IMDb
Movie Analysis

TABLE OF CONTENTS

01 PROJECT DESCRIPTION

Project Description:

- **Data description:**
  This project is done using IMDB_Movies dataset containing 28 features related to different aspects of 1 IMDb movie.
- **Expected Project Deliverables:**
  i. Project report on the provided tasks.
  ii. Solving a problem after framing it

TASKS 02

I Tasks

1. Data Cleaning
2. Find movies with highest profit
3. Find Top 250 IMDb movies
4. Top 10 directors (based on avg IMDb rating/score)
5. Find popular genres
6. Find critic favourite and audience favourite actors

Tech stack used: MS Office Excel 2007

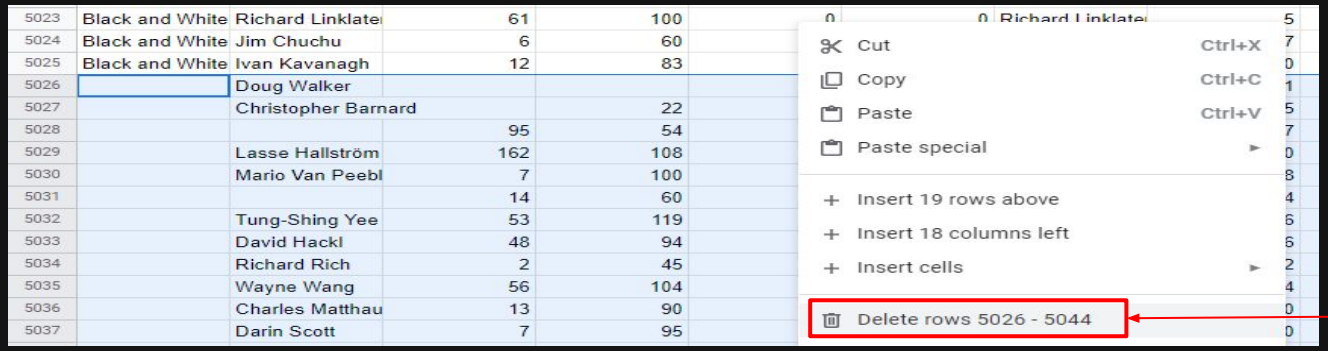
03

Approach & Insights

This section describes upon how the problems/tasks are solved and the results/insights from the tasks.

1. Data Cleaning

- **NULL VALUES:**
  To-Do: Delete each records where values are absent (in each column).
  Approach:
  I. **Make a table:** Select the dataset range and convert to table.
  II. **Group empty records together:** Sort column A-Z OR Z-A.
  III. **Find empty records:** Go to the bottom of the range.
  IV. **Select empty records:** Select the empty records for the respective column.
  V. **Delete empty records:** Right click the selected records and delete entire row.
  VI. **Loop:** Repeat steps I-V for each column.



I 1. Data Cleaning

- **DUPLICATE/REDUNDANT VALUES:**
  To-Do: Edit the table for unique records only.
  Approach:
  > Go to the Data tab.
  > Click remove duplicates button.
  > Check the movie_title box.
  > Click OK.
  > Following dialogue box appears

I 2. Movies with highest profit (Approach)

- **Create new column:**
  - Go to 1 column after the extreme right column(in first row).
  - Enter ‘Profit’ as a header.

- **Enter formula:** under the ‘Profit’ cell type the following text in red **=I2-W2**

  where = : to initiate a formula

  I2 : 1st record of Gross(movie earning),

  W2 : 1st record of Budget

- **Copy formula to entire column:**

  Double click at the bottom right of the formula applied cell to copy down the formula to the entire column.

I 2. Movies with highest profit (Insights)

Row Labels	profit
Avatar	523,505,847
Deadpool	305,024,263
E.T. the Extra-Terrestrial	424,449,459
Jurassic World	502,177,271
Star Wars: Episode I - The Phantom Menace	359,544,677
Star Wars: Episode IV - A New Hope	449,935,665
The Avengers	403,279,547
The Dark Knight	348,316,061
The Hunger Games	329,999,255
The Lion King	377,783,777
Titanic	458,672,302

* Link to the full list of movies with highest profit.

