NAME:ANUSHKA SUR

EMAIL:anushkasur35@gmail.com

IMDB Movie Analysis

Project Description

In order to learn more about movie ratings, genres, top-rated films, directors, cast members, and user reviews, this project will analyse IMDb movie data. We strive to comprehend the trends, preferences, and elements affecting movie ratings on IMDb by conducting a thorough investigation. The Here, we have the IMBD dataset for motion pictures from 1920 to 2010. This includes details on the actors, directors, budget, box office, etc. in the movie. Using Office 365 Excel and the Five Why approach for analytics, we will purge the dataset and determine the answer to the query.

Approach :

I would ensure the data's accuracy and consistency by cleaning and preprocessing it. In doing so, it would be necessary to handle missing values, get rid of duplicates, and standardise formats. Once the data is prepared, I will use my statistical knowledge to carry out numerous calculations and analysis. I'd use Excel formulas to calculate the necessary metrics. I would effectively display the results using data visualisation approaches. we'll using pivot table, various functions, and charts for desired answers for the questions

Tech-Stack Used

The main piece of software I used for this project was Microsoft Excel, namely version 2021. In the business world, Excel is frequently used for data analysis and offers a large range of tools, formulas, and functions that are ideal for analysing hiring patterns. I selected Excel for a number of reasons. First off, it is a well-known and generally accessible technology, making it simpler for hiring department stakeholders to comprehend and make use of the data. The spreadsheet format of Excel makes it simple to manipulate, purge, and alter data. Additionally, it provides a range of statistical formulas and procedures that are essential for drawing conclusions from the data.

Answers

A. **Cleaning the data::** PThis is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

1. **Dropping unnecessary columns.**

   (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)

2. Remove Blank Cell / Null Value.

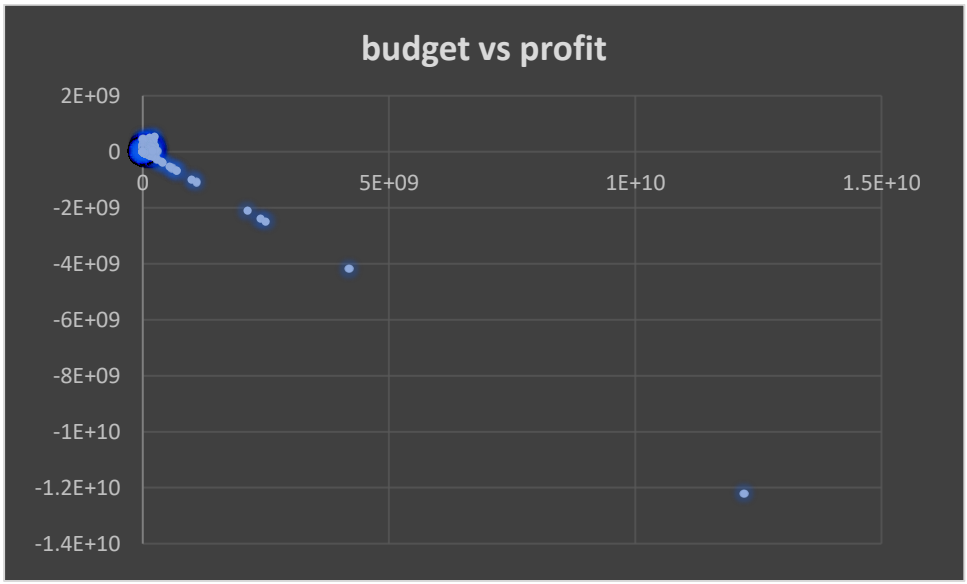
- **Select the column or range of cells that contains the null values.**
- **Go to the "Home" tab in the Excel ribbon.**
- **Click on the "Find & Select" button in the "Editing" group.**
- **From the dropdown menu, select "Go To Special".**
- **In the "Go To Special" dialog box, select the option for "Blanks" and click "OK". This will select all the blank cells in the selected column or range.**
- **Right-click on any of the selected blank cells and choose "Delete" from the context menu.**
- **In the "Delete" dialog box, select the option for "Shift cells up" or "Shift cells left", depending on the orientation of your data, and click "OK". This will delete the selected blank cells and shift the remaining cells up or left to fill the gaps.**
- **Save your modified Excel file.**

