Project5: IMDB Movie Analysis

Project Description: The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Approach:

1. Download the dataset
2. Remove extra columns which are not required.
3. Handle the missing values, removing duplicates if any.
4. Use Excel formulas for analysis.
5. Showcase results in an easy format using the Five “whys” approach.

Tech-Stack Used: I used MS Excel 2019 for analysis.

Insights:

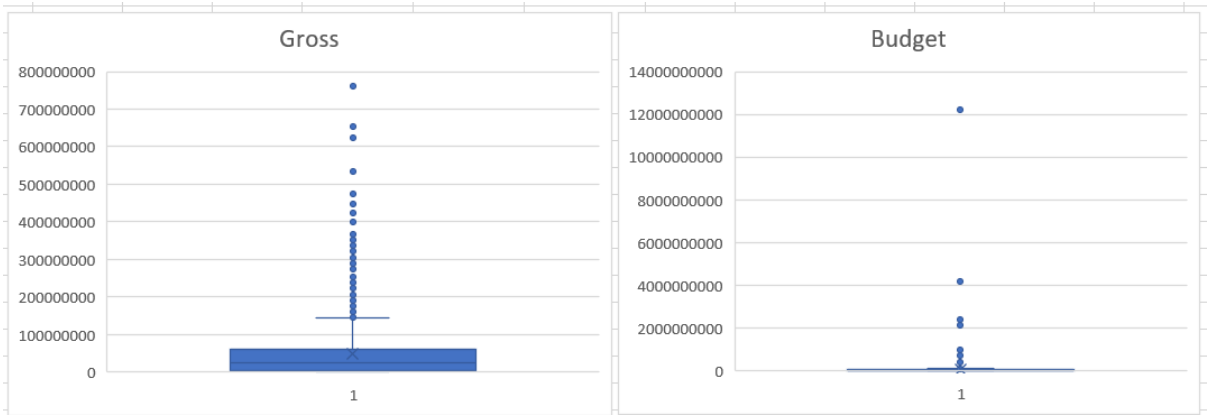
1. Maximum number of movies are made on Drama genre.
2. Documentary and Biography have the highest IMDB rating mean.
3. English is the most used language in movies.
4. Movie with maximum profit in my is “The Blair Witch Project”.

Link to the excel sheet: https://docs.google.com/spreadsheets/d/1BBVgb0DnIutP_rfa91Fz-bPsMgHvplEX/edit?usp=sharing&ouid=115109770037321084146&rtpof=true&sd=true

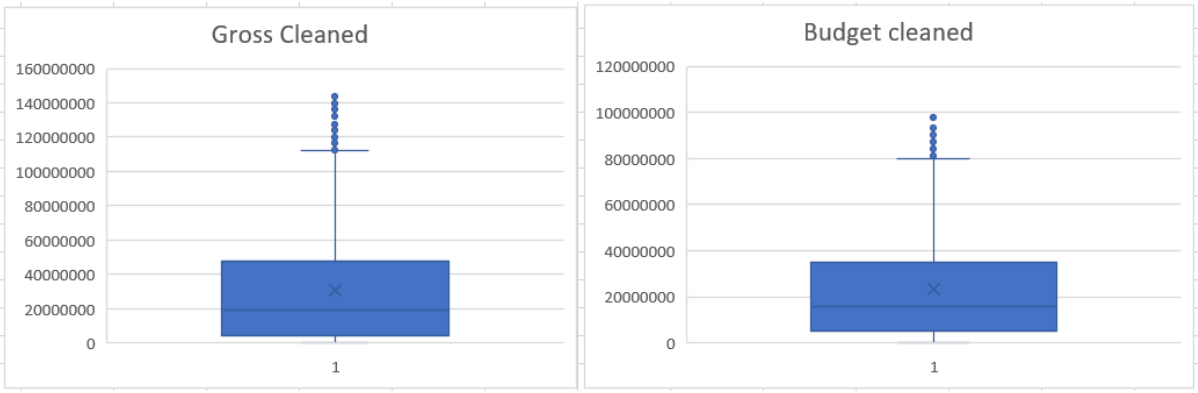
Cleaning the data:

I cleaned the data in the following way:

