Project 5

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:

This project is executed using excel. It involves identifying the relation between the columns of the data from the excel worksheet, removing unnecessary columns that are not required for the analysis, removing rows that contain null values, identifying duplicate rows and deleting duplicate rows, removing unnecessary special characters (like "Â") in any specific columns, creating new column using the existing columns to easily analyze data and identify the insights from the data.

Tech-Stack Used:

The tech-stack used for this project is EXCEL 2019 version.

Insights:

Case 1: Data cleaning

1) Removed unnecessary columns that are not required for analysis:

‘’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, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, num_user_for_reviews, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes’’ columns were deleted from the copy of original data sheet, so that the data can be easily analysed.

2) **Handling Missing Data:**

“director_name, duration, gross, language, budget” columns had some rows in which the data was missing, those rows were deleted, as they affected in analyzing the data. A total of 1156 rows were deleted.

3) Removing duplicates:

Duplicate rows are identified and deleted from the datasheet. A total of 101 rows were deleted from the datasheet.

4) Handling unnecessary special characters:

Unnecessary special character "Â" was removed movie_title column from all rows.

5) New column:

Profit/Loss column was created using gross and budget columns by using the formula.

=gross-budget

6) Outlier Detection:

There are 5 outliers in the Profit/Loss column of the dataset. The outliers are not deleted from dataset as they are not miss entered data, they might be special data points that are important to analysis. Therefore, median can be considered for average calculation as the median dose get not affected by outliers.

-12213298588, -4199788333, -2499804112, -2397701809, -2127109510 are the outliers that are identified in the dataset.

Case 2: Data Analytics Tasks

1) 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:

The 'genres' column was separated to multiple genres for each movie.

genres	imdb_score	genres 1	genres 2	genres 3	genres 4	genres 5	genres 6	genres 7	genres 8
Comedy|Drama|Horror|Sci-Fi	7	Comedy	Drama	Horror	Sci-Fi				
Crime|Drama	7.7	Crime	Drama						
Drama|Romance|War	7.1	Drama	Romance	War					
Adventure|Animation|Fantasy	8.4	Adventure	Animation	Fantasy					
Action|Adventure|Animation|Family|Sci-Fi|Thriller	6.9	Action	Adventure	Animation	Family	Sci-Fi	Thriller		
Action|Animation|Sci-Fi	8.1	Action	Animation	Sci-Fi					
Action|Adventure|Drama|Sci-Fi|Thriller	6	Action	Adventure	Drama	Sci-Fi	Thriller			
Drama|Musical	7.2	Drama	Musical						
Drama	6	Drama							
Action|Drama|Romance|Thriller	6	Action	Drama	Romance	Thriller				
Action|Adventure|Drama|History|War	7.4	Action	Adventure	Drama	History	War			
Action|Adventure|Drama|History|War	6.6	Action	Adventure	Drama	History	War			
Adventure|Biography|Drama|History|War	6.4	Adventure	Biography	Drama	History	War			
Action	6.2	Action							
Action|Adventure|Sci-Fi	6.6	Action	Adventure	Sci-Fi					
Action|Crime|Drama|Thriller	7.1	Action	Crime	Drama	Thriller				
Adventure|Animation|Family|Fantasy	6.6	Adventure	Animation	Family	Fantasy				
Comedy|Drama	7.2	Comedy	Drama						
Documentary	8	Documentary							
Action|Adventure|Sci-Fi|Thriller	5.9	Action	Adventure	Sci-Fi	Thriller				
Comedy|Family|Fantasy|Sci-Fi	6	Comedy	Family	Fantasy	Sci-Fi				
Action|Adventure|Drama|Fantasy	6.3	Action	Adventure	Drama	Fantasy				
Adventure|Fantasy	6.3	Adventure	Fantasy						
Action|Adventure|Sci-Fi	5.4	Action	Adventure	Sci-Fi					
Action|Adventure|Animation|Comedy|Family|Sci-Fi	5.4	Action	Adventure	Animation	Comedy	Family	Sci-Fi		
Action|Adventure|Western	6.5	Action	Adventure	Western					
Action|Adventure|Biography|Drama|History|Romance|War	5.5	Action	Adventure	Biography	Drama	History	Romance	War	
Adventure|Family|Fantasy	5.8	Adventure	Family	Fantasy					
Action|Adventure|Fantasy	7.3	Action	Adventure	Fantasy					
Action|Crime|Romance|Thriller	3.7	Action	Crime	Romance	Thriller				
Adventure|Family|Fantasy	6.1	Adventure	Family	Fantasy					
Action|Adventure|Fantasy|Romance	6.6	Action	Adventure	Fantasy	Romance				
Action|Adventure|Fantasy	5.5	Action	Adventure	Fantasy					
Action|Adventure|Animation|Fantasy|Romance|Sci-Fi	6.4	Action	Adventure	Animation	Fantasy	Romance	Sci-Fi		
Action|Adventure|Sci-Fi|Thriller	5	Action	Adventure	Sci-Fi	Thriller				
Biography|Crime|Drama|War	7.3	Biography	Crime	Drama	War				
Adventure|Animation|Family|Sci-Fi	7.1	Adventure	Animation	Family	Sci-Fi				
Action|Adventure|Drama|Thriller	5.6	Action	Adventure	Drama	Thriller				
Drama|Romance	7.3	Drama	Romance						

