IMDB Movie Analysis

Project Description

The project is about finding out the various insights in the IMDB_Movies dataset. We analyze this data and answered the following questions:
1. Clean the data
2. Movie Genre Analysis
3. Movie Duration Analysis
4. Language Analysis
5. Director Analysis
6. Budget Analysis

Approach

Understanding Data: Before analysis, it is important to understand the data. Look at the structure of data and get a sense of overall content, This helps in identifying any potential issues or challenges that need to be addressed.

Cleaning Data: Once any issue is identified in data it is important to resolve and clean the data for better analysis. This includes handling missing or incomplete data, identifying and handling outliers, and correcting text mistakes.

Analyzing Data: Once the data is cleaned it is analyzed using various formulas and functions for finding out insights and answering the business questions.

Visualizing Data: After doing analysis the insights are visualized using various charts and plots, so that insights can be understood by stakeholders and better business decisions can be taken.

Tech stack

Used Microsoft Excel 2016 for cleaning, analyzing, and visualizing data. And used Google Docs to create reports.

Data Cleaning:

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.
1. First drop 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.

After Cleaning the data we have 3849 Rows and 14 Columns

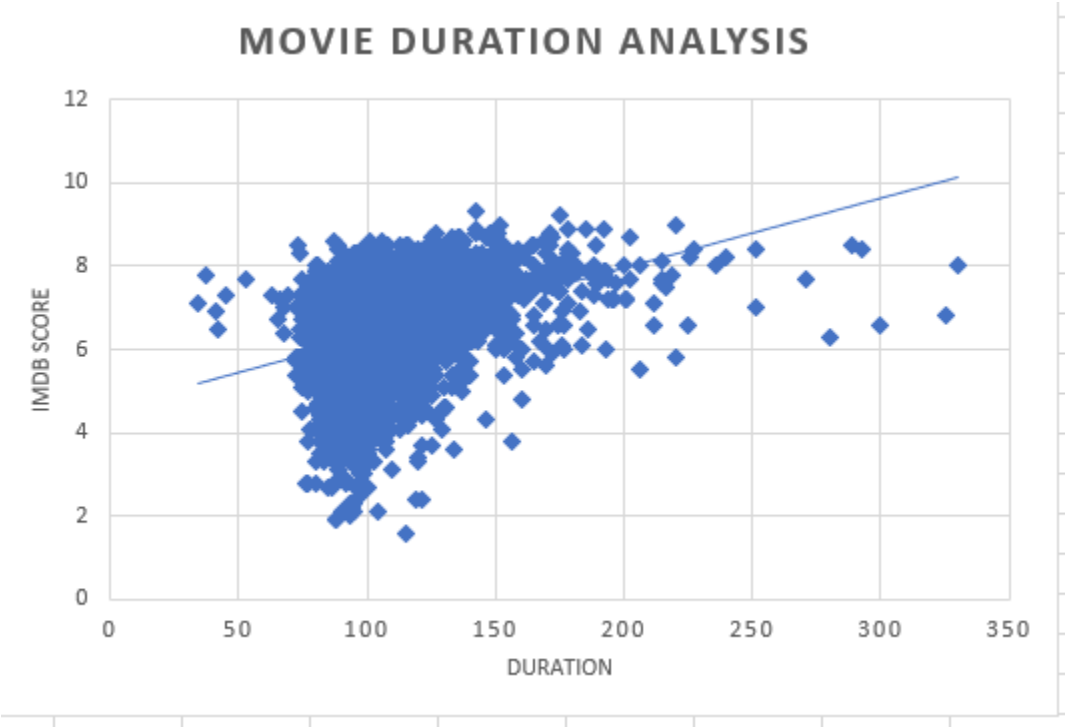
Insights

A) Movie Genre Analysis: Determine the most common genres of movies in the dataset and calculate descriptive statistics of the IMDB scores.

Genre	Count	Mean	Median	Mode	Var	Stdev
Action	962	6.290748441	6.3	6.1	1.063754	1.03138454
Adventure	787	6.458322745	6.6	6.7	1.227980	1.10814299
Animation	199	6.700502513	6.8	6.7	0.977322	0.98859646
Biography	243	7.141563786	7.2	7	0.502686	0.70900409
Comedy	1503	6.184497671	6.3	6.7	1.077409	1.03798329
Crime	714	6.544817927	6.6	6.6	0.960709	0.98015786
Documen	64	6.99375	7.2	6.6	1.499642	1.22459906
Drama	1940	6.786804124	6.9	6.7	0.793740	0.89092096
Family	450	6.210444444	6.3	6.7	1.351583	1.16257615
Fantasy	514	6.290077821	6.4	6.7	1.276607	1.12987035
Horror	391	5.926086957	6	5.9	0.994189	0.99709052
Musical	103	6.559223301	6.7	7.1	1.301850	1.14098657
Romance	877	6.429190422	6.5	6.5	0.935608	0.96726838
Sci-Fi	497	6.322736419	6.4	6.7	1.336558	1.15609628
Thriller	1117	6.378066249	6.4	6.5	0.933470	0.96616256
Western	58	6.765517241	6.8	6.8	0.997035	0.99851674

B) Movie Duration Analysis: Analyze the distribution of movie durations and perform Descriptive Statistics on IMDB score.

