Trainity Assignment-5

Muthiah Sivavelan
Ph:8525021258

Excel document: 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.

Data Cleaning: This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

Data Analysis: Here, we explore the data to understand the relationships between different variables and look at the correlation between movie ratings and other factors like genre, director, budget, etc.

Report and Data Story: After the analysis, I created a report that tells a story with the data. This includes the initial problem, findings, and the insights I have gained. I used visualizations to help tell the story and make my findings more understandable.

The goal is to provide actionable insights that can help stakeholders make informed decisions.

Approach:

First of all, the duplicates are delted then we have to do data cleaning. The following explains how I handled the missing values.

- Color column had like 46 missing values and I deleted these rows.
- Director_name: Missing cells are imputed with the value Unknown
- Duration: 14 missing cells were there and I deleted those rows.
- Director_facebook_likes: 99 cells had missing values and deleting the rows that had missing values were deleting the missing values in Director_name, therefore I decided to delete the rows that had missing values.

- Actor_3_facebook_likes: Very less number of rows which had missing values.
- Facenumber_in_poster: 12 cells had missing values and the rows are deleted.
- Language: 12 cells had missing values and the rows are deleted.
- Country: 1 row had missing value which I deleted.
- Num_critic_users: 14 rows had missing values which I deleted.
- Gross: Contains 745 missing values (Ignored for now: constitutes around 15% of total number of rows, If the problem requires the use of Gross then the column is copied to a separate sheet and the missing values are deleted.)
- Plot_keywords: Contains 126 missing values and these are ignored as it is not that important.
- Content_rating: Contains 225 missing values and these are ignored as it is not that important.
- Budget: Contains 371 missing values which are ignored.
- Aspect Ratio: Contains 268 missing values and these values are imputed with the median value(2.35).

Based on the above data cleaning process, some questions may yield slightly different output but the approach and the formula used is correct. It would have been better if you have provided which columns are more important and which columns are not, as in real life we can discuss this with the teammates and others but here I decided things myself which could vary with person.

Tasks:

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.

Approach:

I solved this problem in a separate sheet(Sheet1). First, I copied the genre column and pasted in this sheet. Then I removed the duplicates. I created a new table with this column. I added some new columns named: Count, Avg_Imdb, Median, Mode, Range, Var_Imdb, STD_dev_Imdb. To get the count I used the countif function in excel and then gave the criteria

as “*”&genre&”*”. This would check if the given genre is present inside the cell. For the Avg_Imdb column, I used the averageif function with range as the genres column, criteria as “*”&genre&”*” and range as the imdb_score column. For the median and mode, I used the median and mode functions respectively and I used an if condition inside this ensures that it finds the median and mode of the particular genre class only. For the range, I used the min function and the max function and I separated these by a hyphen. Here also, I used a nested if and search to group based on genre given and the minimum and maximum in the given genre is being found. Similarly for variance and standard deviation, I used Var, STDEV functions followed by an If and search functions(nested), which helped me find the variance and standard deviation of the imdb_score based on the given genre. I used the formulas given in query sub-section for the respective column to get the output.

Query:
(These queries are calculated based on genre.)

For count: =COUNTIF(IMDB_Movies[genres],"*"&Sheet1!A3&"*")

For Average of imdb_score:
=AVERAGEIF(IMDB_Movies[genres],"*"&Sheet1A3&"*",IMDB_Movies[imdb_score])

For Median: =MEDIAN(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score]))

For Mode: =IFERROR(MODE(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score])), "None")

For Range:
=TEXT(MIN(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score])),"0.0")
&" - "&
TEXT(MAX(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score])),"0.0")

For Variance: =VAR(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score]))

For standard deviation: =STDEV(IF(--ISNUMBER(SEARCH(Sheet1!A3, IMDB_Movies[genres])), IMDB_Movies[imdb_score]))

Output:

1)