The count of number of movies for each genre is calculated. The Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics of the genres.

genres	Count of genres 1	Count of genres 2	Count of genres 3	Count of genres 4	Count of genres 5	Count of genres 6	Count of genres 7	Count of genres 8
Action	935	929	855	515	194	45	11	2
Adventure	367	365	328	220	110	26	9	1
Animation	46	46	46	38	20	3		
Biography	206	206	175	71	15	1		
Comedy	1026	881	460	123	23	4		
Crime	252	252	194	76	5			
Documentan	43	19	4					
Drama	676	524	260	89	8			
Family	3	3	2					
Fantasy	35	35	20	9	1			
Horror	156	115	52	5				
Musical	2	2						
Mystery	23	23	7					
Romance	2	1	1					
Sci-Fi	8	7						
Thriller	3							
Western	3							

Genres	Number_of_movies	Mean	Median	Mode	Maximum	Minimum	Variance	Standard deviation
Action	3486	6.285989305	6.3	6.6	9	2.1	1.077033647	1.037802316
Adventure	1426	6.454960836	6.6	6.6	8.9	2.3	1.245895756	1.116197006
Animation	199	6.700507614	6.8	7.3	8.6	2.8	0.982284006	0.99110242
Biography	674	7.140082645	7.2	7	8.9	4.5	0.502153712	0.708628049
Comedy	2517	6.183310992	6.3	6.7	8.8	1.9	1.080706732	1.039570455
Crime	779	6.548148148	6.6	6.6	9.3	2.4	0.967083465	0.983404019
Documentary	66	7.011940299	7.2	7.6	8.5	1.6	1.418364892	1.190951255
Drama	1557	6.789115646	6.9	6.7	9.3	2.1	0.793581165	0.890831726
Family	8	6.2	6.3	5.4	8.6	1.9	1.364807256	1.168249655
Fantasy	100	6.285080645	6.4	6.7	8.9	2.2	1.297922574	1.139264049
Horror	328	5.903957784	5.9	5.9	8.6	2.3	0.979535787	0.989715003
Musical	4	6.550980392	6.7	7.1	8.5	2.1	1.29485198	1.13791563
Mystery	53	6.469496021	6.5	6.6	8.6	3.1	1.01214643	1.006054884
Romance	4	6.426212471	6.5	6.5	8.5	2.1	0.937869488	0.968436621
Sci-Fi	15	6.327272727	6.4	7	8.8	1.9	1.359504132	1.165977758
Thriller	3	6.372309108	6.4	6.5	9	2.7	0.938248854	0.968632466
Western	3	6.765517241	6.8	6.8	8.9	4.1	0.979845422	0.989871417

Mean of IMDB score for each genre is calculated using the function:

{=AVERAGE(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Median of IMDB score for each genre is calculated using the function:

{=MEDIAN(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Mode of IMDB score for each genre is calculated using the function:

{=MODE(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Maximum of IMDB score for each genre is calculated using the function:

{=MAX(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Minimum of IMDB score for each genre is calculated using the function:

{=MIN(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Variance of IMDB score for each genre is calculated using the function:

{=VAR.P(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Standard deviation of IMDB score for each genre is calculated using the function:

{=STDEV.P(IF(($D$2:$D$3787=M25) + ($E$2:$E$3787=M25) + ($F$2:$F$3787=M25) + ($G$2:$G$3787=M25) + ($H$2:$H$3787=M25) + ($I$2:$I$3787=M25) + ($J$2:$J$3787=M25) + ($K$2:$K$3787=M25), $C$2:$C$3787))}

Insights:

Action genre has the highest number of count of movies of 3486 movies. Thriller and western genres have the least number of count of movies of 3 movies.

