



IMDB Movies Analysis

TECH STACK USED: MICROSOFT EXCEL

Attulaya Singh | attulaya@gmail.com | 04/06/2023

1. Cleaning the data:

This is the most difficult and crucial step of any Data analysis project. The process included in this step varies from question to question and Dataset to Dataset. To clean the dataset we will be:-

1. First dropping the columns which have no use for the analysis that we will be doing

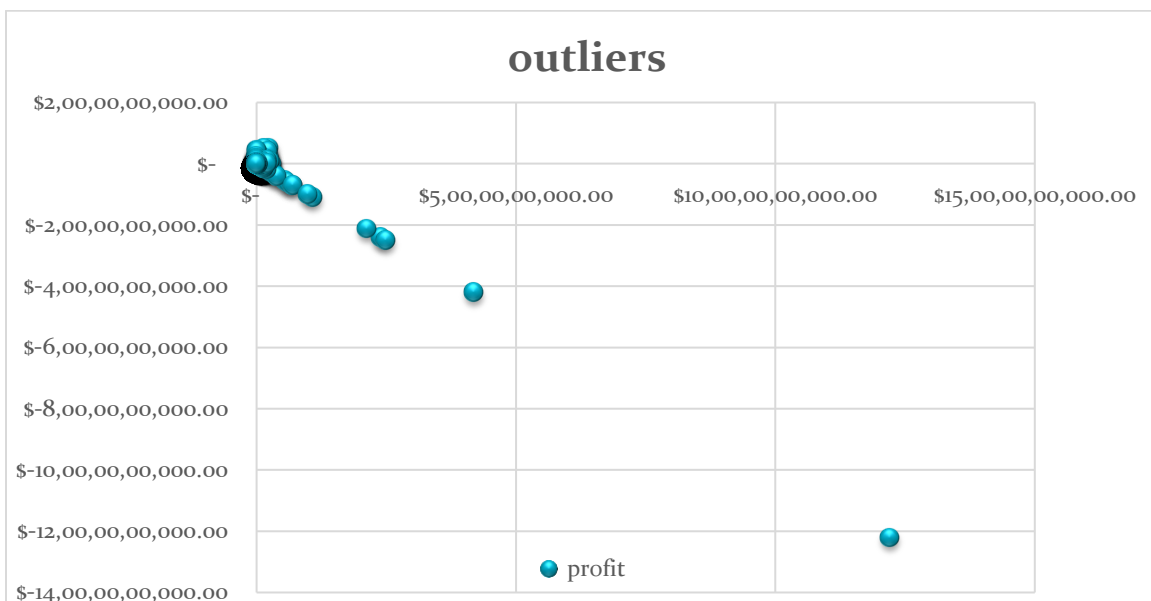
2. Columns like 'Color',
'director_facebook_likes',
'actor_3_facebook_likes', 'actor_2_name',
'actor_1_facebook_likes',
'cast_total_facebook_likes', 'actor_3_name',
'facenumber_in_posts', 'plot_keywords',
'movie_imdb_link', 'content_rating',
, 'actor_2_facebook_likes', 'aspect_ratio',
'movie_facebook_likes' are the columns containing irrelevant data for the analysis tasks provided. So, these columns needs to be dropped.

3. After dropping the irrelevant columns now we need to remove the rows from the dataset having anyone of its column value as blank/NULL

4. Then we need to get rid off the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab

2. Movies with highest profit:

1. First we need to subtract the budget value from the gross value to get the profit.
2. Then, by using the scatter plot option we will plot values of profit(y_axis) and budget(x_axis)
3. Then with the help of graph we will be finding the outliers

