IMDb Movie Analysis

Project Discription

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

Once you have the problem better defined, you can use 5 Whys technique to determine its root cause by repeatedly asking the question “Why”.

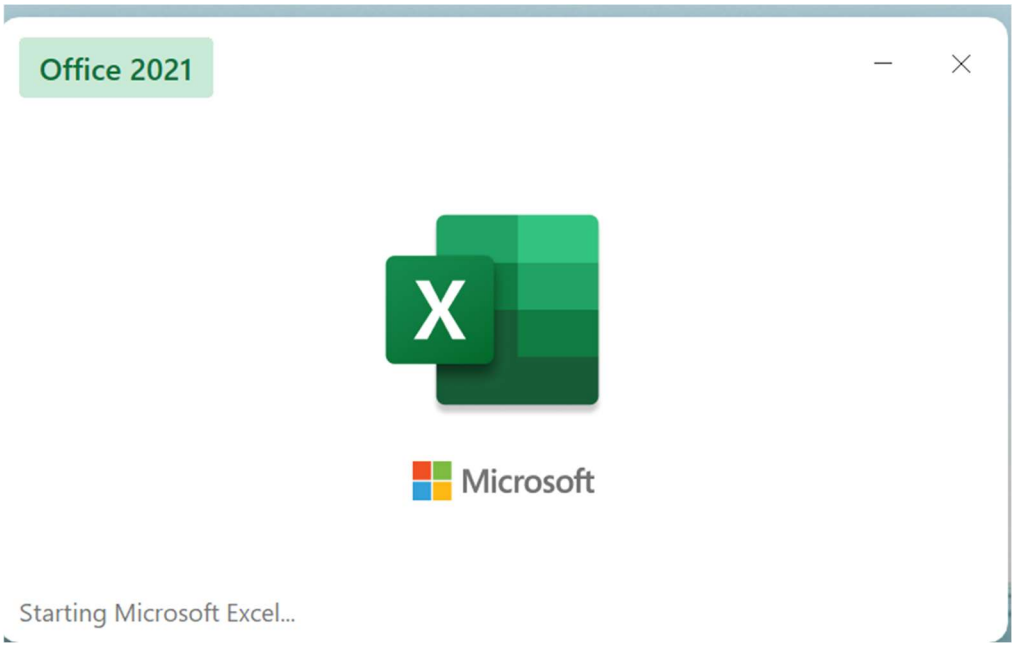
It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here’s an example of how this technique could be used to figure out the cause of the following problems.

Use the below **Steps for EDA:**

1. Understanding data columns and data
2. Checking for missing data
3. Clubbing columns with multiple categories
4. Checking for outliers
5. Removing outliers
6. Drawing Data Summary

Tech Stack Used

Name : Microsoft Excel 2021 and other functionalities of MS Excel like charts, Pivot tables etc.



Insight and Results

A. **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. (Dropping columns, removing null values, etc.)
Your task: Clean the data

Approach:

First approach is to make our data a table and apply filter for what data we want. It is temporary as our data is just filtered not cleaned.

