IMDB MOVIE ANALYSIS

SUBMITTED BY - HASAN AZIZ

FINAL PROJECT - 1

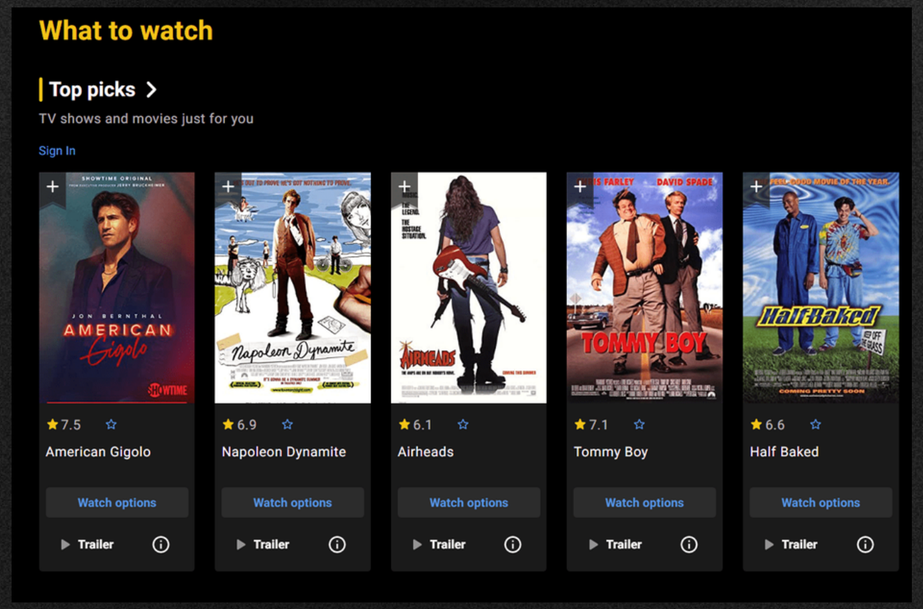
IMDb

trainity

ADVANCE EXCEL | STATISTICS

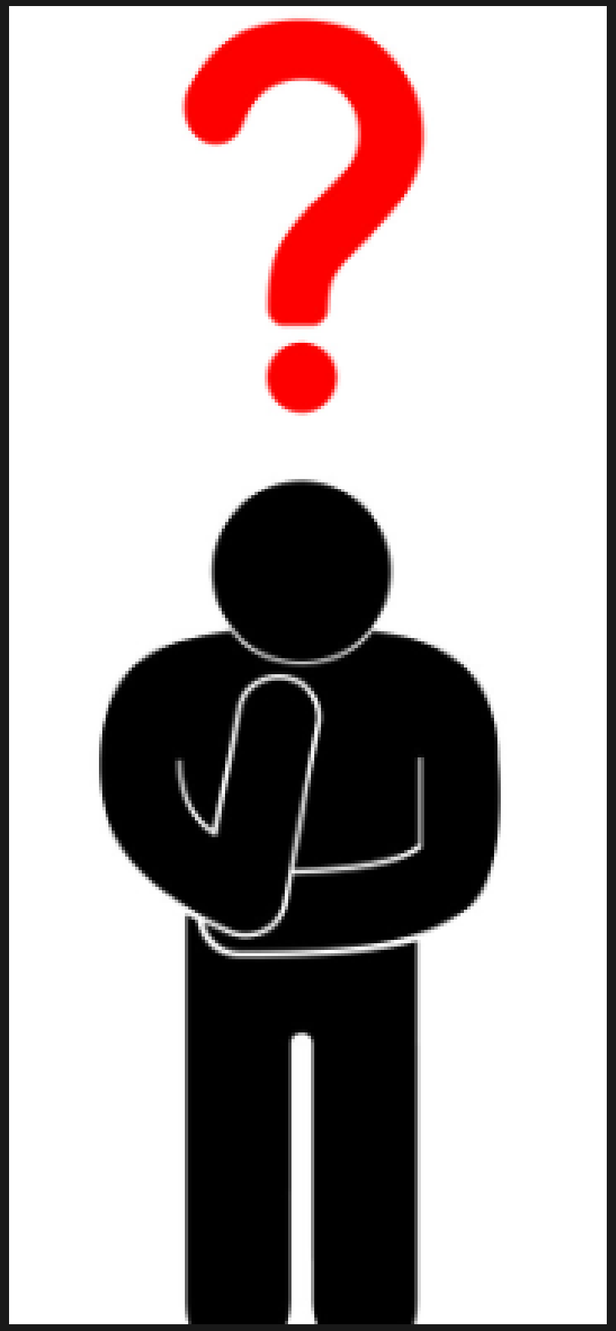
AGENDA

1. INTRODUCTION / PROJECT DESCRIPTION
2. PROBLEM STATEMENT
3. HYPOTHESIS
4. APPROACH AND TECH-STACK USED
5. INISGHTS
6. RESULTS AND CONCLUSION



Introduction

- IMDB or “Internet Movie Database” is a reputed online platform that provides large volume of information on movies, TV shows, reality shows, and OTT content.
- It provides details on plot summaries, IMDB rating, casts and crew details, trailers, critique reviews, etc.
- Our stakeholder(trainity) has provided a large dataset of IMDB movies which involves data related to actor names, director names, genres, gross, IMDb rating of each movie, language of movies, country, and no of critic review.
- As a data analyst I am going to analyze the dataset based on our specific problem statement and Hypothesis and produced some useful insights.



