IMDB Movie Analysis

Submitted By: Miss. Aishwarya Ganapat Desai

Project Description

The Dataset provided for analysis contains information related to IMDB Movies. The success of any movie is determined by its IMDB ratings. In this project my role is to analyze what factors are affecting the success of a movie on IMDB which will help Movie Producers, Directors and Investors to take their decision for future Projects.

Approach

- **Download the Dataset:** This is the first step whille starting the analysis, which is downloading and converting the data in to suitable format.

- **Understanding the data:** After successfully downloading or converting the data, one need to understandthe labels given in the dataset. Also needs to understand problem statement.

- **Cleaning the data:** This is first step of analysis where we need to identify null, duplicate values from the data & cleanit with proper process. Also it involves to delete unnecessary column which will not be used during analysis

- **Analyzing the data:** After the cleaning next is to explore the dataset and find answers to the question.

- **Representation:** This is last but important step which willmake analysis representative and easy to understand to people.

Tech-Stack Used

- Microsoft Excel 2016 (for working, analysis purpose)
- Microsoft Word 2016 (for presentation purpose)

Data Preprocessing

Data cleaning:

Data cleaning is the most important step in analysis. Initially there were 5044 rows in the data and 28 columns, after that I deleted blank rows, some duplicate and null values also some unnecessary columns such as color, num_critic_for_reviews, director_facebook_likes,
actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, actor_1_name, num_voted_users, cost_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, num_user_for_reviews, actor_2_facebook_likes, movie_facebook_likes.

Now for analysis purpose I had data with 3784 rows and 12 columns.

Insights

TASK 1: Movie Genre Analysis

To determine most common genres of movie in the dataset and calculate descriptive statistics (mean, median, range, variance, standard deviation)

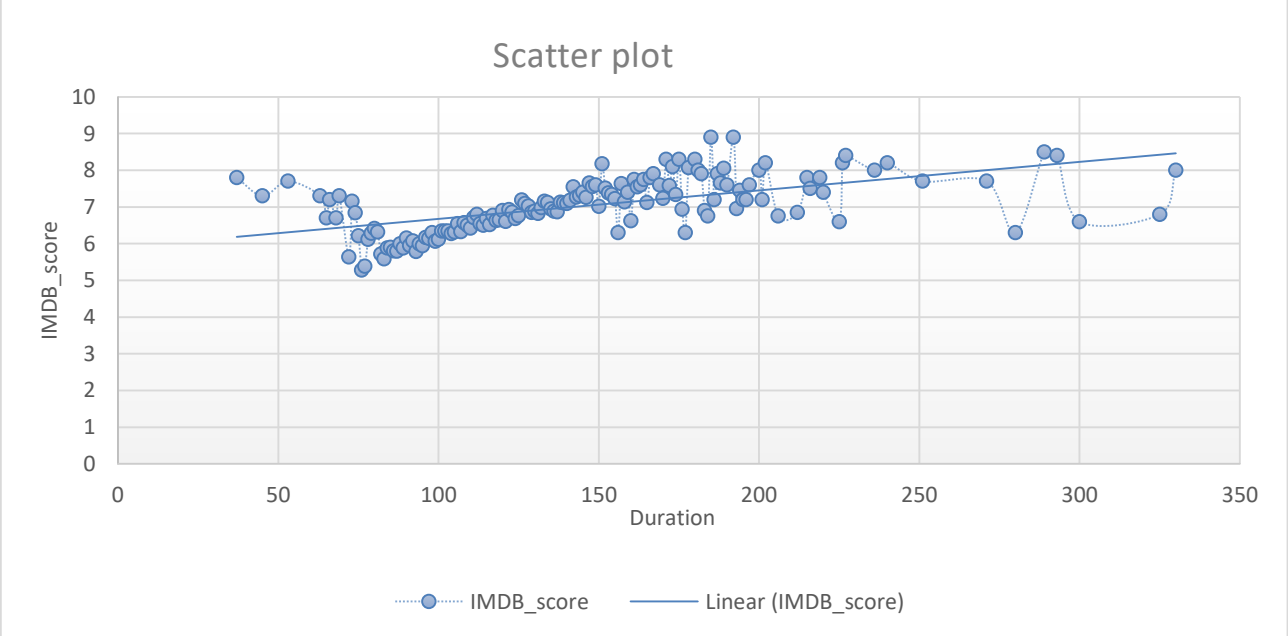
Genres	Count of IMDB_score
Action	962
Adventure	371
Animation	46
Biography	207
Comedy	991
Crime	256
Documentary	29
Drama	679
Family	3
Fantasy	37
Horror	165
Musical	2
Mystery	24
Romance	1
Sci-Fi	7
Thriller	1
Western	3
Grand Total	**3784**

