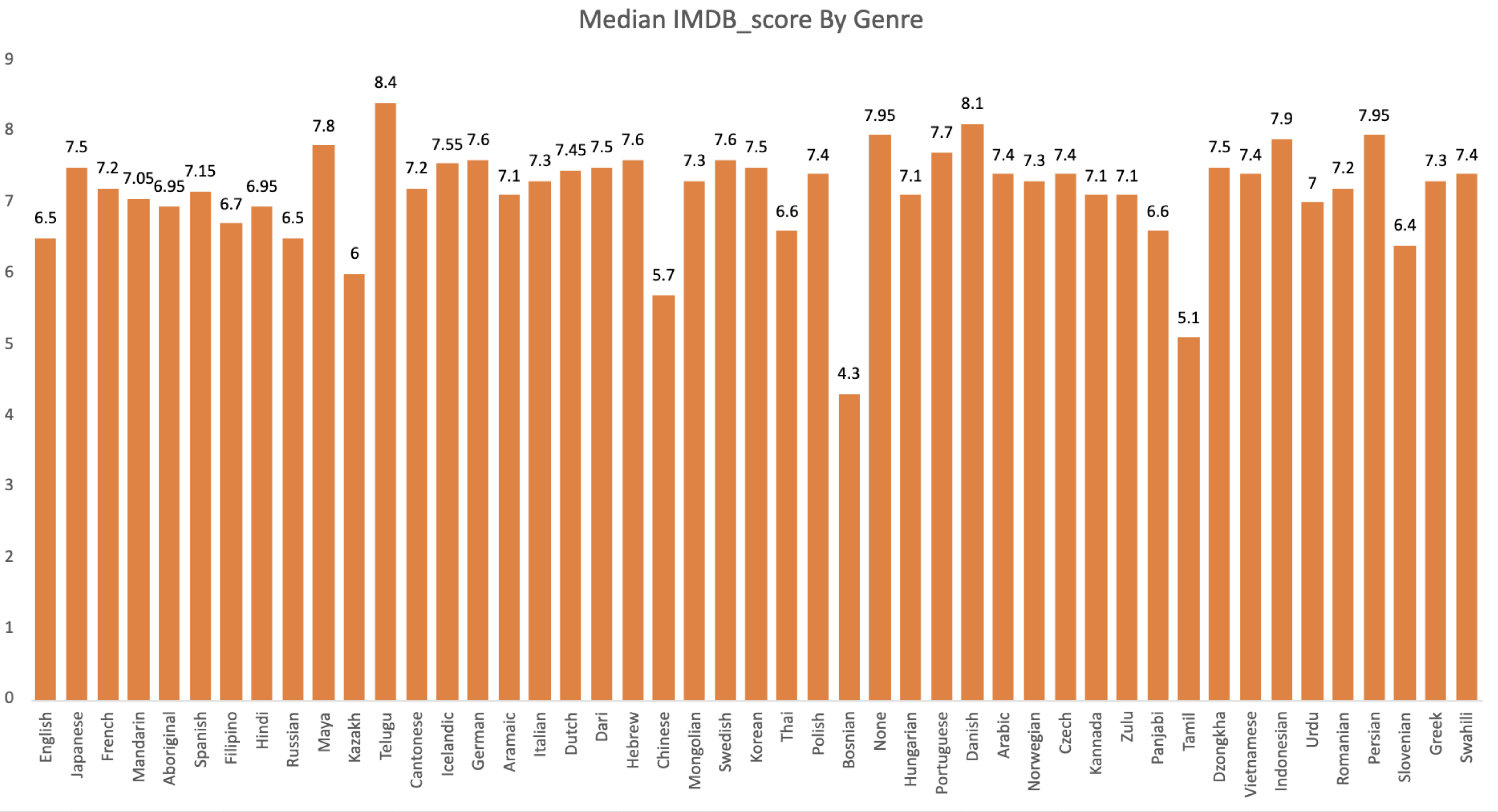
- Use CountBlank function to get the count of blank cells. 12 blank cells are there for language field. Remove the blank cells.
- Count of Movies are removing Blank cells = 4909
- Use UNIQUE function to get the list of all unique languages.
- Use COUNTIF function to get count of movies for each genre.
- Calculate Mean, Median, Variance, and Std deviation to perform descriptive statistics.