PROBLEM STATEMENT

IMDB hosts a large volume of data related to movies and it is very difficult for users to find some specific details such as the top movies according IMDB rating, highest grossing movies, the best directors etc and that's where data analyst comes in. As a data analyst, I am going to answer the following questions provided by our stakeholder(trainity):

1. ***Find the movies with the highest profit?***
2. ***Find IMDB Top 250 movies. Also, find the top foreign language films.***
3. ***Find the top 10 best directors***
4. ***Find the most popular genres in IMDB***
5. ***Find the critic-favorite and audience-favorite actors***

Apart from the previous questions, I have developed some more objectives that I want to shed light on. These are –

1. ***Analyzing the Performance of Movies Across Countries***
2. ***Analyzing the Relationship between IMDb Ratings and Gross Revenue***
3. ***Exploring the relationship between movie Language on IMDb Ratings***

HYPOTHESIS

Hypothesis: Relationship between IMDB rating and Gross Revenue

- **Null Hypothesis (H0):** There is no relationship between IMDB rating and gross revenue.
- **Alternative Hypothesis (HA):** There is a significant relationship between IMDB rating and gross revenue.

Hypothesis: Relationship between language and IMDB ratings

- **Null Hypothesis (H0):** There is no difference in the average IMDb ratings among movies in different languages.
- **Alternative Hypothesis (HA):** The average IMDb ratings differ among movies in different languages.

APPROACH AND TECH-STACK USED

01

Data cleaning-

I have used **MS Excel software** to conduct Data cleaning and entire Analysis.

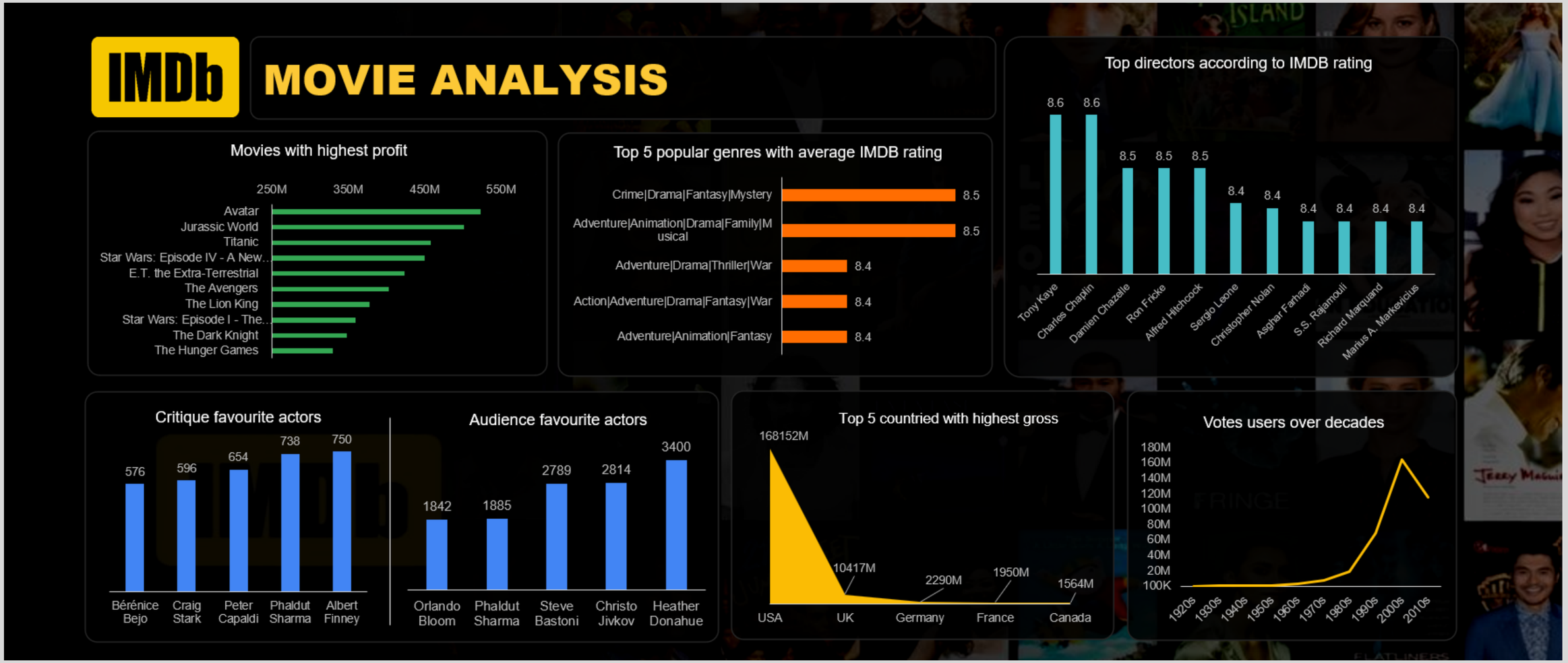
- According to our problem statements some column were not required, So I removed the unnecessary columns such as actor's facebook likes, director's facebook likes, duration, actor 2 and actor 3 names.
- Next important step was to remove the columns with missing values. I have noticed that some values related to gross and budget were missing. As these two columns are vital in our analysis, removing those rows with missing values was necessary to get correct insights.
- Removed the duplicate cells as well. Using conditional formatting I highlighted the movie names which were duplicates. Although, there were some movies which have same name but released in different year, so I have kept those.
- For better understanding of the cleaned dataset, I have highlighted the columns with different colour (Blue for existing columns, green for newly made columns)

02

Data Analysis -

- Once data are cleaned, I have answer each questions step by step using different excel functions (such as IF, AVERAGE) and using Pivot tables.
- I used some statistical analysis to test the hypothesis such as Descriptive statistics, regression test, ANOVA test. I used the MS Excel Data Analysis tool to perform these.
- For visualization I prepared a dashboard for the user to showcase all insights in one place. I prepared various bar charts, column charts, scatter plots for visualizations.

