EXPLORATORY DATA ANALYSIS OF MOVIES ON NETFLIX

By

Ibekwe ebere

Data Analyst | Ebere Ibekwe

Exploratory data analysis of movies on Netflix

Based on 26 attributes updated in April 2021.

This report is a detailed descriptive analysis of the historic data of movies on Netflix updated April 2021. Netflix has been experiencing huge spike in views of movies released in the 90s than recent movies.

My goal is to highlight the issues that caused low number of votes in movies directed in recent years and recommend changes on the kind of movies or genres that the viewers prefer to watch.

Data Description

There are 1000 movies in my Netflix dataset gotten from Kaggle between 1920 to 2020. Each genre is subject to a specific series title. The data set consists of 1000 observations with 8 variables. There are 5 categorical variables and 3 numerical variables. The raw data is shown in the fig. Every series title is directed by a director. Some directors directed three or more series . The directors have 4 different stars in every movie and are rated differently.

The information on both rating and number of votes is used to know the genre that the viewers prefer regardless of the year released.

	Series_Ti	Releas	Certific	Runtime	Genre	IMDB_	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of	Gross
https://m.med	The Shawsh	1994	A	142 min	Drama	9.3	Two imprisor	80	Frank Darabc	Tim Robb	Morgan Fre	Bob Gunto	William Sa	2343110	28341469
https://m.med	The Godfath	1972	A	175 min	Crime, Drama	9.2	An organized	100	Francis Ford	Marlon Br	Al Pacino	James Caa	Diane Keat	1620367	134966411
https://m.med	The Dark Kni	2008	UA	152 min	Action, Crime,	9	When the me	84	Christopher I	Christian	Heath Ledg	Aaron Eckh	Michael Ca	2303232	534858444
https://m.med	The Godfath	1974	A	202 min	Crime, Drama	9	The early life	90	Francis Ford	Al Pacino	Robert De N	Robert Du	Diane Keat	1129952	57300000
https://m.med	12 Angry Me	1957	U	96 min	Crime, Drama	9	A jury holdou	96	Sidney Lume	Henry For	Lee J. Cobb	Martin Bal	John Fiedl	689845	4360000
https://m.med	The Lord of t	2003	U	201 min	Action, Advent	8.9	Gandalf and A	94	Peter Jackso	Elijah Wo	Viggo Mort	Ian McKell	Orlando Bl	1642758	377845905
https://m.med	Pulp Fiction	1994	A	154 min	Crime, Drama	8.9	The lives of t	94	Quentin Tara	John Trav	Uma Thurm	Samuel L. J	Bruce Willi	1826188	107928762
https://m.med	Schindler's L	1993	A	195 min	Biography, Dra	8.9	In German-oc	94	Steven Spiel	Liam Nee	Ralph Fienr	Ben Kingsl	Caroline G	1213505	96898818
https://m.med	Inception	2010	UA	148 min	Action, Advent	8.8	A thief who s	74	Christopher I	Leonardo	Joseph Gor	Elliot Page	Ken Watan	2067042	292576195

Data Analvst | Ebere Ibekwe

Task Performed

- Analyzed data to identify noteworthy patterns or relationships pertaining to genre and the series titles from the viewers perspective.

- Provided for each identified pattern or relationship a plausible underlying rationale and useful insights that can be drawn.

- Recommended changes to specific aspects of directing more movies in the first top ten genre and lower the production of other genre or merge then with the top genre to increase number of views.

Quality of Data

After accessing the data, I decided to fill up empty numerical cells with their average . I also decided to look at the data, and to draw links across the different aspects of the data, as they tell different but equally important stories about the problem in question. Instead removing null cell, I filled them their average .

Descriptive Analysis

The total number of votes directly implies the series title that the viewers watched most and its more interesting. I have identified that in investigating the total votes and rating of the series title , it is important to analyze the following within my data sets.

- Directors and certificate issued for the top ten series title.
- Genre composition of each series title and year released. Directors choice of stars; series title overview for viewers to glance through. Diversity of movies in different genre directed and so on.
- I also anticipate that I will identify some more relevant reasons of how some series title released in the 90s has little or more votes as I go on with my analyses.

Data Analyst | Ebere Ibekwe

Attribute of each variable

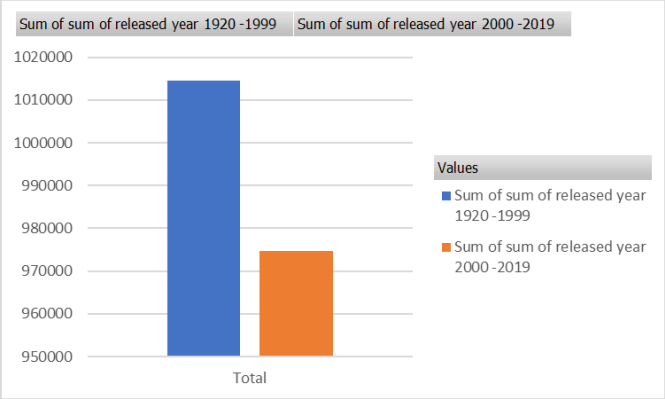
- **Series Tittle**

Comprises of different series title, except dryshwani is a series title that occurred twice in my dataset. This data set contains a total of 1000 series title .

