IMDB Movie Analysis

Trainity Project No. 5

Prepared By –
Nilesh Kulkarni

AGENDA

- Project Description
- Project Approach
- Tech Stack Used
- Data Cleaning/Preparation
- Insights
- Results

-Nilesh Kulkarni

Project Description

We have provided data on movies and their details like title, language, other metadata including IMDB rating. The project involves analysing the data and extract important insights like impact of Language, Director, Movie Duration on IMDB rating and on overall success of the movies like profitability. These insights would help Producers and other stakeholders to make future movies more successful.

The project first need data preparation/cleaning to make data usable for analysis and post that there are tasks involved to extract insights.

The tasks in this project are as outlined below:

A. Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.
B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.
C. Language Analysis: Examine the distribution of movies based on their language
D. Director Analysis: Influence of directors on movie ratings
E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

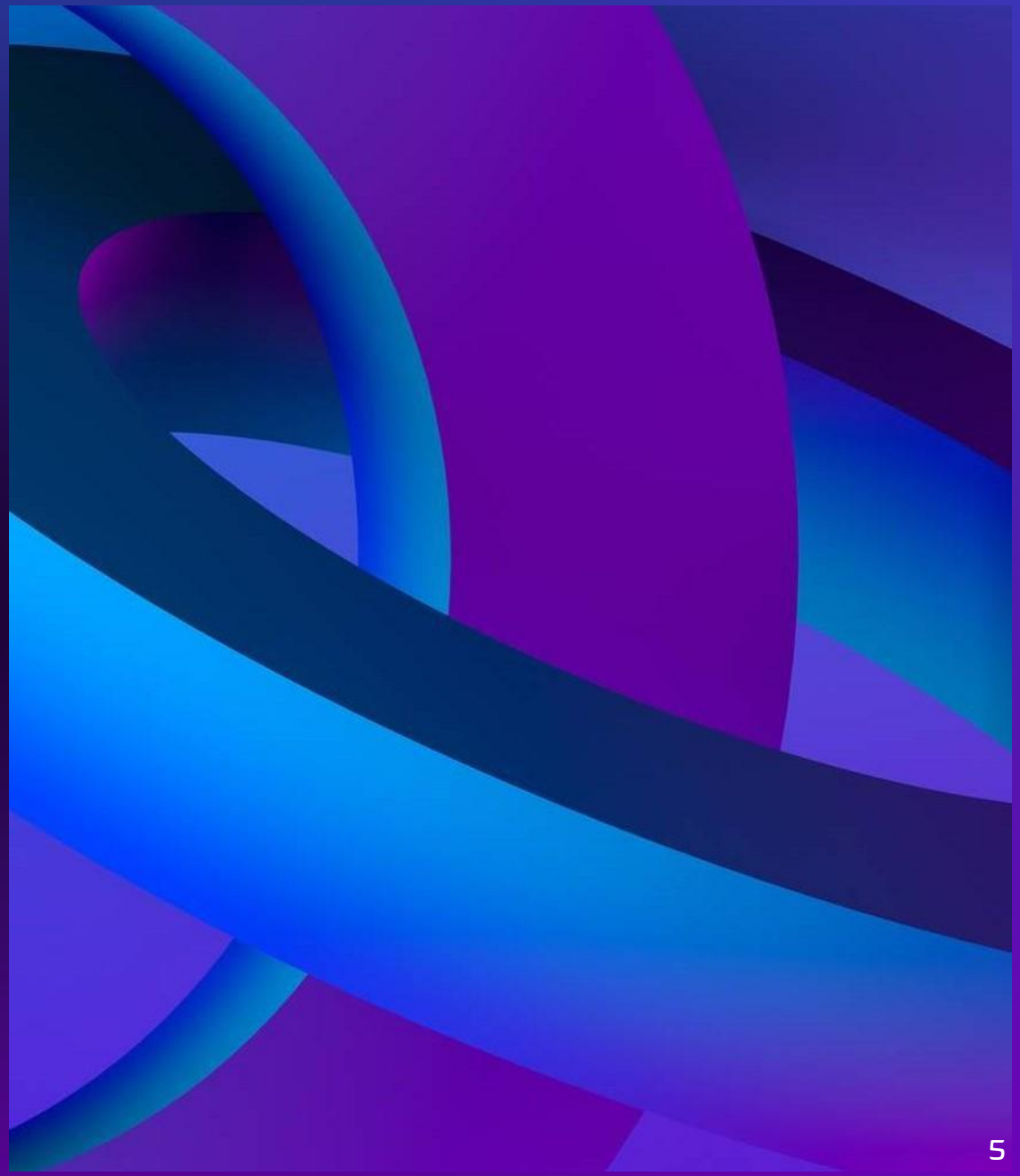
Project Approach

High level steps for the Project approach are as outlined below:

- Data cleaning/preparation – Work on identifying missing data, duplicate data, data errors etc. Additionally work on removing irrelevant columns and adding extra calculated columns.
- Data Analysis: Understand the given data set and it's attributes
- Insights Analysis: Analyse each insights requirement in detail and prepare MS Excel formula or functions to extract insights. Select optimal and efficient approach.
- Extract insights: Use MS Excel as a tool to extract new insights as required including visuals/charts.
- Review: Review and cross check output to verify it matches with the requirements/insights required
- Document: Document the insights and results to be shared across business teams

Tech Stack Used

- **Data Analytics tool:** Microsoft Excel (Office 365) IMDB movies data has been provided in Excel format and excel is further used for data cleaning, analysis and creating visuals/charts to demonstrate insights. Excel is user friendly and functionally rich tool to analyze, visualize and report the data insights.
- **Operating System:** Microsoft Windows 11 Version 22H2
- **Documentation**: Microsoft office 365 (Power Point) & Acrobat PDF