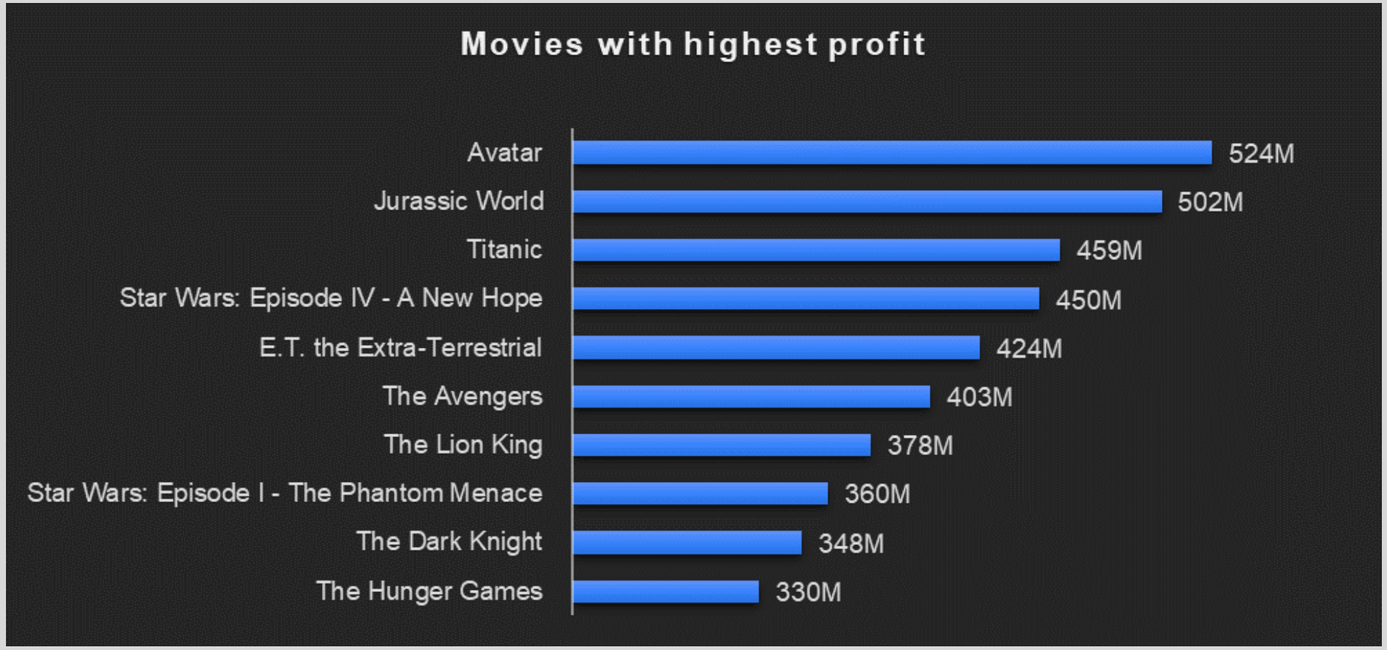
INSIGHTS


IMDb
MOVIE ANALYSIS
Movies with highest profit
250M
350M
450M
550M
Avatar
Jurassic World
Titanic
Star Wars: Episode IV - A New...
E.T. the Extra-Terrestrial
The Avengers
The Lion King
Star Wars: Episode I - The...
The Dark Knight
The Hunger Games
Top 5 popular genres with average IMDB rating
Crime|Drama|Fantasy|Mystery
8.5
Adventure|Animation|Drama|Family|Musical
8.5
Adventure|Drama|Thriller|War
8.4
Action|Adventure|Drama|Fantasy|War
8.4
Adventure|Animation|Fantasy
8.4
Top directors according to IMDB rating
8.6
8.6
8.5
8.5
8.5
8.4
8.4
8.4
8.4
8.4
8.4
Tony Kaye
Charles Chaplin
Damien Chazelle
Ron Fricke
Alfred Hitchcock
Sergio Leone
Christopher Nolan
Asghar Farhadi
S.S. Rajamouli
Richard Marquand
Marius A. Markevicius
Critique favourite actors
576
596
654
738
750
Bérénice Bejo
Craig Stark
Peter Capaldi
Phaldut Sharma
Albert Finney
Audience favourite actors
1842
1885
2789
2814
3400
Orlando Bloom
Phaldut Sharma
Steve Bastoni
Christo Jivkov
Heather Donahue
Top 5 countried with highest gross
168152M
10417M
2290M
1950M
1564M
USA
UK
Germany
France
Canada
Votes users over decades
180M
160M
140M
120M
100M
80M
60M
40M
20M
100K
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s

Movies with the highest profit

INSIGHTS (CONTD...)

- As per our dataset, **Avatar** has made the highest profit of **524M.**
- The below bar chart has depicted the list of top 10 movies with acsending order of their profit.
- To conduct this analysis I used Pivot tables as it is easier to quickly analyze a dataset and find some insights.



IMDB TOP 250 MOVIES

- The popular American Drama **Shawshank redemption** is the highest rated (9.3) IMDB film of all time.
- The data in the right side also shows the entire list of top 250 movies based on IMDB ratings.
- To prepare this column firstly I sorted the IMDB_score column in descending order. Then, I created this new column and used the following excel formula to extract the top 250:

=IF(N2>25000,B2," ")

IMDB_Top_250	Rank
The Shawshank Redemption	1
The Godfather	2
The Dark Knight	3
The Godfather: Part II	4
The Lord of the Rings: The Return o	5
Pulp Fiction	6
Schindler's List	7
The Good, the Bad and the Ugly	8
Forrest Gump	9
Star Wars: Episode V - The Empire	10
The Lord of the Rings: The Fellowsh	11
Inception	12
Fight Club	13
Star Wars: Episode IV - A New Hope	14
The Lord of the Rings: The Two Tov	15
The Matrix	16
One Flew Over the Cuckoo's Nest	17
Goodfellas	18
City of God	19
Seven Samurai	20
Saving Private Ryan	21
The Silence of the Lambs	22
Se7en	23
Interstellar	24
The Usual Suspects	25
American History X	26
Modern Times	27
Spirited Away	28
The Lion King	29
Raiders of the Lost Ark	30
The Dark Knight Rises	31
Back to the Future	32
Terminator 2: Judgment Day	33
Gladiator	34
The Green Mile	35
Django Unchained	36
Apocalypse Now	37
The Departed	38
Psycho	39
Memento	40
The Prestige	41
Whiplash	42
The Lives of Others	43
The Pianist	44
Star Wars: Episode VI - Return of th	45
American Beauty	46
Aliens	47
WALL·E	48
A Separation	49
Braveheart	50