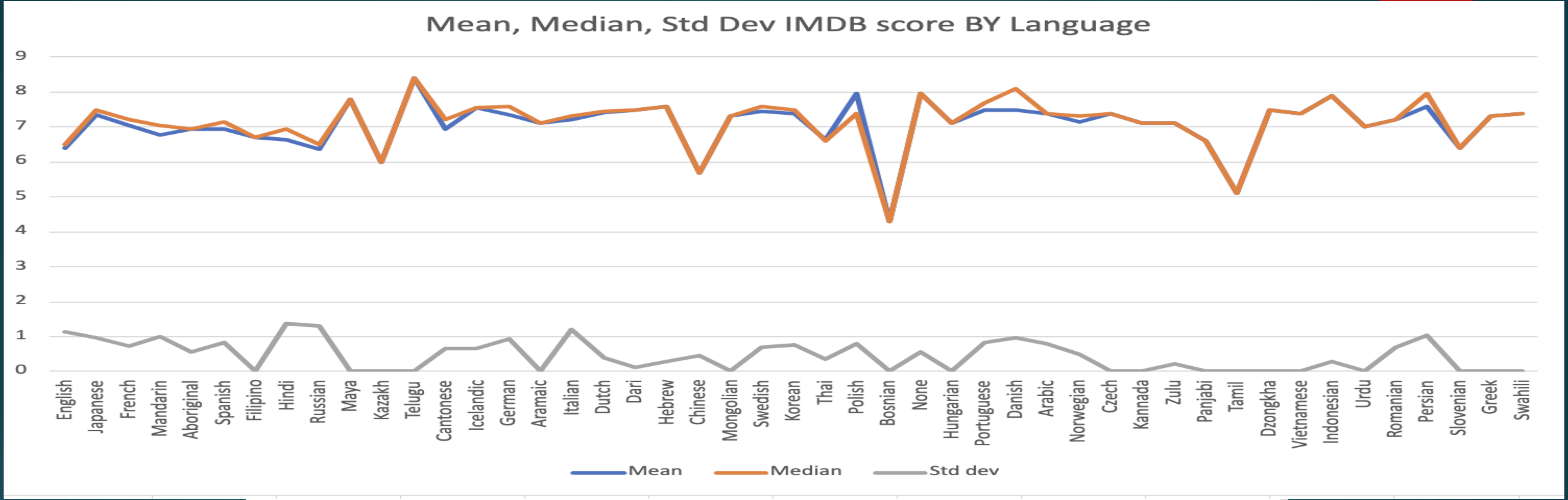


Language	Count of Movies	Mean	Median	Variance	Std deviation
English	4586	6.39	6.5	1.2654	1.1249
Japanese	17	7.35	7.5	0.9413	0.9702
French	73	7.04	7.2	0.5213	0.722
Mandarin	24	6.79	7.05	1.0303	1.015
Aboriginal	2	6.95	6.95	0.3025	0.55
Spanish	40	6.94	7.15	0.7128	0.8443
Filipino	1	6.7	6.7	0	0
Hindi	28	6.63	6.95	1.8872	1.3738
Russian	11	6.36	6.5	1.7405	1.3193
Maya	1	7.8	7.8	0	0
Kazakh	1	6	6	0	0
Telugu	1	8.4	8.4	0	0
Cantonese	11	6.95	7.2	0.4516	0.672
Icelandic	2	7.55	7.55	0.4225	0.65
German	19	7.34	7.6	0.8624	0.9287
Aramaic	1	7.1	7.1	0	0
Italian	11	7.23	7.3	1.4074	1.1863
Dutch	4	7.43	7.45	0.1419	0.3767
Dari	2	7.5	7.5	0.01	0.1
Hebrew	5	7.58	7.6	0.0896	0.2993
Chinese	3	5.67	5.7	0.2022	0.4497
Mongolian	1	7.3	7.3	0	0
Swedish	5	7.44	7.6	0.4584	0.6771
Korean	8	7.39	7.5	0.5961	0.7721
Thai	3	6.63	6.6	0.1356	0.3682