I 3. IMDb Top 250 (Approach)

- Make new worksheet with existing records of:
  - i. Movie_title (A)
  - ii. Num_voted_users (B)
  - iii. Imdb_score (C)

- Insert two new empty columns namely:
  - i. Movie_title
  - ii. Imdb_score

- Enter formula under the empty records:
  - =IF(B2>25000,A2) for movie_title
  - =IF(B2>25000,C2) for imdb_score

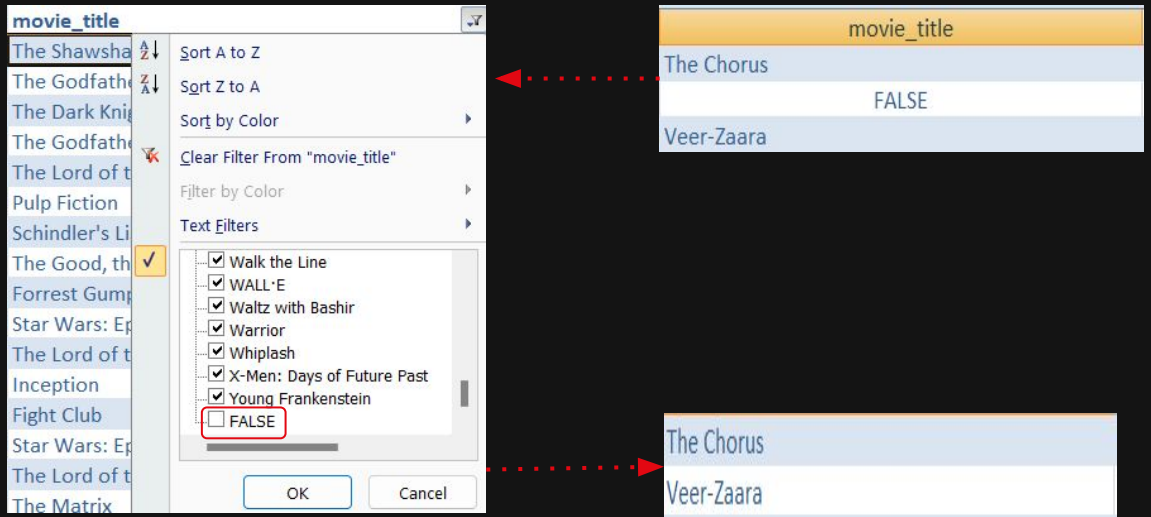
	A	B	C
1	movie_title	num_voted_users	imdb_score
2	The Shawshank Redemption	1689764	9.3
3	The Godfather	1155770	9.2
4	The Dark Knight	1676169	9

movie_title	imdb_score

	A	B
1	movie_title	imdb_score
2	The Shawshank Redemption	9.3
3	The Godfather	9.2
4	The Dark Knight	9

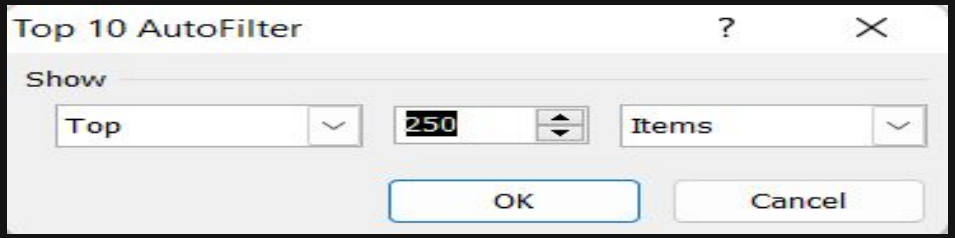
I 3. IMDb Top 250 (Approach)

- In the text filters, uncheck FALSE text under movie_title.



I 3. IMDb Top 250 (Approach)

- Apply a number filter of Top 10, edit it to Top 250 for imdb_score column.



- The above filter ensures there's only movies with num_voted_users>25000 in the range.

