IMDB Movie Analysis

# Project Description:

This is detailed overview of complete analysis of imdb\_movies. The dataset which used is imdb\_movie dataset. I was required to create a problem which I want to shed some light on.

# Tech Used:

**Microsoft excel** : excel is used for complete analysis done in this project.

**google drive**: drive is used for storing this ﬁle in it.

**canva**: canva is used to make this report.

# Approach:

In this project I followed 5 Whys technique to determine its root cause by repeatedly asking the question “Why”. It's also called root cause analysis. which is developed by sakichi toyoda founder of toyata industires. This approach is basically asking why in every step.

# Insights and reports:

I was required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

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. (Dropping columns, removing null values, etc.)

**Your task:** Clean the data

## Insights:

In this step I removed all the blank cells.

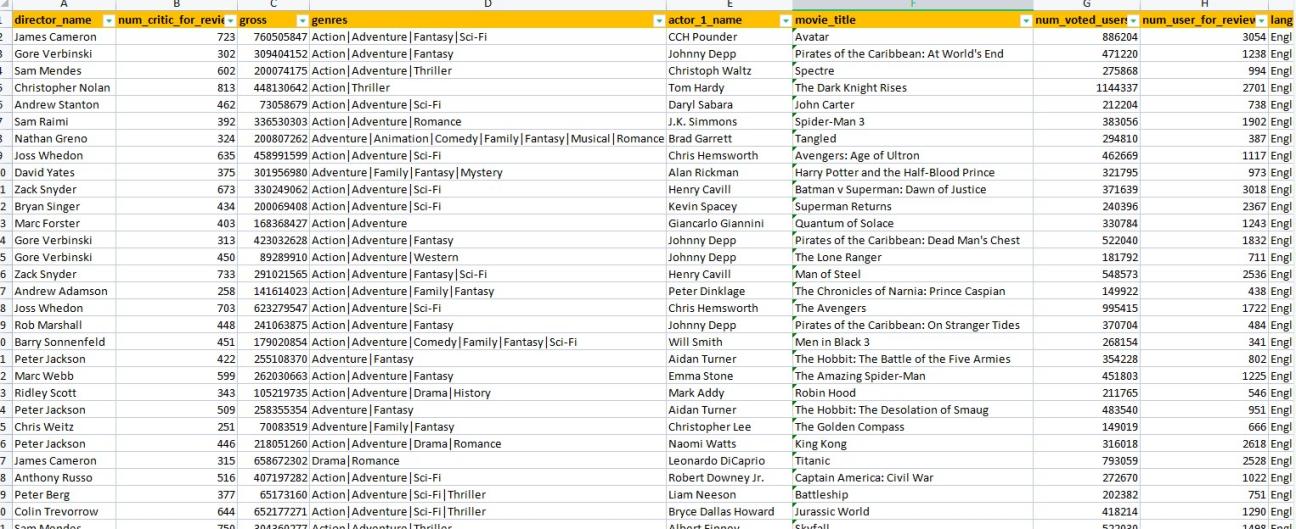
there are lots of dropping columns present in this dataset which has no use at all, so I removed all the dropping columns.

I even removed the duplicate value and made the data set look more appealing.

## Before cleaning:



**After Cleaning:**



1. **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?

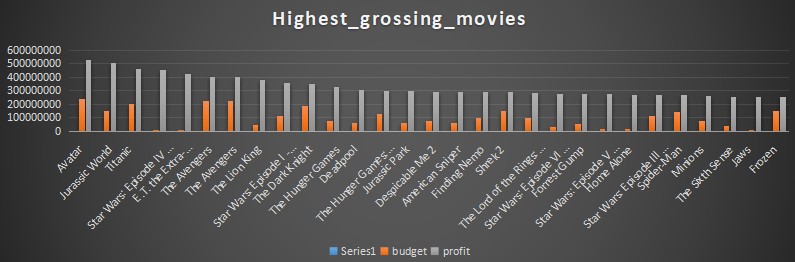
## Insights:

So basically proﬁt is all about gross - budget. In this step I simply reduce the gross from budget and kept them in separate column.

on the analysis of this movies, we came to know about the top highest grossed movies ever.

we can able to see the difference between the budget and proﬁt for these very clearly

with the help of charts.