Genre	Count	Avg_imdb	Median	Mode	Range	Var_imdb	STD_dev_imdb
Action	1109	6.217583408	6.3	6.1	1.7 - 9.0	1.19437466	1.092874494
Adventure	894	6.425279642	6.55	6.7	1.9 - 8.9	1.261017559	1.122950381
Drama	2452	6.746492659	6.8	7.2	2.0 - 9.3	0.869530737	0.93248632
Animation	233	6.539914163	6.7	6.7	1.7 - 8.6	1.296287924	1.138546408
Comedy	1805	6.168199446	6.3	6.4	1.7 - 8.8	1.144243153	1.069693018
Mystery	463	6.432181425	6.5	6.4	2.2 - 8.6	1.122923153	1.059680685
Crime	839	6.522884386	6.6	6.6	2.4 - 9.3	0.992148711	0.99606662
Biography	288	7.147569444	7.2	7	4.5 - 8.9	0.516370378	0.718589158
Fantasy	581	6.271772806	6.4	6.7	1.7 - 8.9	1.323994955	1.150649797
Documentary	103	7.169902913	7.4	7.5	1.6 - 8.5	1.168595089	1.081015767
Sci-Fi	592	6.236824324	6.3	6.7	1.9 - 8.8	1.428861641	1.195350008
Horror	539	5.825788497	5.9	6.2	2.2 - 8.6	1.209649702	1.099840762
Romance	1069	6.444434051	6.5	6.5	2.1 - 8.6	0.952134253	0.975773669
Thriller	1346	6.3	6.4	6.1	2.2 - 9.0	1.048475836	1.023951091
Family	524	6.209923664	6.35	6.7	1.7 - 8.6	1.414968254	1.189524382
Music	320	6.4628125	6.7	7.1	1.6 - 8.5	1.408675451	1.186876342
Western	91	6.703296703	6.8	6.5	3.8 - 8.9	1.138100122	1.066817755
Musical	132	6.507575758	6.7	7	2.1 - 8.5	1.502384918	1.225718123
Film-Noir	6	7.633333333	7.65	None	7.1 - 8.2	0.186666667	0.43204938
History	197	7.071573604	7.2	7.5	2.0 - 8.9	0.778983736	0.882600553
War	203	7.054679803	7.1	7.1	2.7 - 8.6	0.764668585	0.874453306
Sport	179	6.587709497	6.8	7.2	2.0 - 8.4	1.205803151	1.098090684
News	3	7.533333333	7.4	None	7.1 - 8.1	0.263333333	0.513160144
Short	2	6.65	6.65	None	6.2 - 7.1	0.405	0.636396103

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.

Approach:

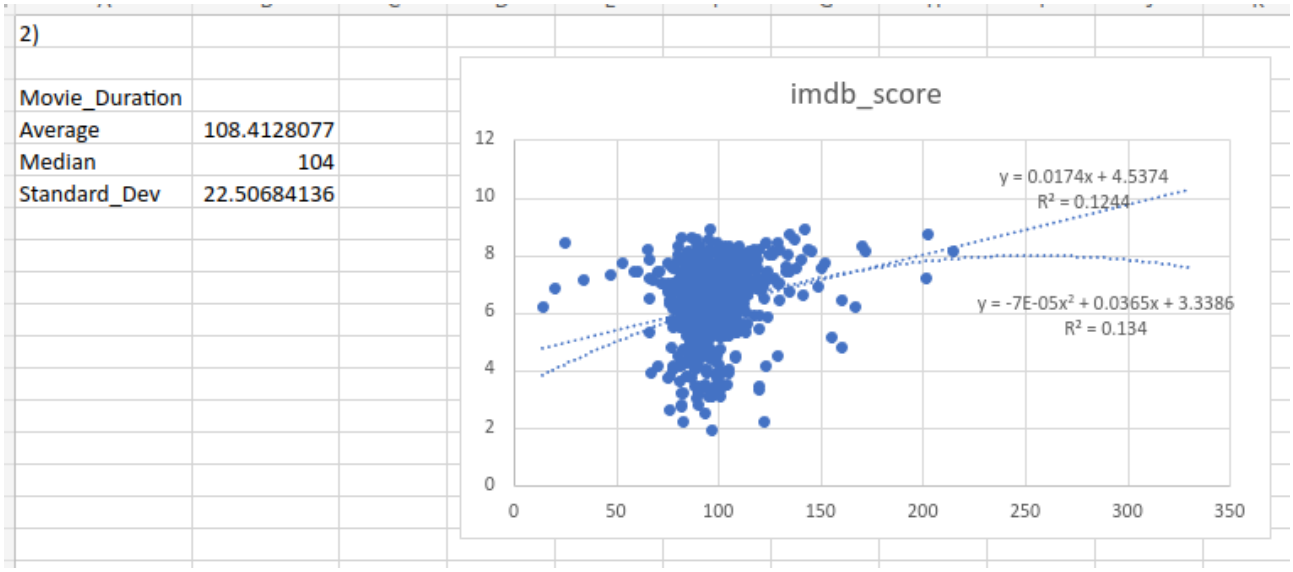
First, I created a new sheet for this problem(Sheet6). To analyze the distribution of movie durations, I first caculated the average, median and standard deviation using the excel functions: average, median and stdev. Then, I selected the imdb_score and the duration column and then clicked on scatter plot for to get the scatter plot chart. I moved the chart to Sheet6. I found the relationship between the movie duration and the imdb_score but fitting a curve in the scatter plot. I right-clicked a data point in the

chart and then in the pop up box, I chose the linear curve which gave the R squared output as 0.1244. I also tried the second degree polynomial curve which gave a better R squared(0.134). Overall, I found that as the movie duration increases the imdb_score increases as well upto an extent(upward trend).