- Put a rank column next to the extreme right for S.No.

Insights

movie_title	imdb_score	rank
The Shawshank Redemption	9.3	1
The Godfather	9.2	2
The Dark Knight	9	3
The Godfather: Part II	9	4
The Lord of the Rings: The Return of the King	8.9	5
Pulp Fiction	8.9	6
Schindler's List	8.9	7
The Good, the Bad and the Ugly	8.9	8
Forrest Gump	8.8	9
Star Wars: Episode V - The Empire Strikes Back	8.8	10

*view full list here

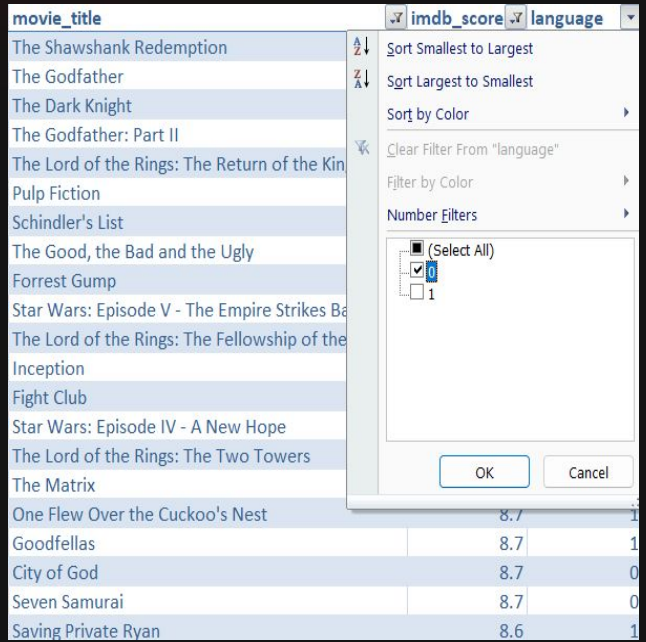
3. Top Movies in languages other than english

Approach:

- Add a third column “language” to the top 250 movies table.
- Enter the following formula under the language column

  `=IF(AND(Table1[[#This Row],[movie_title]]='working data'!$L$2:$L$3790,'top 250 movies'!$D$2:$D$266="English"),1,0)`

  to output a binary value for the movie language if English.
- Apply filter to language column and deselect 1 to edit the data frame for movies other than English only.



| 3. Top Movies in languages other than english

INSIG HTS

rank	movie_title	imdb_score
1	The Good, the Bad and the Ugly	8.9
2	City of God	8.7
3	Seven Samurai	8.7
4	Spirited Away	8.6
5	The Lives of Others	8.5
6	Children of Heaven	8.5
7	A Separation	8.4
8	Oldboy	8.4
9	Das Boot	8.4
10	Baahubali: The Beginning	8.4

*view full list here

4. Top 10 Directors (Approach)

- **Create pivot table:**
  > Insert pivot table to the data range.
- **Add row labels:**
  > Drage director_name to row labels under field list.
- **Add a value column:**
  > Drag imdb_score to values field under field list.
  > Select average of values in value field settings.
- **Sorting:**
  > Sort the imdb_score column Z-A/descending.
- **Value filter:**
  > Apply value filter to movie_title.
  > Select Top 10, click OK.

Insights

director_name	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4

4. Top 10 Directors (Approach)

- Convert pivot table to a table.
- Custom sort the data in the following order:

  I. imdb_score by descending order.

  II. director_name by ascending order.
- Add a rank column to extreme right:

  Enter ‘1,2’ to first two columns and 2x click on bottom right of the cell to copy sequence.
- Apply Top10 filter to the table.