- **Released Year**

From my dataset more movies were released between the year 1920 to 1999 while less movies were released between 2000 to 2019.

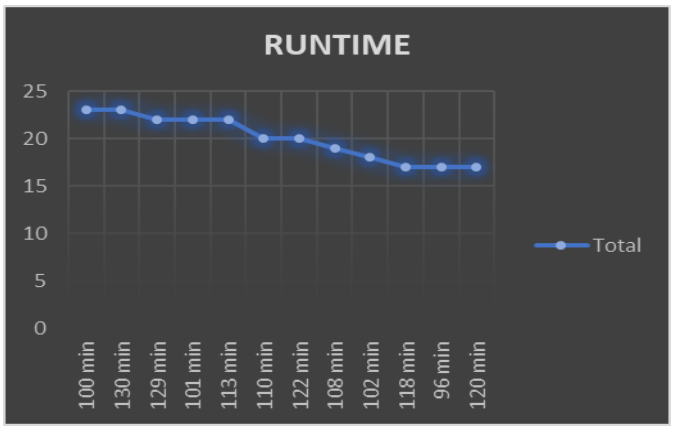
sum of released year 1920 -1999	sum of released year 2000 -2019
1014495	974731



- **Runtime**

From the table below ,100min and 130 min runtime has 23 series title each while 96,118,120min runtime has a series title of 17.

Row Labels	Count of Runtime
100 min	23
130 min	23
129 min	22
101 min	22
113 min	22
110 min	20
122 min	20
108 min	19
102 min	18
118 min	17
96 min	17
120 min	17



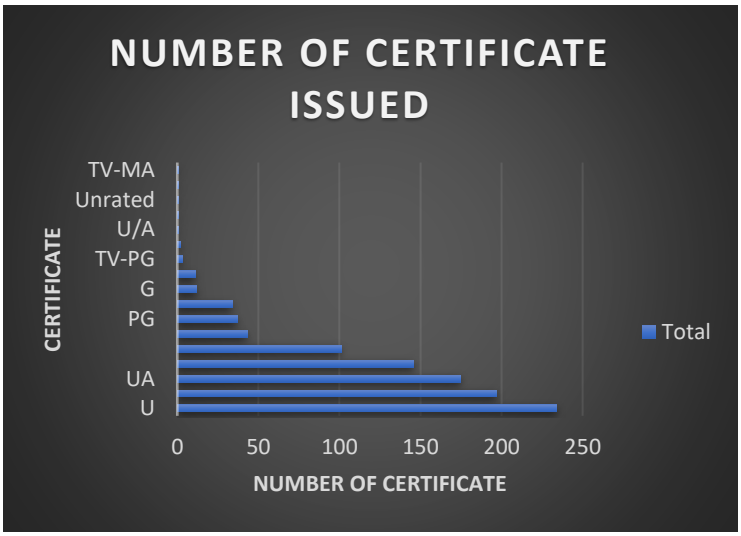
Data Analyst | Ebere Ibekwe

Certificate

The table below shows that certificate U was issued to more series title, but 101 series title were not issued certificate. While certificate U/A ,16, TV-14,

TV-MA were the lowest issued certificate.

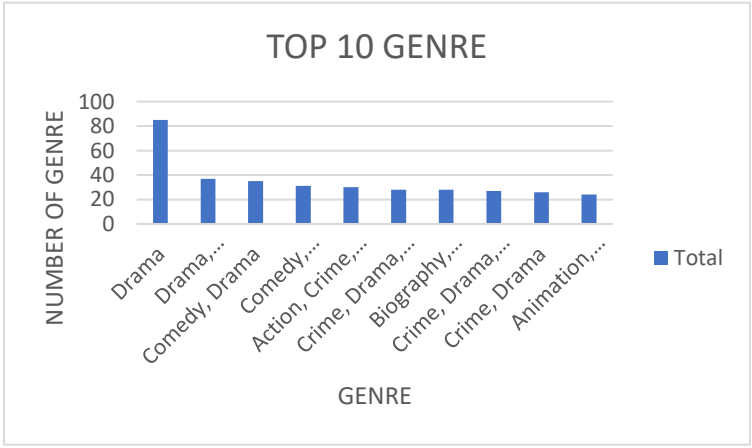
Row Labels	Count of Certificate
U	234
A	197
UA	175
R	146
	101
PG-13	43
PG	37
Passed	34
G	12
Approved	11
TV-PG	3
GP	2
U/A	1
16	1
Unrated	1
TV-14	1
TV-MA	1