3. Removing Duplicate.

(Link to excel data):https://1drv.ms/x/s!AhPlDIVl_coYgSYytJ4syiztzoKe?e=vcUlPa

B. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.
Your task: Find the movies with the highest profit?

SCATTERPLOT:-



OUTLIERS:-

-12213298588 -4199788333 -2499804112 -2397701809 -2127109510

Top 5 Profitable Movies: SORT THE PROFIT ROW FROM LARGEST TO SMALLEST TO FIND OUT TOP 5 MOVIES

	A	B	C	D	E
	movie_title	gross	budget	PROFIT	
	Avatar	760505847	237000000	523505847	
	Titanic	658672302	200000000	458672302	
	Jurassic World	652177271	150000000	502177271	
5	Star Wars: Episode IV - A New Hope	460935665	11000000	449935665	
9	E.T. the Extra-Terrestrial	434949459	10500000	424449459	
0					

THEY ARE AVATAR,JURASSIC WORLD,TITANIC,STARS WARS:EPISODE1V-A NEW HOPE , E.T THE EXTRA-TERRESTIAL

Link to EXCEL data- https://1drv.ms/x/s!AhPlDIV1_coYgSYcs9JB9XoYKiVR?e=NC750J

Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score).

Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

Top 250 Movies:

1.Filter out data where num_voted_users > 25,000 using filter.

2. Sort the data using imbd_score column in descending order

3. Use first 250 entry for our analysis.

4. Give rank using ROW()-1 AND FLASH FILL .

When all Language is considered Shawshank Redemption is on the top with rating of 9.3 and list goes on till LITTLE MISS SUNSHINE with rating 7.9.

A	B	C	D	E	F	G
movie_title	language	imdb_score	num_voted_users	RANK	Top 250 movies	TOP FOREIGN LANG FILM
The Shawshank Redemption	English	9.3	1689764	1	The Shawshank Redemption	The Good, the Bad and the Ugly
The Godfather	English	9.2	1155770	2	The Godfather	City of God
The Dark Knight	English	9	1676169	3	The Dark Knight	Seven Samurai
The Godfather: Part II	English	9	790926	4	The Godfather: Part II	Spirited Away
The Lord of the Rings: The Return of the King	English	8.9	1215718	5	The Lord of the Rings: The Return of the King	The Lives of Others
Schindler's List	English	8.9	865020	6	Schindler's List	Children of Heaven
Pulp Fiction	English	8.9	1324680	7	Pulp Fiction	Amélie
The Good, the Bad and the Ugly	Italian	8.9	503509	8	The Good, the Bad and the Ugly	Baahubali: The Beginning
Inception	English	8.8	1468200	9	Inception	Princess Mononoke
The Lord of the Rings: The Fellowship of the Ring	English	8.8	1238746	10	The Lord of the Rings: The Fellowship of the Ring	Das Boot
Fight Club	English	8.8	1347461	11	Fight Club	Oldboy
Forrest Gump	English	8.8	1251222	12	Forrest Gump	A Separation
Star Wars: Episode V - The Empire Strikes Back	English	8.8	837759	13	Star Wars: Episode V - The Empire Strikes Back	Metropolis
The Lord of the Rings: The Two Towers	English	8.7	1100446	14	The Lord of the Rings: The Two Towers	Downfall
The Matrix	English	8.7	1217752	15	The Matrix	The Hunt
Goodfellas	English	8.7	728685	16	Goodfellas	Howl's Moving Castle

- Top 250 Foreign Language Movies:

IN language column DESELECT the English language using filter and sort.