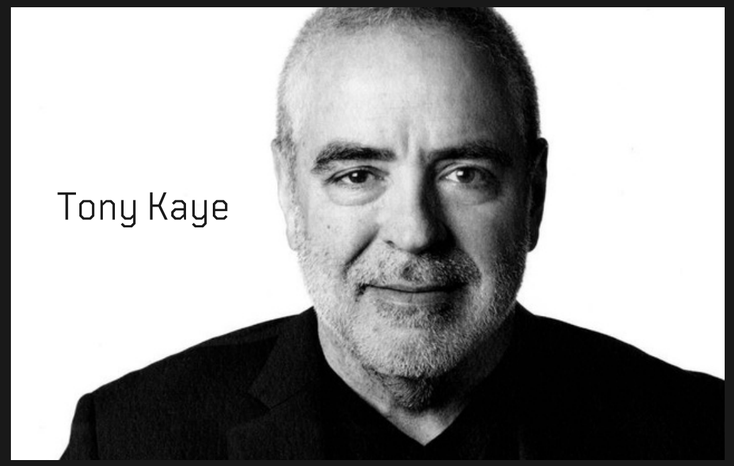
IMDB_Top_250	Rank
Reservoir Dogs	51
Oldboy	52
Requiem for a Dream	53
Das Boot	54
Lawrence of Arabia	55
Baahubali: The Beginning	56
Once Upon a Time in America	57
Amélie	58
Princess Mononoke	59
Toy Story 3	60
Inside Out	61
Toy Story	62
The Sting	63
Indiana Jones and the Last Crusade	64
Good Will Hunting	65
Up	66
Unforgiven	67
Batman Begins	68
Inglourious Basterds	69
2001: A Space Odyssey	70
Amadeus	71
L.A. Confidential	72
Snatch	73
Some Like It Hot	74
Scarface	75
Eternal Sunshine of the Spotless Mind	76
Room	77
Monty Python and the Holy Grail	78
The Hunt	79
Metropolis	80
Downfall	81
Raging Bull	82
Finding Nemo	83
Gone with the Wind	84
Captain America: Civil War	85
Gran Torino	86
A Beautiful Mind	87
Die Hard	88
How to Train Your Dragon	89
The Bridge on the River Kwai	90
Pan's Labyrinth	91
The Secret in Their Eyes	92
The Wolf of Wall Street	93
V for Vendetta	94
Trainspotting	95
On the Waterfront	96
Into the Wild	97
Lock, Stock and Two Smoking Barrels	98
The Big Lebowski	99
Incendies	100

IMDB_Top_250	Rank
Blade Runner	101
The Thing	102
Casino	103
Warrior	104
Howl's Moving Castle	105
The Avengers	106
Deadpool	107
Jurassic Park	108
The Sixth Sense	109
Monsters, Inc.	110
Pirates of the Caribbean: The Curse of the	111
Guardians of the Galaxy	112
The Help	113
Platoon	114
The Martian	115
The Bourne Ultimatum	116
Rocky	117
Gone Girl	118
Butch Cassidy and the Sundance Kid	119
The Imitation Game	120
Million Dollar Baby	121
The Truman Show	122
Groundhog Day	123
No Country for Old Men	124
The Revenant	125
Shutter Island	126
Stand by Me	127
Kill Bill: Vol. 1	128
12 Years a Slave	129
Annie Hall	130
Sin City	131
The Grand Budapest Hotel	132
The Terminator	133
Spotlight	134
The Best Years of Our Lives	135
The Wizard of Oz	136
There Will Be Blood	137
Prisoners	138
The Princess Bride	139
Hotel Rwanda	140
Mad Max: Fury Road	141
Amores Perros	142
Before Sunrise	143
The Celebration	144
Donnie Darko	145
Elite Squad	146
The Sea Inside	147
Rush	148
Tae Guk Gi: The Brotherhood of War	149
Akira	150

