PROJECT 5

IMDB Movie Analysis

Project description:

This project aims to understand the factors that contribute to a movie's success on IMDB, with success defined by high ratings. This project will dive into a dataset of IMDB movies to explore how different factors like genre, director, budget, release year, and actors impact a movie's rating.

To start, the data will be cleaned and prepped, fixing any missing values, duplicates, or inconsistencies so it is ready for analysis. The main goal is to uncover patterns and relationships between these factors and movie ratings. Using the "Five Whys" method, we will go beyond surface-level insights to dig into the deeper reasons behind the trends we find.

This problem holds great importance for movie producers, directors, and investors who are looking to understand what drives a movie's success. By uncovering the key factors that lead to higher ratings, they can make smarter decisions in their future projects, ensuring better outcomes both creatively and financially.

Approach:

- Import the dataset into Excel and clean it by removing duplicates, handling missing values like adding values like in column duration from searching for it on google and filling these values in place of blanks and deleting rows where value cannot be replaced, dropping columns that are not required, splitting the column and ensuring consistent data types.
- Standardize data using Excel functions.
- Once the dataset is clean, I will summarize the data by calculating relevant statistical measures such as averages and medians and generating visualizations to uncover trends related to rejections, interviews, job types, and vacancies.

Tech-stack used:

Microsoft Excel 2019

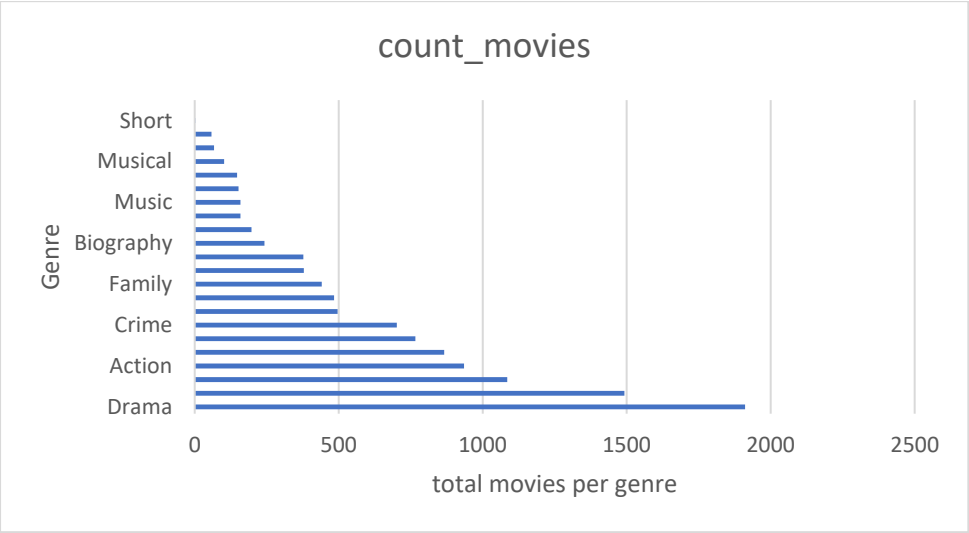
Excel is an excellent tool for handling, analyzing, and visualizing datasets of moderate size. It offers functionalities for data cleaning and statistical analysis. Excel's built-in charts and pivot tables are useful for creating visual representations like bar charts, pie charts, and histograms, which help in understanding trends in the dataset related to rejections, interviews, job types, and vacancies.

Data Analytics Tasks:

A. Movie Genre Analysis: 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.

1. Movie genre analysis

unique_genre	count_movies	mean	median	mode	max	min	Range	variance	st.Dev
Drama	1911	4.137089981	6.9	6.7	9.3	2.1	7.2	0.793996652	0.891064898
Comedy	1492	5.271996951	6.3	6.3	8.8	1.9	6.9	1.081431552	1.039919012
Thriller	1085	2.234142857	6.4	6.5	9	2.7	6.3	0.938931964	0.968985017
Action	935	6.285989305	6.3	6.6	9	2.1	6.9	1.078186788	1.038357736
Romance	866	2.096107383	6.5	6.5	8.5	2.1	6.4	0.940456888	0.969771565
Adventure	766	4.312827225	6.6	6.6	8.9	2.3	6.6	1.247524378	1.116926308
Crime	702	4.211811966	6.6	6.6	9.3	2.4	6.9	0.968463042	0.984105199
Fantasy	496	2.9923125	6.4	6.7	8.9	2.2	6.7	1.30054464	1.140414241
Sci-Fi	484	2.517368421	6.4	7	8.8	1.9	6.9	1.362318841	1.16718415
Family	441	2.138181818	6.3	5.4	8.6	1.9	6.7	1.367909091	1.169576458
Horror	379	4.092354949	5.9	6.2	8.6	2.3	6.3	0.982127152	0.991023285
Mystery	377	2.669591837	6.5	6.6	8.6	3.1	5.5	1.014838309	1.007391835
Biography	242	6.478361345	7.2	7	8.9	4.5	4.4	0.504237338	0.71009671
Animation	197	3.291309524	6.8	7.3	8.6	2.8	5.8	0.987295659	0.993627525
War	159	2.2	7.1	7.1	8.6	4.3	4.3	0.652386753	0.80770462
Music	159	2.03880597	6.5	6.5	8.5	1.6	6.9	1.473940769	1.214059623
History	152	2.147021277	7.2	7.7	8.9	5.5	3.4	0.451578947	0.671996241
Sport	147	2.08025641	6.8	7.2	8.4	2	6.4	1.09876526	1.048220043
Musical	102	2.4945	6.7	7.1	8.5	2.1	6.4	1.307672297	1.143535
Documentary	67	5.682807018	7.2	6.6	8.5	1.6	6.9	1.439855269	1.199939694
Western	58	3.509090909	6.8	6.8	8.9	4.1	4.8	0.997035693	0.998516746