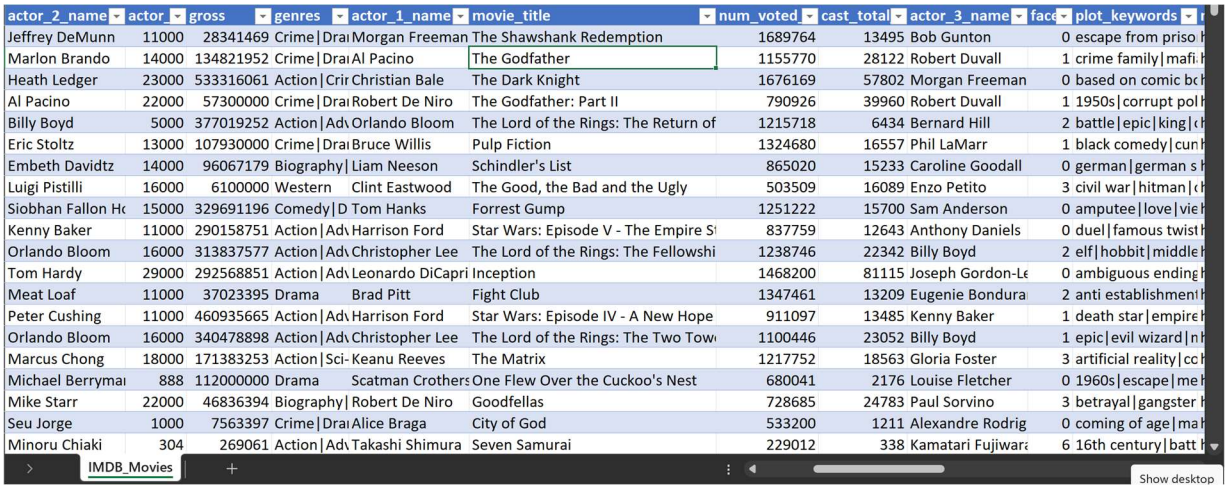
Second approach is to use “Find and Select > Go to Special”. Now we can select only blank cells from our data. Now select Delete rows to clean the data.

After dropping columns and removing null cells our dataset is down to 3811 rows from 5043 rows.

In movie_title column all the movie names have an unwanted ‘^A’ symbol. Now we must have to clean it using:

=LEFT(AK2,LEN(AK2)-2)

Which gives us only the name of movie cleaned.



actor_2_name	actor_	gross	genres	actor_1_name	movie_title	num_voted	cast_total	actor_3_name	face	plot_keywords
Jeffrey DeMunn	11000	28341469	Crime|Drai	Morgan Freeman	The Shawshank Redemption	1689764	13495	Bob Gunton	0	escape from priso
Marlon Brando	14000	134821952	Crime|Drai	Al Pacino	The Godfather	1155770	28122	Robert Duvall	1	crime family|mafi
Heath Ledger	23000	533316061	Action|Crir	Christian Bale	The Dark Knight	1676169	57802	Morgan Freeman	0	based on comic bc
Al Pacino	22000	57300000	Crime|Drai	Robert De Niro	The Godfather: Part II	790926	39960	Robert Duvall	1	1950s|corrupt pol
Billy Boyd	5000	377019252	Action|Adv	Orlando Bloom	The Lord of the Rings: The Return of	1215718	6434	Bernard Hill	2	battle|epic|king|
Eric Stoltz	13000	107930000	Crime|Drai	Bruce Willis	Pulp Fiction	1324680	16557	Phil LaMarr	1	black comedy|cun
Embeth Davidtz	14000	96067179	Biography|	Liam Neeson	Schindler's List	865020	15233	Caroline Goodall	0	german|german s
Luigi Pistilli	16000	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	16089	Enzo Petito	3	civil war|hitman|
Siobhan Fallon Hc	15000	329691196	Comedy|D	Tom Hanks	Forrest Gump	1251222	15700	Sam Anderson	0	amputee|love|vie
Kenny Baker	11000	290158751	Action|Adv	Harrison Ford	Star Wars: Episode V - The Empire St	837759	12643	Anthony Daniels	0	duel|famous twist
Orlando Bloom	16000	313837577	Action|Adv	Christopher Lee	The Lord of the Rings: The Fellowshi	1238746	22342	Billy Boyd	2	elf|hobbit|middle
Tom Hardy	29000	292568851	Action|Adv	Leonardo DiCapri	Inception	1468200	81115	Joseph Gordon-Le	0	ambiguous ending
Meat Loaf	11000	37023395	Drama	Brad Pitt	Fight Club	1347461	13209	Eugenie Bondura	2	anti establishment
Peter Cushing	11000	460935665	Action|Adv	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	13485	Kenny Baker	1	death star|empire
Orlando Bloom	16000	340478898	Action|Adv	Christopher Lee	The Lord of the Rings: The Two Tow	1100446	23052	Billy Boyd	1	epic|evil wizard|n
Marcus Chong	18000	171383253	Action|Sci-	Keanu Reeves	The Matrix	1217752	18563	Gloria Foster	3	artificial reality|cc
Michael Berryman	888	112000000	Drama	Scatman Crothers	One Flew Over the Cuckoo's Nest	680041	2176	Louise Fletcher	0	1960s|escape|me
Mike Starr	22000	46836394	Biography|	Robert De Niro	Goodfellas	728685	24783	Paul Sorvino	3	betrayal|gangster
Seu Jorge	1000	7563397	Crime|Drai	Alice Braga	City of God	533200	1211	Alexandre Rodrig	0	coming of age|ma
Minoru Chiaki	304	269061	Action|Adv	Takashi Shimura	Seven Samurai	229012	338	Kamatari Fujiwara	6	16th century|batt