Query:

Average: =AVERAGE(IMDB_Movies[duration])
Median: =MEDIAN(IMDB_Movies[duration])
Standard deviation: =STDEV(IMDB_Movies[duration])

Output:



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.

Approach:

First, I copied the language column to a new spreadsheet(Sheet4). I deleted the duplicate rows. Then, I created a table containing the language column. I added the columns: Count, Mean, Median and Std_dev. I used the count, averageif functions to calculate the count and mean. I used the median and nested if query to find the median for each language. I used

the STDEV and nested if function to calculate the standard deviation for each language.

Query:

For Count: =COUNTIF(IMDB_Movies[language],Sheet4!A3)

For Mean: =AVERAGEIF(IMDB_Movies[language],A3,IMDB_Movies[imdb_score])

For Median: =MEDIAN(IF(IMDB_Movies[language]=A3,IMDB_Movies[imdb_score]))

For Standard deviation: =IF(COUNTIF(IMDB_Movies[language],Sheet4!A3)=1,"Undefined", STDEV(IF(IMDB_Movies[language]=A3,IMDB_Movies[imdb_score])))

Output:

(45 rows were returned)

3)

Language	Count	Mean	Median	Std_dev
English	4484	6.375959	6.5	1.1053465
Japanese	16	7.36875	7.6	1.028733
French	72	7.020833	7.2	0.716395
Mandarin	23	6.773913	7	1.0579643
Aboriginal	2	6.95	6.95	0.7778175
Spanish	40	6.9375	7.15	0.8550566
Filipino	1	6.7	6.7	Undefined
Hindi	27	6.774074	7	1.2027509
Russian	11	6.363636	6.5	1.383671
Maya	1	7.8	7.8	Undefined
Kazakh	1	6	6	Undefined
Telugu	1	8.4	8.4	Undefined
Cantonese	11	6.954545	7.2	0.7047888
German	19	7.342105	7.6	0.9541231
Aramaic	1	7.1	7.1	Undefined
Italian	10	7.08	7.15	1.2062799
Dutch	4	7.425	7.45	0.4349329
Dari	2	7.5	7.5	0.1414214
Hebrew	4	7.675	7.7	0.2986079
Chinese	3	5.666667	5.7	0.5507571
Mongolian	1	7.3	7.3	Undefined
Swedish	4	7.275	7.15	0.7632169
Korean	8	7.3875	7.5	0.8253787
Thai	3	6.633333	6.6	0.450925
Bosnian	1	4.3	4.3	Undefined
None	2	7.95	7.95	0.7778175
Hungarian	1	7.1	7.1	Undefined
Portuguese	8	7.4875	7.7	0.8838835
Icelandic	1	6.9	6.9	Undefined
Danish	5	7.5	8.1	1.077033
Arabic	4	7.175	7.3	0.8732125

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.

Approach:

First, I copied the director_name column to a new sheet(Sheet5) and then I removed the duplicates. I created a new table and then added a new column Avg_IMDB. I used the averageif function to find the average of the imdb_score for each director. Then, I used the percentile function in imdb_score column to find the imdb rating of the top 5% movies and I found out it was greater than 8%. Then I found out the number of directors whose average imdb_score was greater than 8. I found out that 59 directors had their average imdb_score greater than 8. For finding this, I used countif of Avg_IMDB and gave the criteria as greater than equal to cell which contained the value 8.

Query:

For filling the Avg_IMDB column:

```
=AVERAGEIF(IMDB_Movies[director_name],A4,IMDB_Movies[imdb_score])
```

For finding the top 5% in the imdb_score:

```
=PERCENTILE(IMDB_Movies[imdb_score],0.95)
```

For finding the number of directors whose average imdb_score is greater than the top5% in the imdb_score(This gives the top 5% of directors):

```
=COUNTIF(Table4[Avg_IMDB],">="&PERCENTILE(IMDB_Movies[imdb_score],0.95))
```

Output:

(2325 rows with unique directors where present)

4)

Director_name	Avg_IMDB
James Cameron	7.914285714
Gore Verbinski	6.985714286
Sam Mendes	7.5
Christopher Nolan	8.425
Andrew Stanton	7.733333333
Sam Raimi	6.907692308
Nathan Greno	7.8
Joss Whedon	7.866666667
David Yates	7.2
Zack Snyder	7.175
Bryan Singer	7.2875
Marc Forster	7.15
Andrew Adamson	7.08
Rob Marshall	6.6
Barry Sonnenfeld	6.457142857
Peter Jackson	7.654545455
Marc Webb	7.133333333
Ridley Scott	7.070588235
Chris Weitz	6.08
Anthony Russo	7
Peter Berg	6.666666667
Colin Trevorrow	7
Shane Black	7.4
Tim Burton	6.93125
Brett Ratner	6.455555556
Dan Scanlon	7.3
Michael Bay	6.638461538
Joseph Kosinski	6.866666667
John Lasseter	7.38
Martin Campbell	6.711111111

