IMDB Movie Analysis

Final Project-1

Project Description:

The dataset provided is related to IMDB Movies.
Here, success can be defined by high IMDB ratings.
This impact 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:

We start by downloading the provided dataset and loading it on excel.
During the Data Cleaning process, we first delete the columns we dont require.
Next, we check for any blanks or discrepancies in the data and take care of it.
We delete duplicate records from certain columns.
We insert the data into a table.
Further we create various pivot tables according to our requirements.
We make use of Excel functions to derive insights from the data and show them graphically using charts.

Tech-Stack Used:

Microsoft 360

Version : 18.2411.1163.0
Correlation ID : dac98643-f7f0-449f-aebb-fd9c99011baf
Session ID : 13fff215-657d-4760-ae33-ccab8bee0d16

Excel

Insights:

Loading the dataset in excel file:

We start by opening Microsoft 360, then click on Excel.
When the window opens, we click on Upload → Locate our file → Press Okay.

Deleting Unwanted Columns:

We delete columns like Color, Facebook likes, Movie link, Actor Name, etc.
In the end, we're left with columns: Director Name, Duration, Gross, Genre, Language, Budget, imdb score.
These are the important columns that we will use for analysis.

Handling Missing Data:

Select a column header and apply Sort & Filter function to filter out all blank values.
Apply this technique to each column.
You can either delete the rows with missing values or replace those cells with MEAN or MODE values of the column.

Inserting a Table:

Select the complete data, then go to the insert tab, then click on the table option.
Your table has been created.
You can rename and customize your table according to your desire.

Dataset Table:IMDB_Movies.xlsx

Pivot Table:

A PivotTable is an interactive way to quickly summarize large amounts of data. You can use a PivotTable to analyze numerical data in detail, and answer unanticipated questions about your data. A PivotTable is especially designed for: Querying large amounts of data in many user-friendly ways.
To make a Pivot table, select your table or columns that you need, go to Insert, click on PivotTable.
You get the option to create the Pivot table on a new spreadsheet or existing sheet. In that case you have to mention the location on the sheet.
Once the Pivot table has been created, you can rename and customize it according to requirement.

Charts:

An Excel chart or graph is a visual representation of a Microsoft Excel worksheet's data. These graphs and charts allow you to see trends, make comparisons, pinpoint patterns, and glean insights from within the raw numbers. Excel includes countless options for charts and graphs, including bar, line, and pie charts.

To create a chart

   Select the data for which you want to create a chart.

   Click INSERT > Recommended Charts.

   On the Recommended Charts tab, scroll through the list of charts that Excel recommends for your data, and click any chart to see how your data will look. ...

   When you find the chart you like, click it > OK.

Excel Functions:

Excel has many premade formulas, called functions.

Functions are typed by = and the functions name.

Once you have typed the function name you need to apply it to a range.

The range is always inside of parentheses.

A)Movie Genre Analysis:

Analyze the distribution of movie genres and their impact on the IMDB score.

Your 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.
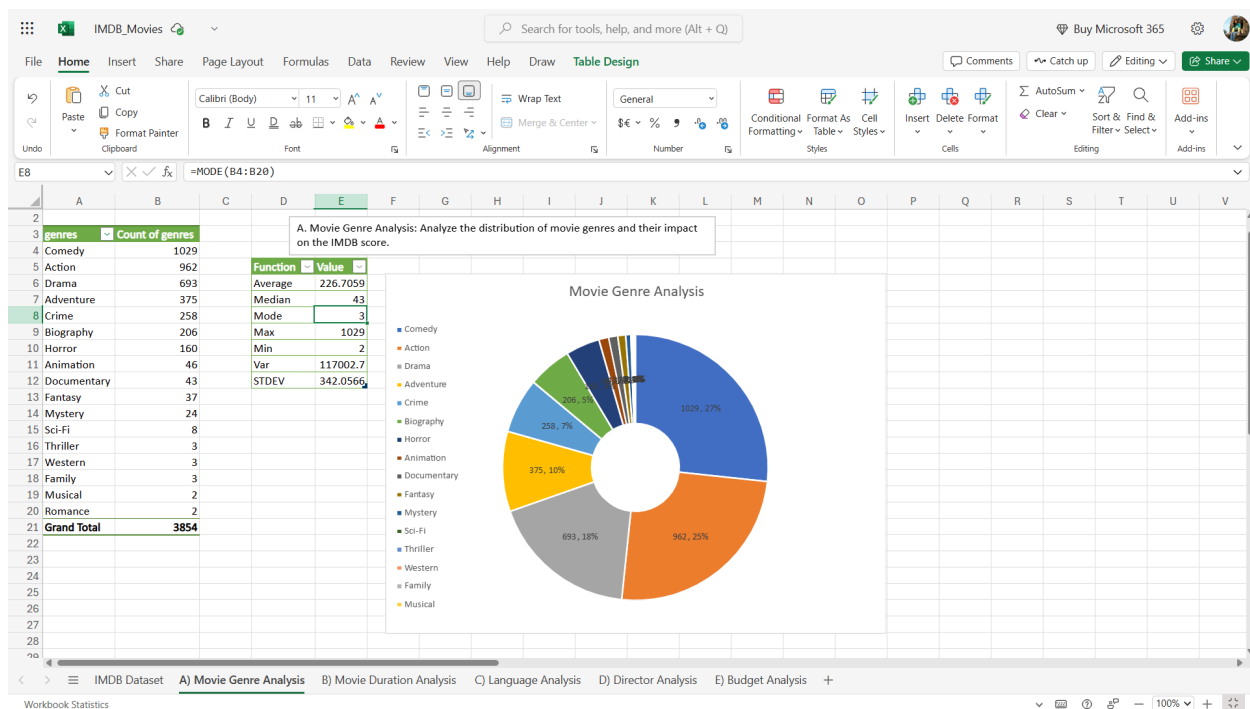
As the dataset contains multiple genres for each movie entry in a single column, we select Genre column then click on Split Text To Column option, in the Data ribbon option.