Biography genre has the highest mean and median of 7.14 and 7.2 respectively. Comedy genre has the least mean and median of 6.18 And 6.3 respectively.

Crime and drama genres have the maximum IMDB scores of 9.3. Comedy and sci-fi has the least IMDB score of 1.9.

Documentary genre has the highest standard deviation of 1.19. Drama has least IMDB score of 0.89.

2) 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.

Results:

Descriptive statistics of duration of movies is:

Mean	Median	Mode	Maximum	Minimum	Variance	Standard Deviation
109.809	105	101	330	34	518.03	22.76019457

Mean of duration of movies is calculated using the function:

=AVERAGE(A:A)

Median of duration of movies is calculated using the function:

=MEDIAN(A:A)

Mode of duration of movies is calculated using the function:

=MODE(A:A)

Maximum duration of movies is calculated using the function:

=MAX(A:A)

Minimum duration of movies is calculated using the function:

=MIN(A:A)

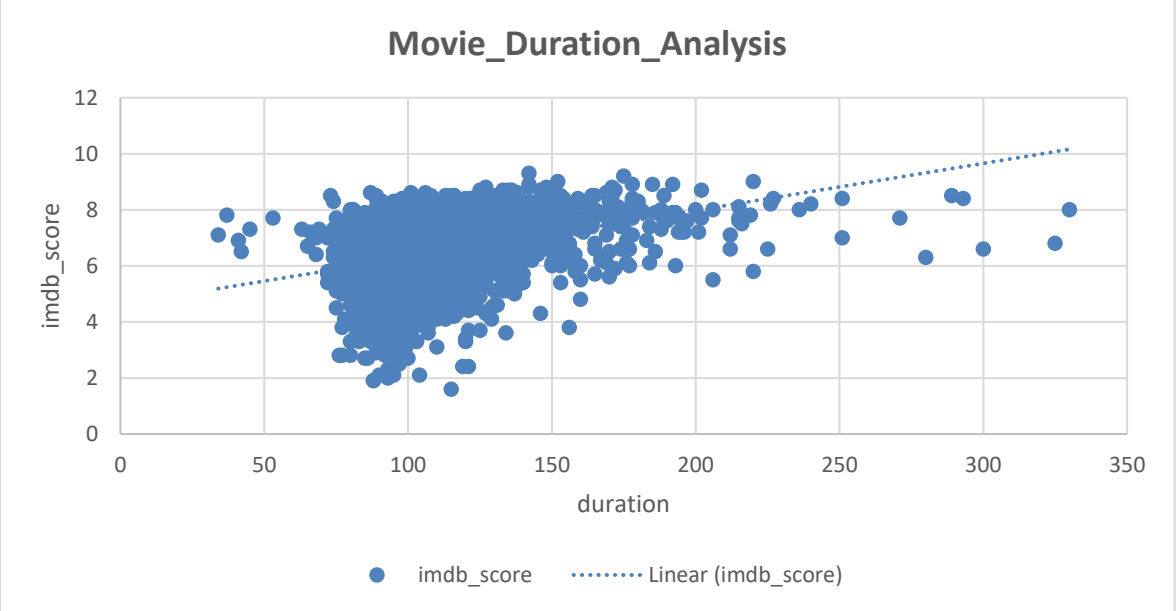
Variance of duration of movies is calculated using the function:

=VAR.P(A:A)

Standard deviation of duration of movies is calculated using the function:

=STDEV.P(A:A)

scatter plot is created to visualize the relationship between movie duration and IMDB score.



Insights:

The movies have the mean, median, mode duration of 109.8, 105, 101 respectively.

The maximum and minimum duration of movies is 330 and 34 respectively.

The standard deviation of duration of movies is 22.76.