- **Genre**

From the top ten genre drama has the highest number of series title while

Row Labels	Count of Genre
Drama	85
Drama, Romance	37
Comedy, Drama	35
Comedy, Drama, Romance	31
Action, Crime, Drama	30
Crime, Drama, Thriller	28
Biography, Drama, History	28
Crime, Drama, Mystery	27
Crime, Drama	26
Animation, Adventure, Comedy	24
Grand Total	**351**

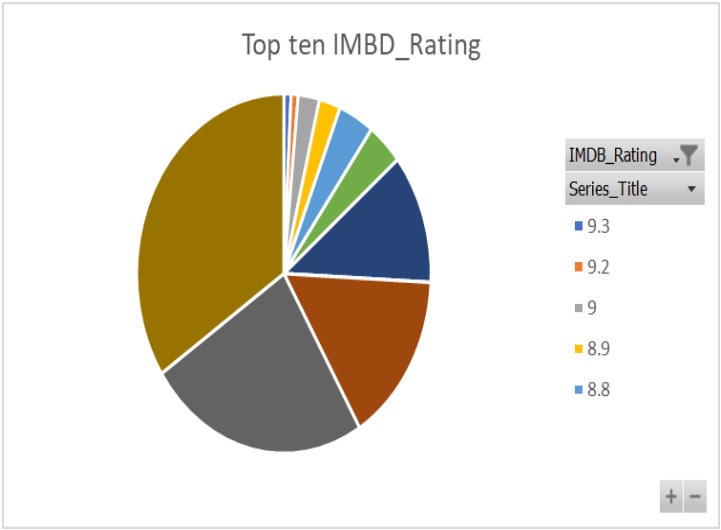


Data Analyst | Ebere Ibekwe

- **Rating**

One series title were rated 9.3(highest rating) .44 series title were rated 8.3 while 123 series title were rated 7.7 which has the lowest rating .

IMDB_Rating	
Mean	7.9493
Standard Error	0.0087118
Median	7.9
Mode	7.7
Standard Deviat	0.27549121
Sample Varianc	0.07589541
Kurtosis	1.432727
Skewness	1.01696445
Range	1.7
Minimum	7.6
Maximum	9.3
Sum	7949.3
Count	1000



- **Overview**

The overview gives an insight of the plot in the series title and they are arranged in alphabetical order.

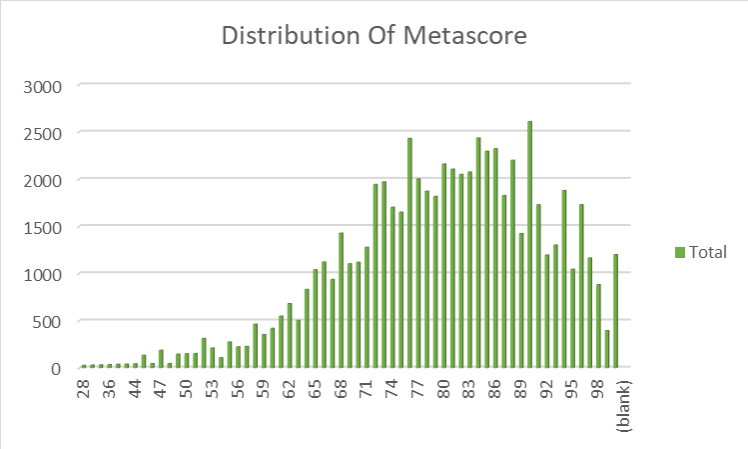
- **Metascore**

The distribution of the Metascore Variable is left-skewed. From our variable summary table, I see that the mean Metascore is 77. The spread of the distribution is 11.3.

The maximum Metascore is 100 while the minimum Metascore is 28.

Data Analyst | Ebere Ibekwe

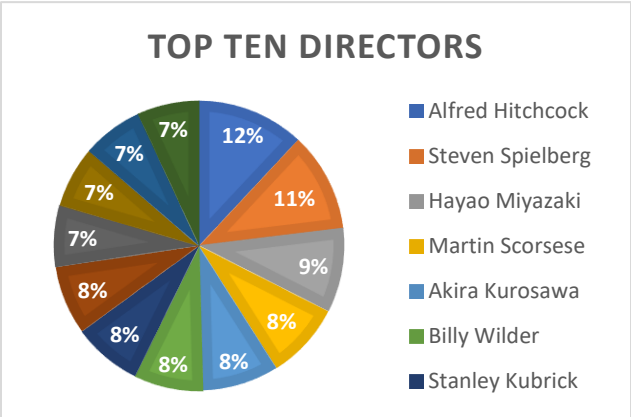
Meta_score	
Mean	77.82
Standard Error	0.359473
Median	77
Mode	77
Standard Deviation	11.36753
Sample Variance	129.2208
Kurtosis	1.015833
Skewness	-0.61807
Range	72
Minimum	28
Maximum	100
Sum	77820
Count	1000



- **Director**

From the pie chart 12% of the series title were directed by Alfred Hitchcock.