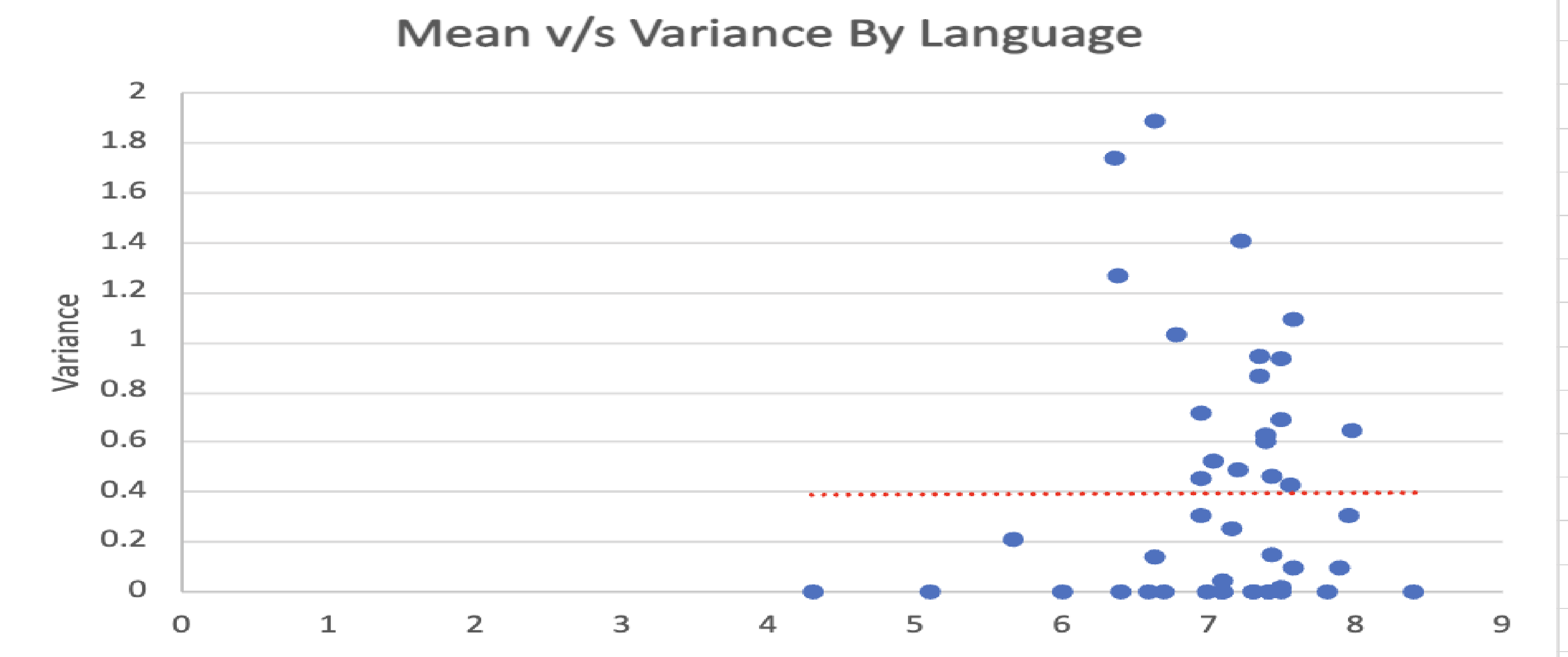


Language	Count of Movies	Mean	Median	Variance	Std Deviation
Polish	3	7.97	7.4	0.6422	0.8014
Bosnian	1	4.3	4.3	0	0
None	2	7.95	7.95	0.3025	0.55
Hungarian	1	7.1	7.1	0	0
Portuguese	8	7.49	7.7	0.6836	0.8268
Danish	5	7.5	8.1	0.928	0.9633
Arabic	5	7.38	7.4	0.6256	0.7909
Norwegian	4	7.15	7.3	0.2475	0.4975
Czech	1	7.4	7.4	0	0
Kannada	1	7.1	7.1	0	0
Zulu	2	7.1	7.1	0.04	0.2
Panjabi	1	6.6	6.6	0	0
Tamil	1	5.1	5.1	0	0
Dzongkha	1	7.5	7.5	0	0
Vietnamese	1	7.4	7.4	0	0
Indonesian	2	7.9	7.9	0.09	0.3
Urdu	1	7	7	0	0
Romanian	2	7.2	7.2	0.49	0.7
Persian	4	7.58	7.95	1.0869	1.0425
Slovenian	1	6.4	6.4	0	0
Greek	1	7.3	7.3	0	0
Swahili	1	7.4	7.4	0	0


Median IMDB_score By Genre
9
8
7
6
5
4
3
2
1
0
English
6.5
Japanese
7.5
French
7.2
Mandarin
7.05
Aboriginal
6.95
Spanish
7.15
Filipino
6.7
Hindi
6.95
Russian
6.5
Maya
7.8
Kazakh
6
Telugu
8.4
Cantonese
7.2
Icelandic
7.55
German
7.6
Aramaic
7.1
Italian
7.3
Dutch
7.45
Dari
7.5
Hebrew
7.6
Chinese
5.7
Mongolian
7.3
Swedish
7.6
Korean
7.5
Thai
6.6
Polish
7.4
Bosnian
4.3
None
7.95
Hungarian
7.1
Portuguese
7.7
Danish
8.1
Arabic
7.4
Norwegian
7.3
Czech
7.4
Kannada
7.1
Zulu
7.1
Panjabi
6.6
Tamil
5.1
Dzongkha
7.5
Vietnamese
7.4
Indonesian
7.9
Urdu
7
Romanian
7.2
Persian
7.95
Slovenian
6.4
Greek
7.3
Swahili
7.4