When ENGLISH Language is NOT considered The Good, the Bad and the Ugly is on the top with rating of 8.9 and list goes on till Amour with rating 7.9.

E	F	G
RANK	Top 250 movies	TOP FOREIGN LANG FILM
1	The Shawshank Redemption	The Good, the Bad and the Ugly
2	The Godfather	City of God
3	The Dark Knight	Seven Samurai
4	The Godfather: Part II	Spirited Away
5	The Lord of the Rings: The Return of the King	The Lives of Others
6	Schindler's List	Children of Heaven
7	Pulp Fiction	Amélie
8	The Good, the Bad and the Ugly	Baahubali: The Beginning
9	Inception	Princess Mononoke
10	The Lord of the Rings: The Fellowship of the Ring	Das Boot
11	Fight Club	Oldboy
12	Forrest Gump	A Separation
13	Star Wars: Episode V - The Empire Strikes Back	Metropolis
14	The Lord of the Rings: The Two Towers	Downfall
15	The Matrix	The Hunt
16	Goodfellas	Howl's Moving Castle
17	Star Wars: Episode IV - A New Hope	Pan's Labyrinth
18	One Flew Over the Cuckoo's Nest	Incendies
19	City of God	The Secret in Their Eyes
20	Seven Samurai	The Sea Inside
21	Interstellar	Tae Guk Gi: The Brotherhood of War
22	Saving Private Ryan	Akira
23	Se7en	Elite Squad
24	The Silence of the Lambs	Amores Perros
25	Spirited Away	The Celebration
26	American History X	My Name Is Khan
27	The Usual Suspects	Persepolis
28	Modern Times	Central Station
29	The Dark Knight Rises	Waltz with Bashir
30	Gladiator	A Fistful of Dollars
31	Terminator 2: Judgment Day	Hero
32	Django Unchained	Crouching Tiger, Hidden Dragon
33	The Departed	Letters from Iwo Jima
34	The Lion King	Amour

LINK TO EXCEL SHEET –

https://1drv.ms/x/s!AhPlDIV1_coYgSYcs9JB9XoYKiVR?e=gjIh4B

C. **Best Directors:** Group the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

BY USING PIVOT TABLE:-

TOP 10 DIRECTORS	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

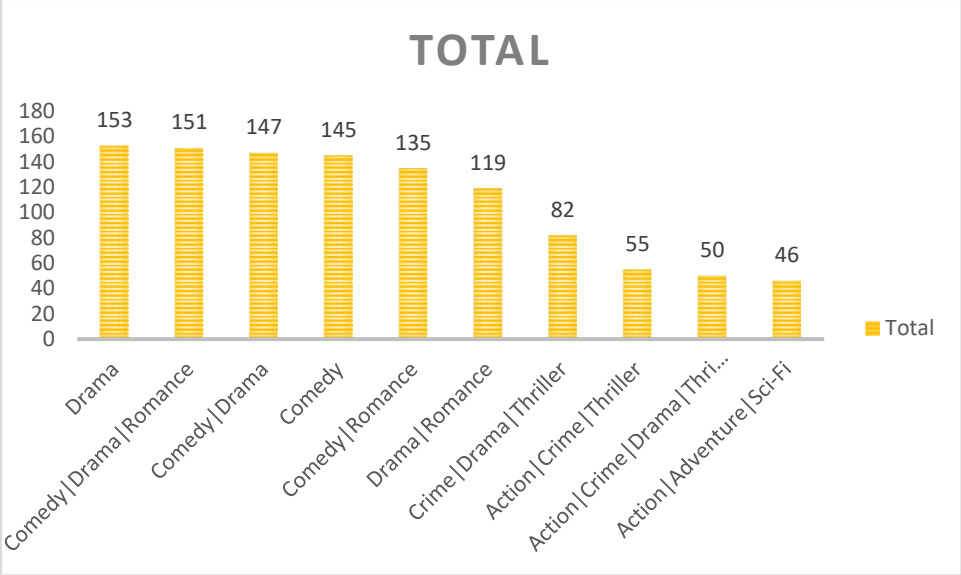
LINK TO EXCEL SHEET-
https://1drv.ms/x/s!AhPlDIVl_coYgSYcs9JB9XoYKiVR?e=QMTBom

D. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
Your

task: Find popular genres

BY USING PIVOT TABLE:-

TOP 10 GENRES	Count of genres
Drama	153
Comedy|Drama|Romance	151
Comedy|Drama	147
Comedy	145
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46
Grand Total	**1083**



LINK TO EXCEL SHEET - https://1drv.ms/x/s!AhPlDIVl_coYgSYcs9JB9XoYKiVR?e=cn4av9

E. **Charts:**

1. Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
   Append the rows of all these columns and store them in a new column named Combined. Group the combined column using the actor_1_name column. Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

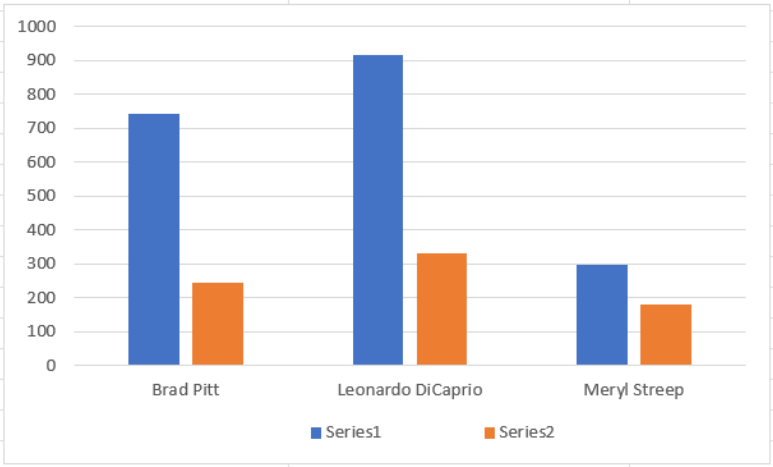
LINK TO EXCEL SHEET---
https://1drv.ms/x/s!AhPlDIVl_coYgSYcs9JB9XoYKiVR?e=z9aeTB

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: The Vampire Chronicles	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 Seven Seas	Shutter Island	The River Wild
Spy Game	The Aviator	
The Assassination of Jesse James by the Coward Robert Ford	The Beach	
The Curious Case of Benjamin Button	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	

COMBINE(Actor)	Movies
Meryl Streep	The Hours
Meryl Streep	Hope Springs
Meryl Streep	One True Thing
Meryl Streep	The Devil Wears Prada
Meryl Streep	Out of Africa
Meryl Streep	A Prairie Home Companion
Meryl Streep	The River Wild
Meryl Streep	Lions for Lambs
Meryl Streep	The Iron Lady
Meryl Streep	It's Complicated
Meryl Streep	Julie & Julia
Brad Pitt	Fight Club
Brad Pitt	True Romance
Brad Pitt	By the Sea
Brad Pitt	Killing Them Softly
Brad Pitt	The Assassination of Jesse James by the Coward Robert Ford
Brad Pitt	Spy Game
Brad Pitt	Ocean's Eleven
Brad Pitt	Interview with the Vampire: The Vampire Chronicles
Brad Pitt	The Tree of Life
Brad Pitt	Troy
Brad Pitt	Fury
Brad Pitt	Babel