Row Labels	Count of Director
Alfred Hitchcock	14
Steven Spielberg	13
Hayao Miyazaki	11
Martin Scorsese	10
Akira Kurosawa	10
Billy Wilder	9
Stanley Kubrick	9
Woody Allen	9
Clint Eastwood	8
Quentin Tarantino	8
Christopher Nolan	8
David Fincher	8

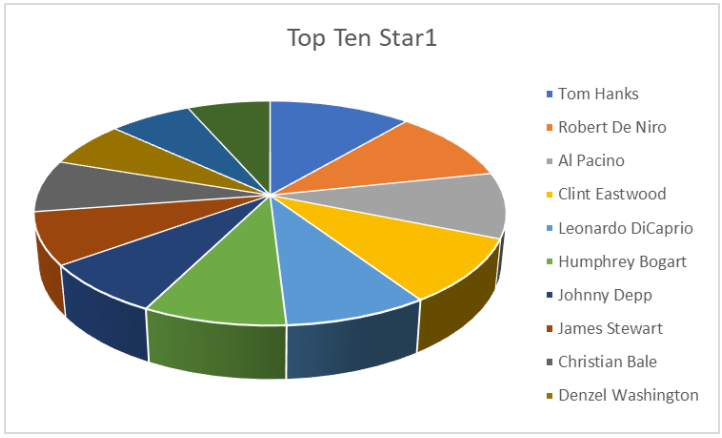


Data Analyst | Ebere Ibekwe

Star1.

Tom Hanks stared in 12 series title .

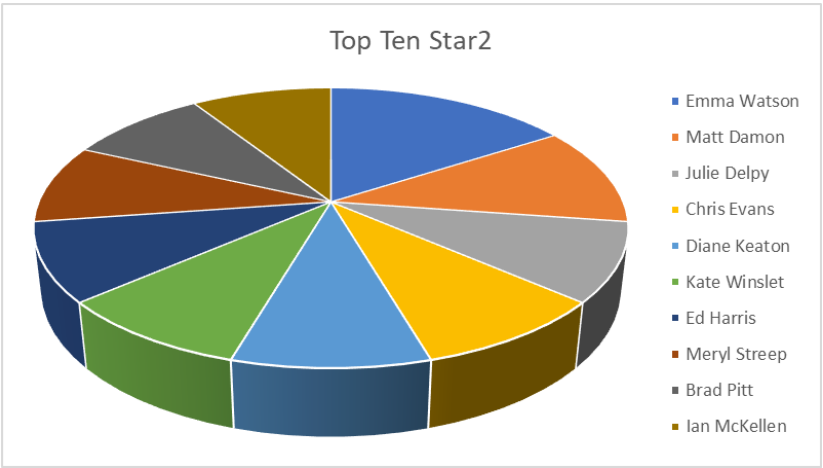
Row Labels	Count of Star1
Tom Hanks	12
Robert De Niro	11
Al Pacino	10
Clint Eastwood	10
Leonardo DiCaprio	9
Humphrey Bogart	9
Johnny Depp	8
James Stewart	8
Christian Bale	8
Denzel Washington	7
Toshirô Mifune	7
Aamir Khan	7



- **Star2**

Emma Watson starred in 7 series title.

Row Labels	Count of Star2
Emma Watson	7
Matt Damon	5
Julie Delpy	4
Chris Evans	4
Diane Keaton	4
Kate Winslet	4
Ed Harris	4
Meryl Streep	4
Brad Pitt	4
Ian McKellen	4

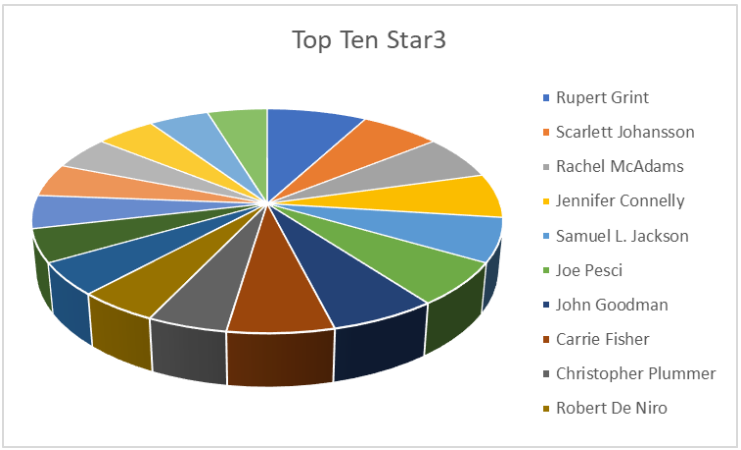


Data Analyst | Ebere Ibekwe

- Star3

Rupert Grint starred in 5 series title .