IMDB_Movies

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?

Approach:

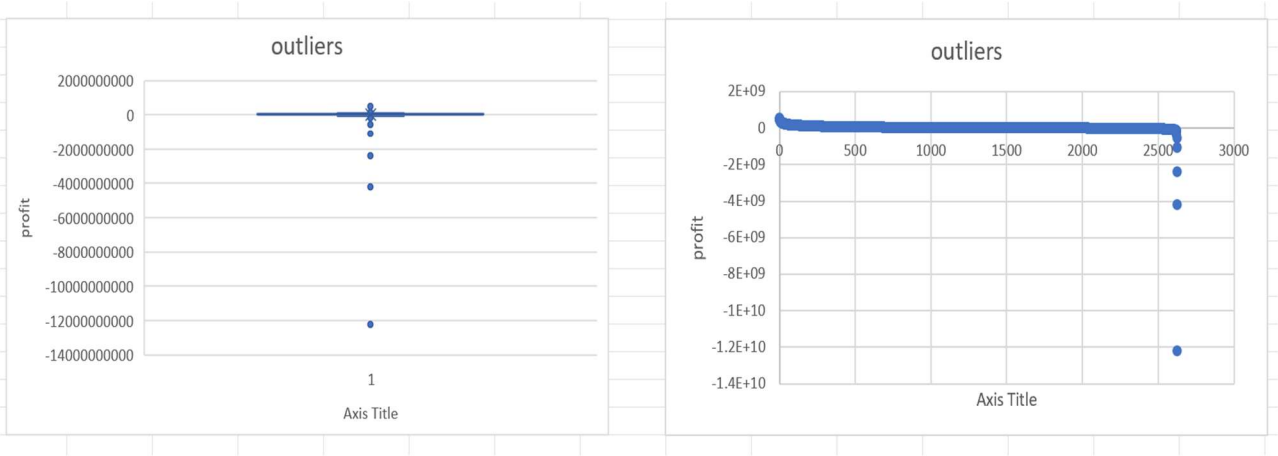
Creating new column named Profit using formula:

Profit = gross – budget

Top-10 movies with the highest profit.

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

Sorted column is now visualized using excel chart for outlier analysis.
Avatar movie made the highest profit of 523505847$.



Here are some huge negative outliers are detected.

C. **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

Approach:

New column is created for IMDb top 250 movies list. All the filters like num_voted_users > 25000 and Largest IMDb rating are applied. Here top 250 results are selected for IMDb_top_250.

