IMDB MOVIE ANALYSIS
-YAADHAV R

PROJECT DESCRIPTION :

THE OBJECTIVE OF THIS PROJECT IS TO INVESTIGATE THE FACTORS THAT CONTRIBUTE TO THE SUCCESS OF MOVIES ON IMDB, WITH SUCCESS DEFINED BY HIGH IMDB RATINGS. THIS ANALYSIS AIMS TO PROVIDE VALUABLE INSIGHTS FOR MOVIE PRODUCERS, DIRECTORS, AND INVESTORS TO MAKE INFORMED DECISIONS IN THEIR FUTURE PROJECTS. BY UNDERSTANDING THE RELATIONSHIPS BETWEEN VARIOUS VARIABLES AND EMPLOYING A 'FIVE WHYS' APPROACH, I SEEK TO UNCOVER THE UNDERLYING FACTORS DRIVING HIGH IMDB RATINGS AND, CONSEQUENTLY, A MOVIE'S SUCCESS.

TECH-STACK USED :

MICROSOFT EXCEL 2023 - VERSION 16.80 IS USED IN THIS PROJECT AS IT IS :

- IT A SIMPLE AND EASY TO USE SOFTWARE.
- ALL THE TOOLS FOR DATA ANALYSIS IS AVAILABLE.

DATA PREPROCESSING

CLEANING & ANALYSIS :

FIRTS THE FOLLOWING ARE DELETED:

- IRRELEVANT COLUMNS.
- ROWS WITH DUPLICATE MOVIE_TITLE.
- ROWS WITH MISSING DIRECTOR_NAME OR DURATION CELL.
- ROWS WITH MISSING BOTH GROSS AND BUDGET CELLS.

THE REMAINING MISSING CELLS ARE HANDLED AS FOLLOWS:

- A ‘EST_GROSS’ COLUMN IS CREATED TO FILL THE MISSING CELLS OF GROSS WITH THE AVERAGE GROSS OF ITS RESPECTIVE DIRECTOR.
- SIMILARLY A ‘EST_BUDGET IS CREATED FOR BUDGET.
- FINALLY THE MISSING CELLS OF LANGUAGE IS FILLED WITH ENGLISH AS IT WAS THE MOST POPULAR LANGUAGE IN USA.

DRIVE LINK OF PREPROCESSED EXCEL SHEET : 🔗

TASKS

A)MOVIE GENRE ANALYSIS :

dist_genres	frequency	mean_score	medain_score	stdev_score	var_score	max_score	min_score
Drama	1799	6.8	6.9	0.9	0.81	9.3	2.1
Comedy	1443	6.2	6.3	1	1	8.8	1.9
Action	975	6.3	6.3	1	1	9.1	2.1
Adventure	801	6.5	6.6	1.1	1.21	8.9	2.3
Thriller	707	6.4	6.5	1	1	9	2
Crime	645	6.6	6.6	1	1	9.3	2.4
Romance	532	6.5	6.5	1	1	8.8	2.7
Horror	339	5.9	6	1	1	8.5	2.3
Fantasy	335	6.3	6.4	1	1	8.3	2.1
Family	302	6.2	6.3	1.3	1.69	8.7	1.9
Mystery	279	6.4	6.5	1.1	1.21	9.3	2.8
Sci-Fi	263	6.5	6.5	1.1	1.21	8.7	2.3
Biography	261	7.2	7.2	0.7	0.49	8.9	4.5
Animation	173	6.7	6.8	1	1	8.6	2.8
War	119	6.7	6.7	1.1	1.21	8.6	3.2
Music	115	6.4	6.5	1.1	1.21	8.5	1.6
History	98	6.8	6.9	1	1	8.2	2.7
Sport	86	6.5	6.5	1.2	1.44	9.2	2
Musical	80	6.5	6.7	1.2	1.44	8.5	2.1
Documentary	70	7.1	7.3	1.2	1.44	8.5	1.6
Western	48	6.6	6.8	1.2	1.44	8.9	4
Film-Noir	3	7.9	8	0.3	0.09	8.2	7.6
Short	1	6.3	6.3	0	0	6.3	6.3

APPROACH:

A NEW WORKSHEET ‘GENRE ANALYSIS’ IS CREATED AND THE FOLLOWING STEPS ARE CARRIED OUT :