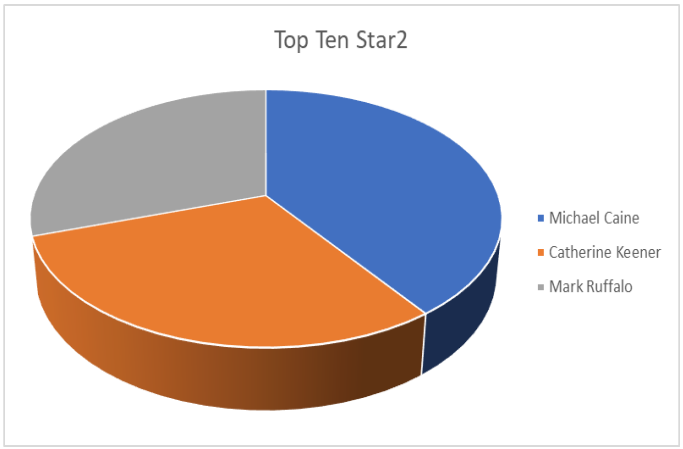
Row Labels	Count of Star3
Rupert Grint	5
Scarlett Johansson	4
Rachel McAdams	4
Jennifer Connelly	4
Samuel L. Jackson	4
Joe Pesci	4
John Goodman	4
Carrie Fisher	4
Christopher Plummer	3
Robert De Niro	3
Vera Miles	3
Michael Madsen	3
Chiwetel Ejiofor	3
Morgan Freeman	3
Edward Norton	3
Nawazuddin Siddiqui	3
Frances McDormand	3
Kevin Bacon	3

- 

- **Star4**

Michael Caine starred in four serries title.

Row Labels	Count of Star4
Michael Caine	4
Catherine Keener	3
Mark Ruffalo	3

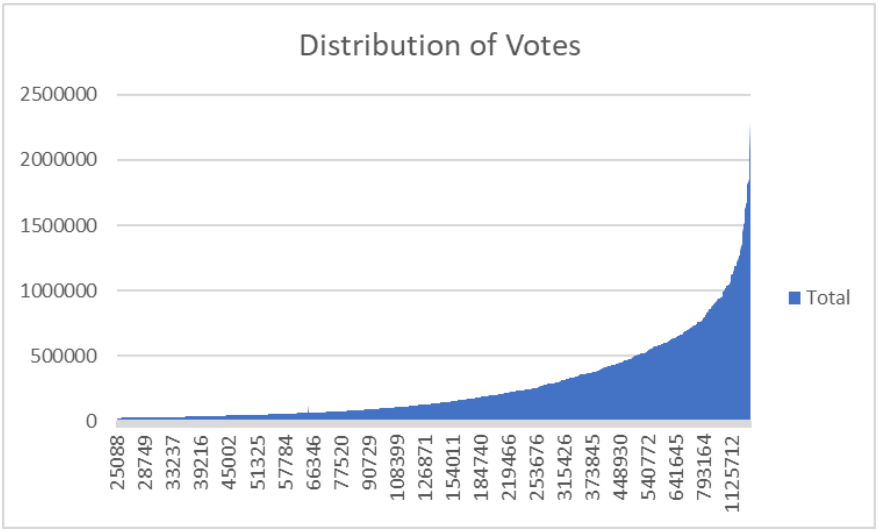


Data Analyst | Ebere Ibekwe

Votes

The distribution of the number of votes is left-skewed. From our variable summary table, I see that the mean vote is 271621.4. The spread of the distribution is 320912.6. The median score is 138356.