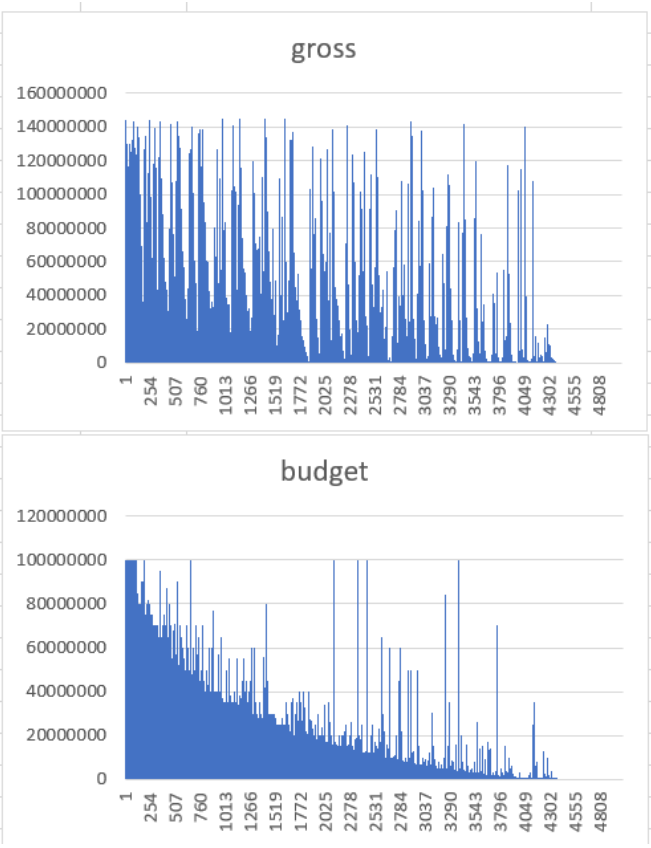
1. Deleted the unwanted columns that are not required in our analysis.
2. Deleted the duplicate rows by using the delete duplicates option in excel. Excel caught 122 duplicate rows that were deleted, 4922 rows remain.
3. Removed the characters like Ã‰, Â from Directors’s name and Movie name respectively.
4. Removed the blank values from language column, director name column 12 rows and 101 rows deleted respectively, 4809 rows left after deletion.
5. Removed outliers from Gross(292) and Budget(298) columns but after removing Gross outliers, Budget outliers reduced to 158 rows, hence the data rows left are now 4359.
   Plot of Gross and Budget with outliers:



Plot of Gross and Budget with outliers removed:



6. There are 755 blank values in Gross column after removing outliers hence it can't be deleted, thus filling in the value with median of gross as the plot of values is not a normal plot but a skewed one.
7. There are 388 blank values in Budget column after removing outliers hence it can't be deleted, thus filling in the value with median of Budget values as the plot of values is not a normal plot but a skewed one.



Data Analytics tasks:

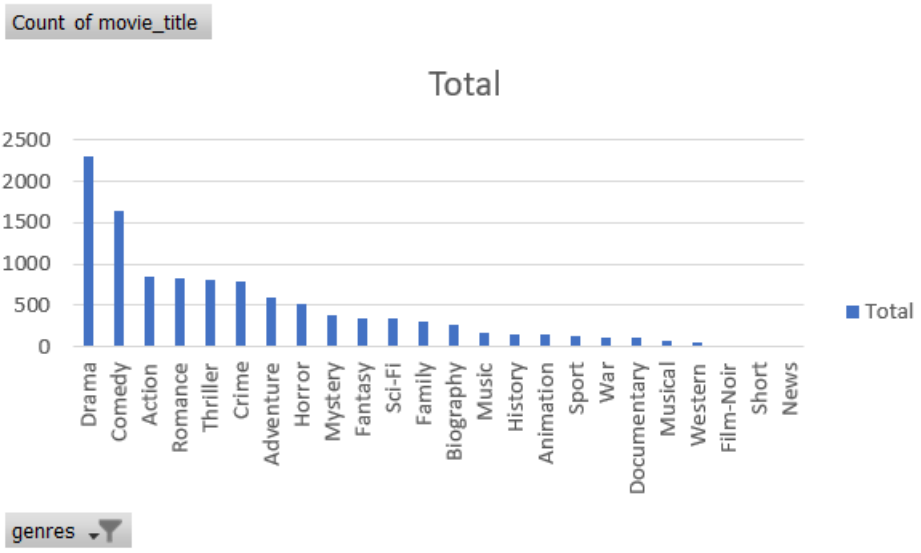
A. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.

- **Task:** Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Results and Insights:

According to the analysis the most common genre of movies in the dataset is “Drama” with a total count of 2336 movies.