Insights

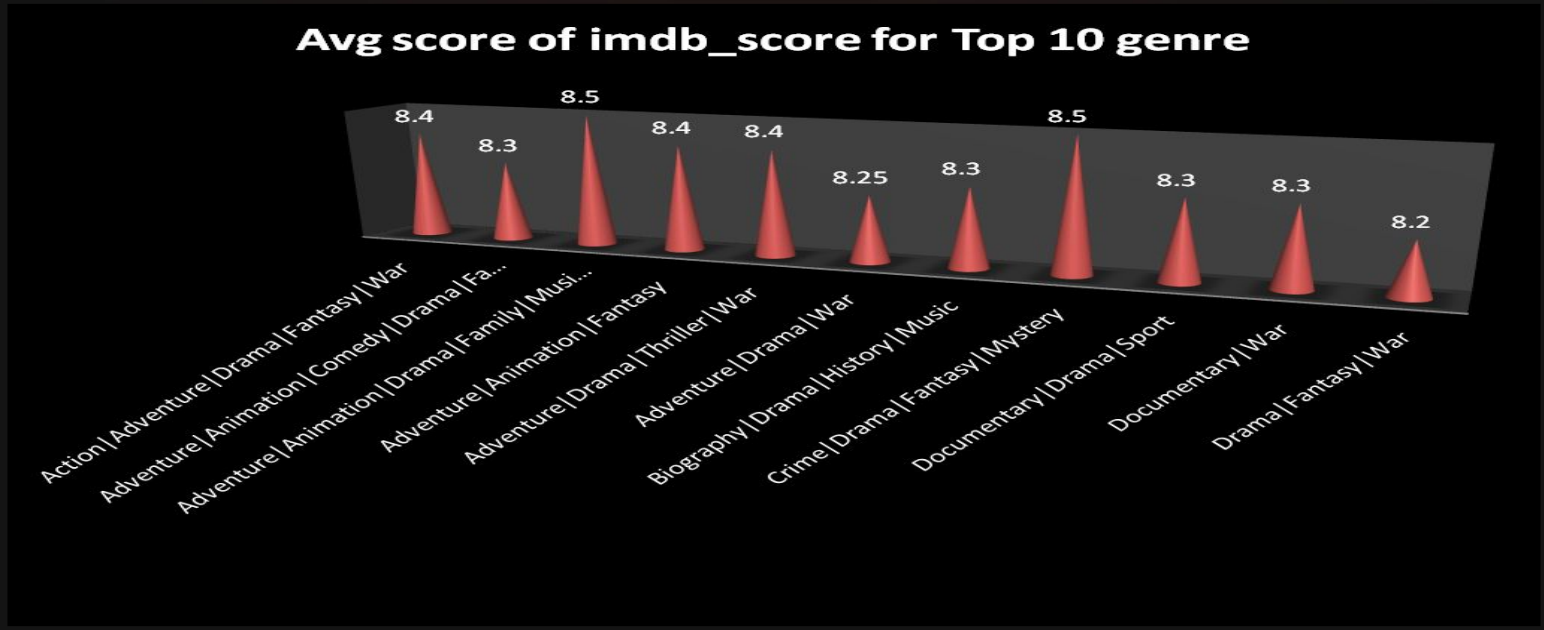
top10 directors	imdb_score	rank
Tony Kaye	8.6	1
Charles Chaplin	8.6	2
Alfred Hitchcock	8.5	3
Ron Fricke	8.5	4
Damien Chazelle	8.5	5
Majid Majidi	8.5	6
Sergio Leone	8.433333333	7
Christopher Nola	8.425	8
S.S. Rajamouli	8.4	9
Richard Marquai	8.4	10

5. Popular genres(Approach)

- Get the name of all the distinct genre records:
  On the genre column, apply the advanced filter and select unique records only.
- Use the AVERAGEIF function to get the average imdb_score of each genre:
  =AVERAGEIF(Table2[genres],D2,Table2[imdb_score])
- Copy the formula to each cell and then sort the avgscore column by Z-A.

genres	avgscore
Adventure|Animation|Drama|Family|Musical	8.500
Crime|Drama|Fantasy|Mystery	8.500
Adventure|Drama|Thriller|War	8.400
Action|Adventure|Drama|Fantasy|War	8.400
Adventure|Animation|Fantasy	8.400
Adventure|Animation|Comedy|Drama|Family|F	8.300
Biography|Drama|History|Music	8.300
Documentary|Drama|Sport	8.300