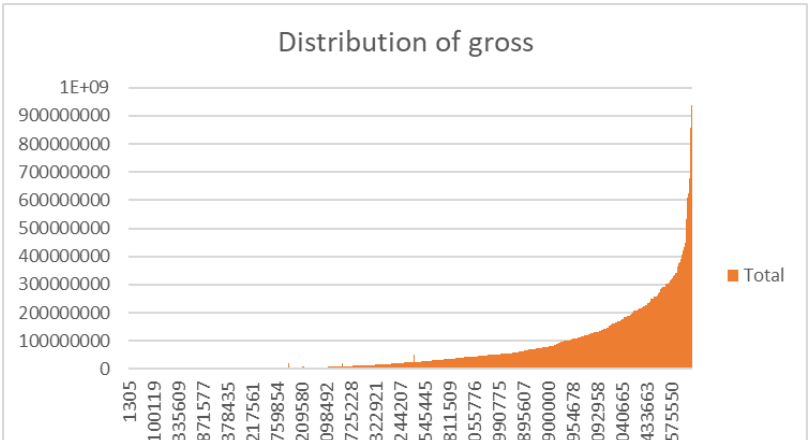
summary of votes	
Mean	271621.4
Standard Error	10153.23
Median	138356
Mode	65341
Standard Deviation	320912.6
Sample Variance	1.03E+11
Kurtosis	6.041324
Skewness	2.194351
Range	2278144
Minimum	25088
Maximum	2303232
Sum	2.71E+08
Count	999



Gross

The distribution of the number of gross is left-skewed. From our variable summary table, I see that the mean is 68074484 . The median score is 42438300.

Summary of Gross	
Mean	68074484
Standard Error	3166369
Median	42438300
Mode	68034751
Standard Deviation	1E+08
Sample Variance	1E+16
Kurtosis	17.31833
Skewness	3.430621
Range	9.37E+08



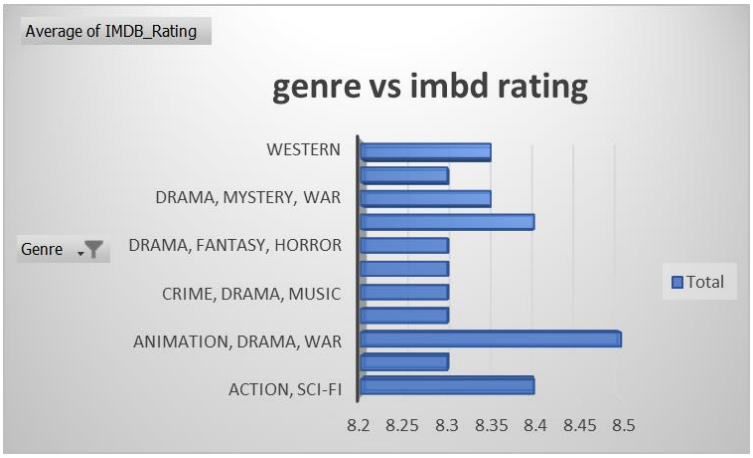
Data Analyst | Ebere Ibekwe

Relationship between categorical and numerical data

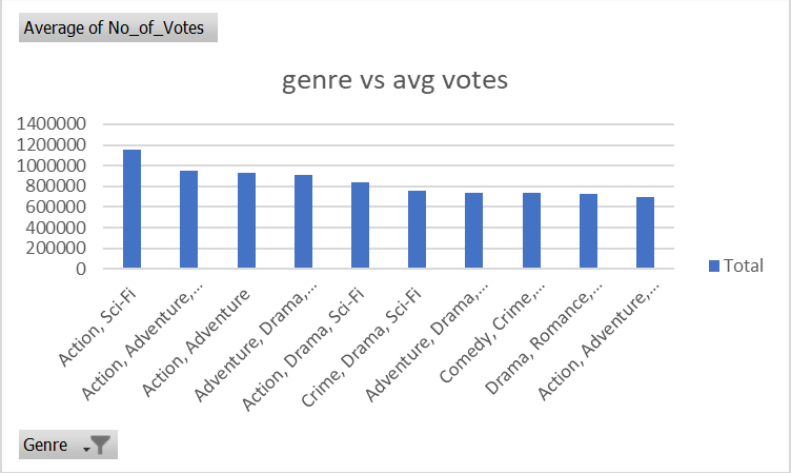
Genre

Genre vs rating and average number of votes have two similar genre. But total number of votes has similar 8 genre. Action,sci-Fi has the highest number of votes and rating .

genre	Average of IMDB_Rating
Action, Sci-Fi	8.4
Adventure, Mystery, Thriller	8.3
Animation, Drama, War	8.5
Comedy, Musical, Romance	8.3
Crime, Drama, Music	8.3
Crime, Drama, Sci-Fi	8.3
Drama, Fantasy, Horror	8.3
Drama, Musical	8.4
Drama, Mystery, War	8.35
Mystery, Romance, Thriller	8.3
Western	8.35
Grand Total	**8.352941176**



genre	Average of No_of_Votes
Action, Sci-Fi	1157242.333
Action, Adventure, Fantasy	954767.6667
Action, Adventure	925533.4
Adventure, Drama, Sci-Fi	912522.3333
Action, Drama, Sci-Fi	840316.5
Crime, Drama, Sci-Fi	757904
Adventure, Drama, War	735666
Comedy, Crime, Sport	732620
Drama, Romance, Sci-Fi	726218
Action, Adventure, Sci-Fi	696942.7619
Grand Total	**810788.913**



Data Analyst | Ebere Ibekwe

• I can deduce from this chart that the movie industry has heavily directed movies in drama over other genre but it did not attain highest vote.More adverts should be done to increase the rate of viewers.