Data Cleaning/Preparation

- Duplicate Data:
  - 47 rows were deleted which were duplicate based on all columns
  - 79 movie titles were found as duplicates and has been removed
- Missing data:
  - 102 rows had director name as blank and deleted
  - 13 rows with duration as blank has been deleted
  - 752 rows with gross as blank has been deleted
  - 262 rows with budget as blank has been deleted
  - Languages for 3 movies was blank and has been set to English based on internet search/data
  - After cleaning data, there are 3788 rows in the datasheet – Dataset-LINK-Click-Here
- Features/Data enhancement
  - Genre details extracted from combination to separate columns using text-to-column and pipe as delimiter-new columns (Genre_1 to Genre_8)
  - Changed data type of budget, gross and duration to numeric
  - Added new column as Profit (difference between gross & budget)
  - Deleted few columns not relevant to the analysis so only remaining columns are as listed – director_name, duration, gross, genres, actor1_name, movie_title, movie_imdb_link, language, country, budget, title_year & imdb_score, genre separated columns & profit

Insights

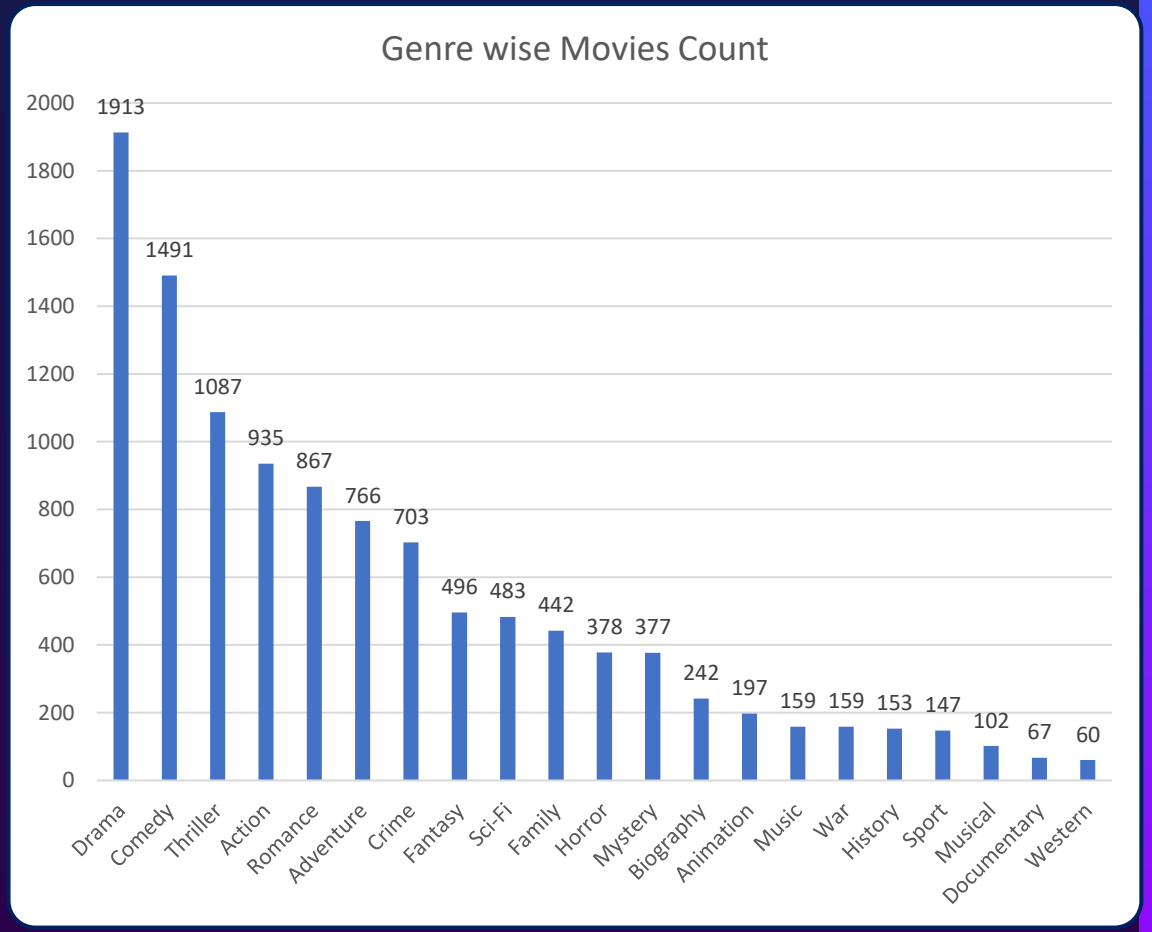
-Nilesh Kulkarni

Task 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.