Insights:

1. The above table shows that “Drama” followed by “comedy” and “thriller” are the most common genres in the dataset. By identifying which genres are most common in the dataset, we can understand the genre distribution and popularity within the dataset.

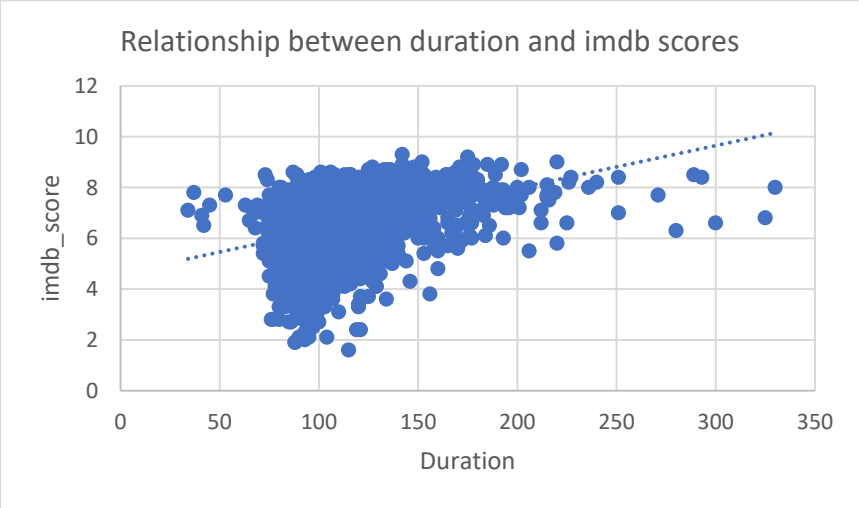
2. The average rating helps identify which genres tend to have higher overall ratings. Here "Action" has a higher average rating compared to "Drama" which suggests that on average, actions are rated more favourably by viewers.
3. The median rating provides a measure of central tendency that is less affected by extreme values. The median rating is significantly different from the average rating, indicating that the ratings for that genre might be skewed by a few very high or very low ratings.
4. Knowing which genres are most common helps understand dataset composition.

B. Movie Duration Analysis: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Distribution of duration

Distribution of movie duration	
Mean	109.8175
Median	105
st.Dev	22.76698

Relationship between movie duration and the imdb scores by scatter plot



Insights:

1. **Distribution:** The descriptive analysis shows the average duration of movies is 109 minutes and the standard deviation of 22.7 reflects the amount of variation or dispersion from that mean.
2. This shows that most values fall within the range of approximately 86.3 to 131.7 as mean ± st.dev.

3. **Relationship:** The above scatter plot shows the relationship between movie duration and imdb is positive correlation as points trend upwards from left to right suggesting that as movie duration increases, imdb scores tend to increase.
4. Also, it shows longer movies may have higher imdb scores but relationship may not be very strong as points are kind of dispersed around the line.

Now if we ask "five whys"

- Why do longer movies receive higher IMDb scores?
  Because they often have more developed plots and characters.

- Why do more developed plots and characters lead to higher scores?
  Because audiences may feel more invested in the story and characters.

- Why do audiences feel more invested in longer films?
  Because they have more time to connect with the characters and plot.

- Why does having more time to connect matter to audiences?
  Because emotional engagement can enhance overall enjoyment of the film.

- Why does emotional engagement enhance enjoyment?
  Because it creates a more immersive experience, making the film memorable.

This helps to dig deeper into the problem.

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

distinct_language	count_movies	mean	median	st.dev
English	3607	6.42107	6.5	1.052583
French	37	7.286486	7.2	0.561329
Spanish	26	7.05	7.15	0.826196
Mandarin	14	7.021429	7.25	0.765786
German	13	7.692308	7.7	0.640913
Japanese	12	7.625	7.8	0.899621
Hindi	10	6.76	7.05	1.111755
Cantonese	8	7.2375	7.3	0.440576
Italian	7	7.185714	7	1.155319
Korean	5	7.7	7.7	0.570088
Portuguese	5	7.76	8	0.978775
Norwegian	4	7.15	7.3	0.574456
Dutch	3	7.566667	7.8	0.404145
Thai	3	6.633333	6.6	0.450925
Danish	3	7.9	8.1	0.52915
Hebrew	3	7.5	7.3	0.43589
Persian	3	8.133333	8.4	0.550757
Aboriginal	2	6.95	6.95	0.777817
Dari	2	7.5	7.5	0.141421
Indonesian	2	7.9	7.9	0.424264

Insights:

The table above shows the distribution of movies based on their language.

