MOVIE ANALYSIS

Trainity Project Report

Rohit kumar

rohitk,.ug20.ce@nitp.ac.in

DESCRIPTION

The Project is based on IMDb data records to gain meaningful insights from raw data of movies which can be useful for movie recommendation process and to study nature of audience and their preference. Furthermore, it answer the critical questions which help in finding of user & critics favorite genres, actor, top directors etc.

APPROACH

- To perform Analysis on data records I begin with choosing right column and extracting it for further cleaning process
- After cleaning, filtering and sorting the dataset, it time to process the data,
- Here I use excel function such as advance filter, Text to column, mathematical function, Pivot Table etc. and apply different combinations and visualize it to understand it better.
- Expect cleaning process all analysis is done by using Pivot table and basic excel function

TECH-STACK USED

‘Microsoft Excel 2013’ was used to perform Analysis
‘MS Power Point 2013’ was used to prepare to report.

Click below to view excel sheet containing steps & solution

IMDb_movies_analysis.xlxs

CLEANING THE DATA

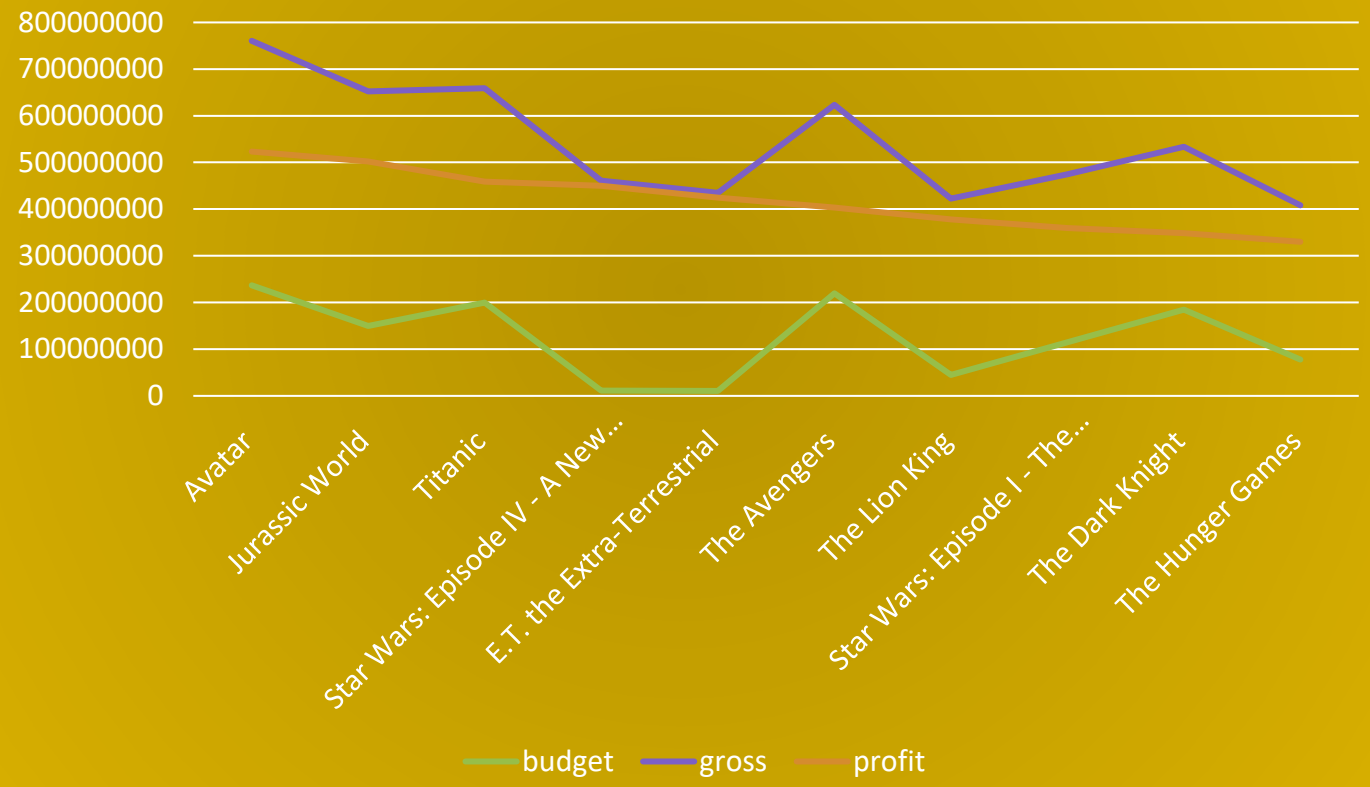
- Extracted & arranged required column by using cut , copy & paste function
- Removed extra 'Â' character from movie title column
- By using 'Remove Duplicates' function drop 112 duplicate rows (to check duplicate I use movie title, director name, imdb score, actor_1, Actor_2, actor_3, budget column)
- Checked Missing values and took appropriate action to each column
- Calculated absolute number of missing values with ‘Countblank' function for each column and calculate percentage of null values per column