5. Popular genres(Insights)



popular genres

I 6. Favourite actors

Brad Pitt	Leonardo DiCaprio	Meryl Streep
Babel	Blood Diamond	A Prairie Home Companion
By the Sea	Body of Lies	Hope Springs
Fight Club	Catch Me If You Can	It's Complicated
Fury	Django Unchained	Julie & Julia
Interview with the Vampire:	Gangs of New York	Lions for Lambs
Killing Them Softly	Inception	One True Thing
Mr. & Mrs. Smith	J. Edgar	Out of Africa
Ocean's Eleven	Marvin's Room	The Devil Wears Prada
Ocean's Twelve	Revolutionary Road	The Hours
Seven Years in Tibet	Romeo + Juliet	The Iron Lady
Sinbad: Legend of the Sever	Shutter Island	The River Wild
Spy Game	The Aviator	
The Assassination of Jesse J	The Beach	
The Curious Case of Benjam	The Departed	
The Tree of Life	The Great Gatsby	
Troy	The Man in the Iron Mask	
True Romance	The Quick and the Dead	
	The Revenant	
	The Wolf of Wall Street	
	Titanic	

Apply filter to actor_1_name = 'Brad Pitt' and copy paste to column T

Combined

- **Brad Pitt**
  - Babel
  - By the Sea
  - Fight Club
  - Fury
  - Interview with the Vampire: The Vampire Chronicles
  - Killing Them Softly
  - Mr. & Mrs. Smith
  - Ocean's Eleven
  - Ocean's Twelve
  - Seven Years in Tibet
  - Sinbad: Legend of the Seven Seas
  - Spy Game
  - The Assassination of Jesse James by the Coward Robert Ford
  - The Curious Case of Benjamin Button
  - The Tree of Life
  - Troy
  - True Romance
- **Leonardo DiCaprio**
  - Blood Diamond
  - Body of Lies
  - Catch Me If You Can
  - Django Unchained
  - Gangs of New York
  - Inception
  - J. Edgar
  - Marvin's Room
  - Revolutionary Road
  - Romeo + Juliet
  - Shutter Island
  - The Aviator
  - The Beach
  - The Departed
  - The Great Gatsby
  - The Man in the Iron Mask
  - The Quick and the Dead
  - The Revenant
  - The Wolf of Wall Street
  - Titanic
- **Meryl Streep**
  - A Prairie Home Companion
  - Hope Springs
  - It's Complicated
  - Julie & Julia
  - Lions for Lambs
  - One True Thing
  - Out of Africa

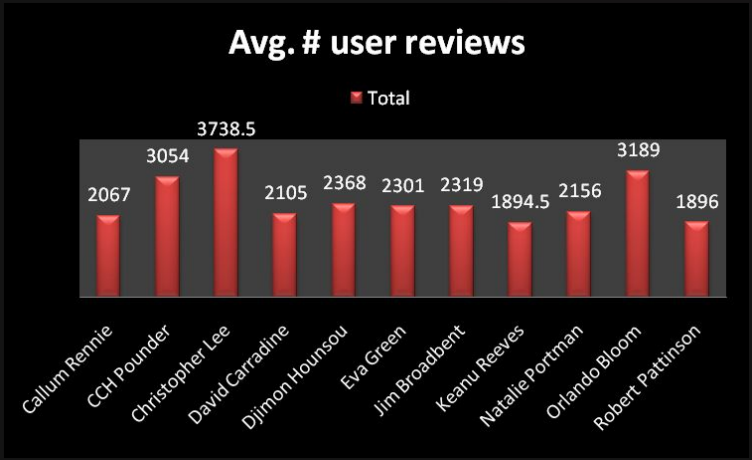
6. Critic-Favourite actors (mean of critic reviews)

- **Create pivot table:**
  On columns:
1. actor_1_name
2. Average of num_crtitc_reviews

- **Create pivot chart:**
  On pivot table for first 10 rows.

*Click to view the full table.