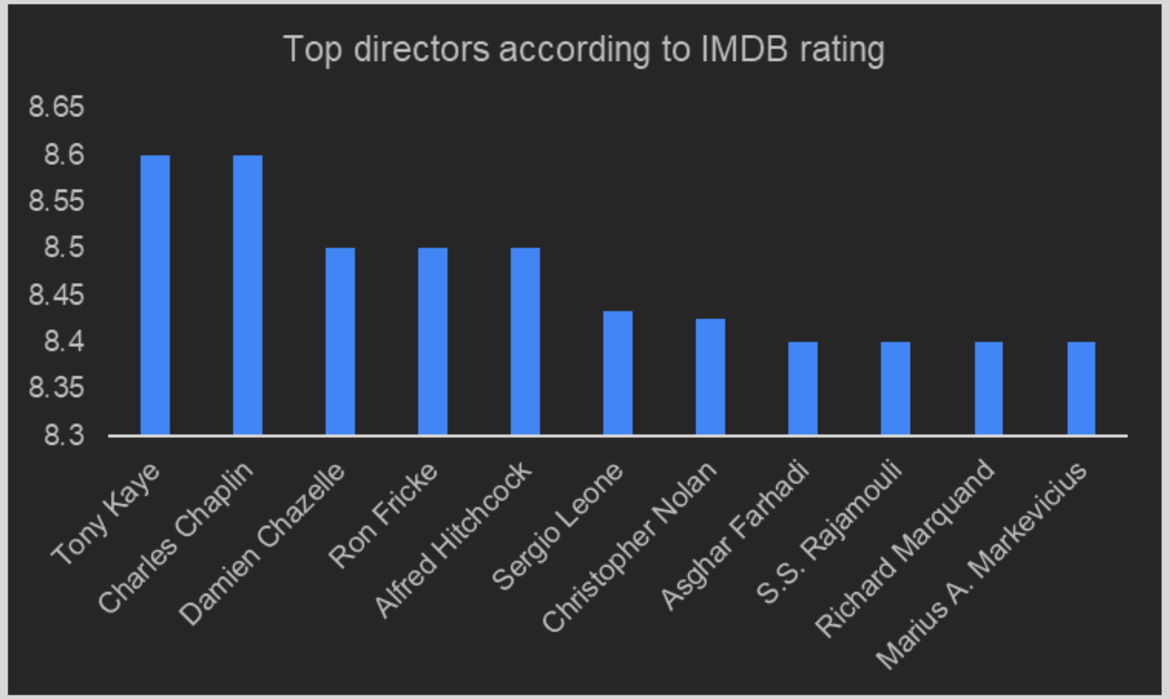
IMDB_Top_250	Rank
Jaws	151
The Exorcist	152
Aladdin	153
The Incredibles	154
Dances with Wolves	155
The Sound of Music	156
Rain Man	157
Slumdog Millionaire	158
The King's Speech	159
Catch Me If You Can	160
Star Trek	161
The Pursuit of Happyness	162
Doctor Zhivago	163
Black Swan	164
District 9	165
Young Frankenstein	166
Dead Poets Society	167
Mystic River	168
Ratatouille	169
Fiddler on the Roof	170
Kill Bill: Vol. 2	171
X-Men: Days of Future Past	172
JFK	173
The Artist	174
Sling Blade	175
Dallas Buyers Club	176
Boyhood	177
Bowling for Columbine	178
Casino Royale	179
Sicko	180
Shaun of the Dead	181
Life of Pi	182
The Perks of Being a Wallflower	183
A Fistful of Dollars	184
Before Sunset	185
Central Station	186
Her	187
Waltz with Bashir	188
True Romance	189
Persepolis	190
Big Fish	191
The Straight Story	192
Brazil	193
In Bruges	194
Mulholland Drive	195
My Name Is Khan	196
Dancer in the Dark	197
Magnolia	198
Serenity	199
Cinderella Man	200

IMDB_Top_250	Rank
Blood Diamond	201
The Iron Giant	202
Avatar	203
E.T. the Extra-Terrestrial	204
Shrek	205
Iron Man	206
Toy Story 2	207
Straight Outta Compton	208
The Hobbit: An Unexpected Journey	209
Taken	210
Crouching Tiger, Hidden Dragon	211
Walk the Line	212
The Fighter	213
The Bourne Identity	214
Big Hero 6	215
My Fair Lady	216
Captain Phillips	217
Little Miss Sunshine	218
The Untouchables	219
Crash	220
Halloween	221
Edward Scissorhands	222
The Hobbit: The Desolation of Smau	223
How to Train Your Dragon 2	224
The Blues Brothers	225
Nightcrawler	226
Do the Right Thing	227
The Wrestler	228
Hot Fuzz	229
The Remains of the Day	230
Boogie Nights	231
The Hateful Eight	232
Once	233
Glory	234
Before Midnight	235
4 Months, 3 Weeks and 2 Days	236
Moon	237
Nine Queens	238
The Chorus	239
Veer-Zaara	240
Letters from Iwo Jima	241
The Right Stuff	242
Amour	243
Ed Wood	244
The World's Fastest Indian	245
Almost Famous	246
The Notebook	247
Hero	248
The Insider	249
Children of Men	250