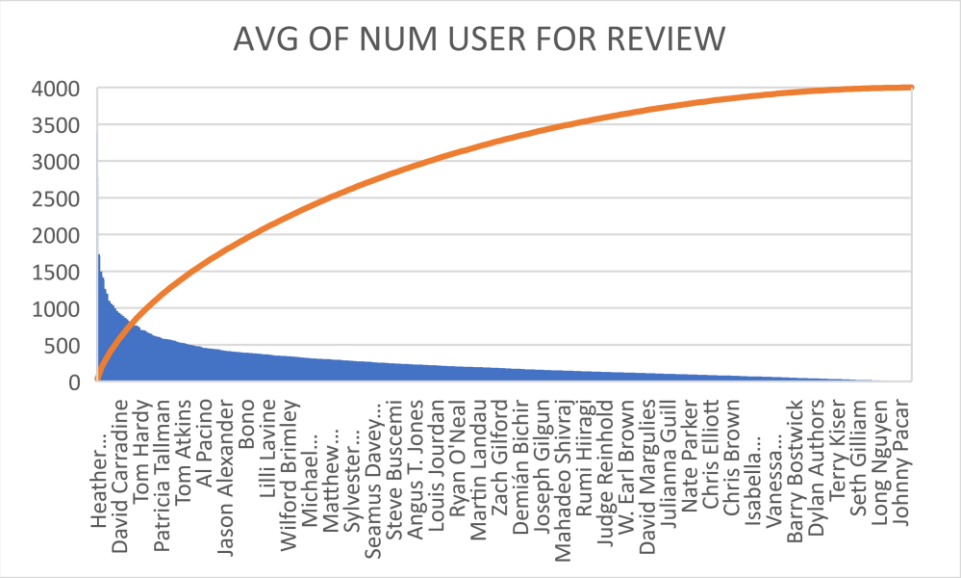
- In case of names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' We can see that Leonardo DiCaprio is audience's and Critic's favourite actor.

actor_1_name	Mean of num_user_for_reviews	Mean of num_critic_for_reviews
Brad Pitt	742.35	245
Leonardo DiCaprio	914.48	330.19
Meryl Streep	297.18	181.45



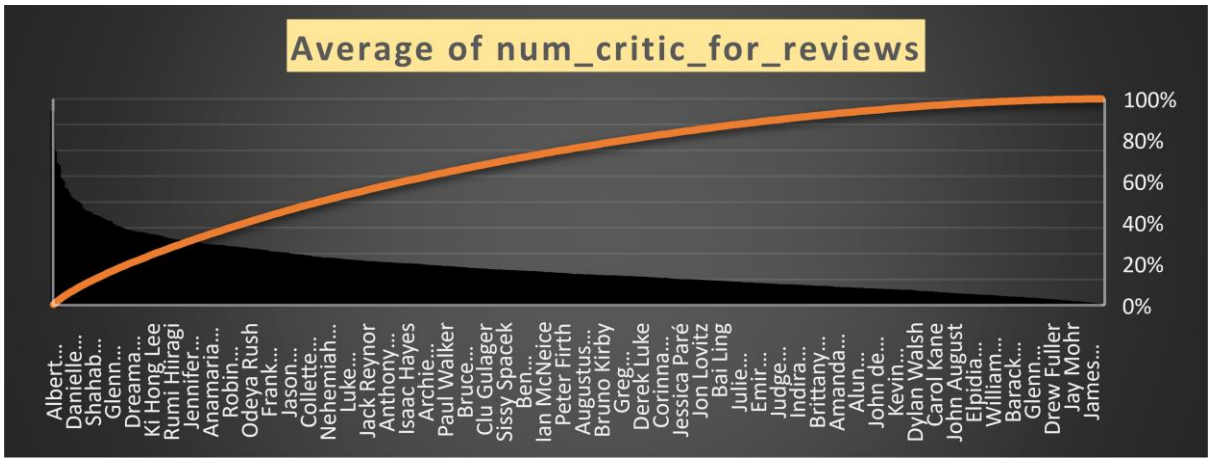
- According to only User Reviews Heather Donahue is the favourite.