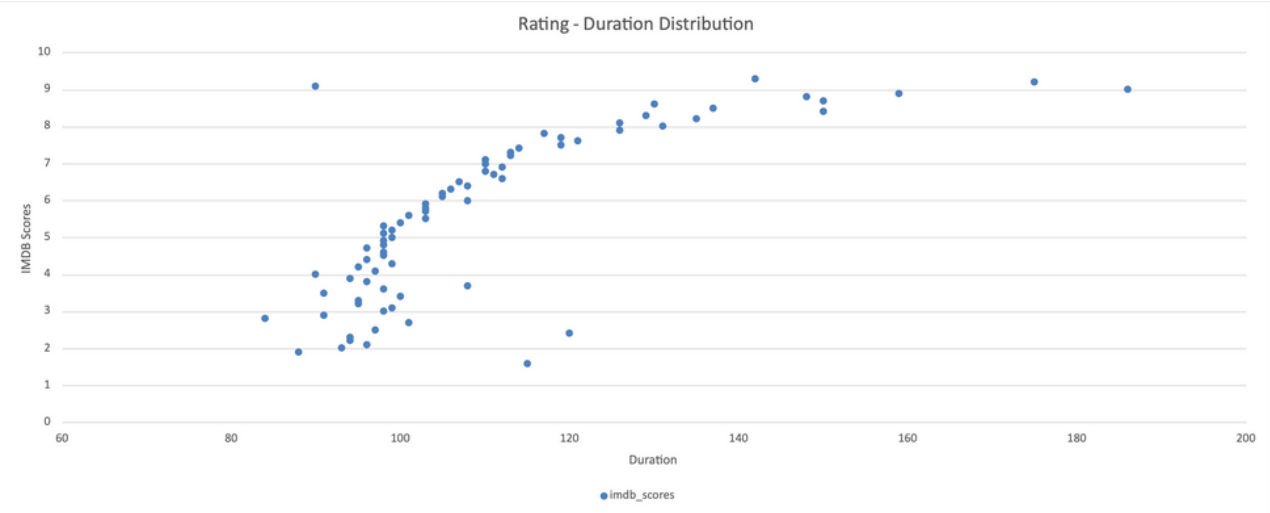
- COLUMNS ‘GENRE’ & ‘IMDB_SCORE’ IS COPIED FROM ‘DATA’.
- THE GENRE COLUMN IS SPLIT AND THESE SPLIT COLUMNS ARE ADDED TO BOTTOM OF ‘GENRE’.
- THEN A TABLE IS CREATED WITH COLUMNS AS SHOWN ABOVE.
- ‘DIST_GENRE’ IS OBTAINED BY REMOVING DUPLICATES FROM ‘GENRE’.
- THE FOLLOWING FORMULAE COUNTIF(), AVERAGEIF(), MEDIAN(FILTER()), STDEV(FILTER()), VAR(FILTER()), MAX(FILTER(), MIN(FILTER()) ARE USED TO FORM RESPECTIVE COLUMN DATAS.

INSIGHTS:

- THE RANGE OF MEN_SCORE IS 2 BUT MOST OF THE MEAN_SCORES ARE AROUND 6.8.
- THIS SUGGESTS THAT AUDIENCE ARE NOT COMMITED ONLY GENRES AND THEY ARE DIVERSE IN CHOOSING GENRES.
- HENCE, GENRE HAS A LESSER IMPACT ON THE IMDB SCORE .

B)MOVIE DURATION ANALYSIS :

imdb_scores	avg_duration	medain_duration	stdev_duration
9.3	142	142	0
9.2	175	175	0
9.1	90	90	0
9	186	186	48.1
8.9	159	178	0
8.8	148	148	15.9
8.7	150	136	27.5
8.6	130	127	0
8.5	137	123	47.1
8.4	150	134	57.3



APPROACH:

A NEW WORKSHEET 'DURATION ANALYSIS' IS CREATED AND THE FOLLOWING STEPS ARE CARRIED OUT :

- COLUMN 'IMDB_SCORE' IS COPIED FROM 'DATA'.
- THE FOLLOWING FORMULAE AVERAGEIF(), MEDIAN(FILTER()), STDEV(FILTER()) ARE USED TO FORM RESPECTIVE COLUMN DATA.
- AFTER SELECTING 'IMDB_SCORES' & 'AVG_DURATION' A SCATTER PLOT IS INSERTED AND FORMATED AS SHOWN.

INSIGHTS:

- AS THE DURATION INCREASES, IT IS MOST LIKELY THAT THE SCORE MIGHT ALSO INCREASE.
- THIS SHOWS THAT AUDIENCE PREFER MOVIES WITH LONGER DURATION.