Mean, Median, Std Dev IMDB score BY Language



Mean v/s Variance By Language

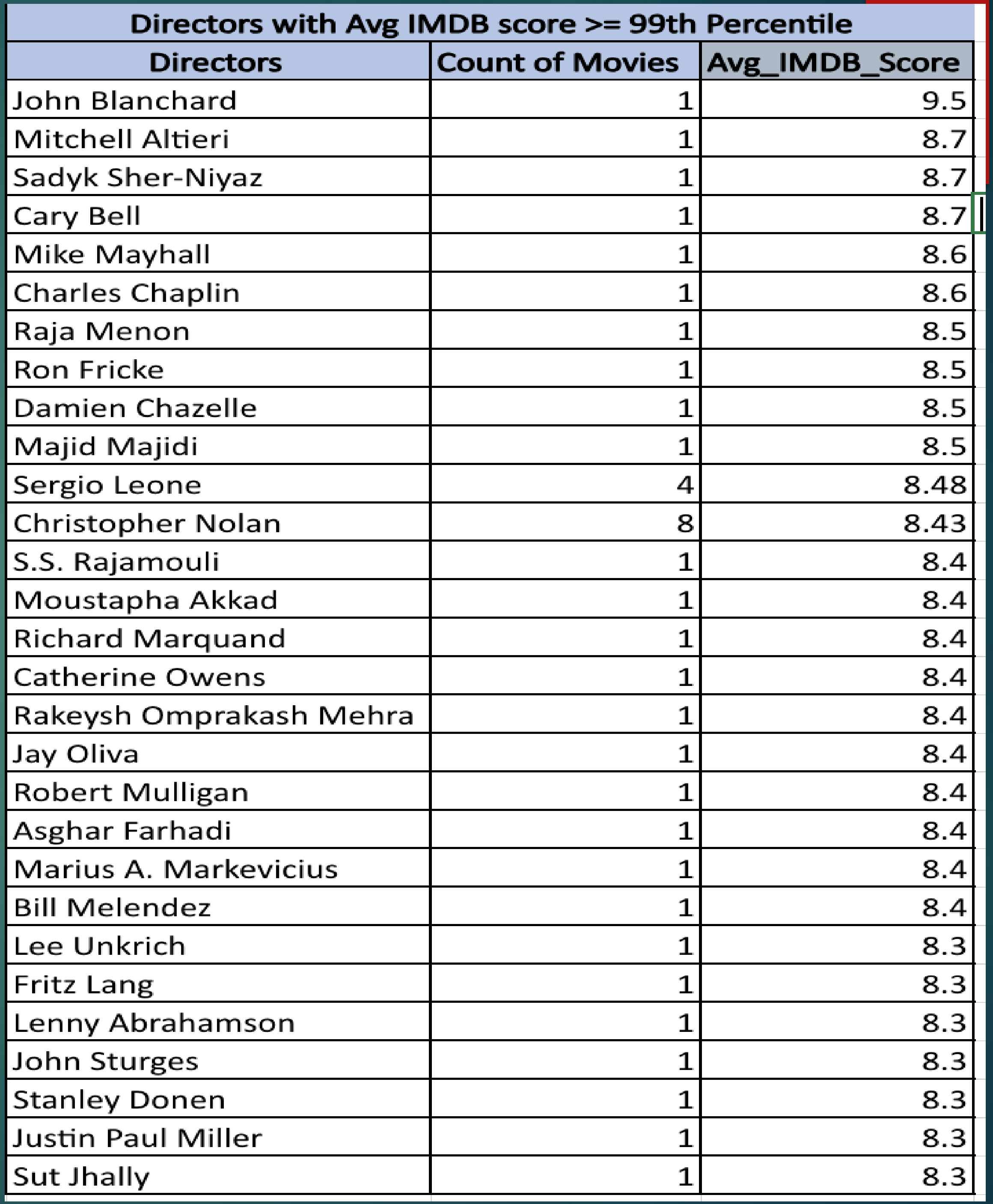
Summary of Above table & charts :

- Descriptive Statistics Table shows that 93.4 % of the Movies in the dataset are made in English Language and Average imdb score of such movies is 6.39 with 1.12 std deviation.
- Median IMDB score chart shows that most of the movies have median score of more than 6.5. However, Telugu Language tops the chart with 8.4 followed by Portuguese with 8.1 median score.
- Mean, Median, Std deviation chart shows that movies made in Hindi, Perisan, Russian, English, and Italian have std variation of more than 1 in their IMDB score. Mean & Median values for all languages except few overlaps which shows insignificant number of outliers for IMDB scores.
- Mean score v/s Variance scatter plot shows that average IMDB score for movies is likely to have a variance of roughly 0.4.

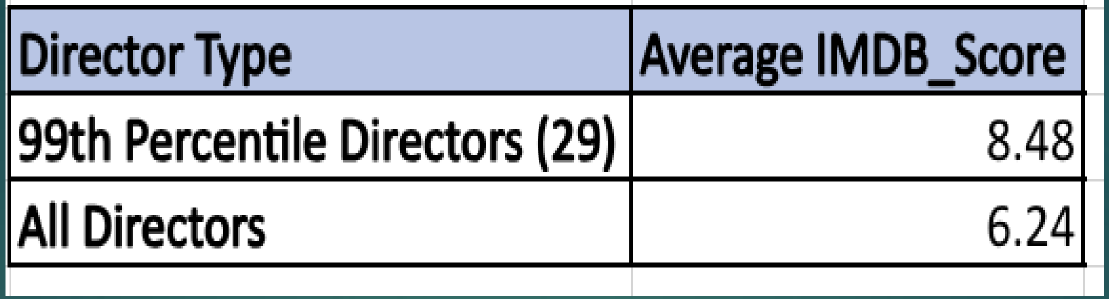
Task – D : Director Analysis: Influence of directors on movie ratings. Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

- Copy-paste the director and IMDB_score fields to another excel sheet.
- Count of Blank cells for director is 102. Remove the rows with blank cells.
- Use UNIQUE function to get list of all the distinct directors (2398)
- Use COUNTIF and AVERAGEIF functions to get the director wise count of movies and average IMDB score.
- Use PERCENTILE function to get the 99th percentile mean IMDB score i.e. 8.3.
- Use FILTER function to get the table of Directors with mean IMDB score >= 99th percentile and their respective count of movies and mean score.
- Apply SORT function on the filtered table range to get the values in descending order of mean IMDB score.