Next we create a pivot table, and enter Genres in row, and Count of Genres in ∑ Values.

Using this Pivot Table, we insert a Pie Chart to visualize distribution of various movie genres.

Lastly, we use Excel functions =AVERAGE(B4:B20) to find average, =MEDIAN(B4:B20) to find median, =MODE(B4:B20) to find mode value, =MAX(B4:B20) to find maximum value,

=MIN(B4:B20)to find the minimum value, =VAR(B4:B20) for finding variance and =STDEV(B4:B20)to find Standard Deviation.

Result:

Most movies had been in the comedy genre, which is 1029 movies, that's 27% of total distribution and lowest is 2 in Musical and Romance genre.

Average count of genres is 226.7, Median is 43, Mode is 3, Variance is 117002.7 and Standard Deviation is 342.

B) Movie Duration Analysis:

Analyze the distribution of movie durations and its impact on the IMDB score.

Your Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
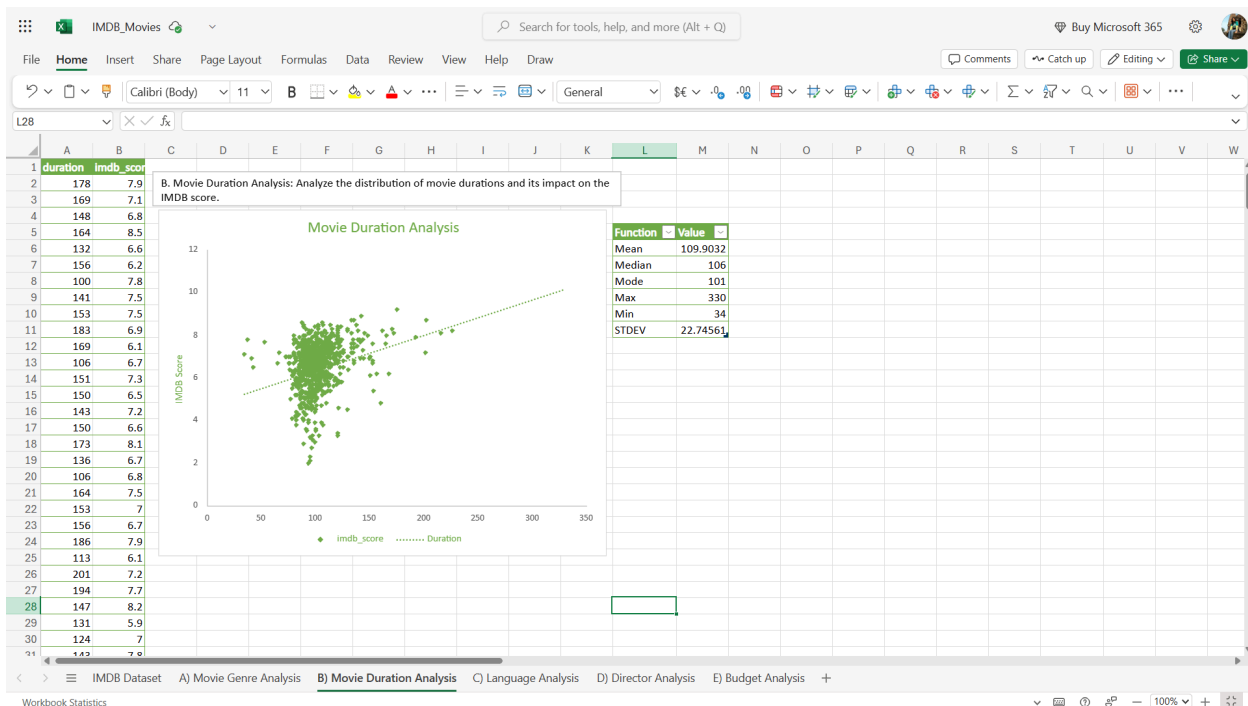
First we copy paste Duration and imdb_score columns in another sheet, then we insert them in a table.
Then we use this table to create a chart. We are using a column chart over here.
We also find values of Mean, Median, Mode, Min, Max and Stdev for movie duration.