C)LANGUAGE ANALYSIS :

language	count	mean_score	median_score	stdev_score
English	3906	6.4	6.5	1.1
French	43	7.3	7.2	0.6
Spanish	31	7.1	7.2	0.8
Mandarin	18	6.9	7.1	0.8
German	14	7.6	7.7	0.7
Japanese	12	7.6	7.8	0.9
Hindi	11	6.8	7	1.1
Cantonese	9	7.2	7.3	0.5
Italian	8	7.3	7.4	1.1
Portuguese	6	7.7	8	0.9
Norwegian	4	7.2	7.3	0.6
Korean	4	7.9	7.9	0.5
Hebrew	3	7.5	7.3	0.4
Persian	3	8.1	8.4	0.6
Thai	3	6.6	6.6	0.5
Dutch	3	7.6	7.8	0.4
Danish	3	7.9	8.1	0.5
Swedish	2	7.2	7.2	0.6
Indonesian	2	7.9	7.9	0.4
None	2	8	8	0.8

APPROACH:

A NEW WORKSHEET ‘LANGUAGE ANALYSIS’ IS CREATED AND THE FOLLOWING STEPS ARE CARRIED OUT :

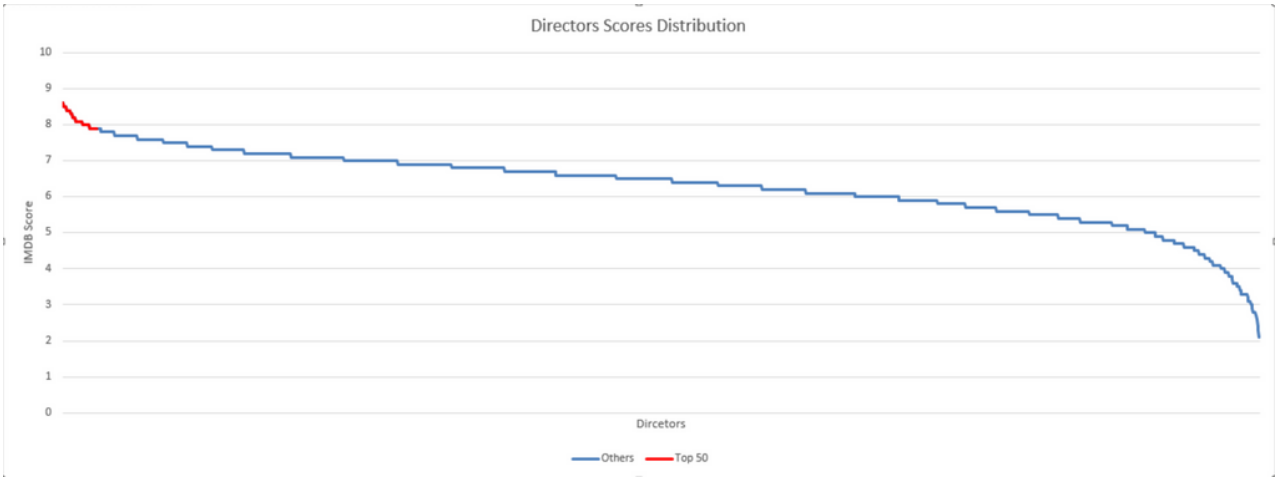
- COLUMN ‘LANGUAGE’ IS COPIED FROM ‘DATA’ AND THE DUPLICATES ARE REMOVED.
- THE FOLLOWING FORMULAE COUNTIF(), AVERAGEIF(), MEDIAN(FILTER()), STDEV(FILTER()) ARE USED TO FORM RESPECTIVE COLUMN DATA.

INSIGHTS:

- ENGLISH IS THE MOST COMMONLY USED LANGUAGE.
- IT CONSTITUTES OVER 95% OF TOTAL MOVIES, WHICH SHOWS THE DIRECTORS ARE PRIORITIZING ENGLISH OVER THE NATIVE LANGUAGES.
- BUT ENGLISH HAS THE LOWEST MEAN_SCORE AS IT IS USED BY MOST OF THE DIRECTORS.