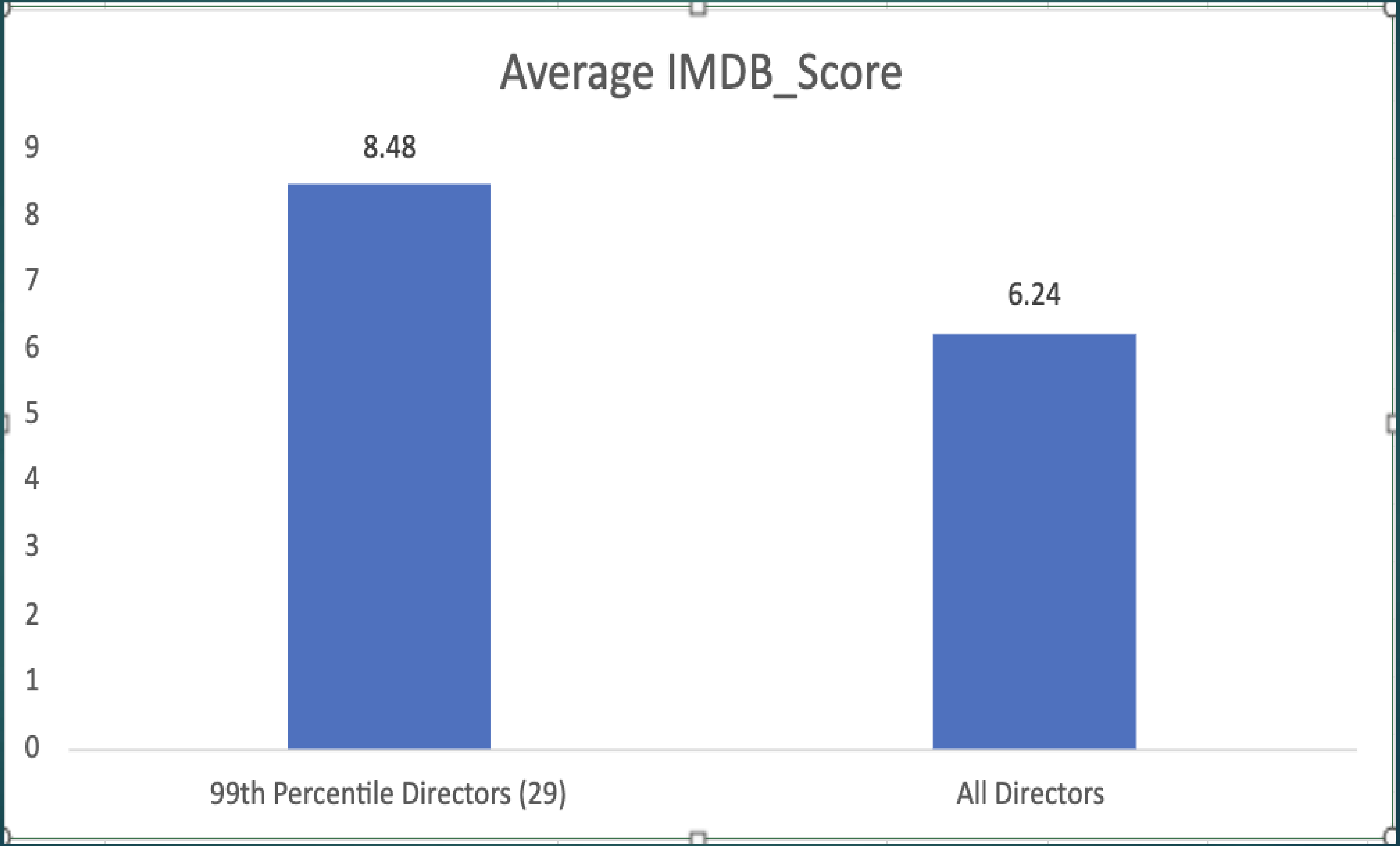
Top 29 Directors with Mean IMDB_score >= 8.3 (99th percentile)



Directors with Avg IMDB score >= 99th Percentile		
Directors	Count of Movies	Avg_IMDB_Score
John Blanchard	1	9.5
Mitchell Altieri	1	8.7
Sadyk Sher-Niyaz	1	8.7
Cary Bell	1	8.7
Mike Mayhall	1	8.6
Charles Chaplin	1	8.6
Raja Menon	1	8.5
Ron Fricke	1	8.5
Damien Chazelle	1	8.5
Majid Majidi	1	8.5
Sergio Leone	4	8.48
Christopher Nolan	8	8.43
S.S. Rajamouli	1	8.4
Moustapha Akkad	1	8.4
Richard Marquand	1	8.4
Catherine Owens	1	8.4
Rakeysh Omprakash Mehra	1	8.4
Jay Oliva	1	8.4
Robert Mulligan	1	8.4
Asghar Farhadi	1	8.4
Marius A. Markevicius	1	8.4
Bill Melendez	1	8.4
Lee Unkrich	1	8.3
Fritz Lang	1	8.3
Lenny Abrahamson	1	8.3
John Sturges	1	8.3
Stanley Donen	1	8.3
Justin Paul Miller	1	8.3
Sut Jhally	1	8.3



