Vikash Kumar Maheshwari

Project- IMDB Movie Analysis

Description :

The dataset having various columns of different IMDB Movies. We are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.

Approach :

Firstly we will go through the data and check the missing data or null one. After that we will remove the outliers and using charts we will show our results.

Tech-Stack Used :

Google sheets

Insights:

1. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this.

Your task: Clean the data

There are 5043 rows and 28 columns

19 blanks in colors

104 blanks in directors name -4939

45 in number critic

9 in duration

732 in gross

37 in plot_keywords

221 in budget

3816 rows after data cleaning

17 columns

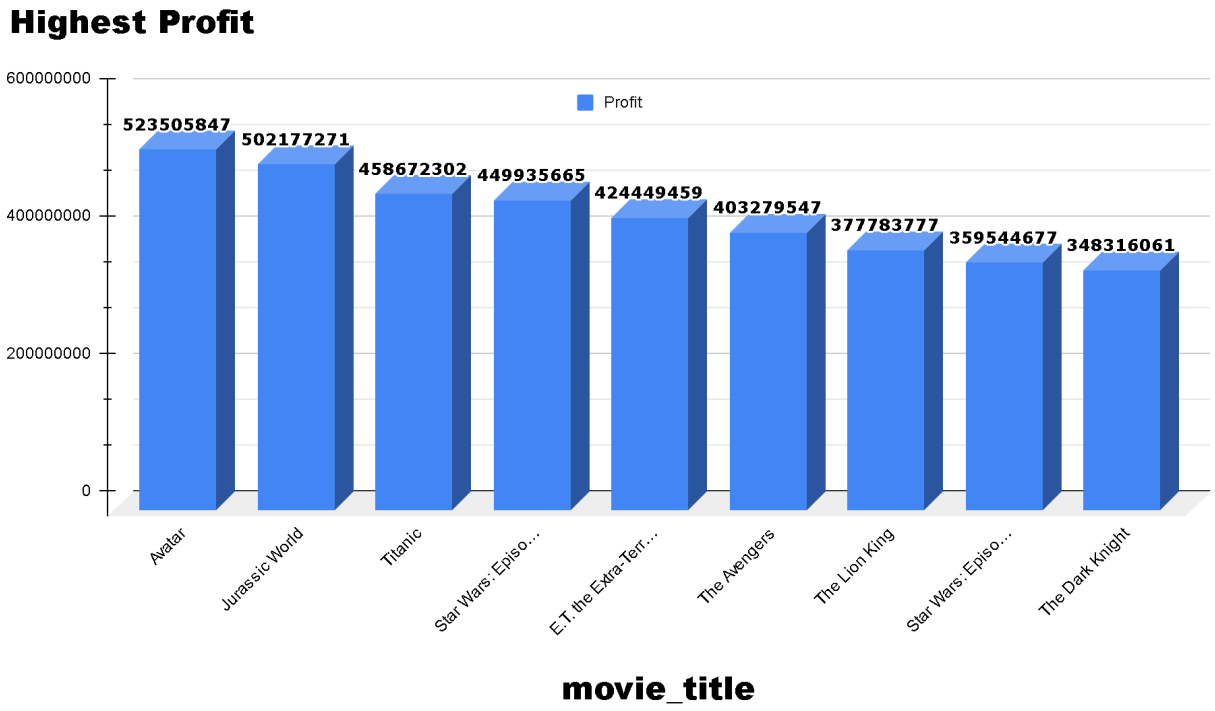
Vikash Kumar Maheshwari

Insights :

2. 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?

Result :

movie_title	Profit
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Avengers	403279547
The Lion King	377783777
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061