1. **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!

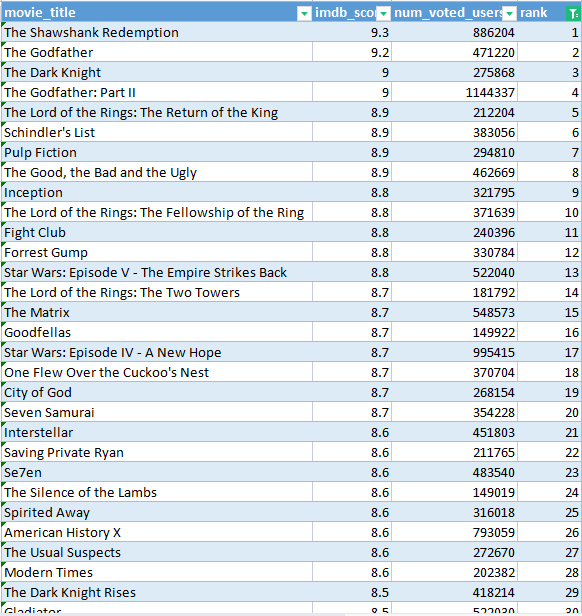
**Your task:** Find IMDB Top 250

## Insights:

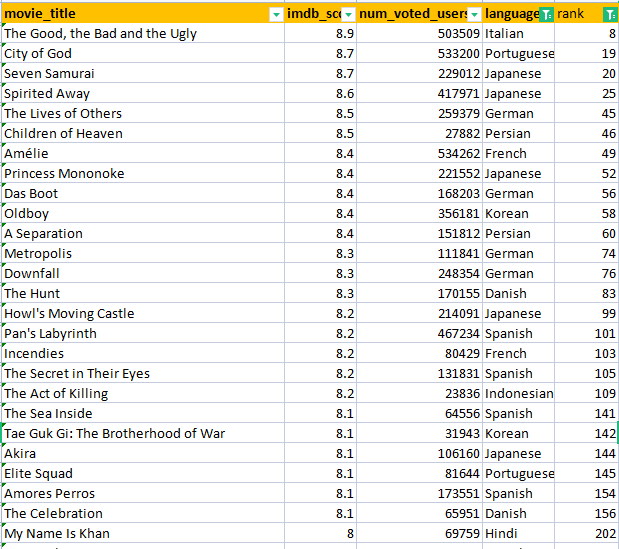
In this task we came to know about the imdb top 250 movies. about that movies we only group movies only in English language and rest of them in other column.

These top movies have filtered based on the num\_voted\_users. the movies which has num\_voted\_users is more than 25000 had made into this list.

## Top\_imdb\_movies:



**Top\_imdb\_movies in foreign languages:**

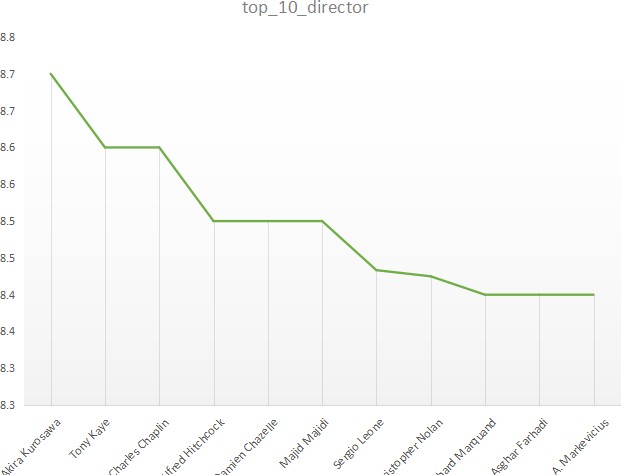




1. **Best Directors:** Group the column using the director\_name column. Find out the top 10 directors for whom the mean of imdb\_score is the highest and store them in a new column top 10 director.