By narrowing my focus on these genres, I might find top four genre .

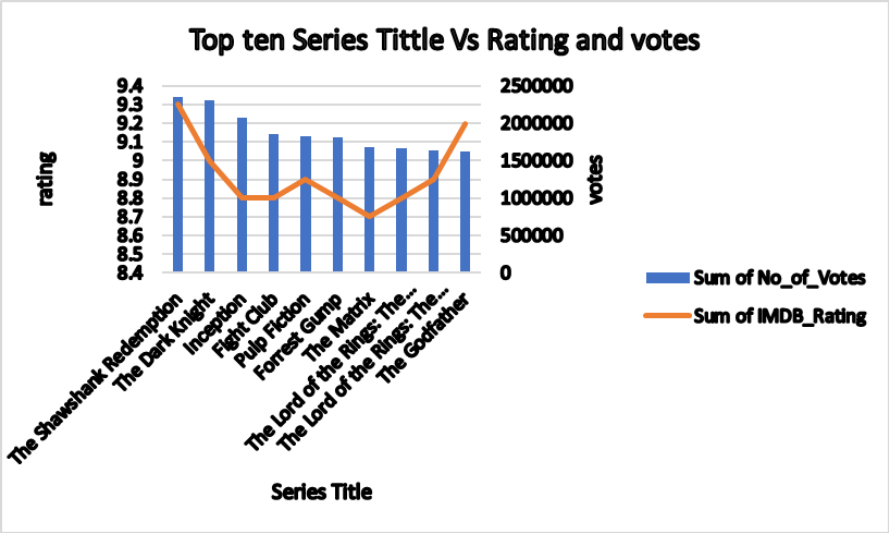
- Movies in other genre should be advertised reguary to increase visibility and votes.
- Overall, the highest top 10 genre with the highest vote and rating can be seen in the table above .

Top ten Series Tittle vs total number of votes and rating.

I observed that Shawshank Redemption has the highest no of votes and the highest rating. It is a series tittle under drama .The Godfather was the tenth voted series tittle but is the second rated of 9.2 as seen in the pivot table below.

Row Labels	Sum of No_of_Votes	Sum of IMDB_Rating
The Shawshank Redemption	2343110	9.3
The Dark Knight	2303232	9
Inception	2067042	8.8
Fight Club	1854740	8.8
Pulp Fiction	1826188	8.9
Forrest Gump	1809221	8.8
The Matrix	1676426	8.7
The Lord of the Rings: The Fellowship of the Ring	1661481	8.8
The Lord of the Rings: The Return of the King	1642758	8.9
The Godfather	1620367	9.2
Grand Total	**18804565**	**89.2**