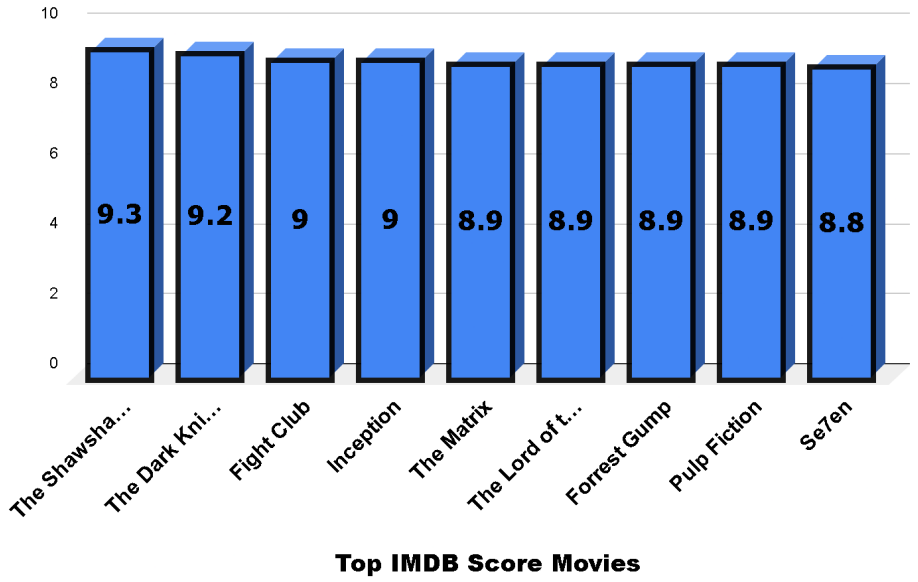
Vikash Kumar Maheshwari

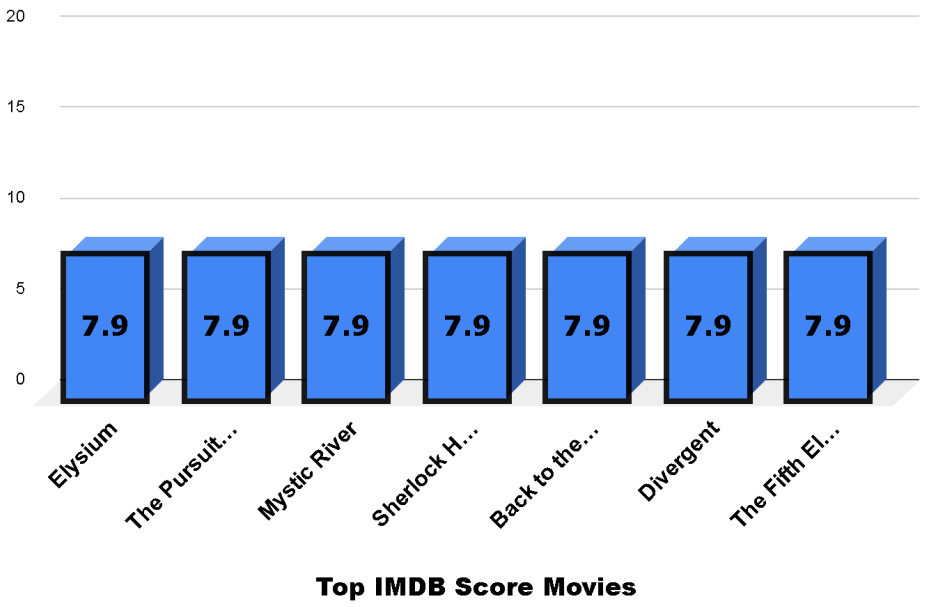
Insights:

3. 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.

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.