**Your task:** Find the best directors

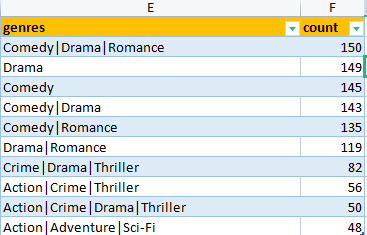


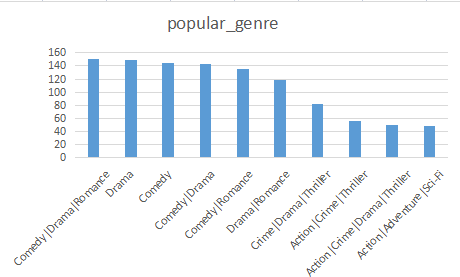




**E:Popular Genres:** Perform this step using the knowledge gained while performing previous steps.

**Your task:** Find popular genres





**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.

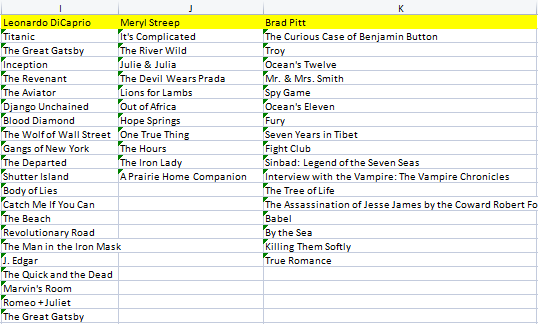
Find the mean of the num\_critic\_for\_reviews and num\_users\_for\_review and identify the

actors which have the highest mean.

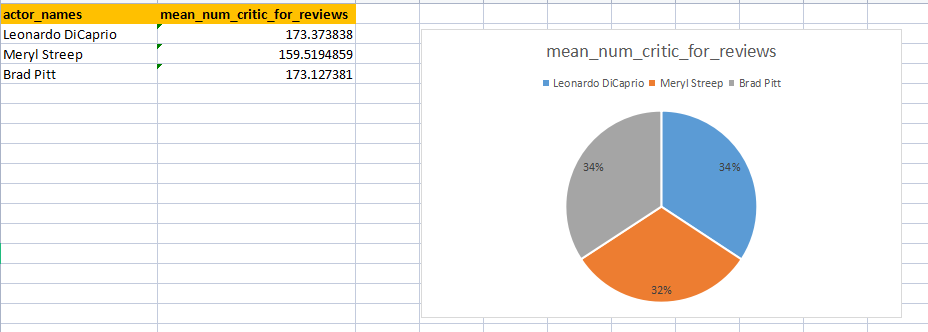
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title\_year year 1923, 1925 should be stored as 1920s. Sort the column based on

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

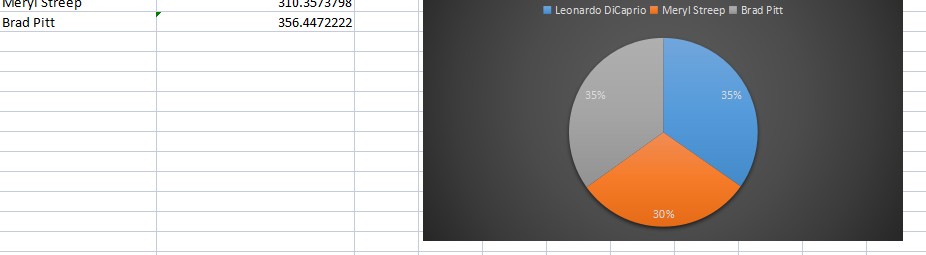


## critic favorite actors:



**Audience favorite actors:**





## Num\_voted \_user by decade:

**insights:**

In this graph we can clearly able to see the count of vote by users. In over period of time we can able to see the step rise.