The scatter plot shows visualization of the relationship between movie duration and IMDB score.

A positive trendline can be seen in the scatter plot which shows that as the duration of the movies increases the IMDB score is also higher. It is also seen that movies with duration 0 to 60 and duration greater than 160 are having IMDB scores of greater than 5.

3) **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:

Language	count_of_movies	Mean	Median	Standard Deviation
Aboriginal	2	6.95	6.95	0.55
Arabic	1	7.20	7.20	0.00
Aramaic	1	7.10	7.10	0.00
Bosnian	1	4.30	4.30	0.00
Cantonese	8	7.24	7.30	0.41
Czech	1	7.40	7.40	0.00
Danish	3	7.90	8.10	0.43
Dari	2	7.50	7.50	0.10
Dutch	3	7.57	7.80	0.33
Dzongkha	1	7.50	7.50	0.00
English	3606	6.42	6.50	1.05
Filipino	1	6.70	6.70	0.00
French	37	7.29	7.20	0.55
German	13	7.69	7.70	0.62
Hebrew	3	7.50	7.30	0.36
Hindi	10	6.76	7.05	1.05
Hungarian	1	7.10	7.10	0.00
Icelandic	1	6.90	6.90	0.00
Indonesian	2	7.90	7.90	0.30
Italian	7	7.19	7.00	1.07
Japanese	12	7.63	7.80	0.86
Kazakh	1	6.00	6.00	0.00
Korean	5	7.70	7.70	0.51
Mandarin	14	7.02	7.25	0.74
Maya	1	7.80	7.80	0.00
Mongolian	1	7.30	7.30	0.00
None	1	8.50	8.50	0.00
Norwegian	4	7.15	7.30	0.50
Persian	3	8.13	8.40	0.45
Portuguese	5	7.76	8.00	0.88
Romanian	1	7.90	7.90	0.00
Russian	1	6.50	6.50	0.00
Spanish	26	7.05	7.15	0.81
Swedish	1	7.60	7.60	0.00
Telugu	1	8.40	8.40	0.00
Thai	3	6.63	6.60	0.37
Vietnamese	1	7.40	7.40	0.00
Zulu	1	7.30	7.30	0.00

The figure shows common languages used in movies with their mean, median and standard deviation of their IMDB scores.

Count of each language is calculated using the function:

=COUNTIF(IMDB_Movies_cleaned!$J:$J,Task_3!A4)

Mean of each language is calculated using the function:

=AVERAGEIF(IMDB_Movies_cleaned!$J:$J,Task_3!A4,IMDB_Movies_cleaned!$K:$K)

Median of each language is calculated using the function:

{=MEDIAN(IF(IMDB_Movies_cleaned!$J:$J=Task_3!A4,IMDB_Movies_cleaned!$K:$K))}

Standard deviation of each language is calculated using the function:

{=STDEV.P(IF(IMDB_Movies_cleaned!$J:$J=Task_3!A4,IMDB_Movies_cleaned!$K:$K))}

Insights:

Language English has the highest count of 3606 movies followed by French and Spanish with 37 and 26 movies.

Language Telugu has largest mean and median of 8.40 each but since, there are many languages in which count of movies is less than 10. Therefore, if we consider languages with count of movies greater than 10, then German has largest mean and median of 7.69 and 7.70 respectively. English has the largest standard deviation of 1.05.

4) 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:

There are a total of 1751 directors. The figure below shows the list of directors based on their average IMDB scores. The directors are arranged in descending order of their average IMDB scores.

director_name	average_imdb_score
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40
Richard Marquand	8.40
S.S. Rajamouli	8.40
Billy Wilder	8.30
Fritz Lang	8.30
Lee Unkrich	8.30
Lenny Abrahamson	8.30
Pete Docter	8.23
Hayao Miyazaki	8.23
Quentin Tarantino	8.20
Elia Kazan	8.20
George Roy Hill	8.20
Joshua Oppenheimer	8.20
Juan JosÃ© Campanella	8.20
Victor Fleming	8.15
Milos Forman	8.13
Akira Kurosawa	8.10
David Sington	8.10
Je-kyu Kang	8.10
Michael Roemer	8.10
Michael Wadleigh	8.10
Terry George	8.10
Tim Miller	8.10
William Wyler	8.10
Ari Folman	8.00

Average IMDB scores of each director is calculated using the function:

=AVERAGEIF(IMDB_Movies_cleaned!$A:$A,A4,IMDB_Movies_cleaned!$K:$K)

Top 5% percentile of average_imdb_score directors are assumed to be top directors.