director_name	movie_title	num_voted_user	language	country	imdb_score	IMDb_Top_250
Frank Darabont	The Shawshank Redemption	1689764	English	USA	9.3	Yes
Christopher Nolan	The Dark Knight	1676169	English	USA	9.2	Yes
David Fincher	Fight Club	1347461	English	USA	9	Yes
Christopher Nolan	Inception	1468200	English	USA	9	Yes
Lana Wachowski	The Matrix	1217752	English	USA	8.9	Yes
Peter Jackson	The Lord of the Rings: The Fellowshi	1238746	English	New Zealar	8.9	Yes
Robert Zemeckis	Forrest Gump	1251222	English	USA	8.9	Yes
Quentin Tarantino	Pulp Fiction	1324680	English	USA	8.9	Yes
David Fincher	Se7en	1023511	English	USA	8.8	Yes
Peter Jackson	The Lord of the Rings: The Two Towe	1100446	English	USA	8.8	Yes
Christopher Nolan	The Dark Knight Rises	1144337	English	USA	8.8	Yes
Francis Ford Coppola	The Godfather	1155770	English	USA	8.8	Yes
Peter Jackson	The Lord of the Rings: The Return of	1215718	English	USA	8.8	Yes
George Lucas	Star Wars: Episode IV - A New Hope	911097	English	USA	8.7	Yes
Christopher Nolan	Interstellar	928227	English	USA	8.7	Yes
Quentin Tarantino	Django Unchained	955174	English	USA	8.7	Yes
Christopher Nolan	Batman Begins	980946	English	USA	8.7	Yes
Ridley Scott	Gladiator	982637	English	USA	8.7	Yes
Joss Whedon	The Avengers	995415	English	USA	8.7	Yes
Joss Whedon	The Avengers	995415	English	USA	8.7	Yes
Christopher Nolan	The Prestige	844052	English	USA	8.6	Yes
Christopher Nolan	Memento	845580	English	USA	8.6	Yes
Steven Spielberg	Schindler's List	865020	English	USA	8.6	Yes
Martin Scorsese	The Departed	873649	English	USA	8.6	Yes
Steven Spielberg	Saving Private Ryan	881236	English	USA	8.6	Yes
Quentin Tarantino	Inglourious Basterds	885175	English	USA	8.6	Yes
James Cameron	Avatar	886204	English	USA	8.6	Yes
Jonathan Demme	The Silence of the Lambs	887467	English	USA	8.6	Yes
Gary Ross	The Hunger Games	701607	English	USA	8.5	Yes
M. Night Shyamalan	The Sixth Sense	704766	English	USA	8.5	Yes
Andrew Stanton	WALL·E	718837	English	USA	8.5	Yes
Martin Scorsese	Goodfellas	728685	English	USA	8.5	Yes
Robert Zemeckis	Back to the Future	732212	English	USA	8.5	Yes





Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Vikash Kumar Maheshwari

	A	B	C	D	E	F	G	H	I
1	director_name	movie_title	num_voted_user	language	country	imdb_score	IMDb_Top_250	>25000	Rank
2	Frank Darabont	The Shawshank Redemption	1689764	English	USA	9.3	Yes	true	
3	Christopher Nolan	The Dark Knight	1676169	English	USA	9.2	Yes	true	
4	David Fincher	Fight Club	1347461	English	USA	9	Yes	true	
5	Christopher Nolan	Inception	1468200	English	USA	9	Yes	true	
6	Lana Wachowski	The Matrix	1217752	English	USA	8.9	Yes	true	
7	Peter Jackson	The Lord of the Rings: The Fellowshi	1238746	English	New Zealar	8.9	Yes	true	
8	Robert Zemeckis	Forrest Gump	1251222	English	USA	8.9	Yes	true	
9	Quentin Tarantino	Pulp Fiction	1324680	English	USA	8.9	Yes	true	
10	David Fincher	Se7en	1023511	English	USA	8.8	Yes	true	
11	Peter Jackson	The Lord of the Rings: The Two Towe	1100446	English	USA	8.8	Yes	true	
12	Christopher Nolan	The Dark Knight Rises	1144337	English	USA	8.8	Yes	true	

	A	B	C	D	E	F	G	H	I	J
1	director_name	movie_title	num_voted_user	language	country	imdb_score	IMDb_Top_250	>25000	Rank	
2	Matt Reeves	Dawn of the Planet of the Apes	317542	English	USA	7.9	Yes	true	216	
3	Richard Curtis	Love Actually	318634	English	UK	7.9	Yes	true	216	
4	Tate Taylor	The Help	318955	English	USA	7.9	Yes	true	216	
5	Tim Burton	Charlie and the Chocolate Factory	320284	English	USA	7.9	Yes	true	216	
6	Marc Webb	The Amazing Spider-Man 2	321227	English	USA	7.9	Yes	true	216	
7	Sofia Coppola	Lost in Translation	321283	English	USA	7.9	Yes	true	216	
8	David Yates	Harry Potter and the Half-Blood Prin	321795	English	UK	7.9	Yes	true	216	
9	Michael Bay	Armageddon	322395	English	USA	7.9	Yes	true	216	
10	Michael Bay	Transformers: Revenge of the Fallen	323207	English	USA	7.9	Yes	true	216	
11	Paul Greengrass	Captain Phillips	323353	English	USA	7.9	Yes	true	216	
12	Christopher Nolan	The Dark Knight Rises	1144337	English	USA	8.8	Yes	true		

