Trainity Assignments

Project 5 – IMDB Movie Analysis

Task:

Project Description:

The aim of this project is to examine a film's commercial success by looking at its imdb scores. IMBD is a well-known website that rates films and television shows and offers financial information, reviews, and series profiles. The IMDB Movies dataset is provided, which may be utilised to look into what aspects affect a film's success. For the purpose of making well-informed decisions about their next projects, movie producers, directors, and investors need to have high IMDB ratings. Determining a film's success may be made easier with an understanding of these elements. In order to gain insights, we must first complete these steps:

Data Cleaning: In order to improve data analysis, the IMDB Movies dataset is pre-processed in this stage to eliminate duplicate values, missing values, and convert data types.

Data Analysis: In order to determine the correlation between the variables that affect movie ratings, we must compare the statistical relationships between the variables in this stage and comprehend the relationships between the various columns.

Five Why's Approach: By this method, I can find root cause of the problems by digging deeper into the problems using 'Why?'

I am able to deduce the problem's insights after doing data analysis. My task is to make the findings easier to understand through the use of visualisations. In addition to providing answers, a data analyst's role also includes offering relevant insights that can influence choices and guarantee that interested parties make well-informed choices. The following queries need to be answered:

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.
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.
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.

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.

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.

Approach:

I worked on the IMDB Movie Analysis in Microsoft Excel using pivot tables and complex functions in order to finish the task. First, the IMDB Movies database is imported into Excel, after which the dataset's data is examined and an analysis is done. The stages I took in the EDA procedure were:

Recognising the links between the records
Cleaning up data:
1. Remove the unnecessary columns, such as colour, Plot_keywords, movie_imdb_link, content_rating, actor2_facebook_likes, aspect_ratio, director_facebook_likes, actor3_facebook_likes, actor2_name, actor1_facebook_likes, actor1_name, cast_total_facebook_likes, actor3_name, facenumber_in_poster, and actor3_name
2. Next, the missing values need to be examined. The records must be deleted if any row value is empty.
3. Next, duplicate values must be eliminated. 35 duplicate values were discovered in the dataset and removed.
· Recognising and managing the anomalies
. Recapitulating the results

Followed by data cleaning is done, the dataset comprises 3851 distinct records utilized for addressing inquiries. Employing suitable formulas, functions such as average, median, mode, count, minimum, maximum, sum, etc., alongside pivot tables and graphical representations, I've successfully visualized and extracted insights from the data.

Tech-Stack Used:

- Microsoft Excel 2019 version → For analysis and visualization
- Microsoft Word 2019 version → To prepare report

Insights:

Although this project provided me with a thorough and practical understanding of the Excel capabilities and visualisations, I already had some basic hands-on experience with Microsoft Excel and datasets. Through this assignment, I was able to upgrade my Excel skills from beginner to expert. I can now write and assess the functions and formulas, and I can improve the visualisations by using tables, pivot tables, and charts. As a result, this

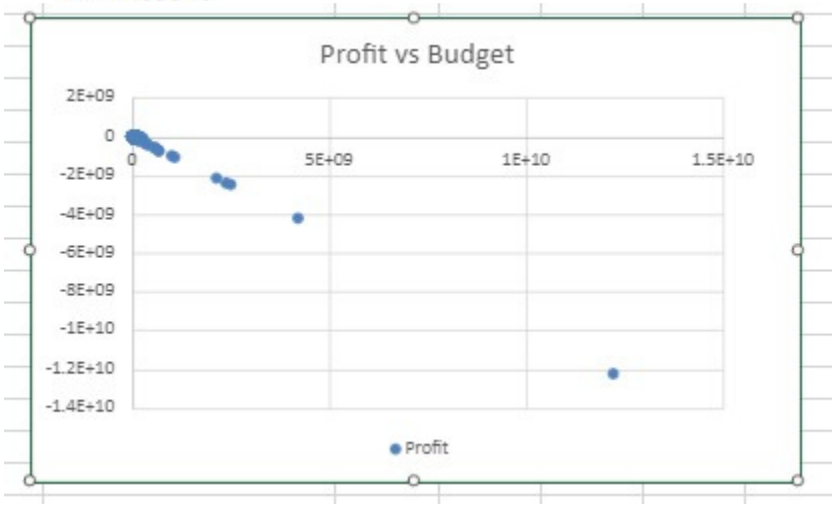
endeavour improved my ability to precisely analyse data, which could help the movie succeed.

Results:

Data Cleaning:

Total records before cleaning	5043
Total records after cleaning	3851

Outliers
-12213298588
-4199788333
-2499804112
-2397701809
-2127109510



Interpretation:

Analysis can be done with the pre-processed IMDB Movies dataset that is provided.

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.