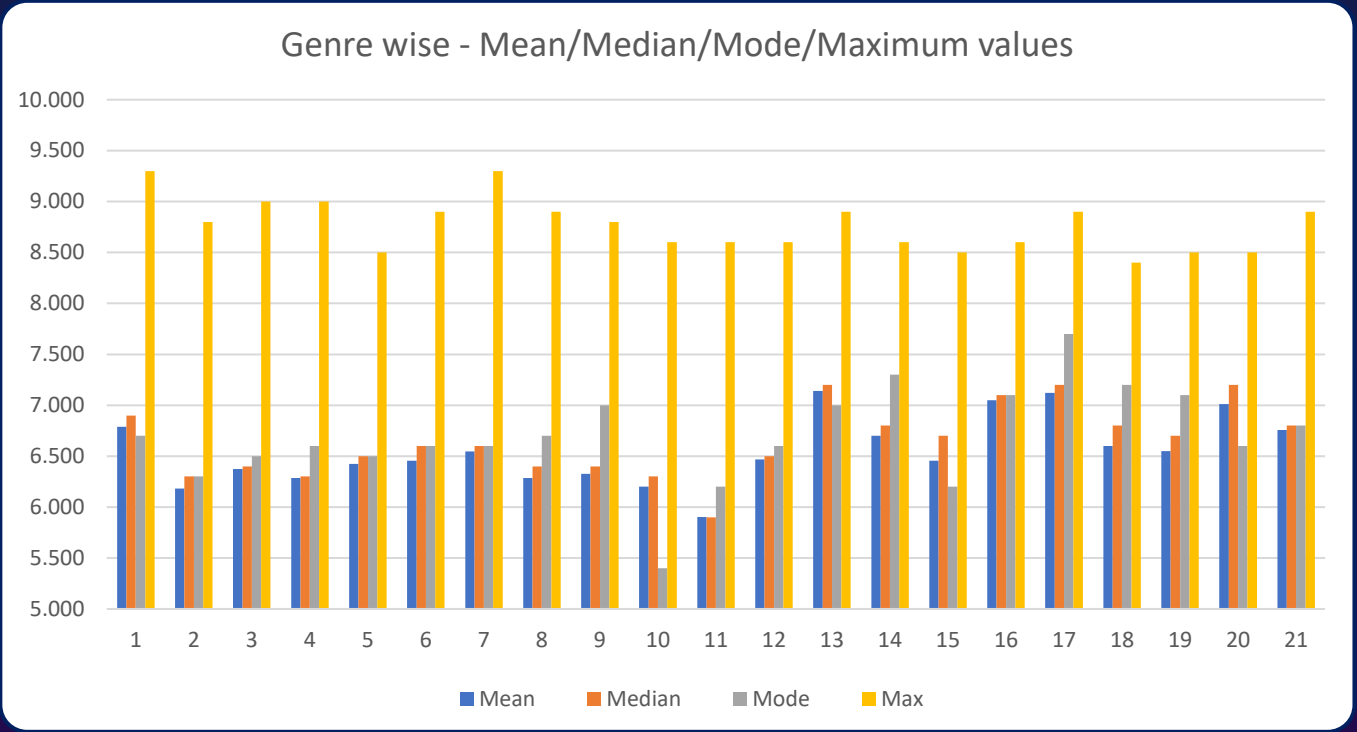
Genre wise descriptive statistics are as shown:

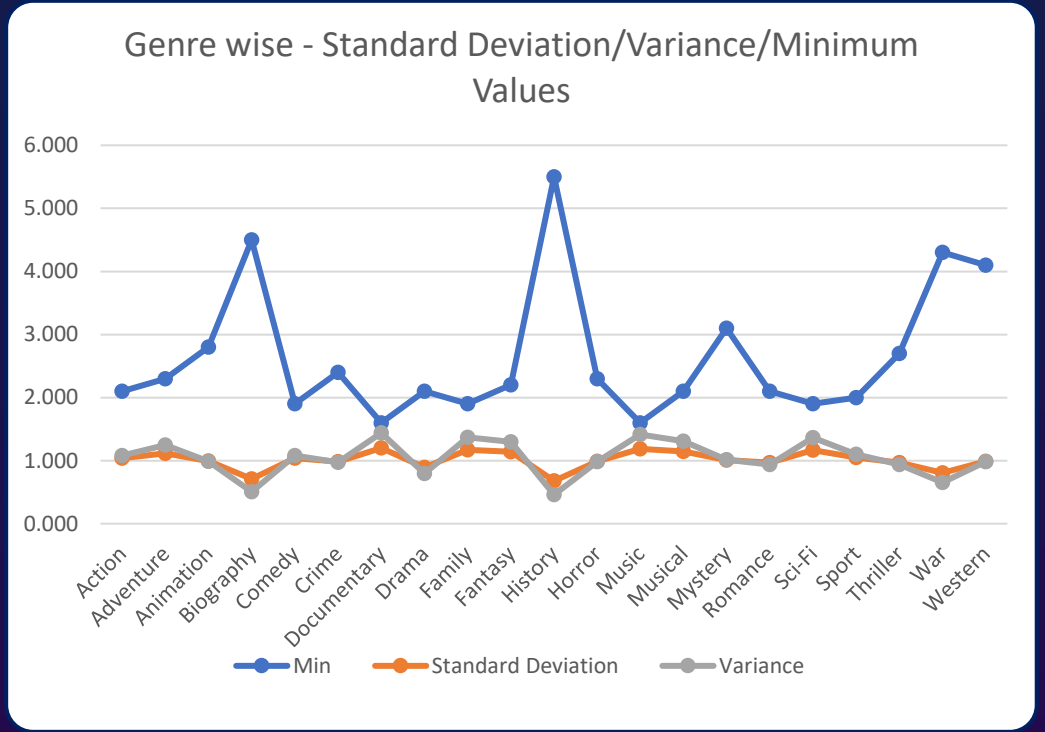
Genre	Count	Mean	Median	Mode	Max	Min	Standard Deviation	Variance
Drama	1913	6.788	6.900	6.700	9.300	2.100	0.892	0.796
Comedy	1491	6.183	6.300	6.300	8.800	1.900	1.040	1.082
Thriller	1087	6.372	6.400	6.500	9.000	2.700	0.969	0.939
Action	935	6.286	6.300	6.600	9.000	2.100	1.038	1.078
Romance	867	6.425	6.500	6.500	8.500	2.100	0.969	0.938
Adventure	766	6.455	6.600	6.600	8.900	2.300	1.117	1.248
Crime	703	6.546	6.600	6.600	9.300	2.400	0.986	0.971
Fantasy	496	6.285	6.400	6.700	8.900	2.200	1.140	1.301
Sci-Fi	483	6.326	6.400	7.000	8.800	1.900	1.168	1.364
Family	442	6.202	6.300	5.400	8.600	1.900	1.169	1.367
Horror	378	5.901	5.900	6.200	8.600	2.300	0.991	0.982
Mystery	377	6.469	6.500	6.600	8.600	3.100	1.007	1.015
Biography	242	7.140	7.200	7.000	8.900	4.500	0.710	0.504
Animation	197	6.701	6.800	7.300	8.600	2.800	0.994	0.987
Music	159	6.457	6.700	6.200	8.500	1.600	1.189	1.413
War	159	7.048	7.100	7.100	8.600	4.300	0.808	0.652
History	153	7.123	7.200	7.700	8.900	5.500	0.678	0.460
Sport	147	6.601	6.800	7.200	8.400	2.000	1.048	1.099
Musical	102	6.551	6.700	7.100	8.500	2.100	1.144	1.308
Documentary	67	7.012	7.200	6.600	8.500	1.600	1.200	1.440
Western	60	6.757	6.800	6.800	8.900	4.100	0.991	0.982

Genre wise Movies count is as shown:



Task 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.