6. Audience-Favourite actors (mean of user reviews)



- **Create pivot table:**
  On columns:
1. actor_1_name
2. Average of num_user_for_reviews

- **Create pivot chart:**
  On pivot table for first 10 rows.

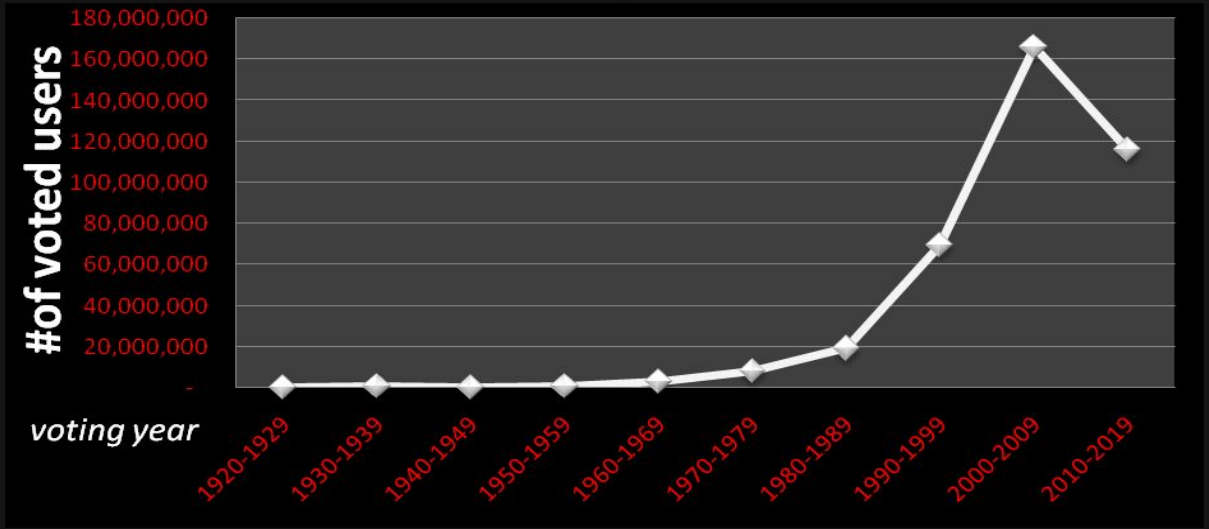
*Click to view the full table.

Timeline analysis (number of voted users)

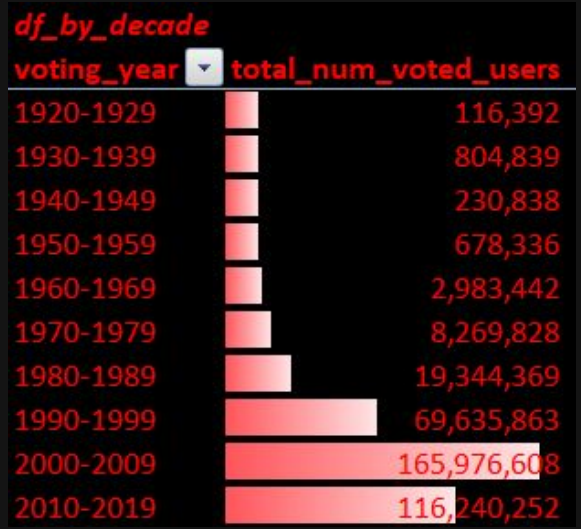
- Columns used: title_year, num_voted_users.
- Convert into pivot tables.
- Group the title_year column by 10 to group the column by decade.
- Make a line pivot chart of the following table.
- Check the following link for the insights:
  number of voted users (timeline analysis)

Timeline analysis (number of voted users by decade)

