Trainity Project 5

Project Description

The project involves analyzing IMDB movie data to gain insights into various aspects of the movies, such as genre, duration, language, directors, and budgets. The dataset contains information on genres, durations, languages, directors, budgets, gross earnings, and IMDB scores.

Approach

Approach to the project and how each task was executed :

A) Data Cleaning - The first step involved cleaning the IMDB movie dataset to handle missing values, duplicates, and ensure consistency in data formats.

B) Analysis -

1) Movie Genre Analysis :

- TEXT TO COLUMN was used to separate the different genres for a movie.
- UNIQUE function was used to get unique genre values.
- COUNTIF function was used to count the number of movies for each genre.
- AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV functions were used to calculate descriptive statistics for top genres.

2) Movie Duration Analysis:

- AVERAGE, MEDIAN, and STDEV functions were used to calculate descriptive statistics for movie durations.
- Scatter plot was created to visualize the relationship between movie duration and IMDB scores. A trendline was added to assess the direction and strength of this relationship.

3) Language Analysis:

- UNIQUE function was used to get unique language values.
- COUNTIF function was used to count the number of movies for each language.
- AVERAGE, MEDIAN, STDEV functions were used for IMDB scores within each language, allowing for an analysis of language's impact.

4) Director Analysis:

- A pivot table was used to calculate the average IMDB score for each director.
- PERCENTILE function was used to identify the directors with the highest scores and compare.

5) Budget Analysis:

- CORREL function was used to calculate the correlation coefficient between movie budgets and gross earnings.
- The profit margin (gross earnings - budget) was calculated for each movie, and MAX function was used to identify the movie with the highest profit margin.

Tech-Stack Used

Microsoft Excel - Excel was the primary software used for data manipulation, analysis, and visualization due to its familiarity and versatility in handling tabular data. Its functions and features were employed for calculations, data cleaning, and creating visualizations. Excel's built-in functions, such as COUNTIFS, AVERAGE, MEDIAN, MODE, STD, VAR were extensively used for various calculations.

Insights

Insights and knowledge gained during the IMDB Movie Analysis :

1) Genre Analysis - Certain genres consistently receive higher IMDB scores, like Drama, Comedy, Romance indicating a preference for these genres among audiences. Comedy genre has the highest rating of 9.5 followed by Drama with 9.3 rating.

2) Movie Duration Analysis - The average movie duration is around 107 minutes. Movies with duration ranging from 100-200 mins have high IMDB ratings. The trendline shows the increase in ratings when the duration increases.

3) Language Analysis - English is the dominant language in movies. A large number of movies are in French, Spanish, Hindi and Mandarin. French movies have the highest average IMDB ratings.

4) Director Analysis - A few directors consistently achieve high average IMDB scores, indicating their talent in delivering well-received movies. Examples include John Blanchard, Mitchell Altieri and Cary Bell.

5) Budget Analysis - The positive correlation indicates that as movie budgets increase, so do gross earnings. But it is a weak correlation meaning this relationship is not affecting the financial status much. The highest profit made by a movie was 523505847.

Result

The project contributed significantly to my understanding of IMDB Movie Analysis by uncovering patterns and relationships within the dataset. It provided insights into the various factors that influence IMDB movie ratings, such as genre, duration, language, and directing. Overall, this project has provided valuable insights into the dynamics of the movie industry, enabling informed decision-making

Excel sheet link - Excel_P5.xlsx

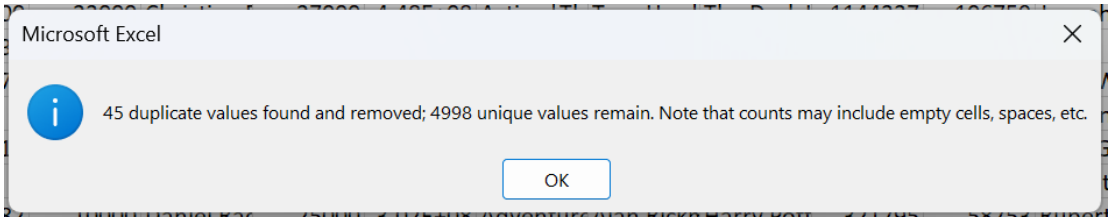
Video link - https://drive.google.com/file/d/153MiB53P9DnwWngwKwDHgeGyNTmgl-pV/view?usp=sharing

Tasks :

DATA CLEANING -

1) Removing duplicates

- Go to the data tab and in the data tools go to remove duplicates.
- The duplicates are removed and 4998 unique rows remain for analysis.