THE BEST DIRECTORS

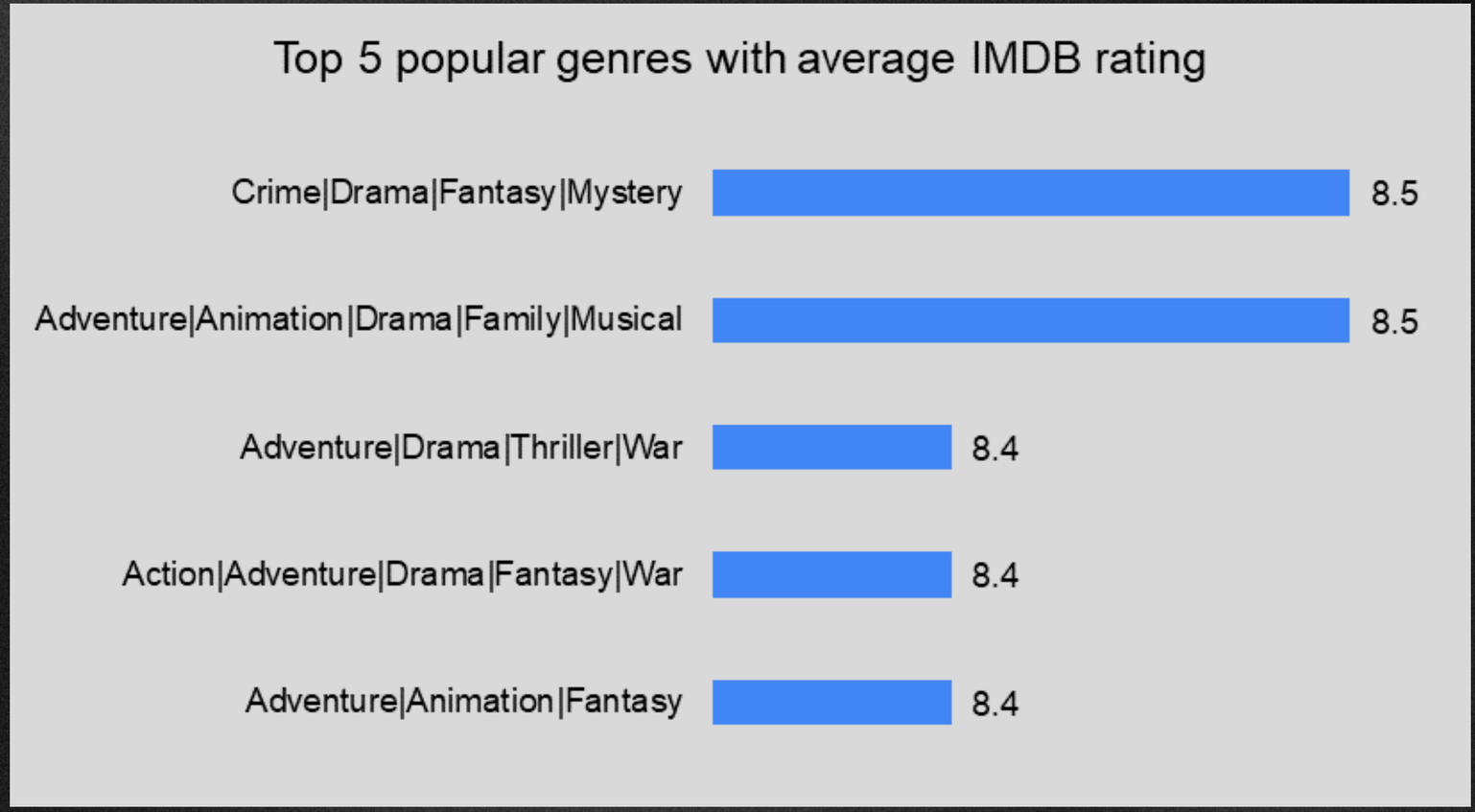


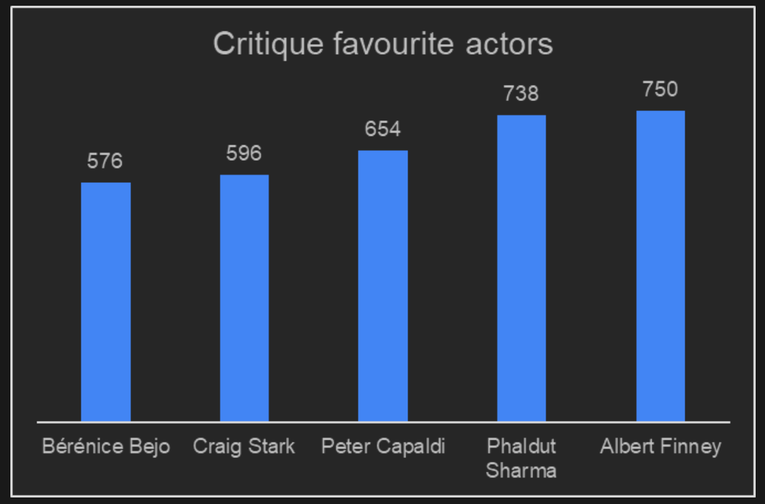
- As we can see from the below column chart, the top 2 directors according average rating of IMDB are **Tony Kaye and the famous Charles Chaplin.**
- The list of top 10 also includes some famous directors such as - Alfred Hitchcock, Damien Chazelle, Ron Fricke, Sergio Leone, Asghar Farhadi, Christopher Nolan, Richard Marquand, our Indian director S.S. Rajamouli and Marius A. Markevicius
- I used pivot tables to find this insight as it is easy to group some data and find their corresponding data using pivot tables.

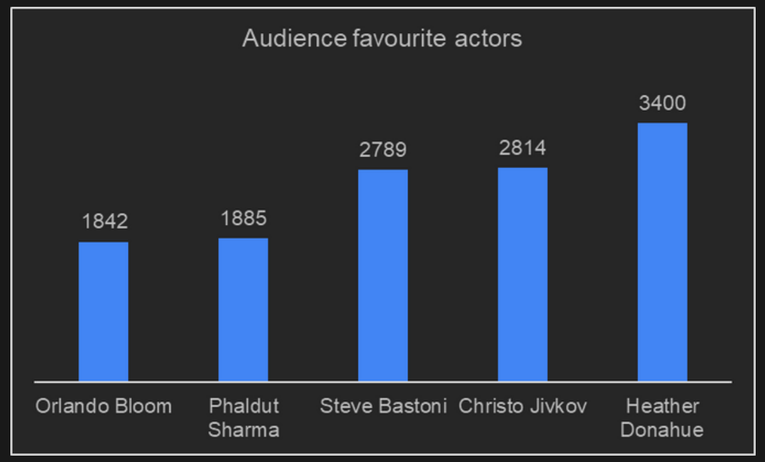


TOP 5 POPULAR GENRES

- According to our findings, the most popular genre as per average IMDB rating is **Crime-Drama-Fantasy-Mystery.** This genre is getting an average of 8.4 IMDB rating.
- Apart from that - Adventure, Animation, Family, War, Fantasy movies are also getting high rating (8.4 or above
- For this analysis also I used Pivot tables to group the genres and sorted them according to average IMDB scores.