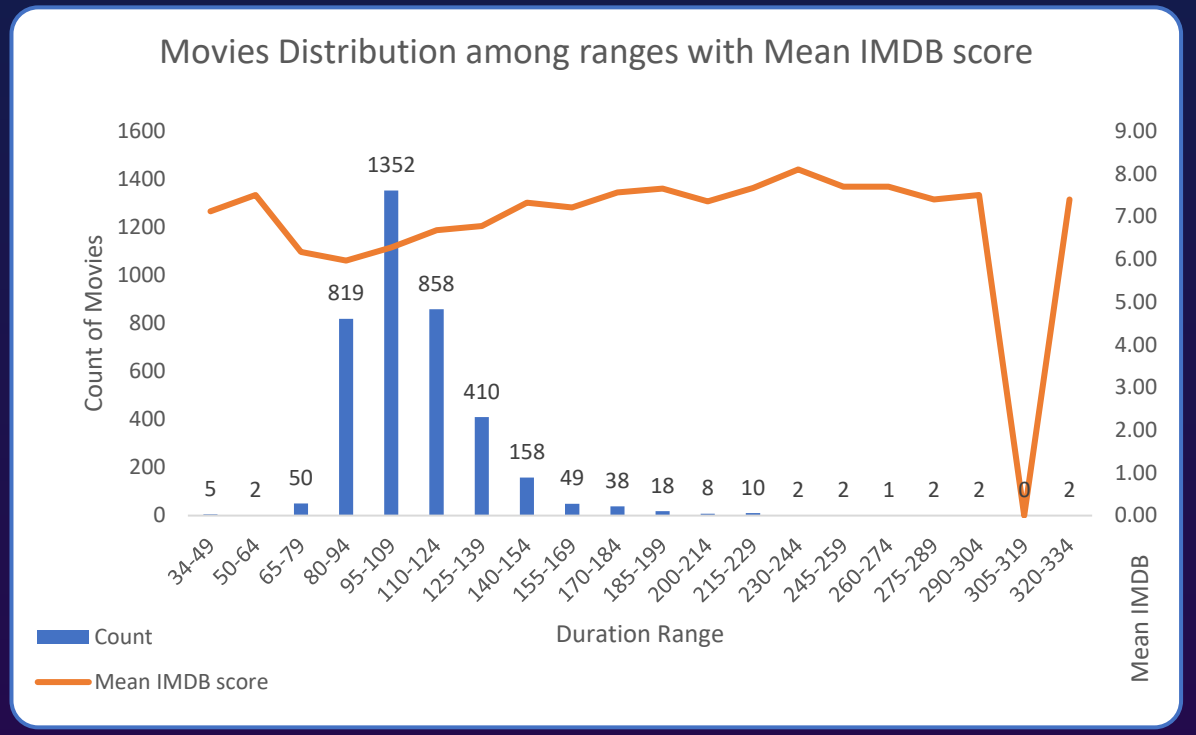
Insights

- More Than 50% movies belong to Drama genre followed by comedy, thriller & action
- Most of the mean IMDB score for all genre is in the range of 6 to 7 with few exceptions
- Mean IMDB score of 3 genres is above 7 – History, Biography and Documentary.
- Horror Genre movies have lowest mean IMDB score of 5.9
- Standard deviation is in the range of 0.9 to 1.1 for almost all genre movies

-Nilesh Kulkarni

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

Duration range	Count	Mean IMDB score
34-49	5	7.12
50-64	2	7.50
65-79	50	6.17
80-94	819	5.97
95-109	1352	6.27
110-124	858	6.68
125-139	410	6.77
140-154	158	7.32
155-169	49	7.21
170-184	38	7.56
185-199	18	7.66
200-214	8	7.35
215-229	10	7.67
230-244	2	8.10
245-259	2	7.70
260-274	1	7.70
275-289	2	7.40
290-304	2	7.50
305-319	0	0.00
320-334	2	7.40