Data Analyst | Ebere Ibekwe



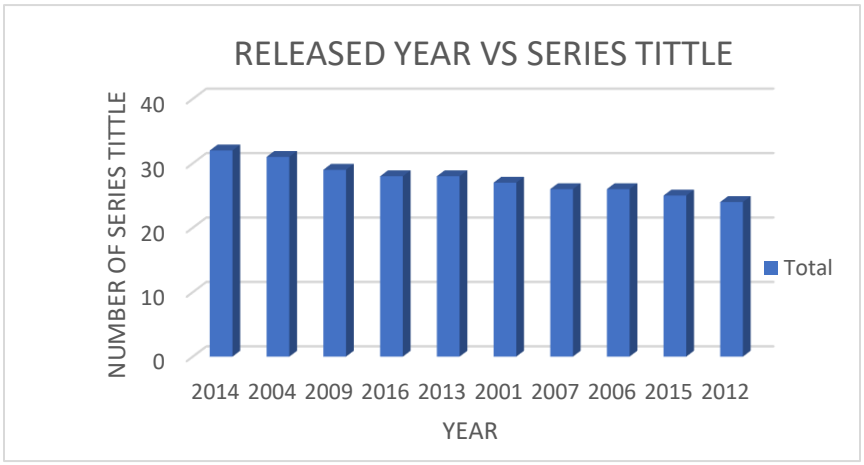
- **Series Tittle and year released**

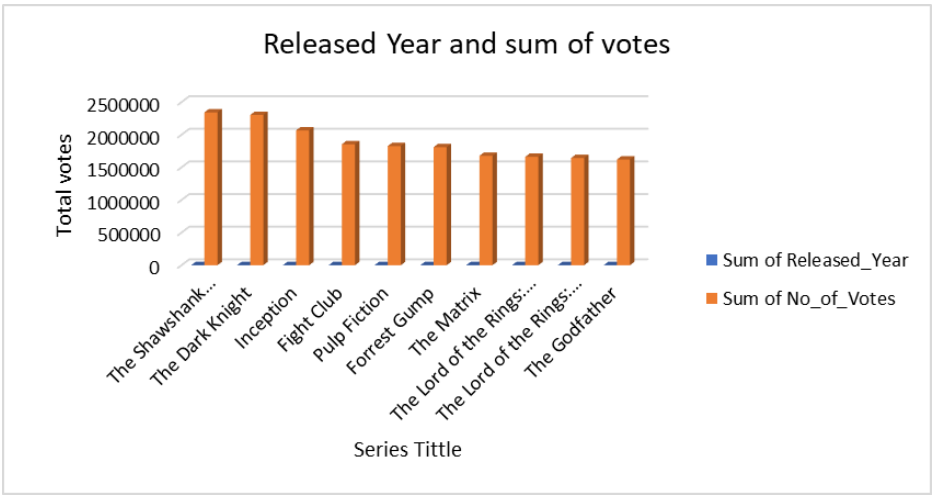
I decided that the overview and rating is an important factor that attracts the viewers. Series title in the 90s are more entertaining to the viewers and reveals useful patterns and helps us understand the viewers preference thereby increasing the number of votes and rating. In the table below 6 series title in the 90s out of the top ten were voted .

Row Labels	Sum of Released Year	Sum of No_of_Votes
The Shawshank Redemption	1994	2343110
The Dark Knight	2008	2303232
Inception	2010	2067042
Fight Club	1999	1854740
Pulp Fiction	1994	1826188
Forrest Gump	1994	1809221
The Matrix	1999	1676426
The Lord of the Rings: The Fellowship of the Ring	2001	1661481

The Lord of the Rings: The Return of the King	2003	1642758
The Godfather	1972	1620367

Data Analyst | Ebere Ibekwe





- The Table above shows the total number of votes for the first top ten 10 movies

- The graph reveals a clear difference in the total number of votes and their released year .This clearly shows that 60% of movies in the 90s and 40% of the movies in the 2000 were has the highest votes..

Data Analyst | Ebere Ibekwe

- At first sight , I see that about a handful of the directed in the 2000 has higher votes in my dataset

Released Year Analysis :

I am curious to know the year with the highest series Tittle but none series tittle attained the top ten number of votes.

The chart below shows that 2014 accounts for the highest number of series tittle released

- **Series Title Vs genre vs director vs number of votes.**

The table below consist of the top 10 series title, its genre, and directors.

Frank Darabont directed the series title with the highest vote. Christopher Nolan and Peter Jackson directed two series title each which were among the top ten.

Data Analyst | Ebere Ibekwe

Row Labels	Sum of No_of_Votes
The Shawshank Redemption	**2343110**
Drama	**2343110**
Frank Darabont	2343110
The Dark Knight	**2303232**
Action, Crime, Drama	**2303232**
Christopher Nolan	2303232
Inception	**2067042**
Action, Adventure, Sci-Fi	**2067042**
Christopher Nolan	2067042
Fight Club	**1854740**
Drama	**1854740**
David Fincher	1854740
Pulp Fiction	**1826188**
Crime, Drama	**1826188**
Quentin Tarantino	1826188
Forrest Gump	**1809221**
Drama, Romance	**1809221**
Robert Zemeckis	1809221
The Matrix	**1676426**
Action, Sci-Fi	**1676426**
Lana Wachowski	1676426
The Lord of the Rings: The Fellowship of the Ring	**1661481**
Action, Adventure, Drama	**1661481**
Peter Jackson	1661481
The Lord of the Rings: The Return of the King	**1642758**
Action, Adventure, Drama	**1642758**
Peter Jackson	1642758
The Godfather	**1620367**
Crime, Drama	**1620367**

Data Analyst | Ebere Ibekwe

Conclusion

Our analysis of data on genre performance is mostly useful to help the film producers and directors to produce more series title on drama . This project has reviewed data showing series title, released year, certificate, runtime ,genre ,imbd rating ,overview ,Metascore ,directors ,star(1,2,3,4),gross ,no of votes. I found out that as the no of votes for drama increases, the rating also increases. This tells us that the genre directly influences the no of votes.

In conclusion, entertainment reveals cultural practices of different people, science and creates space for learning and research and quality of a directors series title has influence on rating and this can help for improvement of new series title .from my analysis ,we can see that the top first directors Alfred Hitchcock,Stevenn Spielberg 13,Hayao Miyazaki 11 directed more of dramas and action, crime and drama .

I suggest that there should be an increase in the production of mores series title with their genre as drama and action , drama and romance .