D)DIRECTOR ANALYSIS :

rank	imbd_score	director
1	8.6	Charles Chaplin
2	8.5	Damien Chazelle
3	8.5	Majid Majidi
4	8.5	Ron Fricke
5	8.5	Sergio Leone
6	8.5	Tony Kaye
7	8.4	Asghar Farhadi
8	8.4	Christopher Nolan
9	8.4	Marius A. Markevicius
10	8.4	Richard Marquand



APPROACH:

A NEW WORKSHEET 'DIRECTIOR ANALYSIS' IS CREATED AND THE FOLLOWING STEPS ARE CARRIED OUT :

- COLUMN 'DIRECTOR' IS COPIED FROM 'DATA' AND THE DUPLICATES ARE REMOVED.
- AVG_SCORE OF EACH DIRECTOR IS CALCULATED BY AVERAGEIF().
- THEN A TABLE IS CREATED WITH COLUMN 'RANK' WHICH CONTAINS NUMBERS 1 TO 50.
- THEN PERCENTILE() IS USED TO CALCULATE THE TOP50 SCORES AND ITS RESPECTIVE DIRECTORS ARE MATCHED.
- AFTER SELECTING 'IMDB_SCORE' & 'DIRECTOR' A LINE GRAPH IS INSERTED AND FORMATED AS SHOWN.

INSIGHTS:

- DIRECTORS HAVE THE HIGHEST IMPACT ON THE IMDB SCORE.
- THIS IS BECAUSE, THE VARIATION OF SCORE FOR EVERY DIRECTOR IS LESS.

E)BUDGET ANALYSIS :

movie_title	est_gross	est_budget	profit margin
Non-Stop	760505847	237000000	523505847
Run All Night	658672302	200000000	458672302
The Grace Card	652177271	150000000	502177271
Hocus Pocus	623279547	220000000	403279547
Jawbreaker	533316061	185000000	348316061
Analyze That	474544677	115000000	359544677
Bedazzled	460935665	11000000	449935665
This Is It	458991599	250000000	208991599
Alleluia! The Devil's Carnival	448130642	250000000	198130642
Zoolander	436471036	150000000	286471036
The Ice Pirates	434949459	10500000	424449459
Star Wars: Episode III - Revenge of the Sith	424645577	130000000	294645577
Police Academy	423032628	225000000	198032628
The Pursuit of D.B. Cooper	422783777	45000000	377783777
Signs	414984497	200000000	214984497
The Pink Panther	408992272	200000000	208992272
Practical Magic	407999255	78000000	329999255
Tomorrowland	407197282	250000000	157197282
Screwed	403706375	139000000	264706375
Collateral	402076689	200000000	202076689

correl_coeff	0.2266
max_profit	523505847
movie	Non-Stop

APPROACH:

A NEW WORKSHEET ‘BUDGET ANALYSIS’ IS CREATED AND THE FOLLOWING STEPS ARE CARRIED OUT :

- COLUMNS ‘MOVIE_TITLE’, ‘EST_GROSS’, ’EST_BUDGET’ IS COPIED FROM ‘DATA’.
- ‘PROFIT_MARGIN’ IS CALCULATED BY ‘EST_GROSS’-’EST_BUDGET’.
- CORREL(), MAX() IS USED TO CALCULATE ‘CORREL_COEFF’ AND ‘MAX_PROFIT’ RESPECTIVELY AND THE CORRESPONDING MOVIE IS MATCHED.

INSIGHTS:

- SINCE THE CORREL_COEFF IS POSITIVE, IT SUGGEST THAT BOTH GROSS AND BUDGET GO HAND IN HAND AS EXPECTED.
- BUT THE MAGNITUDE IMPLIES THE RELATIONSHIP IS NOT STRONG AND HAS LESSER IMPACT ON EACH OTHER.
- THAT IS, FOR A MOVIE TO GET A VERY HIGH GROSS, THE BUDGET NOT TO BE VERY HIGH AND VICE-VERSA.
- NON-STOP MOVIE SECURED THE HIGHEST PROFIT MARGIN.

RESULT:

- FIRST OF ALL THIS PROJECT HELPED TO UNDERSTAND THE CONCEPTS I LEARNED IN BETTER AND INTRESTING WAY.
- SO NOW FEEL CONFIDENT IN APPLYING EXCEL AND STATISTICS SKILLS AS I WAS ABLE TO COMPLETE ALL THE GIVEN TASKS.
- IT WAS SO EXICTING TO ANALYZE THE OUTPUTS AND DERIVE INSIGHTS FROM IT.
- OVERALL IT WAS GREAT TO EXPERIENCE TO APPLY THE SKILLS AND LEARN ALONG THE WAY.
- LOOKING FORWARD TO FACE THE UPCOMING CHALLENGES WITH CONFIDENCE AND EXICTMENT.

EXCEL FILE DRIVE LINK: 🔗