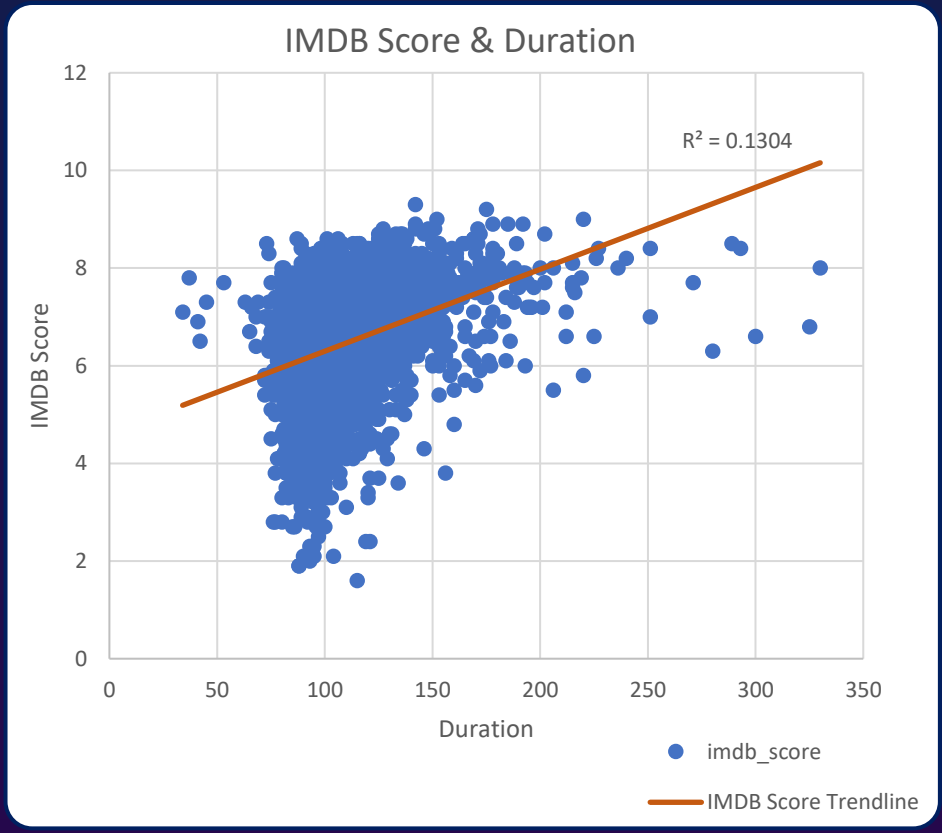


Movies data is grouped in the bins(ranges) of 15 minutes and the mean IMDB score for such groups is calculated as shown above. Majority of movies are in the duration range of 95-109 minutes (count 1352). Other predominant ranges are 80-91 and 110-124 minutes. Mean IMDB score for these 3 dominant ranges is around 6.5.

-Nilesh Kulkarni

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

Mean/Average	Median	Mode	Variance	Standard Deviation	Minimum	Maximum
109.803	105.000	101.000	518.027	22.760	34.000	330.000



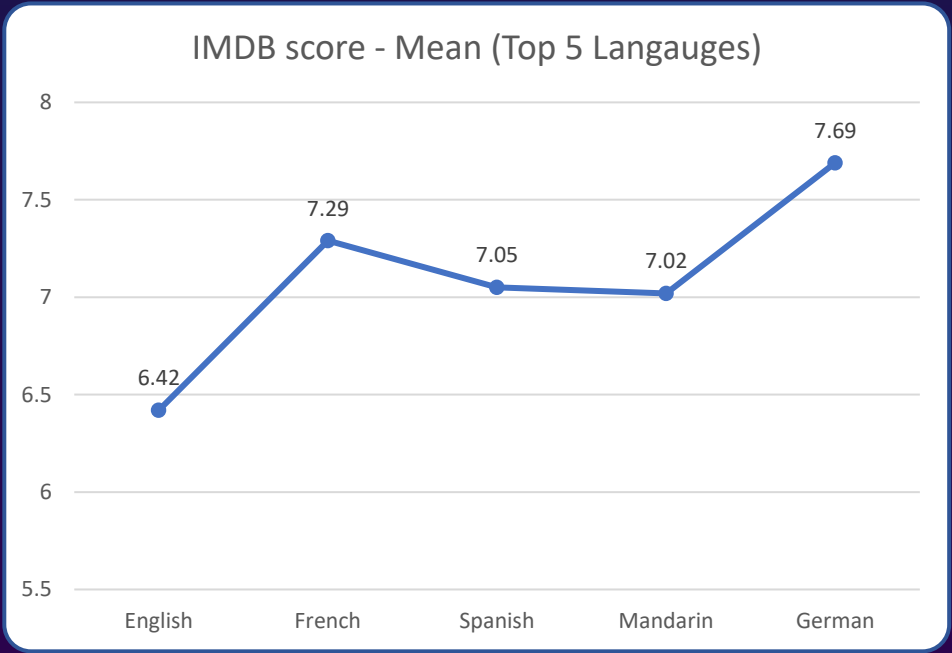
Insights

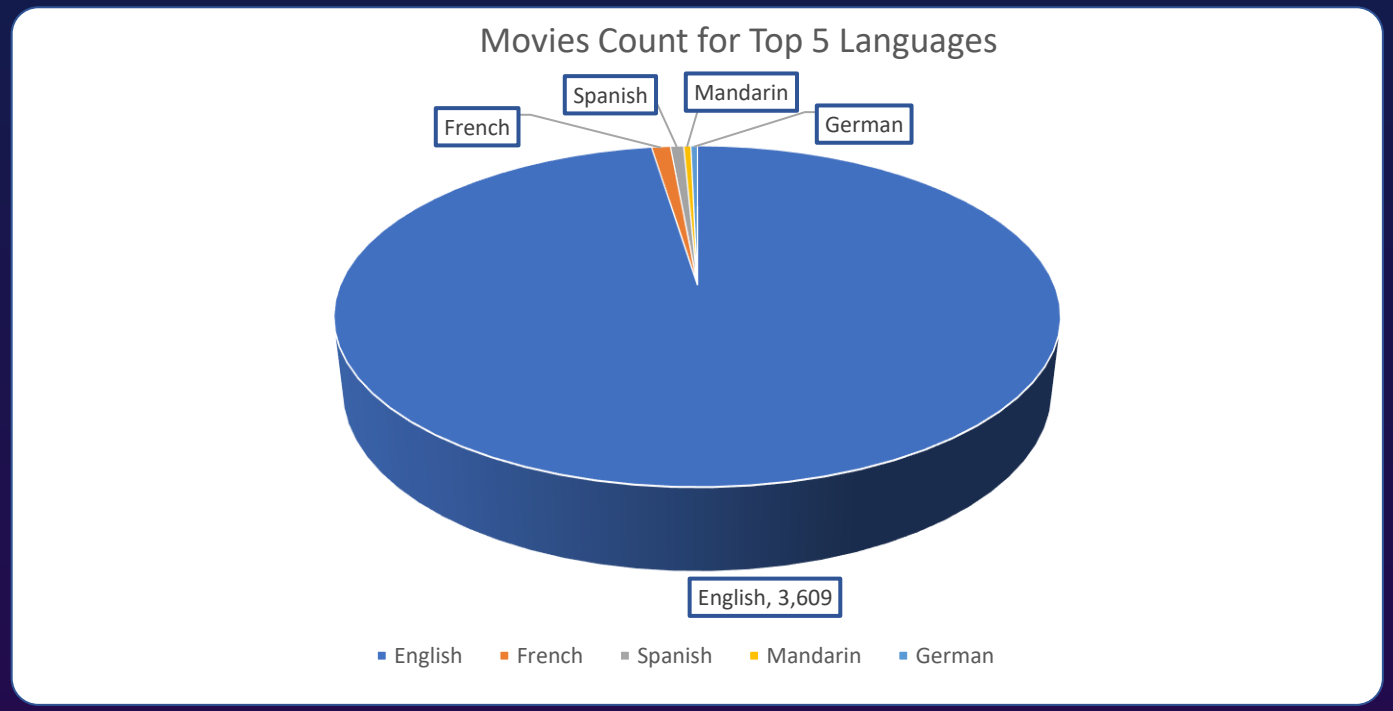
- Mean duration of all the movies is close to 110 with median nearby at 105. So the duration distribution is close.
- Standard deviation is 22 which denotes high variations in movie duration with minimum 34 and maximum 330
- As shown in scatter chart for relation between IMDB score and duration, the value of coefficient of determination (R^2) is low at 0.13. It denotes that IMDB score do not have strong relation or dependency on duration of movies although the value is positive.
- Most occurring movie duration is 101 (9 units from mean)