2) Converting data types

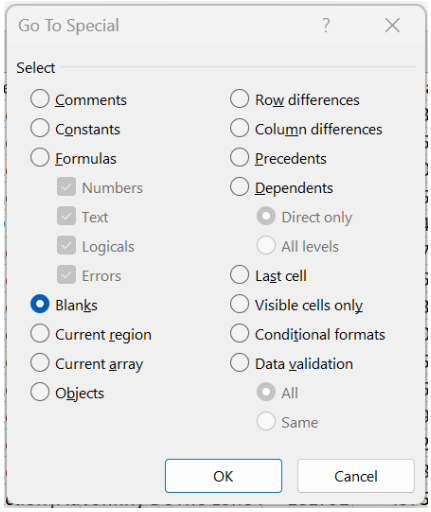
- The data type of gross and budget is changed to number to convert it into a readable format

I	W
gross	budget
7.61E+08	2.37E+08
3.09E+08	3E+08
2E+08	2.45E+08
4.48E+08	2.5E+08

I	W
gross	budget
760505847	237000000
309404152	300000000
200074175	245000000
448130642	250000000

3) Blanks

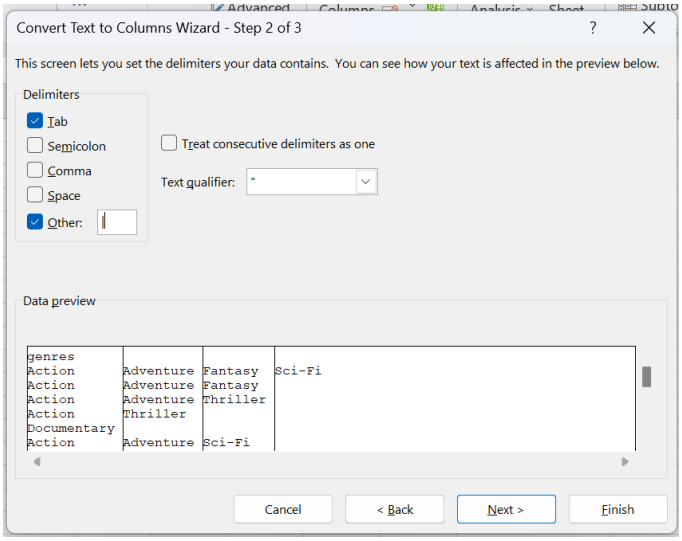
- The blanks can be removed by using Find & Select > Special > Blanks.
- The blanks are removed according to the columns required for analysis at each step.



ANALYSIS -

Q1) 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.

Using Text to columns separate the genres for each movie.



Genre	Count
Drama	2571
Comedy	1862
Thriller	1393
Action	1140
Romance	1097
Adventure	911
Crime	883
Sci-Fi	609
Fantasy	601
Horror	556
Family	542
Mystery	492
Biography	292
Animation	241
Music	212
War	211
History	205
Sport	181
Musical	131
Documentary	120
Western	93
Film-Noir	6
Short	5
News	3
Reality-TV	2
Game-Show	1

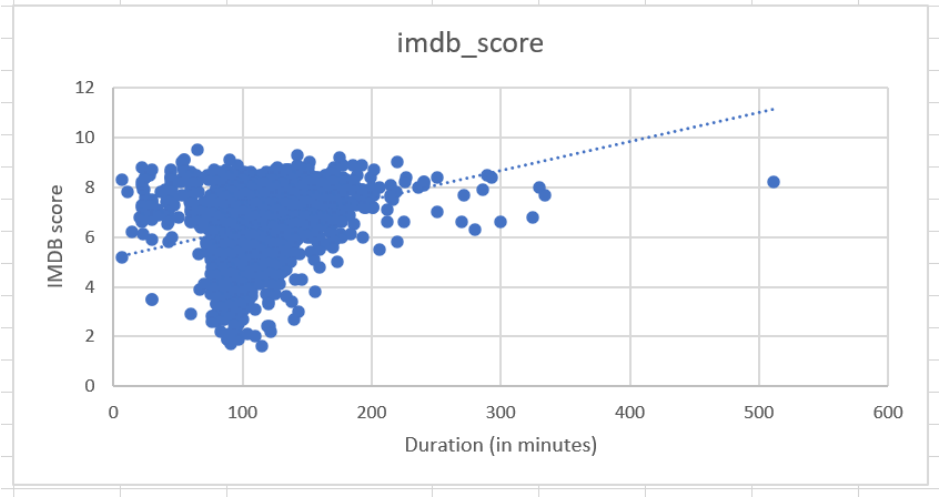
By finding the count we get the most common genres.

Descriptive statistics for the common genres :

IMDB rating statistics for most common genres							
Genres	Average	Median	Mode	Max	Min	Var	Std
Drama	6.76492	6.9	7.2	9.3	2	0.90894	0.95339
Comedy	6.19468	6.3	6.7	9.5	1.7	1.18534	1.08873
Thriller	6.31289	6.4	6.1	9	2.2	1.10598	1.05166
Action	6.23657	6.3	6.1	9.1	1.7	1.23705	1.11223
Romance	6.44791	6.5	6.5	8.6	2.1	0.99468	0.99734

Q2) Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Movie Duration		
Mean	Median	Standard Deviation
107.2133	103	25.24624159



Q3) Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Language	Count		Language	Count
Japanese	17		Hungarian	1
French	73		Portugues	8
Mandarin	24		Danish	5
Aboriginal	2		Arabic	5
Spanish	40		Norwegia	4
Filipino	1		Czech	1
Hindi	28		Kannada	1
Russian	11		Zulu	2
Maya	1		Panjabi	1
Kazakh	1		Tamil	1
Telugu	1		Dzongkha	1
Cantonese	11		Vietnames	1
Icelandic	2		Indonesia	2
German	19		Urdu	1
Aramaic	1		Romanian	2
Italian	11		Persian	4
Dutch	4		Slovenian	1
Dari	2		Greek	1
Hebrew	5		Swahili	1
Chinese	3		Korean	8
Mongolian	1		Thai	3
Swedish	5		Polish	4
English	4662		Bosnian	1

By finding the count we get the most common languages.