Director Type	Average IMDB_Score
99th Percentile Directors (29)	8.48
All Directors	6.24

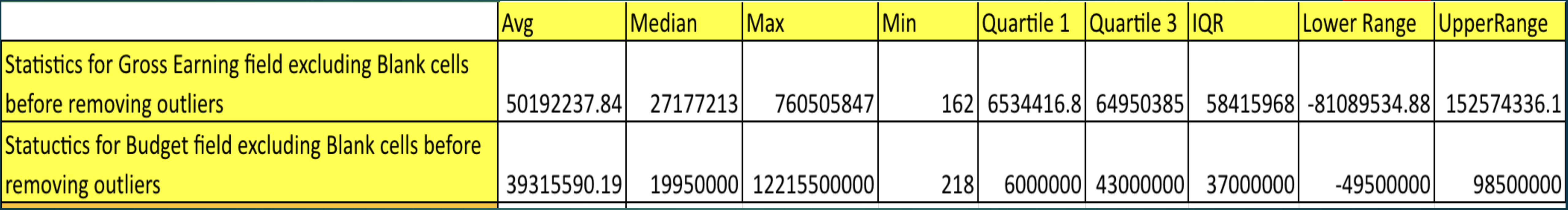


Summary of the above table & chart:

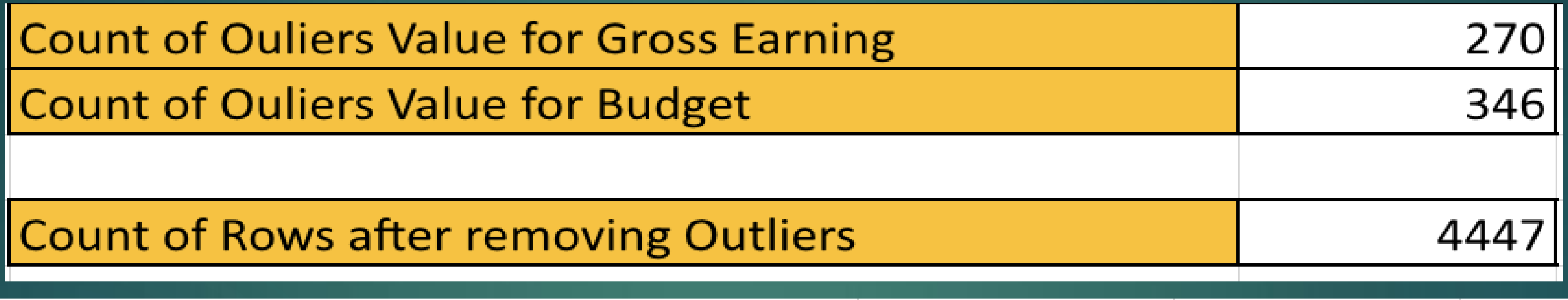
- The table with Name of directors and their respective average IMDB score greater than or equal to 99th Percentile score(8.3) shows that **John Blanchard** tops the chart with 9.5 score followed by **Mitchell Altieri** and **Sadyk Sher** 8.7 score(although, only 1 movie made by these directors)
- Only directors with multiple movies in top director's list : **Christopher Nolan** (8 movies) and **Sergio Leone**(4 movies) and have average rating of 8.43 and 8.48 respectively.
- The comparison column chart of Average IMDB_score of 99th percentile Directors V/S All Directors clearly shows that a movie by directors in 99th percentile is more likely to have higher Average IMDB score as compared to other directors.

Task – E : Budget Analysis- Explore the relationship between movie budgets and their financial success.
Analyze the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

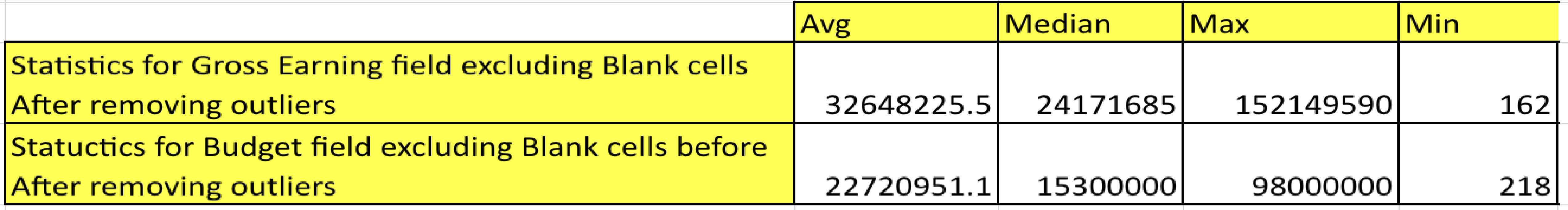
- Copy-Paste the Budget, Gross Earning, and IMBD score fields to another excel sheet.
- Count of Blank cells for Gross earning field is 865 and for Budget field is 485.
- Identify outliers using Quartile function & IQR for Budget & Gross earning field. Remove outliers after identification.
- Perform descriptive statistics excluding blank cells for both the fields.
- Calculate corelation coefficient using CORREL function excluding blank cells.
- Replace blank cells of both the fields with corresponding Median value.
- Add a new Profit field to get the net profit. (Top 10 movies by Profit are calculated w/o removing any outliers)
- Calculate corelation coefficient after replacing blank cells with median values.