-Nilesh Kulkarni

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

Top 5 Langauges by count		
Langauge	**Count of Movies**	**IMDB score - Mean**
English	3609	6.42
French	37	7.29
Spanish	26	7.05
Mandarin	14	7.02
German	13	7.69





Insights

- Movie distribution by language is as shown, English has more than 95% movies with French and Spanish following as distant second and third.
- IMDB score of top 5 languages by count is as shown.

*Standard Deviation is marked as NA if only 1 record exists

Langauge	Count	Mean	Median	Standard Deviation
English	3609	6.42	6.5	1.05
Mandarin	14	7.02	7.25	0.77
Aboriginal	2	6.95	6.95	0.78
Spanish	26	7.05	7.15	0.83
French	37	7.29	7.2	0.56
Filipino	1	6.70	6.7	NA
Maya	1	7.80	7.8	NA
Kazakh	1	6.00	6	NA
Telugu	1	8.40	8.4	NA
Cantonese	8	7.24	7.3	0.44
Japanese	12	7.63	7.8	0.90
Aramaic	1	7.10	7.1	NA
Italian	7	7.19	7	1.16
Dutch	3	7.57	7.8	0.40
Dari	2	7.50	7.5	0.14
German	13	7.69	7.7	0.64
Mongolian	1	7.30	7.3	NA
Thai	3	6.63	6.6	0.45
Bosnian	1	4.30	4.3	NA
Korean	4	7.88	7.9	0.48
Hungarian	1	7.10	7.1	NA
Hindi	10	6.76	7.05	1.11
Icelandic	1	6.90	6.9	NA
Danish	3	7.90	8.1	0.53
Portuguese	5	7.76	8	0.98
Norwegian	4	7.15	7.3	0.57
Czech	1	7.40	7.4	NA
Russian	1	6.50	6.5	NA
None	1	8.50	8.5	NA
Zulu	1	7.30	7.3	NA
Hebrew	3	7.50	7.3	0.44
Dzongkha	1	7.50	7.5	NA
Arabic	1	7.20	7.2	NA
Vietnamese	1	7.40	7.4	NA
Indonesian	2	7.90	7.9	0.42
Romanian	1	7.90	7.9	NA
Persian	3	8.13	8.4	0.55
Swedish	1	7.60	7.6	NA

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

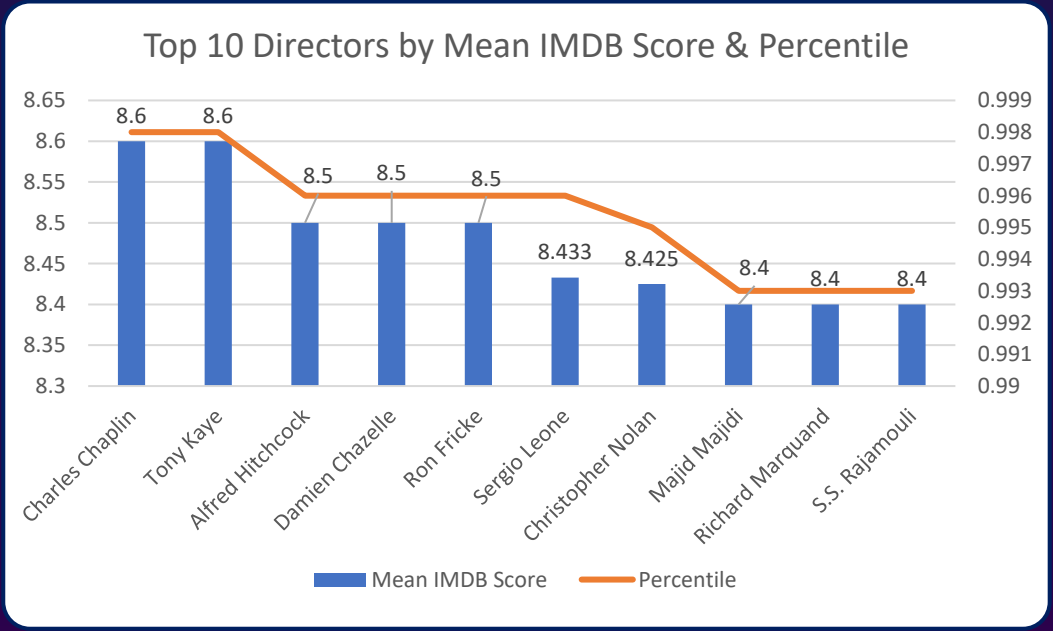
Language wise details of Mean IMDB score, Median, Count and standard deviation is as shown in table

Insights

- English having the most count of movies has mean IMDB score of 6.42 with standard deviation of 1.05
- Among top 5 languages by movie count, German language has the highest mean IMDB score of 7.69
- Many languages have mean IMDB score above 7 but the such movie count is very low
- Looking at the language wise share of movies count, there seems to be no link between language and IMDB score.