	IMDB_Rating
Top 5% directors	8
Count	59

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:

First I copied the movie_title, gross and budget columns to a new sheet(Sheet7). I created a new table consisting of the above mentioned columns. I added a new column named Profit_Margin. For each row of the Profit_Margin column, I added the formula to subtract the given row's budget from its gross. To find the correlation between gross and budget, I used the correl function in excel. To find the maximum in Profit_Margin column I used the max function in excel. To find the movie with the

maximum Profit_Margin, I used the Index and match function to get the value of the movie_title from the row that has the maximum Profit_Margin.

Query:

For Profit_Margin column: =[@gross]-[@budget]

For correlation: =CORREL(Table5[gross],Table5[budget])

For finding the maximum Profit_Margin: =MAX(Table5[Profit_Margin])

For finding the movie with the maximum Profit_Margin: =INDEX(Table5[movie_title], MATCH(MAX(Table5[Profit_Margin]), Table5[Profit_Margin], 0))

Output:

Qn_5)

movie_title	gross	budget	Profit_Margin			
Avatar	760505847	237000000	523505847		Correlation	0.095971486
Pirates of the Caribbean: At World's End	309404152	300000000	9404152			
Spectre	200074175	245000000	-44925825		Max Profit Margin	523505847
The Dark Knight Rises	448130642	250000000	198130642			
John Carter	73058679	263700000	-190641321		Movie with Max Profit Margin	Avatar
Spider-Man 3	336530303	258000000	78530303			
Tangled	200807262	260000000	-59192738			
Avengers: Age of Ultron	458991599	250000000	208991599			
Harry Potter and the Half-Blood Prince	301956980	250000000	51956980			
Batman v Superman: Dawn of Justice	330249062	250000000	80249062			
Superman Returns	200069408	209000000	-8930592			
Quantum of Solace	168368427	200000000	-31631573			
Pirates of the Caribbean: Dead Man's Chest	423032628	225000000	198032628			
The Lone Ranger	89289910	215000000	-125710090			
Man of Steel	291021565	225000000	66021565			
The Chronicles of Narnia: Prince Caspian	141614023	225000000	-83385977			
The Avengers	623279547	220000000	403279547			
Pirates of the Caribbean: On Stranger Tides	241063875	250000000	-8936125			
Men in Black 3	179020854	225000000	-45979146			
The Hobbit: The Battle of the Five Armies	255108370	250000000	5108370			
The Amazing Spider-Man	262030663	230000000	32030663			
Robin Hood	105219735	200000000	-94780265			
The Hobbit: The Desolation of Smaug	258355354	225000000	33355354			
The Golden Compass	70083519	180000000	-109916481			
King Kong	218051260	207000000	11051260			
Titanic	658672302	200000000	458672302			
Captain America: Civil War	407197282	250000000	157197282			
Battleship	65173160	209000000	-143826840			

Tech Stack Used:
I have used Microsoft Excel 2019, since that is the one that came preinstalled in my laptop. When some functions are missing I would upload the file in OneDrive and open the file using microsoft365.com to execute certain functions.

Insights:

- From the median of imdb_scores based on genre, we can tell that making movies with Film-Noir(also has least standard deviation) can guarantee more towards success(high imdb_score) followed by news and then documentary[also considering variance]. It is also important to note that the movies with the highest imdb rating are in drama and crime and since drama has lesser standard deviation comparitively, drama would be preferred in case, but to be on the safe side, it is better to go with the Film-Noir genre.
- We see an upward trend between movie duration and imdb_score. Hence, it is safer to say that if the duration of movie is longer, the imdb_score would be higher, but the imdb_score doesn't only depend on this factor. We might see a negative trend if the movie is too long so it is best to create a movie with duration between 180 to 210 minutes.
- The director can make the movie in Danish as it has the highest median and also the number of movies in Danish is just 5 so there is more scope in that language as well and since 5 movies have been hit in that language, we can rely that the new movie could also be hit. Cantonese or Portuguese could also be an option considering the standard deviation, count and the median.
- Top 5% directors' movies have the average imdb_score of 8, so to be in the top 5% of the directors, the new director has to make a movie that scores at least 8 in the imdb rating.
- Correlation between gross and the budget is very very slightly positive. This means that higher budget of the movie doesn't necessarily guarantee larger profits[may be very slight chance]. Avatar movie has highest profit margin with a value of 523505847.

Result:

From this project, I have learnt to use my knowledge of excel and statistics for data analysis to provide actionable insights that can help stakeholders(directors) make informed decisions on movie making to maximise their chance of making the movie a success(higher imdb rating).