NOTE : Misleading values present in gross & budget column (difference in currency such as….. budgets col recorded in Yuro, INR, won, etc. and gross col recorded in US dollars)

PROFITABLE MOVIE

Top 10 most profitable movies					
Rank	movie_title	budget	gross	profit	profit %
1	Avatar	237000000	760505847	523505847	220.89
2	Jurassic World	150000000	652177271	502177271	334.78
3	Titanic	200000000	658672302	458672302	229.34
4	Star Wars: Episode IV - A New Hope	11000000	460935665	449935665	4090.3
5	E.T. the Extra-Terrestrial	10500000	434949459	424449459	4042.4
6	The Avengers	220000000	623279547	403279547	183.31
7	The Lion King	45000000	422783777	377783777	839.52
8	Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677	312.65
9	The Dark Knight	185000000	533316061	348316061	188.28
10	The Hunger Games	78000000	407999255	329999255	423.08

PROFITABLE MOVIE



IMDb TOP 250

- Extract movie title , language , content type, num_of_voted_users
- Use filter on num_of_voted_users col and only consider those movies which have more than 25000 user votes
- Sort imdb_score column by largest to smallest
- Create column rank and use flash fill to give rank to films

Outcomes of this question is very large to attached here

Hence, attached answer is sample contain first 10 rows of outcome

IMDb TOP 250

Top 250 Movie					
Rank	movie_title	imdb_score	language	content_rating	num_voted_users
1	The Shawshank Redemption	9.3	English	R	1689764
2	The Godfather	9.2	English	R	1155770
3	The Dark Knight	9	English	PG-13	1676169
4	The Godfather: Part II	9	English	R	790926
5	The Lord of the Rings: The Return of the King	8.9	English	PG-13	1215718
6	Schindler's List	8.9	English	R	865020
7	Pulp Fiction	8.9	English	R	1324680
8	The Good, the Bad and the Ugly	8.9	Italian	Approved	503509
9	Inception	8.8	English	PG-13	1468200
10	The Lord of the Rings: The Fellowship of the Ring	8.8	English	PG-13	1238746

BEST DIRECTOR

Rank	Director Name	No_of_Movie	Average of imdb_score
1	Charles Chaplin	4	8.6
2	Alfred Hitchcock	8	8.5
3	Christopher Nolan	4	8.4
4	Asghar Farhadi	6	8.4
5	Billy Wilder	5	8.3
6	Akira Kurosawa	12	8.1
7	Ari Folman	6	8.0
8	Anna Muylaert	7	7.9
9	Christophe Barratier	4	7.9
10	Alejandro G. IÃ±Ã¡rritu	9	7.8

POPULAR GENRE

Genre	num. of films
Drama	**1915**
Comedy	**1492**
Thriller	**1088**
Action	**932**
Adventure	**763**
Romance	**868**
Crime	**703**
Sci-Fi	**482**
Fantasy	**493**
Mystery	**376**

Genre	num. of films
Family	440
Horror	379
Biography	242
Animation	196
War	159
Sport	147
Musical	101
Western	59
Documentary	67
Film-Noir	1