Line chart analysis

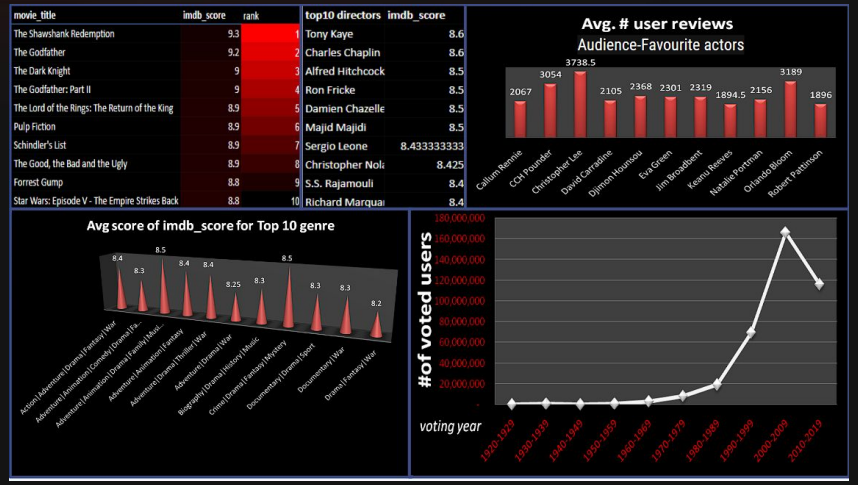


Bar chart analysis



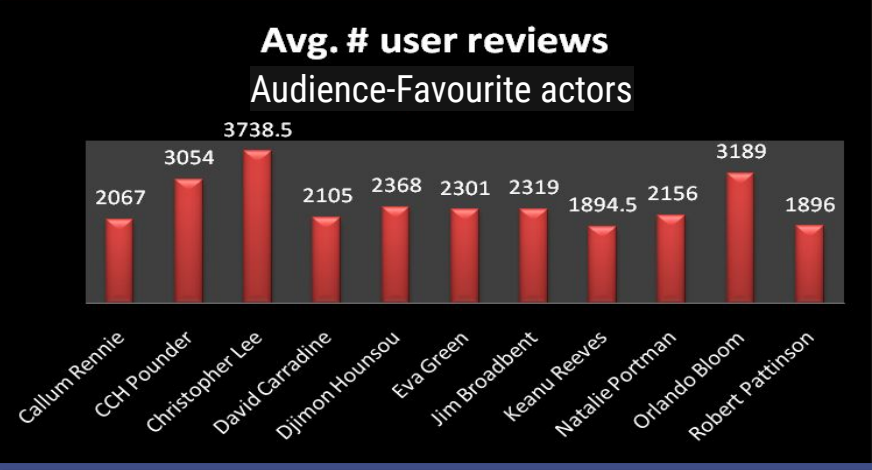
KPIs

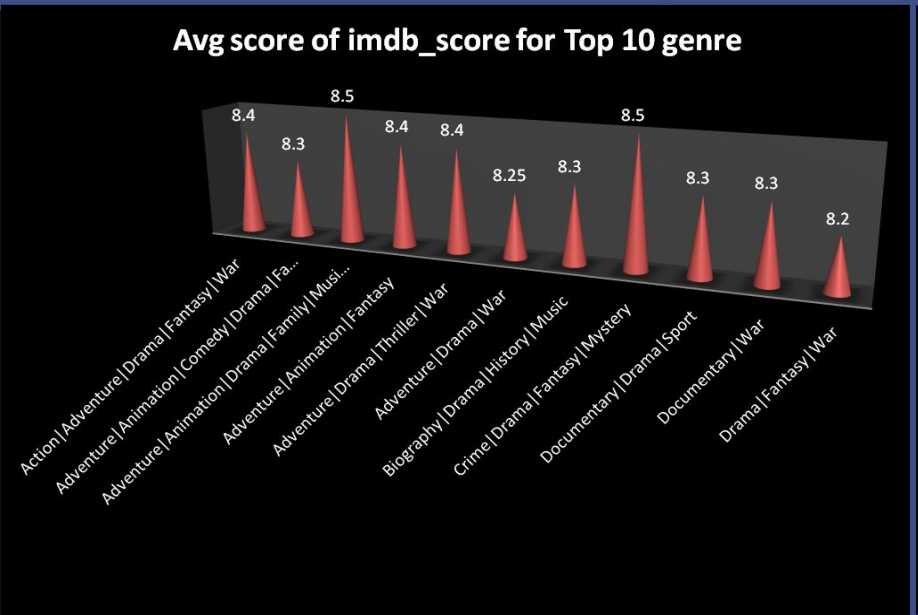
DASH BOARD

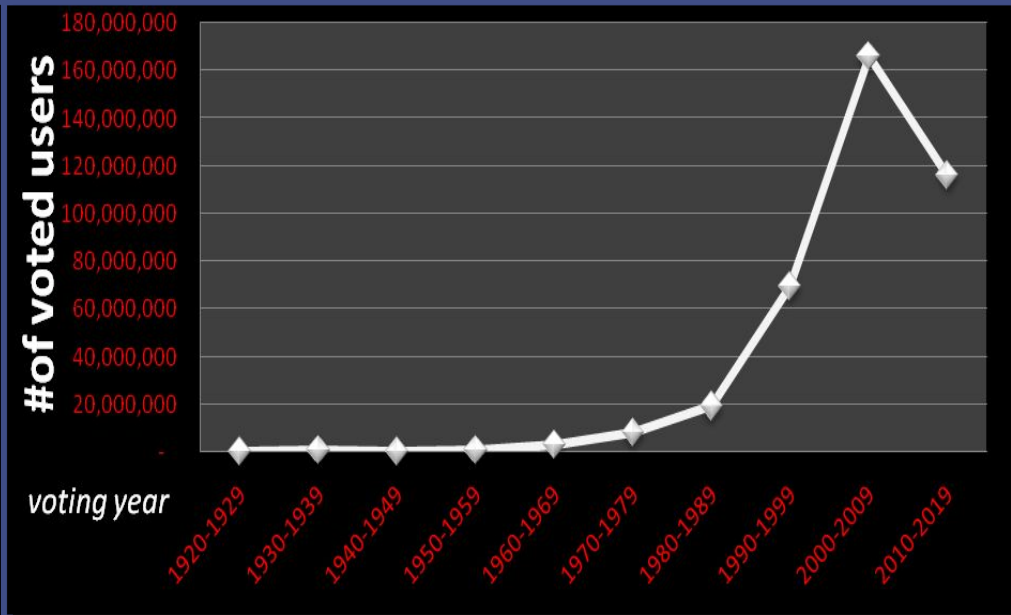


movie_title	imdb_score	rank
The Shawshank Redemption	9.3	1
The Godfather	9.2	2
The Dark Knight	9	3
The Godfather: Part II	9	4
The Lord of the Rings: The Return of the King	8.9	5
Pulp Fiction	8.9	6
Schindler's List	8.9	7
The Good, the Bad and the Ugly	8.9	8
Forrest Gump	8.8	9
Star Wars: Episode V - The Empire Strikes Back	8.8	10

top10 directors	imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nola	8.425
S.S. Rajamouli	8.4
Richard Marquai	8.4







04 RESULT

In conclusion, this project involved performing various data cleaning and analysis tasks on a movie dataset using Excel. The data was cleaned by dropping columns and removing null values. The movies with the highest profit were identified by creating a new column called profit and plotting profit vs budget to observe outliers. The top 250 movies were selected based on their IMDb rating and num_voted_users criteria and a new column was created for non-English movies called Top_Foreign_Lang_Film.

The best directors were identified by grouping the data by director_name and finding the top 10 directors with the highest mean IMDb score. The popular genres were also identified using the knowledge gained from the previous steps. The actors Meryl Streep, Leonardo DiCaprio, and Brad Pitt were extracted from the data and a new column was created called Combined. The actors were grouped and the mean of num_critic_for_reviews and num_users_for_review was calculated to identify the critic-favorite and audience-favorite actors.

Finally, a bar chart was created to observe the change in number of voted users over decades by creating a new column called decade and grouping the data by decade. The sum of users voted in each decade was calculated and stored in a new data frame called df_by_decade. Overall, the project successfully demonstrated various data cleaning and analysis techniques in Excel to gain insights into the movie dataset.

Drive link: IMDB movie analysis