Task 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.

Director-Name	Mean IMDB Score	Percentile
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
Sergio Leone	8.433	0.996
Christopher Nolan	8.425	0.995
Majid Majidi	8.4	0.993
Richard Marquand	8.4	0.993
S.S. Rajamouli	8.4	0.993



Top 10 directors based on mean IMDB score are as shown in the table

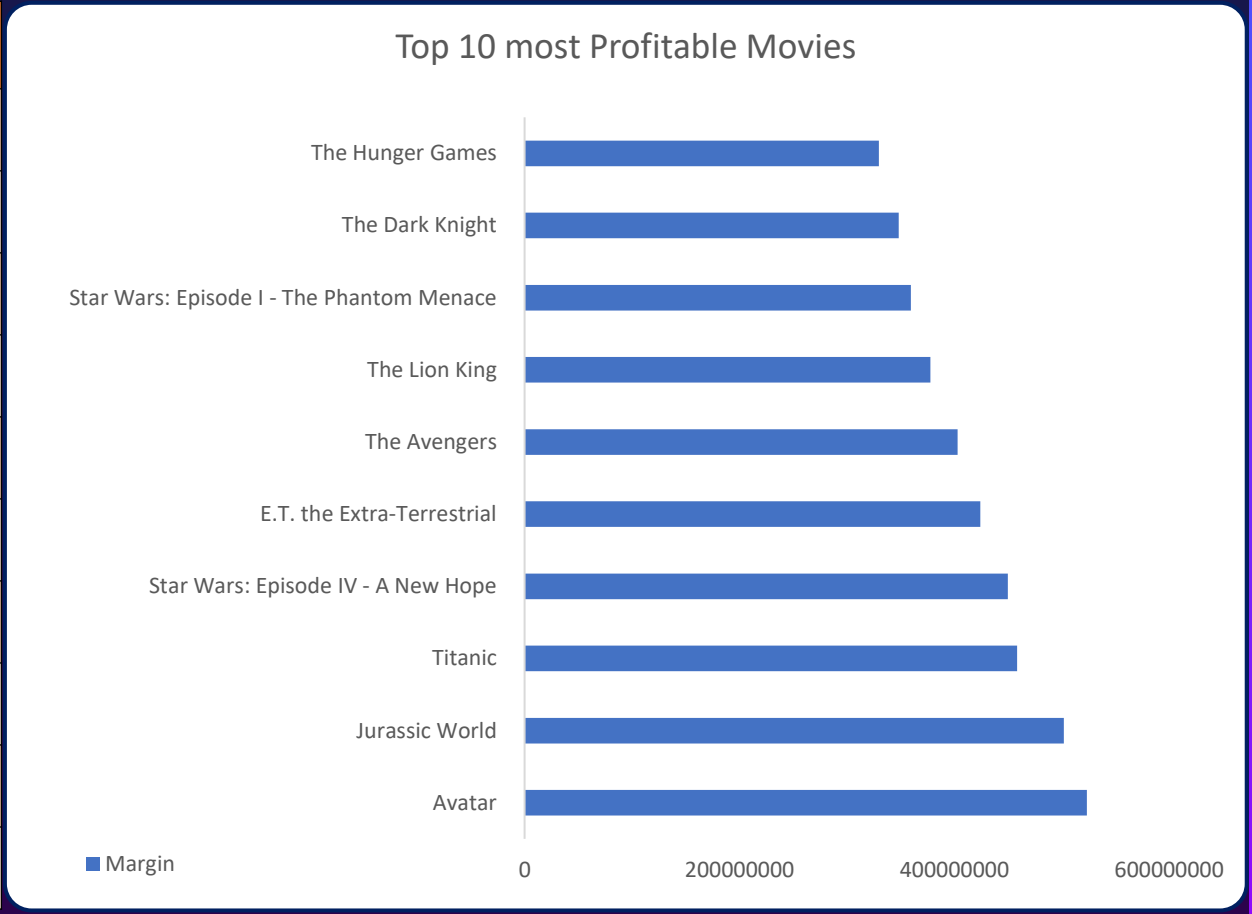
Insights

- Mean IMDB score for top 10 directors is in very close range of 8.4 to 8.6
- Mean IMDB score of all the movies is 6.46 which is 2 points less than the mean IMDB score for top 10 directors, which denotes the influence of top directors on IMDB score and success of movies
- Percentile score for top 10 directors is also very close to each other denoting close range