Genre wise Descriptive Statistics of IMDB Scores

Genres	Count	Mean	Max	Min	Median	StdDev
Action	962	6.288773389	9	2.1	6.3	1.038699035
Adventure	371	6.55309973	8.6	2.3	6.7	1.121923945
Animation	46	6.763043478	8	4.5	7	0.972593028
Biography	207	7.153140097	8.9	4.5	7.2	0.698178834
Comedy	991	6.166801211	8.8	1.9	6.3	1.033703629
Crime	256	6.94140625	9.3	3.3	7	0.867588711
Documentary	29	6.793103448	8.5	1.6	7.4	1.670742144
Drama	679	6.824300442	8.8	2.1	6.9	0.905822727
Family	3	6.5	7.9	5.7	5.9	1.216552506
Fantasy	37	6.281081081	7.9	4.3	6.5	0.894066191
Horror	165	5.850909091	8.5	2.3	5.9	1.032083979
Musical	2	6.75	7.2	6.3	6.75	0.636396103
Mystery	24	6.608333333	8.5	3.3	6.7	1.0898411
Romance	1	7.1	7.1	7.1	7.1	#NUM!
Sci-Fi	7	6.628571429	8.2	5	6.4	1.107119815
Thriller	1	4.8	4.8	4.8	4.8	#NUM!
Western	3	6.766666667	8.9	4.1	7.3	2.444040371
Grand Total	**3784**	**6.464006342**	**9.3**	**1.6**	**6.6**	**1.05654033**

Task 2: Movie Duration Analysis

To Analyze distribution of movie durations and identify relationship between movie duration and IMDB score.



Mean	145.4837
Median	140
Standard Deviation	56.90598

Task 3: Language Analysis

To determine the most common languages used in movies and to analyze their impact on the IMDB score using descriptive statistics.

Language	Count of movie_title
	1
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	7
Czech	1
Danish	3
Dari	2
Dutch	3
English	3624
Filipino	1
French	34
German	11
Hebrew	1
Hindi	5
Hungarian	1
Indonesian	2
Italian	7
Japanese	10
Kazakh	1
Korean	5
Mandarin	15
Maya	1
Mongolian	1

language	Count of movie_title
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	23
Thai	3
Vietnamese	1
Zulu	1
Grand Total	**3784**

Language wise Descriptive Statistics of IMDB Score

Language	Count of imdb_score	Average	StdDevp	Median
	1	5.8	0	5.8
Aboriginal	2	6.95	0.55	6.95
Arabic	1	7.2	0	7.2
Aramaic	1	7.1	0	7.1
Bosnian	1	4.3	0	4.3
Cantonese	7	7.342857143	0.324509048	7.3
Czech	1	7.4	0	7.4
Danish	3	7.9	0.43204938	8.1
Dari	2	7.5	0.1	7.5
Dutch	3	7.566666667	0.329983165	7.8
English	3624	6.425827815	1.05072803	6.5
Filipino	1	6.7	0	6.7
French	34	7.355882353	0.51173935	7.3
German	11	7.763636364	0.644237008	7.8
Hebrew	1	8	0	8
Hindi	5	7.22	0.716658915	7.4
Hungarian	1	7.1	0	7.1
Indonesian	2	7.9	0.3	7.9
Italian	7	7.185714286	1.069617517	7
Japanese	10	7.66	0.939361485	8
Kazakh	1	6	0	6
Korean	5	7.7	0.509901951	7.7
Mandarin	15	7.08	0.745832868	7.4
Maya	1	7.8	0	7.8
Mongolian	1	7.3	0	7.3
None	1	8.5	0	8.5
Norwegian	4	7.15	0.497493719	7.3
Persian	3	8.133333333	0.449691252	8.4
Portuguese	5	7.76	0.875442745	8
Romanian	1	7.9	0	7.9
Russian	1	6.5	0	6.5
Spanish	23	7.082608696	0.841660974	7.2
Thai	3	6.633333333	0.368178701	6.6
Vietnamese	1	7.4	0	7.4
Zulu	1	7.3	0	7.3
Grand Total	**3784**	**6.464006342**	**1.056400714**	**6.6**

Task 4: Director Analysis

To identify top directors on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

10 Directors with highest average of IMDB Score

Director name	Average of IMDB_score
Akira Kurosawa	8.7
Alfred Hitchcock	8.5
Asghar Farhadi	8.4
Charles Chaplin	8.6
Christopher Nolan	8.425
Damien Chazelle	8.5
Majid Majidi	8.5
Richard Marquand	8.4
Ron Fricke	8.5
Sergio Leone	8.433333333
Tony Kaye	8.6
Grand Total	**8.47**

Percentile	1

Task5: Budget Analysis

To analyze the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

Correlation	0.10034

Movie Name	Highest Profit Margin
Avatar	523505847

Results

- Most common genre in the movie is Comedy having highest count of movies i.e. 991
- Film-noir genre has highest average IMDB score ie.7.6 whereas Comedy genre has highest maximum value for IMDB score is 9.5 and Documentary has lowest min value for IMDB Score ie.1.6
- Average duration for the movie is 145.48 with median 140 and std deviation is 56.90
- Scatter plot shows that duration 185 has highest IMDB score ie. 9 while duration 76 has lowest ie.5.28 & it kept fluctuating between them.
- Most common language used in the movie is English with highest count 3624
- Director Akira Kurosawa has highest average IMDB score which is 8.7
- Percentile is 8.7 and Percentilerank is 1 Directors have 100% success rate.
- Correlation value 0.10034 between movie budgets and ross earning shows that there is relatively week positive linear relation between them.
- Movie Avatar has highest profit margin ie. 523505847

Hyperlink of excel worksheet:

https://docs.google.com/spreadsheets/d/1-DT8zLkYZrjKDWVfD2j67iJOYlhaZTzV/edit?usp=sharing&ouid=112713524583719045365&rtpof=true&sd=true

Thank You