CRITIC-FAVORITE AND AUDIENCE-FAVORITE ACTORS

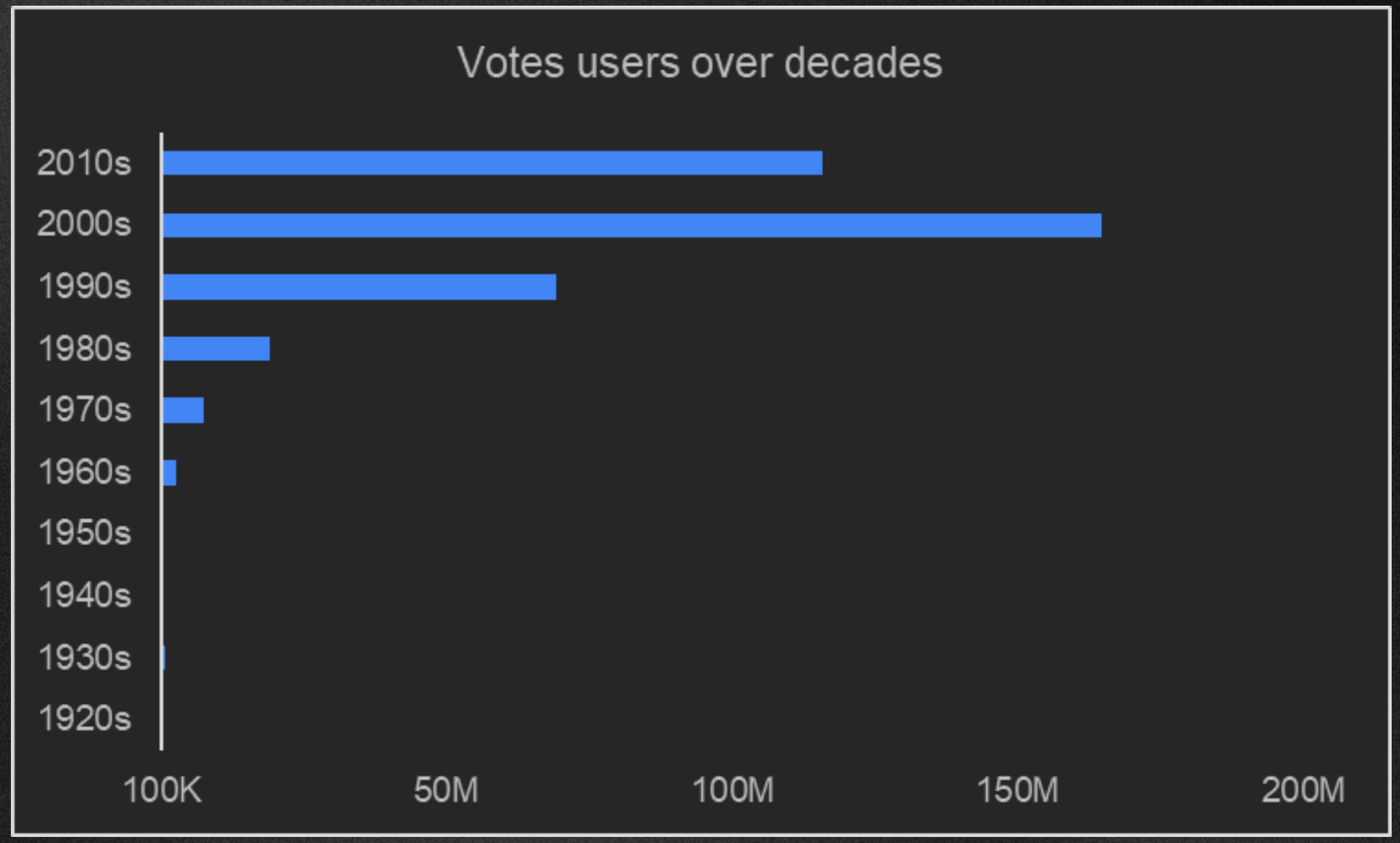
- The two coloumn charts in the right side shows the top 5 critique favourite and top 5 audience favourite actors.
- We can see that, the list is completely different we consider critiques and audiences. According to critiques, Albert Finney got the highest average number of votes of 750. On the other hand, according to audience Heather Donahue got the highest number of mean votes of 3400 votes.
- For this analysis I used Pivot tables and sorting to find insights.

VOTES PATTERN IN DECADES

- We can see that, the number of votes is gradually increasing in every decade. In 2000s decade, IMDB got the highest number of votes on Movies.
- The main reason can be the easy access and the rising use of Internet over the decades. It is expected that the 2020s decade will got more votes from audiences.
- To prepare the column of decades I used the following excel function to extract the values:

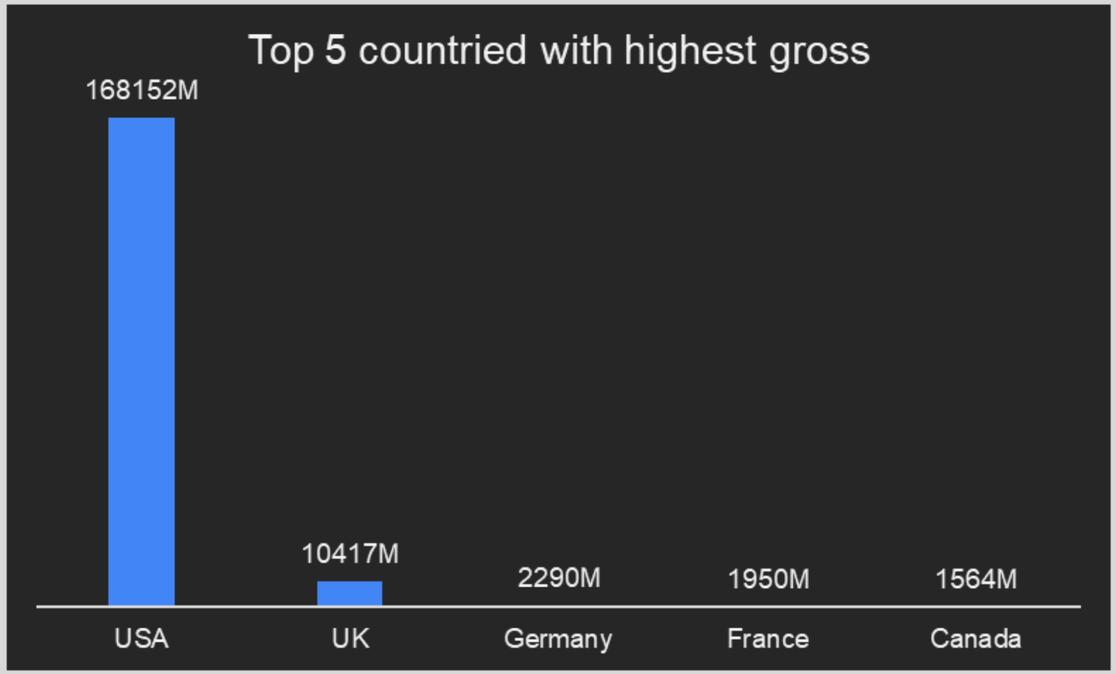
=LEFT(D2,3)&"0s" (where D2 is the year column)

- Once I got the decade column I used Pivot tables and Pivot chart to find the insighrs.