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

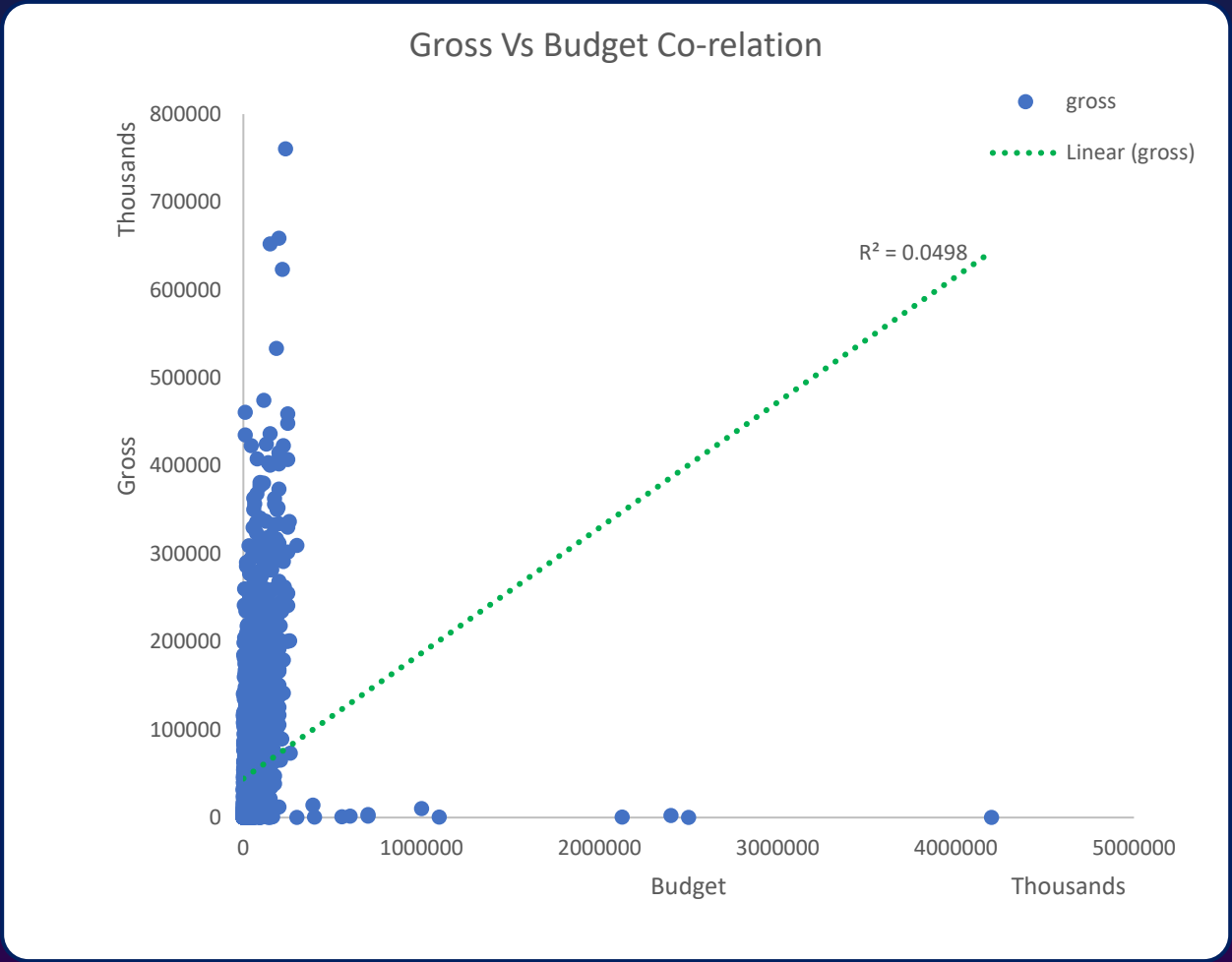
Top 10 movies by profit margin are as shown along with details like what % of budget the profit has been

director_name	movie_title	Profit	Profit as % Of Budget
James Cameron	Avatar	523505847	221%
Colin Trevorrow	Jurassic World	502177271	335%
James Cameron	Titanic	458672302	229%
George Lucas	Star Wars: Episode IV - A New Hope	449935665	4090%
Steven Spielberg	E.T. the Extra-Terrestrial	424449459	4042%
Joss Whedon	The Avengers	403279547	183%
Roger Allers	The Lion King	377783777	840%
George Lucas	Star Wars: Episode I - The Phantom Menace	359544677	313%
Christopher Nolan	The Dark Knight	348316061	188%
Gary Ross	The Hunger Games	329999255	423%

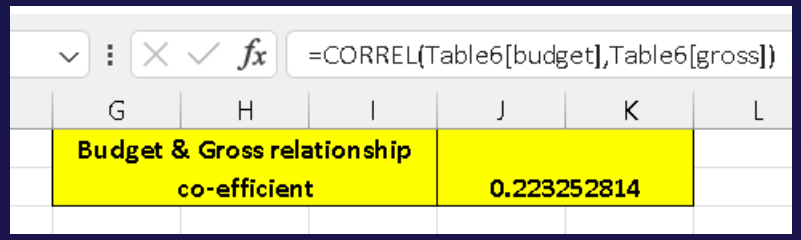


-Nilesh Kulkarni

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



Relationship co-efficient between budget & gross using CORREL function as shown



Insights

- Relationship co-efficient between budget & gross is 0.22 which is reflected in scatter plot as R^2
- The co-efficient of 0.22 indicates that there is positive relationship between budget and gross i.e. as budget increases, gross also increases but the relationship is weak and not strong enough.

-Nilesh Kulkarni

INSIGHTS SUMMARY

Post data cleaning and analysis with visualizations, the data shows impact of various factors on movie's IMDB rating and success. Key insights are as listed below:

- ✓ The data had many data quality issues like blank values, duplicate rows and missing details. so there is certain scope for improving data quality during data collection at source.
- ✓ There is no significant distinction of mean IMDB score across genres. The mean IMDB score lies in the range of 6 to 7 across all genres (except rare exceptions)
- ✓ More than 70% movies have duration of 80-139 minutes with mean IMDB score around 6.5
- ✓ Relationship co-efficient between duration & IMDB score is 0.13 - IMDB score do not have strong relation or dependency on duration of movies
- ✓ English language has more than 95% movies in data. Looking at IMDB scores and dispersed data, no link between language and IMDB score is noticeable.
- ✓ Mean IMDB score for top 10 directors is in very close range of 8.4 to 8.6 (2 points more than mean)
- ✓ Relationship co-efficient between budget & gross is 0.22 indicating weak relationship

RESULTS – PERSONAL UPSKILLING

- The Project has been great learning and value add to my skills
- Firstly, it has been good exposure to data cleaning and data preparation while dealing with huge data in MS-Excel
- The project helped to gain key skills in statistics like Mean/Median/Standard deviation and how to calculate those in MS-Excel
- The project helped me to learn more about specific charts in MS-Excel like combo chart and scatter plot/chart. This has been really value add
- The project has many calculation around averages/counts so I got to know more about functions like AVERAGEIFS, COUNTIFS & CORREL
- I also gained skills to extract business insights and present those using appropriate visuals in MS-Excel
- Overall the project has been great learning and enriching experience

THANK YOU

Nilesh Kulkarni

inileshkulkarni@gmail.com