Likewise…

AZ	BA	BB
Imdb_ranking	Imdb_top_250	Top_foreign_language_film
1	The Shawshank Redemption	The Good, the Bad and the UglyÂ
2	The Godfather	City of GodÂ
3	The Dark Knight	Seven SamuraiÂ
4	The Godfather: Part II	Spirited AwayÂ
5	ord of the Rings: The Return of the	The Lives of OthersÂ
6	Pulp Fiction	Children of HeavenÂ
7	Schindler's List	A SeparationÂ
8	The Good, the Bad and the Ugly	OldboyÂ
9	Forrest Gump	Das BootÂ
10	ars: Episode V - The Empire Strike	AmÃ©lieÂ
11	d of the Rings: The Fellowship of t	Princess MononokeÂ
12	Inception	The HuntÂ
13	Fight Club	MetropolisÂ
14	tar Wars: Episode IV - A New Hop	DownfallÂ
15	e Lord of the Rings: The Two Towe	Pan's LabyrinthÂ
16	The Matrix	The Secret in Their EyesÂ
17	One Flew Over the Cuckoo's Nest	IncendiesÂ
18	Goodfellas	Howl's Moving CastleÂ
19	City of God	Amores PerrosÂ
20	Seven Samurai	The CelebrationÂ
21	Saving Private Ryan	Elite SquadÂ
22	The Silence of the Lambs	The Sea InsideÂ
23	Se7en	Tae Guk Gi: The Brotherhood of WarÂ
24	Interstellar	AkiraÂ
25	The Usual Suspects	A Fistful of DollarsÂ
26	American History X	Central StationÂ
27	Modern Times	Waltz with BashirÂ
28	Spirited Away	PersepolisÂ
29	The Lion King	My Name Is KhanÂ
30	Raiders of the Lost Ark	Crouching Tiger, Hidden DragonÂ
31	The Dark Knight Rises	4 Months, 3 Weeks and 2 DaysÂ

D. **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

Approach:

A pivot table is made for that. Director name and IMDb rating(Average) columns are selected from that top 10 results are chosen as result which are:

Directors	Average of imdb_score
Akira Kurosaw	8.7
Tony Kaye	8.6
Charles Chapli	8.6
Alfred Hitchco	8.5
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chaze	8.5
Sergio Leone	8.433333333
Christopher N	8.425
Marius A. Mar	8.4
Richard Marqu	8.4
Asghar Farhad	8.4

Akira Kurosaw is the best director according to IMDb ratings.

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

Approach:

Here ganre column is like "Action|Adventure|Fantasy|Sci-Fi". So at the first we will do text to column setting for genre to get each genre individual.

Pivot table is used for counting number of movies with a particular genre.

Genre	Movies
Drama	1919
Comedy	1488
Thriller	1121
Action	961
Romance	873
Adventure	787
Crime	712
Sci-fi	499
Family	446
Horror	394
Mystery	386
Biography	240
Animation	197
War	155
Sport	151
Musical	98

The most popular genre is Drama.

F. **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.

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.

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.

Your task: Find the critic-favorite and audience-favorite actors

Approach:

Pivot table is used to find sum of the num_critic_for_reviews and num_users_for_review. Respective chart is also there.

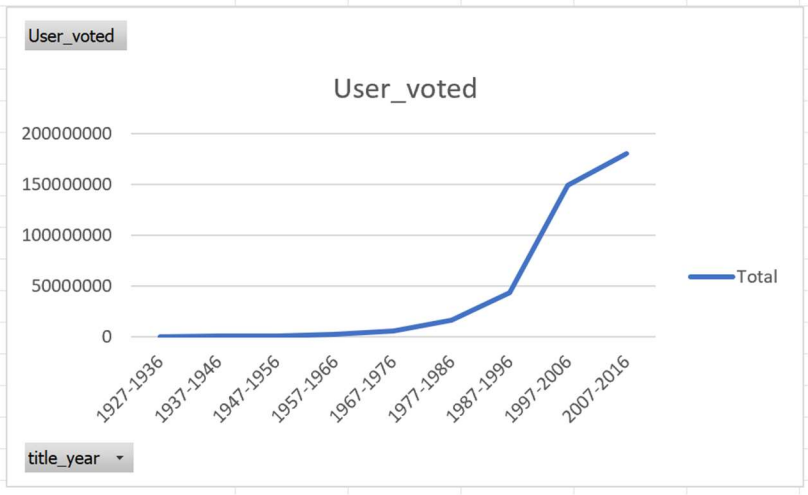
Number od voted uses column (user_voted) and year is selected in pivot table.