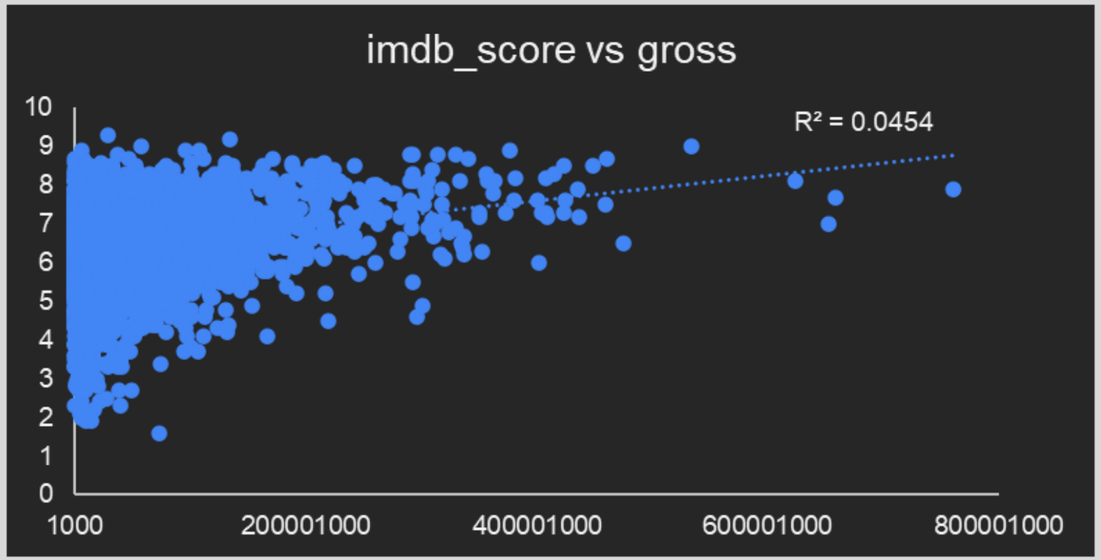
TOP 5 COUNTRIES WITH HIGHEST GROSS

- We all know that, USA has the biggest film studio in the world, for example - Hollywood. Over the years, Hollywood is giving us excellent movies in terms of IMDB ratings and grossing.
- If we consider highest grossing countries, USA wins the game significantly compared to the other countries.
- As shown in the column chart below USA earns way more than the top movies producing countries like UK, Germany, France and Canada.



RELATIONSHIP BETWEEN IMDB RATINGS AND GROSS REVENUE

- The scatter plot is often useful to find relationship between two variables and also to find outliers. The scatter plot between IMDB ratings and Gross revenue provides a general overview that there is no specific trend of higher IMDB rated movies getting higher revenue. To make this test more statistically significant, we performed regression test.
- Regression test is most useful when we have two numerical type variables like in our case. As we can see we got a very high P value. As P value > 0.05, we can accept the **Null Hypothesis (H0)** for this analysis.
- **Therefore, there is no significant relationship between IMDB ratings and gross revenue.**



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.213087
R Square	0.045406
Adjusted R Squ	0.045152
Standard Error	67091592
Observations	3759

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8.04E+17	8.04E+17	178.7043	7.46483E-40
Residual	3757	1.69E+19	4.5E+15		
Total	3758	1.77E+19			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	-4E+07	6816884	-5.8275	6.1E-09	-53090515.2	-2.6E+07	-5.3E+07	-26360211.1
X Variable 1	13915041	1040919	13.36803	7.46E-40	11874219.82	15955862	11874220	15955862.23

RELATIONSHIP BETWEEN LANGUAGE AND IMDB RATING

- The Anova test is best for statistical analysis on two variables where one variable is numerical and the other one is categorical like in this case.
- From the Anova test between the two variables Language and IMDB rating,, we found the P value as absolute 0. It indicates that the P value is significantly lower than 0.01 and its negligible for which excel has returned absolute 0. That means, our result is highly significant statistically.
- As P<0.05, we can reject the null hypothesis.
- Therefore, **there is significant different between IMDB ratings in terms of language of movies.**

H0: There is no difference in the average IMDb ratings among movies in different languages.
HA: The average IMDb ratings differ among movies in different languages.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	3758	24289.6	6.463438	1.104657
Column 2	3758	42766	11.37999	6.534835

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	45420.08	1	45420.08	11890.87	0.00	3.842696691
Within Groups	28701.57	7514	3.819746			
Total	74121.66	7515				

RESULTS AND CONCLUSION

- This project was very interesting for me and it was also very useful to strengthen my data analysis skills using Advance Excel.
- It also helped me to develop skills in statistical analysis and hypothesis testing methods.
- From this project we got the list of the Top 250 highest-rated IMDB movies.
- We found that Avatar made the highest profit of 524 Million dollars.
- Top director who have provided movies with more than an average rating of 8.4 are Charles chaplin and Tony Kaye.
- We also came to know that Crime, Drama, Fantasy, Mystry is the most popular genre.
- The number of Votes in IMDB is increasing every decade which is becoming very helpful for audience to choose high rated movies to watch.


THANK YOU!