Descriptive Statistics	
Mean	109.9241
Median	106
Mode	101
Max	330
Min	34
Var	517.7286
Stdev	22.75365



C) Language Analysis: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Language	Count	Mean	Median	Max	Min	Stdev
Aborigina	2	6.95	6.95	7.5	6.4	0.777817459
Arabic	1	7.2	7.2	7.2	7.2	#DIV/0!
Aramaic	1	7.1	7.1	7.1	7.1	#DIV/0!
Bosnian	1	4.3	4.3	4.3	4.3	#DIV/0!
Cantonese	8	7.2375	7.3	7.8	6.5	0.440575922
Czech	1	7.4	7.4	7.4	7.4	#DIV/0!
Danish	3	7.9	8.1	8.3	7.3	0.529150262
Dari	2	7.5	7.5	7.6	7.4	0.141421356
Dutch	3	7.566667	7.8	7.8	7.1	0.404145188
Dzongkha	1	7.5	7.5	7.5	7.5	#DIV/0!
English	3659	6.423909	6.5	9.3	1.6	1.048750752
Filipino	1	6.7	6.7	6.7	6.7	#DIV/0!
French	37	7.286486	7.2	8.4	5.8	0.561328861
German	13	7.692308	7.7	8.5	6.1	0.640912811
Hebrew	3	7.5	7.3	8	7.2	0.435889894
Hindi	10	6.76	7.05	8	4.8	1.111755369
Hungarian	1	7.1	7.1	7.1	7.1	#DIV/0!
Icelandic	1	6.9	6.9	6.9	6.9	#DIV/0!
Indonesia	2	7.9	7.9	8.2	7.6	0.424264069
Italian	6	7.185714	7	8.9	5.3	1.155318962
Japanese	11	7.625	7.8	8.7	6	0.899621132

D) Director Analysis: identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Director	Imdb score	Average Imdb score	Movie Count	percentile
Tony Kaye	8.6	8.6	1	0.992
Charles Chaplin	8.6	8.6	1	0.992
Alfred Hitchcock	8.5	8.5	1	0.987
Ron Fricke	8.5	8.5	1	0.987
Damien Chazelle	8.5	8.5	1	0.987
Majid Majidi	8.5	8.5	1	0.987
Sergio Leone	8.4	8.433333333	3	0.987
Sergio Leone	8.9	8.433333333	3	0.987
Sergio Leone	8	8.433333333	3	0.987
Christopher Nolan	8.5	8.425	8	0.987
Christopher Nolan	9	8.425	8	0.987
Christopher Nolan	8.6	8.425	8	0.987
Christopher Nolan	8.8	8.425	8	0.987
Christopher Nolan	8.3	8.425	8	0.987
Christopher Nolan	7.2	8.425	8	0.987
Christopher Nolan	8.5	8.425	8	0.987
Christopher Nolan	8.5	8.425	8	0.987
S.S. Rajamouli	8.4	8.4	1	0.983
Richard Marquand	8.4	8.4	1	0.983

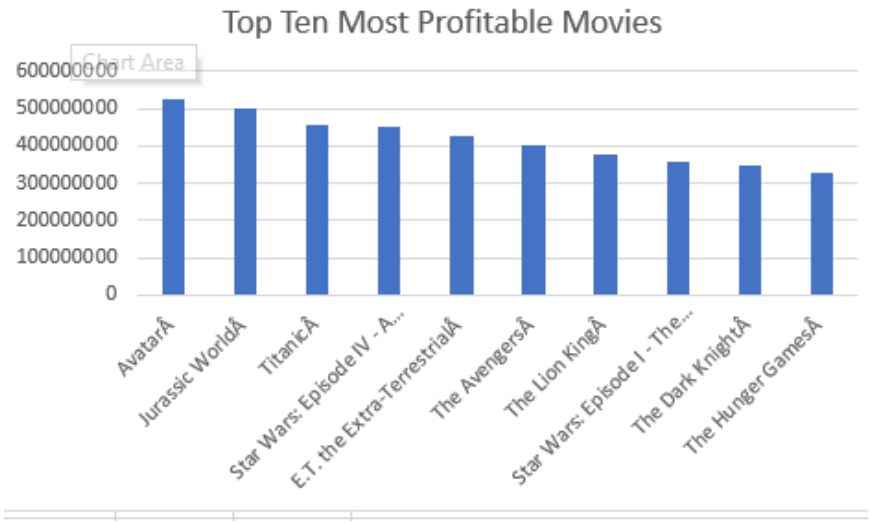
Top Ten Directors	
Directors	Average IMDB Score
Tony Kaye	8.6
Charles Caplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.4
Christopher Nolan	8.4
S.S. Rajamouli	8.4
Richard Marquand	8.4

E) Budget Analysis: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Corr =	0.10085	MAX =	523505847

Top Ten Most Profitable Movies

Movie	Profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255



RESULT

We collected, cleaned and analyzed data to gather useful insights and visualized them for better understanding.

Conclusions from insights:

- The most common Movie genre is **Drama** with a count of **1940** followed by **Comedy** with a count of **1503**.
- The majority distribution of movies is between 90 **min** to **150 min** and it is visible that as the duration increases the IMDb score also increases.
- We can observe that the maximum number of movies are made in the **English** language with a count of **3659**.
- We have analyzed the directors based on average IMDb score and calculated their percentile score. The director with the highest average IMDb score is **Tony Kaye**.
- The **Correlation Coefficient** between movie budget and gross earnings is **0.10085** and the movie with highest profit is **Avatar** with a profit of **523505847$**

Excel file:

IMDB Movie Analysis Trainity Project 5 (Jaskirat Singh).xlsx

Jaskirat Singh