Year column is then grouped as decades and votes during this time is summed. Respective chart is there for better understanding.

Number of user voted grouped by decades.

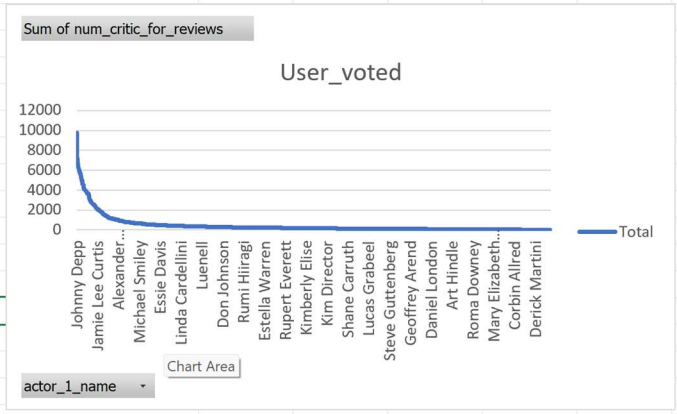
Decades	User_voted
1927-1936	280663
1937-1946	777586
1947-1956	376190
1957-1966	2563721
1967-1976	5896479
1977-1986	16384193
1987-1996	43562999
1997-2006	148869228
2007-2016	179858758
Grand Total	**398569817**

Line chart for better understanding.



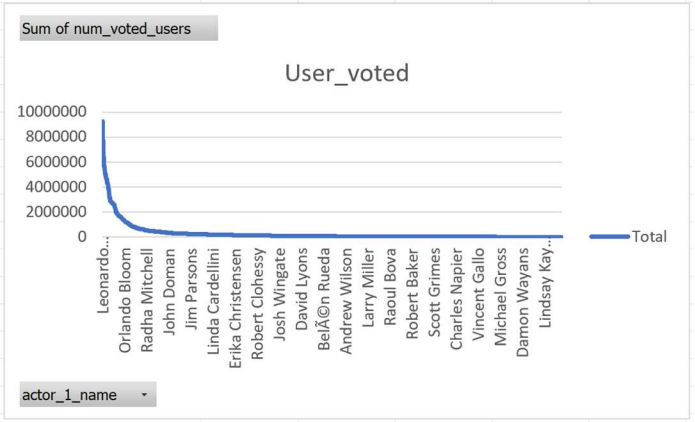
Sum of critic reviews and voted uses for each actor.

Actors	Sum of num_critic_for_reviews
Johnny Depp	9763
J.K. Simmons	7364
Matt Damon	7146
Leonardo DiCaprio	6934
Robert Downey Jr.	6380
Liam Neeson	6312
Chris Hemsworth	6176
Robert De Niro	5958
Jennifer Lawrence	5865
Hugh Jackman	5771
Tom Hanks	5679
Christian Bale	5657
Tom Cruise	5544
Scarlett Johansson	5363
Nicolas Cage	5185
Bruce Willis	5031
Jason Statham	4957
Channing Tatum	4919



Johnny depp is the most critic favourite actor.

	Actors	Sum of num_voted_users
4	Leonardo DiCaprio	9281136
5	Johnny Depp	7833186
6	Tom Hanks	7736413
7	Harrison Ford	6568429
8	Christian Bale	6349003
9	Bruce Willis	5679773
10	Robert De Niro	5675294
11	Tom Cruise	5451105
12	J.K. Simmons	5114349
13	Morgan Freeman	5039175
14	Chris Hemsworth	5005296
15	Brad Pitt	4820925
16	Will Smith	4668784
17	Robert Downey Jr.	4625831
18	Jason Statham	4606482
19	Kevin Spacey	4345418
20	Liam Neeson	4310646
21	Matt Damon	4269181



Leonardo DiCaprio is the most audience favourite actor.

The End