Result:

The chart shows that the majority of movies are around 100 minutes and mostly have around 6.
The linear trendline goes from 30 minutes, 5 score to 340 minutes, 8 score.
Values of Mean, Median, Mode, Min, Max and Stdev for movie duration are 109.9, 106, 101, 330, 34 and 22.74 respectively.

C) Language Analysis:

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

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

We create a pivot table and put language and imdb_score to rows. Then we put ∑Values to Count of Language.
We use Excel functions to find the Average, Median and Standard Deviation distribution of imdb_score for each language.
Then we insert it in a table.

Result:

The table mentions Count, Average, Median and STDEV for languages having more than one count.

Language	Aboriginal	Cantonese	Chinese	Dutse	English	French	German	Hindi	Italian	Japanese	Korean	Mandarin	Portuguese	Russian	Spanish	Swedish	Thai
Count	2	9	3	3	3653	53	11	17	5	13	7	22	2	10	20	3	3
Average	1.75	3.15	3.8	4.8	6.906849315	6.703333333	5.628571429	6.083333333	6.25	7.225	6.183333333	6.686666667	7.55	6	6.625	5.95	7.1
Median	1.75	3.15	3.8	4.8	6.9	6.7	6	6.1	6.25	7.35	6.5	6.8	7.55	6.1	6.5	5.95	7.1
STDEV	0.212132034	0.244948974	0.1	0.1	1.308341695	0.665133294	1.446506465	1.590787874	0.129099445	0.645438554	0.847152092	1.030857248	0.353553391	1.244989996	0.536523199	0.494974747	0.707106781

D) Director Analysis:

Influence of directors on movie ratings.

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

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

director_name	Average of imdb_score	Percentile Rank
Charles Chaplin	8.6	0.998
Tony Kaye	8.6	0.998
Alfred Hitchcock	8.5	0.996
Damien Chazelle	8.5	0.996
Ron Fricke	8.5	0.996
Majid Majidi	8.5	0.996
Sergio Leone	8.433333333	0.996
Christopher Nolan	8.425	0.995
Asghar Farhadi	8.4	0.993
Marius A. Markevicius	8.4	0.993
Richard Marquand	8.4	0.993
S.S. Rajamouli	8.4	0.993
Billy Wilder	8.3	0.99
Lee Unkrich	8.3	0.99
Fritz Lang	8.3	0.99
Lenny Abrahamson	8.3	0.99
Pete Docter	8.233333333	0.99
Hayao Miyazaki	8.225	0.989
Quentin Tarantino	8.2	0.989
George Roy Hill	8.2	0.986
Elia Kazan	8.2	0.986
Joshua Oppenheimer	8.2	0.986
Juan JosÃ© Campanella	8.2	0.986
Victor Fleming	8.15	0.986
Milos Forman	8.133333333	0.985
David Sington	8.1	0.981
Terry George	8.1	0.981

Sheet tabs: IMDB Dataset | A) Movie Genre Analysis | B) Movie Duration Analysis | C) Language Analysis | D) Director Analysis | E) Budget Analysis

We insert a pivot table, put director_name in rows and Average of imdb_score in ∑ Values.

We use `=PERCENTRANK.EXC(`$B$4:$B$1755,B4`)` to find percent rank for each row.

E) Budget Analysis:

Explore the relationship between movie budgets and their financial success.

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

We copy-paste the columns gross and budget columns in a new sheet.
Then we add a third column using the mathematical function '-' subtract to find profit by subtracting gross from budget.
After selecting all three columns we put them in a table.
The **CORREL** function returns the **correlation coefficient of** two cell **ranges.** Use the correlation coefficient to determine the relationship between two properties.
We use excel function `=CORREL(`A:A,B:B`)` to find the correlation coefficient of the table.
We use the excel function `=MAX(`C2:C3855`)` to find the amount of highest profit.

Result:

The value of Correlation Coefficient is 0.10.

The value of movie having maximum profit is
$523,505,847.00

Hiring Process Analytics Workbook: IMDB_Movies.xlsx