Genre	Count of movie_title
Drama	2336
Comedy	1655
Thriller	1217
Romance	989
Action	855
Crime	802
Adventure	596
Horror	525
Sci-Fi	428
Mystery	426
Fantasy	425
Family	390
Biography	276
Music	203
History	183
War	174
Sport	162
Animation	145
Documentary	118
Musical	109
Western	80
Film-Noir	6
Short	5
News	3



Genre	Count of movie_title	MEAN	MEDIAN	MODE	MAX	MIN	Variance	STD DEV
Drama	2336	6.72	6.8	6.7	9.3	2	0.88	0.94
Comedy	1655	6.12	6.2	6.3	9.5	1.7	1.17	1.08
Thriller	1217	6.24	6.3	6.5	8.7	2.2	1.10	1.05
Romance	989	6.42	6.5	6.5	8.6	2.1	0.96	0.98
Action	855	6.04	6.2	6.1	8.7	1.7	1.20	1.10
Crime	802	6.50	6.5	6.3	9.3	2.4	1.01	1.00
Adventure	596	6.21	6.3	6.6	8.7	1.9	1.33	1.15
Horror	525	5.77	5.9	6.2	8.7	2.2	1.22	1.10
Sci-Fi	428	6.04	6.2	6.3	8.5	1.9	1.48	1.22
Mystery	426	6.37	6.45	6.4	8.6	2.2	1.14	1.07
Fantasy	425	6.11	6.2	6.7	8.6	1.7	1.34	1.16
Family	390	6.00	6.1	5.8	8.6	1.7	1.49	1.22
Biography	276	7.13	7.2	7	8.9	4.5	0.52	0.72
Music	203	6.43	6.6	6.7	8.5	1.6	1.38	1.18
History	183	7.07	7.1	7.5	8.9	2	0.80	0.89
War	174	7.06	7.1	7.1	8.6	2.7	0.80	0.90
Sport	162	6.60	6.8	7.2	8.4	2	1.27	1.13
Animation	145	6.27	6.5	7	8.6	1.7	1.44	1.20
Documentary	118	7.18	7.4	7.5	8.7	1.6	1.14	1.07
Musical	109	6.38	6.7	7	8.3	2.1	1.62	1.27
Western	80	6.68	6.8	6.8	8.9	3.8	1.10	1.05
Film-Noir	6	7.63	7.65	#N/A	8.2	7.1	0.19	0.43
Short	5	6.38	6.5	#N/A	7.1	5.2	0.56	0.75
News	3	7.53	7.4	#N/A	8.1	7.1	0.26	0.51

For analysis, I will be considering the genres with **more than 100** and ignoring last four genres:

1. We notice that **Documentary and Biography** have the highest IMDB Mean (7.17,7.13), IMDB Median (7.4,7.2).
2. The genre with the highest consistency in IMDB ratings is **History** with IMDB rating Mode of 7.5.
3. Max rating goes to Comedy with 9.5 after that Crime and Drama with 9.3 rating each.
4. The genre that got the minimum rating is Music and Documentary at a point of time.
5. Biography has lowest Standard Dev of 0.7.

B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

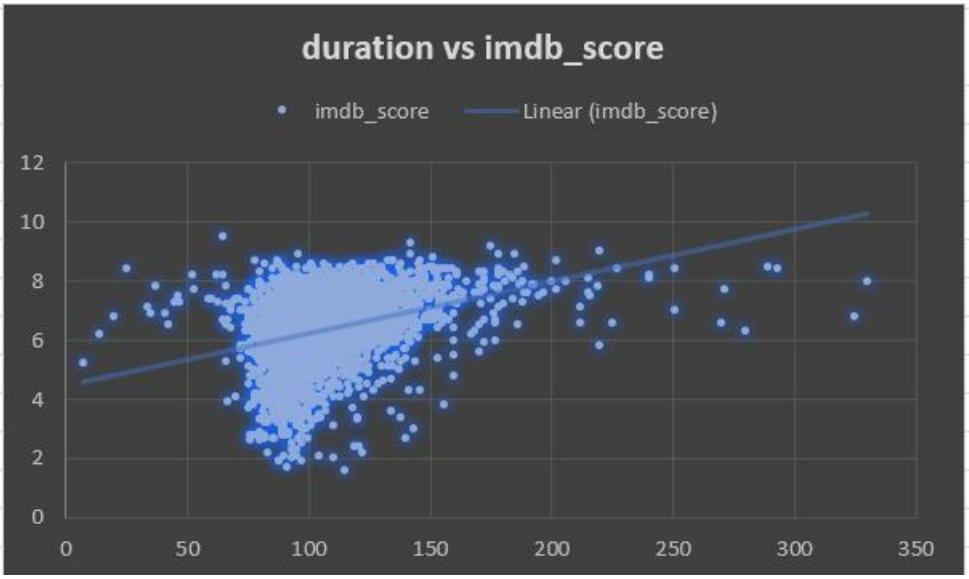
- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Result and Insights:

1. The mean and median don't have much difference hence there are no outliers and the movie duration are normally distributed.
2. Standard deviation is ~21 which is a large value, which shows high deviation in the values from mean.
3. There is little correlation between duration and imdb score as the graph is tilted in the positive direction. For some movie durations the rating increases on increase in duration while for some it doesn't hence it is not a exactly positive correlation.

   The correlation coefficient between duration and imdb score is 0.33.

Mean of movie duration:	106.5451
Median of movie duration:	102
Standard Deviation:	21.4088
Correlation coeff between imdb score and duration:	0.335889



C. Language Analysis: Situation: Examine the distribution of movies based on their language.

- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Results and Insights:

I will be considering the Languages with 10 or more movies made.

1. Considering this we notice that most of the movies are made in English and after that French.
2. The mean and median of English is less and almost same but has a high standard deviation of 1.1 hence it has a high deviation from mean value.
3. The highest IMDB rating is for German and Japanese movies and not much standard deviation from mean.

Languages	Count of imdb_score	MEAN	MEDIAN	Std Dev
English	4058	6.32	6.40	1.12
French	71	7.04	7.20	0.71
Spanish	39	6.93	7.10	0.87
Hindi	25	6.57	6.90	1.44
Mandarin	23	6.76	7.00	1.05
German	19	7.34	7.60	0.95
Japanese	12	7.38	7.60	1.05
Russian	11	6.36	6.50	1.38
Cantonese	11	6.95	7.20	0.70
Italian	10	7.08	7.15	1.21
Portuguese	8	7.49	7.70	0.88
Korean	6	7.40	7.50	0.95
Arabic	5	7.38	7.40	0.88
Swedish	5	7.44	7.60	0.76
Hebrew	5	7.58	7.60	0.33
Danish	5	7.50	8.10	1.08
Norwegian	4	7.15	7.30	0.57
Dutch	4	7.43	7.45	0.43
Persian	4	7.58	7.95	1.20
Chinese	3	5.67	5.70	0.55
Indonesian	2	7.90	7.90	0.42
Romanian	2	7.20	7.20	0.99
Dari	2	7.50	7.50	0.14
Zulu	2	7.10	7.10	0.28
Aboriginal	2	6.95	6.95	0.78
None	2	7.95	7.95	0.78

D. Director Analysis: Influence of directors on movie ratings.

- Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Results and Insights:

I am considering 95th percentile to be the criteria for top directors, hence calculating the percentile value, it comes to be 7.8.

On analyzing the data, 118 directors have an average IMDB rating to be greater than or equal to 7.8.

Director names	Average of imdb_score
John Blanchard	9.50
Sadyk Sher-Niyaz	8.70
Mitchell Altieri	8.70
Cary Bell	8.70
Mike Mayhall	8.60
Charles Chaplin	8.60
Ron Fricke	8.50
Raja Menon	8.50
Majid Majidi	8.50
Damien Chazelle	8.50
Sergio Leone	8.48
Robert Mulligan	8.40
S.S. Rajamouli	8.40
Rakeysh Omprakash Mehra	8.40
Moustapha Akkad	8.40
Marius A. Markevicius	8.40
Jay Oliva	8.40
Catherine Owens	8.40
Bill Melendez	8.40
Asghar Farhadi	8.40
Sut Jhally	8.30
Stanley Donen	8.30
Lenny Abrahamson	8.30
Justin Paul Miller	8.30
John Sturges	8.30

E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

- Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Results and Insights:

Correlation coefficient between gross and budget:	0.5102492
MAX Profit:	140470114
The Blair Witch Project is the movie with highest profit	

The correlation coefficient of 0.5 indicates that there is a slight positive correlation between the movie budget and its success/earnings. If a huge amount of money is invested in making a movie, it will have better visuals and better promotion, the people will enjoy watching the movie and will rate it more and the movie will be a success.

The movie with the maximum profit in my dataset is “The Blair Witch Project”.