95th percentile of average_imdb_score is calculated using the function:

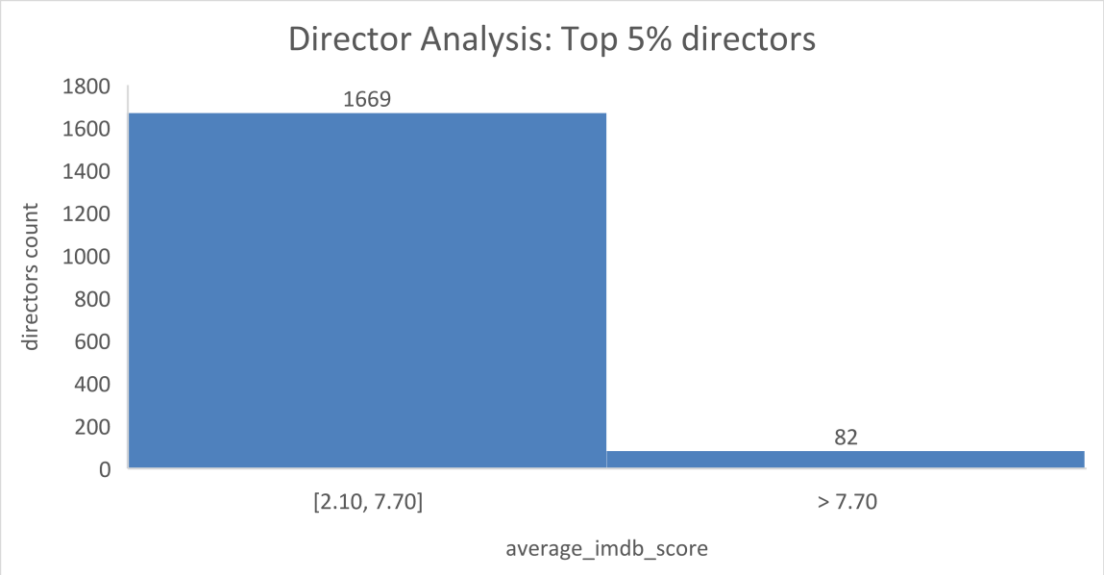
=PERCENTILE(B4:B1754,0.95)

95th percentile of average_imdb_score	
7.7	95% of the average_imdb_score is equal to or below 7.7

Count of top directors is calculated using the function:

=COUNTIF(B4:B1754,">7.7")

Top 5% percentile of average_imdb_score directors are assumed to be top directors	
Count of top directors	82



The above figure shows the histogram of their top 5% percentile of directors based on their average IMDB scores.

Insights:

Charles Chaplin has the highest average IMDB score of 8.60.

Out 1751 directors, if consider top 5% percentile directors as the top directors than 82 directors are top directors based on their average IMDB scores of their movies.

5) 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:

A new column profit/Loss is created from gross and budget.

correlation coefficient between movie budgets and gross earnings	0.0965689
movie with the highest profit margin	523505847
Movie Name	Avatar
Avatar is the movies with the highest profit margin	

Correlation coefficient between movie budgets and gross earnings is calculated using the function:

=CORREL(IMDB_Movies_cleaned!H:H,IMDB_Movies_cleaned!C:C)

Movie with the highest profit margin is found out using the function:

=MAX(IMDB_Movies_cleaned!L:L)

Movie Name of the highest profit margin is found out using the function:

=LOOKUP(B5,IMDB_Movies_cleaned!L:L,IMDB_Movies_cleaned!E:E)

Insights:

The correlation coefficient between movie budgets and gross earning is 0.0965 which indicates that there is weak or no correlation between them.

The movie with the highest profit margin is Avatar with a profit margin of 523505847.

Excel drive link:

The excel sheet has been saved and uploaded to google drive. To access the file following link can be used: https://docs.google.com/spreadsheets/d/1yzPpH3tVz1fAX3bwQxwyPmABeOXFCIho/edit?usp=sharing&ouid=116970069599597283204&rtpof=true&sd=true