Formulas used are:

=COUNTIF(A:A,"*"&D5&"*")

=MAXIFS(B:B,A:A,"*"&D5&"*")

=MINIFS(B:B,A:A,"*"&D5&"*")

=AVERAGEIF(A:A,"*"&D5&"*",B:B)

Output:

Genre	No. of Movies	Max imdb	Min imdb	Avg Imdb
Action	962	9	2.1	6.29
Adventure	787	8.9	2.3	6.46
Animation	199	8.6	2.8	6.70
Biography	243	8.9	4.5	7.14
Comedy	1503	8.8	1.9	6.18
Crime	714	9.3	2.4	6.54
Documentary	67	8.5	1.6	7.01
Drama	1941	9.3	2.1	6.79
Family	450	8.6	1.9	6.21
Fantasy	514	8.9	2.2	6.29
Film-Noir	1	7.7	7.7	7.70
History	153	8.9	5.5	7.14
Horror	391	8.6	2.3	5.93
Music	248	8.5	1.6	6.46
Musical	103	8.5	2.1	6.56
Mystery	383	8.6	3.1	6.48
Romance	877	8.5	2.1	6.43
Sci-Fi	497	8.8	1.9	6.32
Short	2	7.1	6.5	6.80
Sport	151	8.4	2	6.60
Thriller	1117	9	2.7	6.38
War	160	8.6	4.3	7.05
Western	58	8.9	4.1	6.77
(blank)	0	0	0	0.00

Interpretation:

Stats	Values
MAX	1941
MIN	1
AVG	480.04
MEDIAN	315.5
VARIANCE	253524.5
STANDARD DEVIATION	503.51

Comparing genres and ratings reveals Drama flicks dominate. Most movies belong to this type, averaging an impressive 9.3 IMDb score. Clearly, dramas outshine alternatives - the analysis confirms their superior performance.

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

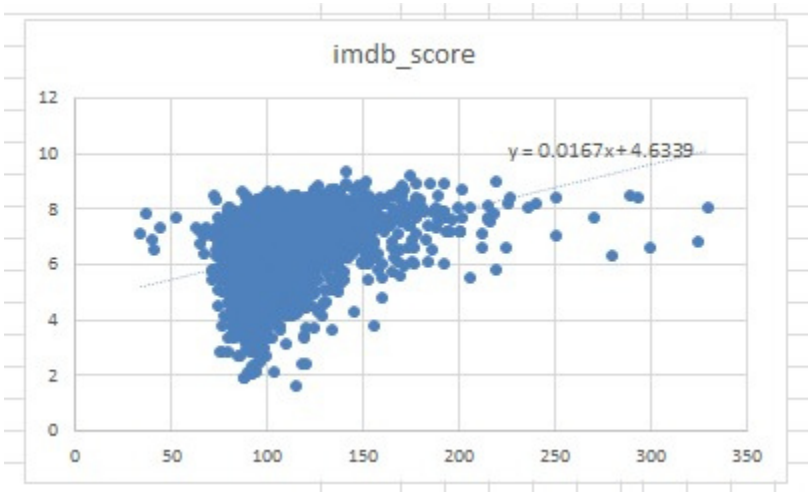
Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score

Pivot table:

Impact on imdb score of the movies based on the duration of the movie:

duration	Average of imdb_score	Max of imdb_score	Min of imdb_score
31-40	7.45	7.8	7.1
41-50	6.90	7.3	6.5
51-60	7.70	7.7	7.7
61-70	6.98	7.3	6.4
71-80	6.19	8.5	2.8
81-90	5.97	8.6	1.9
91-100	6.06	8.4	2
101-110	6.40	8.6	2.1
111-120	6.68	8.5	1.6
121-130	6.85	8.8	2.4
131-140	7.02	8.7	3.6
141-150	7.33	9.3	4.3
151-160	7.20	9	3.8
161-170	7.45	8.6	5.6
171-180	7.69	9.2	5.9
181-190	7.61	8.9	6.1
191-200	7.56	8.9	6
201-210	7.36	8.7	5.5
211-220	7.47	9	5.8
221-230	7.73	8.4	6.6
231-240	8.10	8.2	8
251-260	7.70	8.4	7
271-280	7.00	7.7	6.3
281-290	8.50	8.5	8.5
291-300	7.50	8.4	6.6
321-330	7.40	8	6.8
Grand Total	**6.46**	**9.3**	**1.6**

Relationship scatter plot between duration and imdb scores:



Statistics	Values
Average duration of the movie	109.9
Maximum duration of the movie	330
Minimum duration of the movie	34
Median of duration	106
Mode of duration	101
Variance of duration	517.6
Standard Deviation of duration	22.75

Let's look at movie runtime, which usually spans around 100-150 minutes. Movies within this time frame often have intricate plots. They tend to get higher ratings on sites like IMDb than shorter or longer films with simpler storylines.

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.