Descriptive statistics for the common language:

IMDB rating statistics for most common genres			
Languages	Average	Median	Std
English	6.397404547	6.5	1.120992179
French	7.038356164	7.2	0.721989287
Spanish	6.9375	7.15	0.844300746
Hindi	6.632142857	6.95	1.37374711
Mandarin	6.7875	7.05	1.015017446

Q4) Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Average IMDB score for each director :

Row Labels	Average of imdb_score
A. Raven Cruz	1.9
Ã‰mile Gaudreault	6.7
Ã‰ric Tessier	6.6
Ã‰tienne Faure	4.3
Ãlex de la Iglesia	6.1
Aaron Hann	6
Aaron Schneider	7.1
Aaron Seltzer	2.7
Abel Ferrara	6.6
Adam Brooks	7.2
Adam Carolla	6.1
Adam Goldberg	5.4
Adam Green	5.7
Adam Jay Epstein	3.8
Adam Marcus	4.3
Adam McKay	6.916666667
Adam Rapp	6.4
Adam Rifkin	6.5
Adam Shankman	5.9625
Adrian Lyne	6.4
Adrienne Shelly	7.1
Agnieszka Holland	6.8
Agnieszka Wojtowicz-Vosloo	5.9
AgustÃ-n DÃ-az Yanes	6.1
Aki KaurismÃ¤ki	7.2
Akira Kurosawa	8.1
Akiva Goldsman	6.2
Akiva Schaffer	6.033333333
Al Franklin	4.3
Al Silliman Jr.	4
Alain Resnais	6.3
Alan Alda	7.2

Albert Hughes	7.066666667
Alec Asten	4.3
Alejandro Agresti	6.8
Alejandro AmenÃ¡bar	7.15
Alejandro G. IÃ±Ã¡rritu	7.783333333
Alejandro Monteverde	7.4
Aleksandr Veledinskiy	7.5
Aleksey German	6.7
Alessandro Carloni	7.2
Alex Cox	6.9
Alex Craig Mann	4.6
Alex Garland	7.7
Alex Gibney	7.7
Alex Kendrick	6.775
Alex Proyas	6.82
Alex Ranarivelo	4.8
Alex Rivera	5.9
Alex Smith	6.1
Alex van Warmerdam	7
Alex Zamm	2.3
Alexander Payne	7.42
Alexander Witt	6.2
Alexandre Aja	6.225
Alfonso CuarÃ³n	7.8
Alfred Hitchcock	7.35
Alice Wu	7.6
Alison Maclean	7
Alister Grierson	5.9
Allan Arkush	4.2
Allan Dwan	7.2
Allen Coulter	7.2
Allen Hughes	6.2
Allison Anders	6.4
Allison Burnett	6

Directors with highest scores using percentile calculation :

	Percentile
99%	8.3

Row Labels	Average of imdb_score
John Blanchard	9.5
Sadyk Sher-Niyaz	8.7
Mitchell Altieri	8.7
Cary Bell	8.7
Mike Mayhall	8.6
Charles Chaplin	8.6
Ron Fricke	8.5
Raja Menon	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Sergio Leone	8.475
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Rakeysh Omprakash Mehra	8.4
Robert Mulligan	8.4
Richard Marquand	8.4
Moustapha Akkad	8.4
Marius A. Markevicius	8.4
Jay Oliva	8.4
Catherine Owens	8.4
Bill Melendez	8.4
Asghar Farhadi	8.4
Sut Jhally	8.3
Stanley Donen	8.3
Lee Unkrich	8.3
Justin Paul Miller	8.3
Lenny Abrahamson	8.3
John Sturges	8.3
Fritz Lang	8.3

Q5) Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Correlation coefficient between movie budgets and gross earnings

Correlation coefficient
0.101033478

Profit margin (gross earnings - budget) for each movie :

gross	budget	profit margin
760505847	237000000	523505847
309404152	300000000	9404152
200074175	245000000	-44925825
448130642	250000000	198130642
73058679	263700000	-190641321
336530303	258000000	78530303
200807262	260000000	-59192738
458991599	250000000	208991599
301956980	250000000	51956980
330249062	250000000	80249062
200069408	209000000	-8930592
168368427	200000000	-31631573
423032628	225000000	198032628
89289910	215000000	-125710090
291021565	225000000	66021565
141614023	225000000	-83385977
623279547	220000000	403279547
241063875	250000000	-8936125
179020854	225000000	-45979146
255108370	250000000	5108370
262030663	230000000	32030663
105219735	200000000	-94780265

Max profit :

Max Profit
523505847