The table shows the most common languages for a movie and the average scores for movies in each language. Some languages like Persian, French, German, Dutch have the highest average ratings for score and typical median scores being approximately around 7 for these languages.

Standard deviation shows how spread out scores are for each language and it is low depicting more consistent movie ratings.

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.

Output

unique_directors	avg_imbd	percentile
Charles Chaplin	8.6	1
Majid Majidi	8.5	1
Andrew Haigh	7.7	1
Kevin Jordan	7.6	1
Jafar Panahi	7.5	1
Kiyoshi Kurosawa	7.4	1
Shane Carruth	7	1
Neill Dela Llana	6.3	1
Tony Kaye	8.6	0.998
Christopher Nolan	8.425	0.995
Sergio Leone	8.433333333	0.994
Alfred Hitchcock	8.5	0.994
Richard Marquand	8.4	0.992
Damien Chazelle	8.5	0.992
Lee Unkrich	8.3	0.991
Pete Docter	8.233333333	0.991
S.S. Rajamouli	8.4	0.991
Quentin Tarantino	8.2	0.99
Ron Fricke	8.5	0.99
Hayao Miyazaki	8.225	0.988
Lenny Abrahamson	8.3	0.988
Fritz Lang	8.3	0.987
Milos Forman	8.133333333	0.985
Marius A. Markevicius	8.4	0.985
Billy Wilder	8.3	0.983

Insights:

The above table shows the influence of directors on movie ratings.

The directors in 90th percentile or above are the top 10% of directors based on idmb scores. It can be considered these directors to be consistently delivering high quality movies as per the imdb ratings.

Directors around the 50th percentile produce movies that are typically average in ratings.

Directors in lower percentiles may have few poorly received movies dragging down their average or they might be new.

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

movie_title	budget	gross	Profit
Avatar	237000000	760505847	523505847
Jurassic World	150000000	652177271	502177271
Titanic	200000000	658672302	458672302
Star Wars: Episode IV - A New Hope	11000000	460935665	449935665
E.T. the Extra-Terrestrial	10500000	434949459	424449459
The Avengers	220000000	623279547	403279547
The Lion King	45000000	422783777	377783777
Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677
The Dark Knight	185000000	533316061	348316061
The Hunger Games	78000000	407999255	329999255
Deadpool	58000000	363024263	305024263
The Hunger Games: Catching Fire	130000000	424645577	294645577
Jurassic Park	63000000	356784000	293784000
Despicable Me 2	76000000	368049635	292049635
American Sniper	58800000	350123553	291323553
Finding Nemo	94000000	380838870	286838870
Shrek 2	150000000	436471036	286471036
The Lord of the Rings: The Return of the King	94000000	377019252	283019252
Star Wars: Episode VI - Return of the Jedi	32500000	309125409	276625409
Forrest Gump	55000000	329691196	274691196
Star Wars: Episode V - The Empire Strikes Back	18000000	290158751	272158751
Home Alone	18000000	285761243	267761243
Star Wars: Episode III - Revenge of the Sith	113000000	380262555	267262555
Spider-Man	139000000	403706375	264706375

Output

Correlation coefficient	
0.0965404	
Highest profit margin	
523505847	
Movie with highest profit margin	
1	row number
Avatar	the movie name with highest profit margin

Insights:

A correlation coefficient of 0.09 indicates a very weak positive relationship between movie budgets and gross earnings. The weak correlation shows that spending more on a movie's budget does not relate to significantly higher gross earnings.

The profit margin of some movies is positive showing the movie made more money than it costs to produce meaning the higher the margin, the more profitable the movie and some has negative profits showing they were in loss.

The Avatar movie had the maximum profit margin indication the movie was profitable.

Profit margin showed a clear picture of how well a movie performed financially.

Result:

In this project, using Excel's statistical functions for data analysis provided a powerful tool to extract meaningful insights from the data. Functions such as CORREL enable you to assess the strength and direction of relationships between variables, like movie budgets and gross earnings. Calculating AVERAGE and STDEV helps determine central tendencies and variability, revealing overall performance trends and consistency. The PERCENTILE function allows you to understand data distribution and identify key percentiles, such as the top 10% of movie performances. Overall, these Excel functions facilitate a comprehensive statistical analysis, aiding in decision-making and strategic planning by providing clear, data-driven insights.

Link for the excel workbook:
https://docs.google.com/spreadsheets/d/1lwpwG8zr02ogtAmJoCH-P8pgUhJVJ5zV/edit?usp=sharing&ouid=100957567890552950706&rtpof=true&sd=true