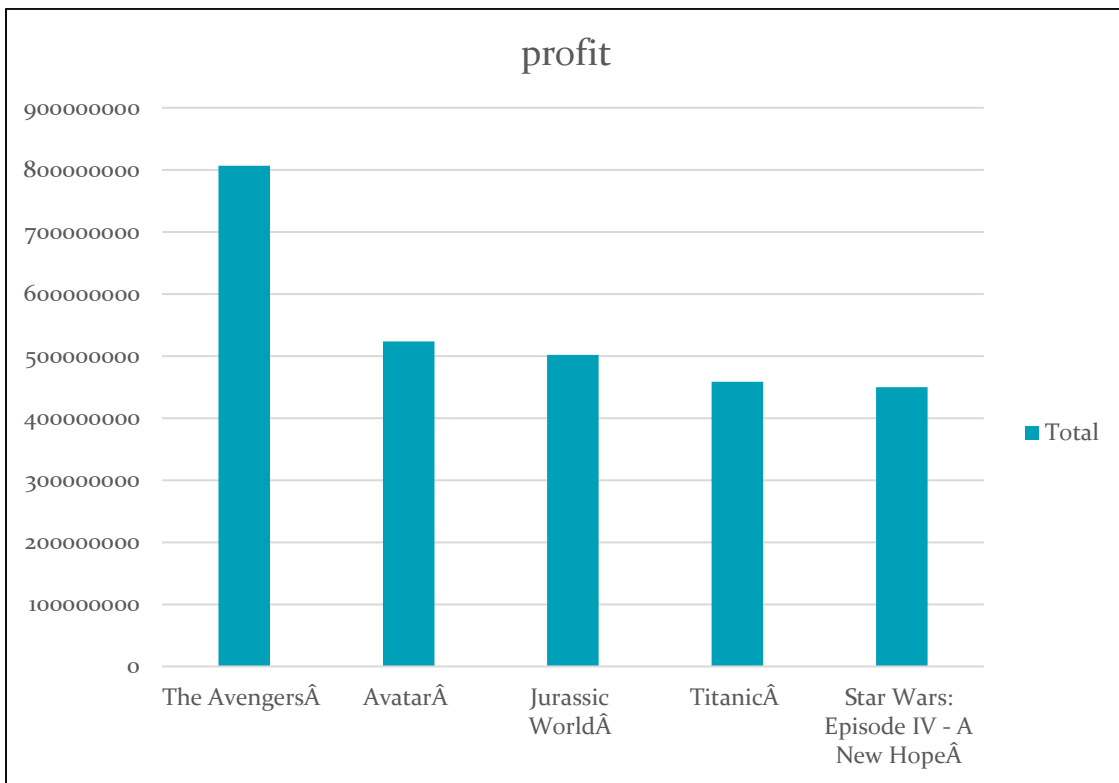


The outliers are:-

- -12213298588
- -4199788333
- -2499804112
- -2127109510
- -1099560838

Top 6 PROFITABLE Movies

movie Title	profit
The Avengers	\$ 80,65,59,094.00
Avatar	\$ 52,35,05,847.00
Jurassic World	\$ 50,21,77,271.00
Titanic	\$ 45,86,72,302.00
Star Wars: Episode IV - A New Hope	\$ 44,99,35,665.00



3. Top 250 IMDB movies:

We will use the pivot table and filter out those rows whose num_voted_users > 25000 then filter the pivot table to top 250 imdb movies

Top 10 English movies:

movie title	imdb_score	Sum of num_voted_users
The Shawshank Redemption	9.3	1689764
The Godfather	9.2	1155770
Dekalog	9.1	25181
Fargo	9	170055
The Dark Knight	9	1676169
The Godfather: Part II	9	790926
12 Angry Men	8.9	447785
Pulp Fiction	8.9	1324680
Schindler's List	8.9	865020
The Good, the Bad and the Ugly	8.9	503509

Top 10 non-english movies:

Movie title	Max of imdb_sco	Sum of num_voted_use
The Good, the Bad and the Ugly	8.9	503509
Seven Samurai	8.7	229012
City of God	8.7	533200
Spirited Away	8.6	417971
Children of Heaven	8.5	27882
The Lives of Others	8.5	259379
Amélie	8.4	534262
A Separation	8.4	151812
Princess Mononoke	8.4	221552
Baahubali: The Beginning	8.4	62756
Das Boot	8.4	168203
Oldboy	8.4	356181

4. Best Directors:

To find the best top 10 directors on the basis of mean of imdb_score we will:-

1. First select the imdb_score column of the cleaned dataset
2. Then we will click on pivot table
3. We will add director_name into the series section of the pivot table
4. Then we will add average imdb_score into the values section of the pivot table
5. Then we will first sort the data on the basis of average of imdb_score indescending order and then on the basis of director name alphabetically

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

5. Popular Genres:

To find the Popular Genres we will:-

1. First select the genres column of the cleaned dataset
2. Then we will go for the pivot table option
3. Then we will Select the genres name as row labels
4. Then we will the values as the count of the number of genres and then sort it in descending order based on count of the number of genres

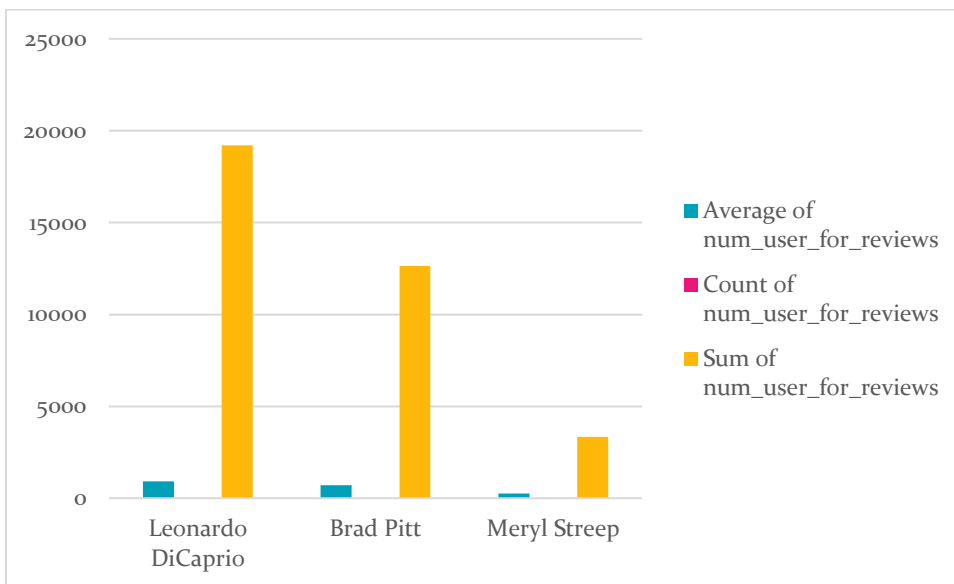
genres	Count of genres
Drama	154
Comedy Drama Romance	151
Comedy Drama	148
Comedy	147
Comedy Romance	136
Drama Romance	120
Crime Drama Thriller	83
Action Crime Thriller	56
Action Crime Drama Thriller	50

6. Charts:

We use the pivot table and select the data for actors brad pitt , leo caprio and meryl and find the critic and crowd favorite

Crowd Favorite

ACTOR	Average of num_user_for_reviews	Count of num_user_for_reviews	Sum of num_user_for_reviews
Leonardo DiCaprio	914.4761905	21	19204
Brad Pitt	702.4444444	18	12644
Meryl Streep	257.3076923	13	3345



Critic Favor:

Row Labels	Sum of num_critic_for_reviews	Average of num_critic_for_reviews2	Count of num_critic_for_reviews2
Leonardo DiCaprio	6934	330.1904762	21
Brad Pitt	4165	245	17
Meryl Streep	1996	181.4545455	11