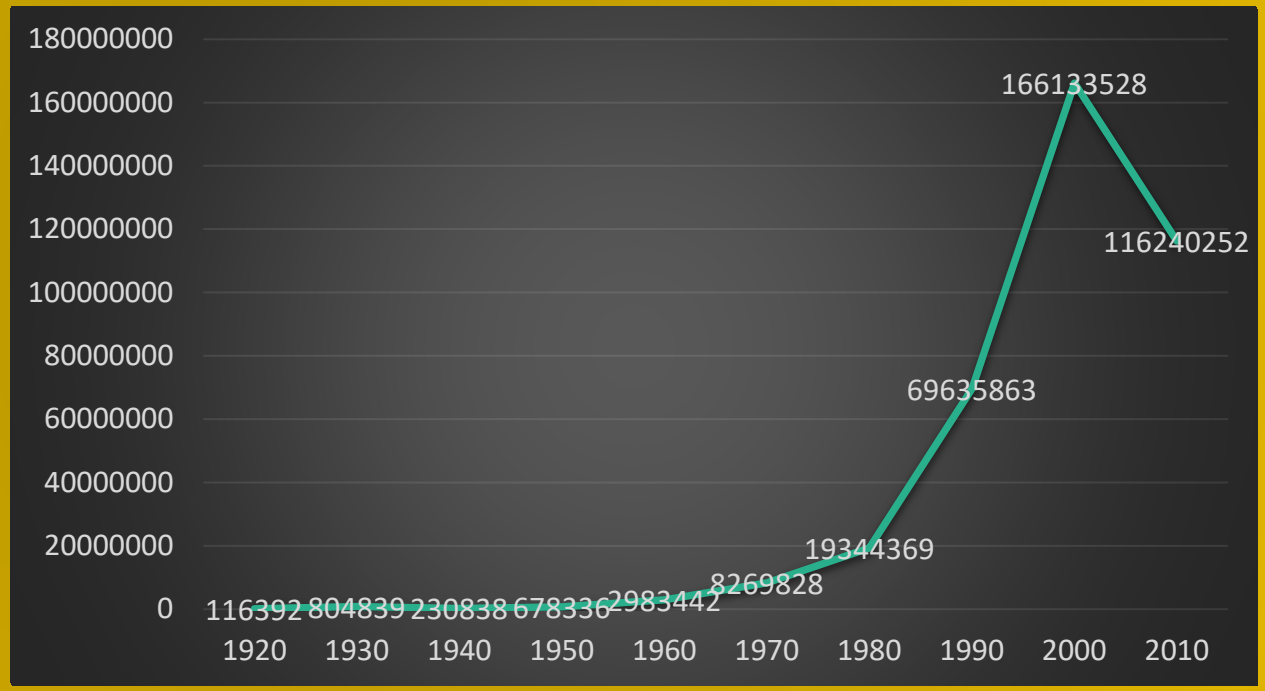
BEST ACTOR

actor name	avg critic review	avg user review	Grand Avg.
Leonardo DiCaprio	245.00	742.35	493.68
Brad Pitt	322.20	922.55	622.38
Meryl Streep	181.45	297.18	239.32

Leonardo DiCaprio is most favourite actor
among users and crictics as well

USER VOTE OVER DECADE

Row Labels	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336
1960	2983442
1970	8269828
1980	19344369
1990	69635863
2000	166133528
2010	116240252
Grand Total	384437687



INSIGHTS

- There are some tv shows' records and web series' records present in dataset
- Content type which related to TV shows and OTT based web series and shows are missing Gross and Budget Amount also Directors' name , may be because of it changes with per episode
- Misleading values present in Gross & Budget column (differences in currency such as...budgets col recorded in Yuro, INR, won...etc.... and gross col rocorded in US dollars..)
- All Top 10 highest grossing movies are high budgets movies
- Most of Top 10 most profitable movies are documentaries or individual movies
- Avatar is highest Grossing movie with 760+ million USD and Paranormal Activity is most profitable movie ever with 719348.55 % profit.
- Top 250 movies in English and other language movies have content type 'R or PG-13'

SUMMARY

- This Project Helps me a lot to learn defining the problem then analysis process, Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries.
- It improve my analytical mindset for defining the problem by the Approach of ‘Five whys’
- I learn different features & uses of Pivot Table and also work with Charts and graphs, apply different formulas

THANK YOU!