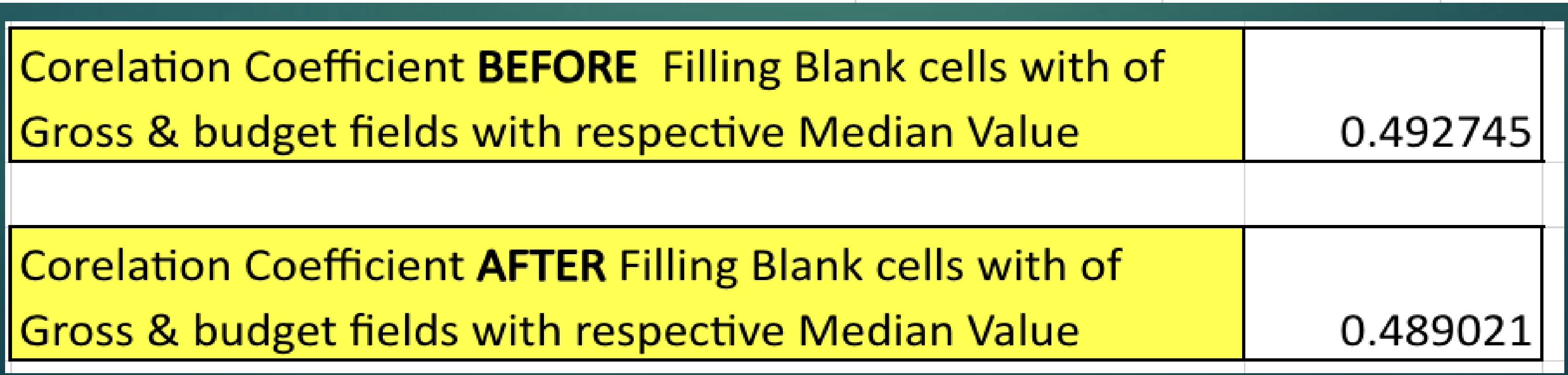
	Avg	Median	Max	Min	Quartile 1	Quartile 3	IQR	Lower Range	UpperRange
Statistics for Gross Earning field excluding Blank cells before removing outliers	50192237.84	27177213	760505847	162	6534416.8	64950385	58415968	-81089534.88	152574336.1
Statuctics for Budget field excluding Blank cells before removing outliers	39315590.19	19950000	12215500000	218	6000000	43000000	37000000	-49500000	98500000



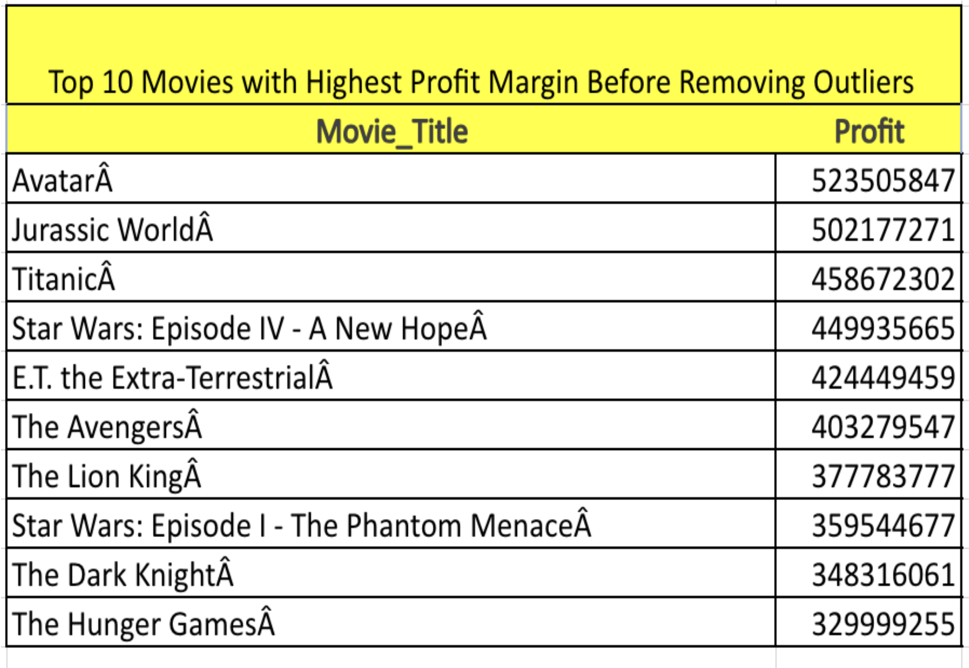
Count of Ouliers Value for Gross Earning	270
Count of Ouliers Value for Budget	346
Count of Rows after removing Outliers	4447