Row Labels	Average of num_user_for_review	Average of num_critic_for_review
Heather Donahue	3400	360
Christo Jivkov	2814	406
Steve Bastoni	2789	275
Phaldut Sharma	1885	738
Orlando Bloom	1842	259
Keir Dullea	1736	285
Eva Green	1708.333333	388.6666667
Chen Chang	1641	287
Nick Stahl	1562	218
Albert Finney	1498	750
Kevin Rankin	1445	267
Noah Huntley	1441	224
Osama bin Laden	1416	288
Seychelle Gabriel	1382	280
Mathieu Kassovitz	1314	242
Essie Davis	1285.5	245
Sharlto Copley	1262	472



- According to Critic Review Albery Finney is the favourite

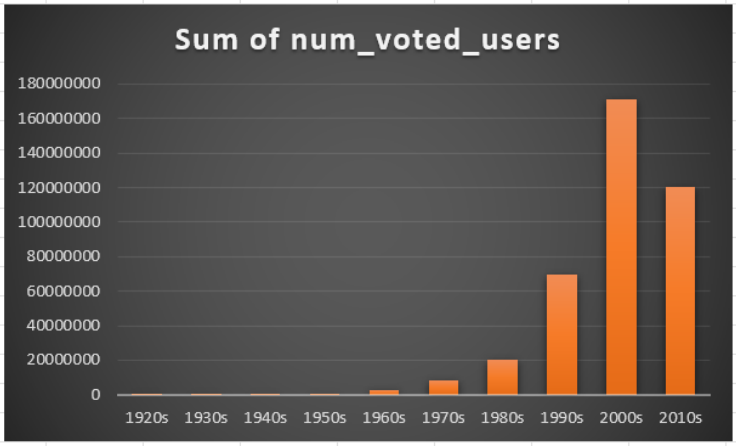
Row Labels	Average of num_user_for_review	Average of num_critic_for_review
Albert Finney	1498	750
Phaldut Sharma	1885	738
Peter Capaldi	995	654
Craig Stark	1018	596
Bérénice Bejo	583	576
Suraj Sharma	755	552
Ellar Coltrane	836	548
Mike Howard	405	546
Lou Taylor Pucci	789	543
Joel Courtney	849	539
Maika Monroe	631	533
Tim Holmes	511	525
Elina Alminas	611	489
Kurt Fuller	509	487
Iko Uwais	316	481
Quvenzhané Wallis	392.3333333	478.6666667
Edgar Arreola	461	478
Sharlto Copley	1262	472
Cory Hardrict	326	452
Matt Frewer	1229	451
Aidan Turner	894.25	447
Elizabeth McGovern	801	447
Michael Fassbender	837.5	434.3333333
Wood Harris	588	432
Jennifer Lawrence	583.2142857	418.9285714



LINK TO SHEET - https://1drv.ms/x/s!AhPlDIV1_coYgSYcs9JB9XoYKiVR?e=Qg8Ked

2. Observe the change in number of voted users over decades using a bar chart.Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

DECADE	Sum of num_voted_users
1920s	116387
1930s	804839
1940s	230838
1950s	678336
1960s	2983442
1970s	8524102
1980s	19987476
1990s	69735679
2000s	170904077
2010s	120640346



From the above table and bar plot I have inferred that most number of votes were in the decade 2001 to 2010 with the count of 178592461.

LINK TO SHEET------
https://1drv.ms/x/s!AhPlDIVl_coYgSYcs9JB9XoYKiVR?e=gnAGdt

INSIGHTS:

IMDb (Internet Movie Database) is a popular online database that provides information about movies, TV shows, actors, and other related content. Analyzing IMDb data can provide valuable insights into various aspects of the film industry, audience preferences, and critical reception.

RESULT:

Through this project, I was able to achieve my objectives and learn how to clean data using excel. I was able to understand a range of visualization options, including charts and graphs, which are essential for presenting the findings in a clear and concise manner. These visualizations aid in understanding trends, identifying patterns, and communicating key insights effectively. I also learnt Fwe lot about Pivot table