Pivot table:

Impact of language of movie on imdb scores:

language	Count of movies	Average of imdb_score	Sum of imdb_score	Max of imdb_score	Min of imdb_score	StdDevp of imdb_score	Varp of imdb_score	Median of imbd_score
Aboriginal	2	6.95	13.9	7.5	6.4	0.55	0.30	6.95
Arabic	1	7.20	7.2	7.2	7.2	0.00	0.00	7.2
Aramaic	1	7.10	7.1	7.1	7.1	0.00	0.00	7.1
Bosnian	1	4.30	4.3	4.3	4.3	0.00	0.00	4.3
Cantonese	8	7.24	57.9	7.8	6.5	0.41	0.17	7.3
Czech	1	7.40	7.4	7.4	7.4	0.00	0.00	7.4
Danish	3	7.90	23.7	8.3	7.3	0.43	0.19	8.1
Dari	2	7.50	15	7.6	7.4	0.10	0.01	7.5
Dutch	3	7.57	22.7	7.8	7.1	0.33	0.11	7.8
Dzongkha	1	7.50	7.5	7.5	7.5	0.00	0.00	7.5
English	3671	6.42	23585.1	9.3	1.6	1.05	1.10	6.5
Filipino	1	6.70	6.7	6.7	6.7	0.00	0.00	6.7
French	37	7.29	269.6	8.4	5.8	0.55	0.31	7.2
German	13	7.69	100	8.5	6.1	0.62	0.38	7.7
Hebrew	3	7.50	22.5	8	7.2	0.36	0.13	7.3
Hindi	10	6.76	67.6	8	4.8	1.05	1.11	7.05
Hungarian	1	7.10	7.1	7.1	7.1	0.00	0.00	7.1
Icelandic	1	6.90	6.9	6.9	6.9	0.00	0.00	6.9
Indonesian	2	7.90	15.8	8.2	7.6	0.30	0.09	7.9
Italian	7	7.19	50.3	8.9	5.3	1.07	1.14	7
Japanese	12	7.63	91.5	8.7	6	0.86	0.74	7.8
Kazakh	1	6.00	6	6	6	0.00	0.00	6
Korean	5	7.70	38.5	8.4	7	0.51	0.26	7.7
Mandarin	14	7.02	98.3	7.9	5.6	0.74	0.54	7.25
Maya	1	7.80	7.8	7.8	7.8	0.00	0.00	7.8
Mongolian	1	7.30	7.3	7.3	7.3	0.00	0.00	7.3
None	1	8.50	8.5	8.5	8.5	0.00	0.00	8.5
Norwegian	4	7.15	28.6	7.6	6.4	0.50	0.25	7.3
Persian	3	8.13	24.4	8.5	7.5	0.45	0.20	8.4
Portuguese	5	7.76	38.8	8.7	6.1	0.88	0.77	8
Romanian	1	7.90	7.9	7.9	7.9	0.00	0.00	7.9
Russian	1	6.50	6.5	6.5	6.5	0.00	0.00	6.5
Spanish	26	7.05	183.3	8.2	5.2	0.81	0.66	7.15
Swedish	1	7.60	7.6	7.6	7.6	0.00	0.00	7.6
Telugu	1	8.40	8.4	8.4	8.4	0.00	0.00	8.4
Thai	3	6.63	19.9	7.1	6.2	0.37	0.14	6.6
Vietnamese	1	7.40	7.4	7.4	7.4	0.00	0.00	7.4
Zulu	1	7.30	7.3	7.3	7.3	0.00	0.00	7.3
Grand Total	**3851**	**6.46**	**24896.3**	**9.3**	**1.6**	**1.05**	**1.11**	**7.3**

Interpretation:

Languages all over the world were analyzed statistically. Most movies—3671—use English. Movies in English tend to get higher ratings on IMDb compared to flicks in other tongues.

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.

Formula for percentile rank:

=PERCENTRANK($I$5:$I$14,I5)

Pivot table:

Top 10 directors based on average imdb scores:

director_name	Average of imdb_score	Percentile Rank
Christopher Nolan	8.9	44%
David Fincher	8.8	0%
Francis Ford Coppol	9.1	89%
Frank Darabont	9.3	100%
Irvin Kershner	8.8	0%
Peter Jackson	8.85	33%
Quentin Tarantino	8.9	44%
Robert Zemeckis	8.8	0%
Sergio Leone	8.9	44%
Steven Spielberg	8.9	44%
Grand Total	**8.93**	**40%**

Interpretation:

Frank Darabont directed movies that earned high scores. His films got perfect ratings, 100 out of 100. We use that data to understand other directors' success. Percentile ranks let us compare how each one's movies were rated by viewers on IMDb.

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.

Formula used are:

Profit margin: =gross-budget(=B2-C2)

Correlation Coefficient: =CORREL(B:B,C:C)

Highest Profit Margin: =MAX(D:D)

Output:

correlation coefficient	0.10

highest Profit Margin	523505847

Using pivot table:

movie_title	Max of profit margin
AvatarÂ	523505847

Interpretation:

Calculating companies' profitability involves a straightforward formula. Five movies faced massive financial losses. However, Avatar experienced an extraordinary windfall, yielding a colossal profit margin—523505847. My Microsoft Excel Hyperlink: Trainity_5.xlsx

Video link : loom vedio

As, the result, The IMDB Mov ies dataset analysed and visualized each question in Microsoft Excel. Overall, the IMDB Mov ie Analysis assignment provided a thorough understanding of Excel fundamentals. It also gave hands-on experience handling and visualizing raw data. The questions increased in depth from intermediate to advanced concepts. This project deepened my understanding of Excel. Directors and investors may utilize these insights to make informed decisions about future projects.