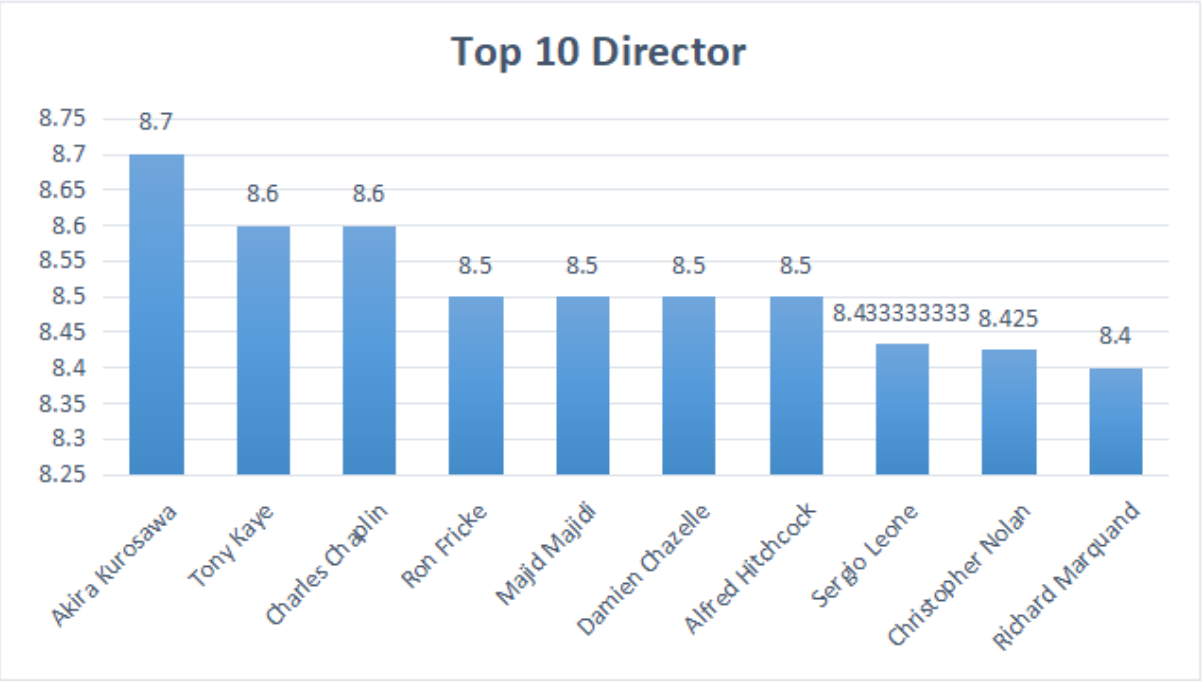
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
Movies other than english are false while english movies are set true

	A	B	C	D	E	F	G	H	I	J
1	director_name	movie_title	num_voted_us	language	country	imdb_sc	IMDb_Top_250	>25000	Rank	Top_Foreign_Lang_Fil
89	Sergio Leone	The Good, the Bad and the Ugly	503509	Italian	Italy	8.2	Yes	true	88	false
90	Fernando Meirelles	City of God	533200	Portuguese	Brazil	8.2	Yes	true	88	false
91	Jean-Pierre Jeunet	Amélie	534262	French	France	8.2	Yes	true	88	false
113	Guillermo del Toro	Pan's Labyrinth	467234	Spanish	Spain	8.1	Yes	true	112	false
162	Hayao Miyazaki	Spirited Away	417971	Japanese	Japan	8	Yes	true	161	false
252										
253										

Insights:

4. 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

Vikash Kumar Maheshwari



Insights:

5. Popular Genres: Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

genres	Count of genres	popular genres
Comedy|Drama|Romance	150	Drama
Drama	145	
Comedy	144	
Comedy|Drama	142	
Comedy|Romance	135	
Drama|Romance	119	
Crime|Drama|Thriller	82	
Action|Crime|Thriller	56	
Action|Crime|Drama|Thriller	50	
Action|Adventure|Sci-Fi	48	
Comedy|Crime	47	
Action|Adventure|Thriller	45	
Horror	44	
Crime|Drama|Mystery|Thriller	42	