	Avg	Median	Max	Min
Statistics for Gross Earning field excluding Blank cells After removing outliers	32648225.5	24171685	152149590	162
Statuctics for Budget field excluding Blank cells before After removing outliers	22720951.1	15300000	98000000	218

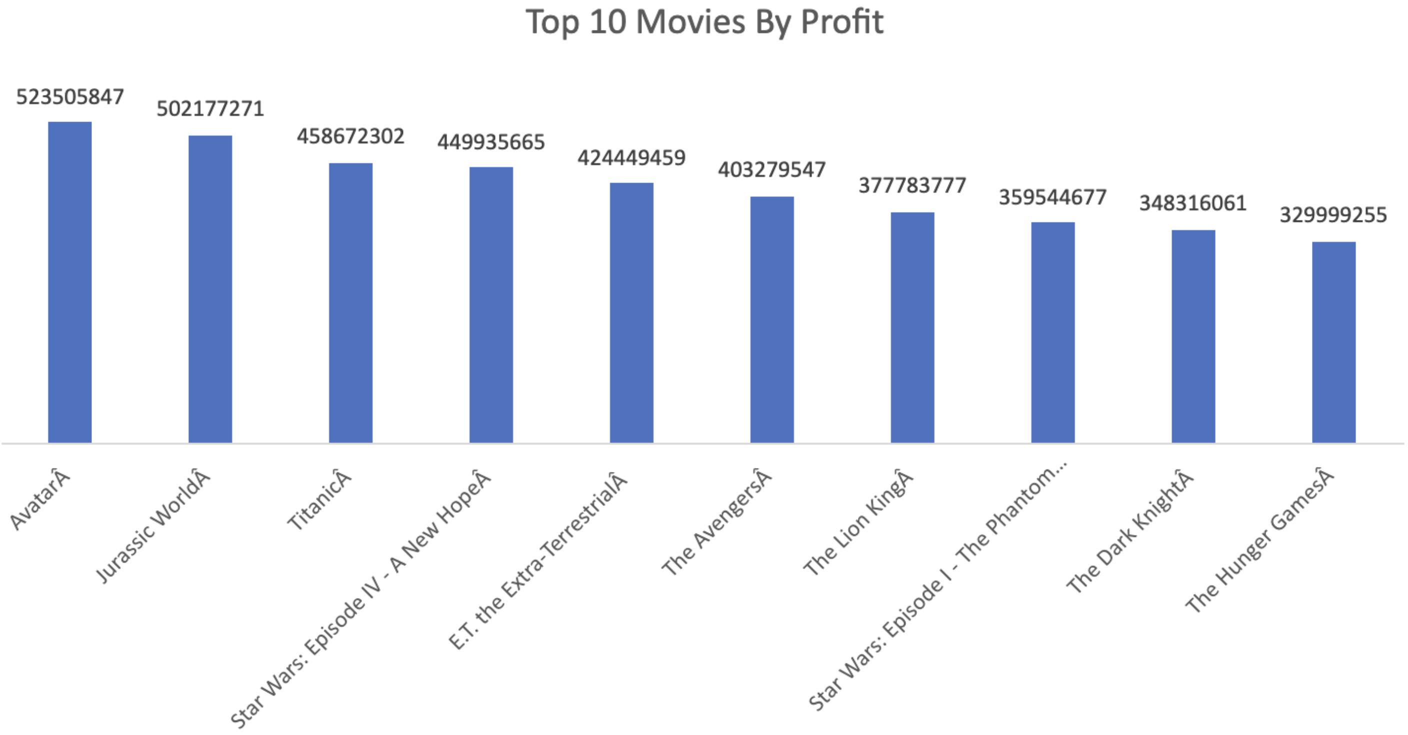


Corelation Coefficient **BEFORE** Filling Blank cells with of Gross & budget fields with respective Median Value	0.492745
Corelation Coefficient **AFTER** Filling Blank cells with of Gross & budget fields with respective Median Value	0.489021



Top 10 Movies with Highest Profit Margin Before Removing Outliers

Movie_Title	Profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255



Top 10 Movies By Profit

Summary for Above Tables and Chart :

- Average earning and Budget of a movie(after removing outliers & replacing blank cells with Median value) is $ 32648225 and $ 22720951.
- Average profit earned by movie(after removing outliers & replacing blank cells with Median value) is $ 7768049
- Corelation coefficient for budget and gross earning (before replacing blank cells with Median value) is 0.492745. However, Corelation coefficient decreases slightly to 0.489021 after replacing blank cells with Median value. This shows that a movie with higher budget is more likely to earn higher gross revenue. Though, the coefficient to not closer to 1.
- Movie with highest profit is **Avatara** followed by **Jurassic World**