Drama is the popular genres

Vikash Kumar Maheshwari

Insights:

6. Charts: 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.
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
Your task: Find the critic-favorite and audience-favorite actors

actor_1_name	Average of num_user_for_reviews	Average of num_critic_for_reviews
Leonardo DiCaprio	914.4761905	330.1904762
Brad Pitt	742.3529412	245
Meryl Streep	297.1818182	181.4545455

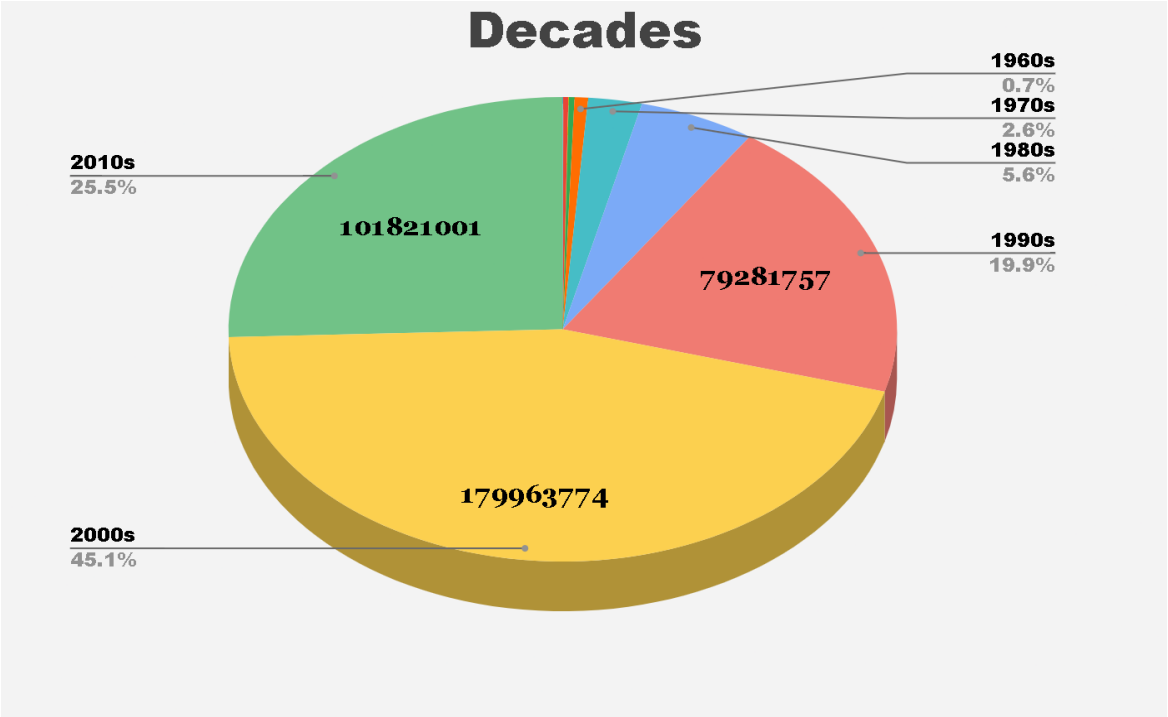
actor_1_name	Average of num_user_for_reviews	Average of num_critic_for_reviews
Heather Donahue	3400	360
Christo Jivkov	2814	406
Steve Bastoni	2789	275
Phaldut Sharma	1885	738
Orlando Bloom	1842	259
Keir Dullea	1736	285
Eva Green	1708.333333	388.6666667
Chen Chang	1641	287

actor_1_name	Average of num_user_for_reviews	Average of num_critic_for_reviews
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

Vikash Kumar Maheshwari

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.

2000s has the highest voting



Decades	df_by_decade
1920s	116387
1930s	966520
1940s	72324
1950s	1097601
1960s	2607791
1970s	10354731
1980s	22445073
1990s	79281757
2000s	179963774
2010s	101821001
Grand Total	**398726959**

Thank you