IMDB Movie Analysis

Created by

Abhishek Adyani

Project Description

Problem Statement

"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.

What I’ll be doing

preprocess the data to make it suitable for analysis then I'll explore the data to understand the relationships between different variables, and crunch out the insights from the cleaned data

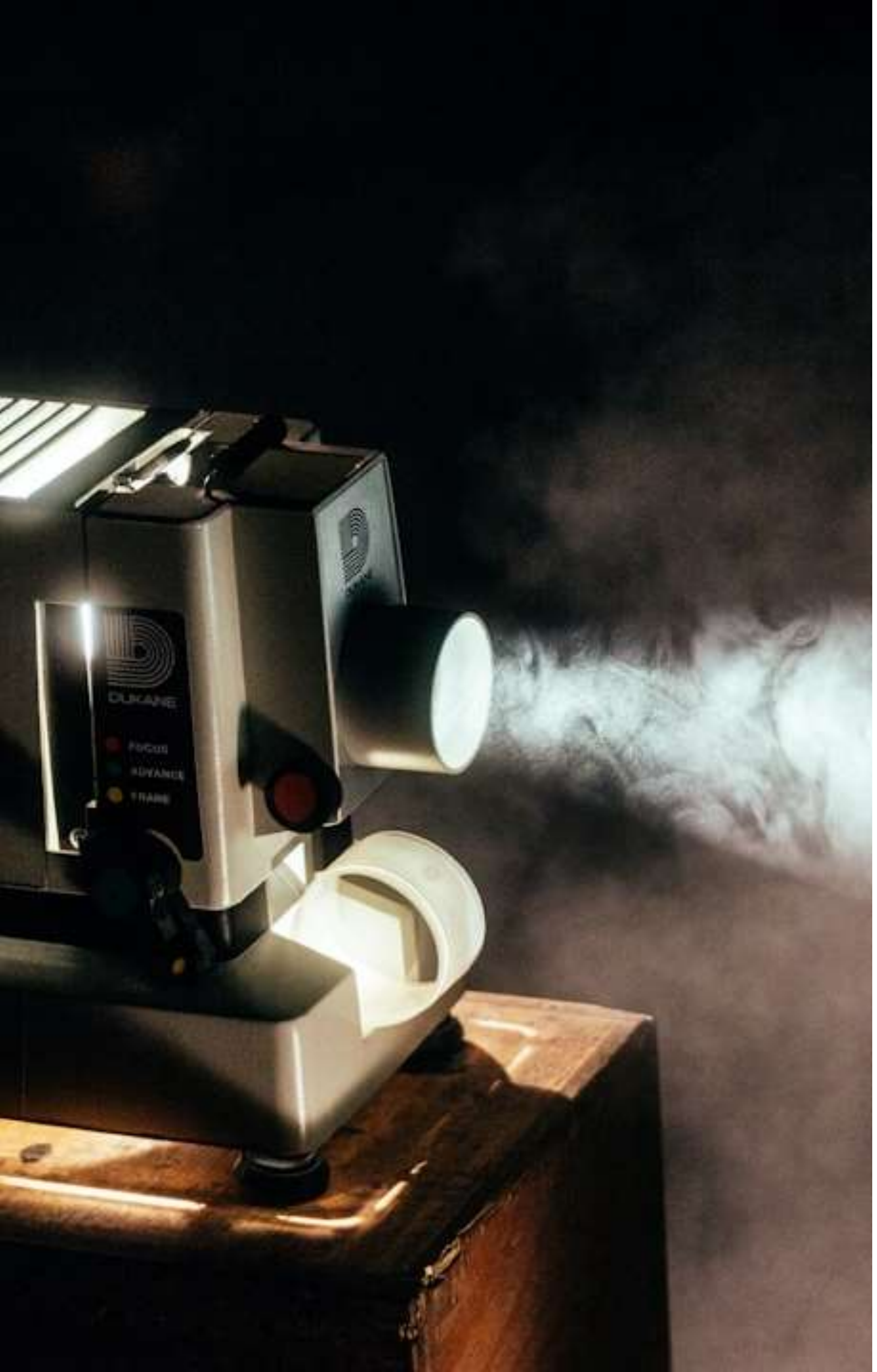
Insights

Explore the correlation between different variables. Identify any patterns in data and providing actionable insights.

Tech-stack used

Microsoft Excel 2021

I used this software as it has various functions that are convenient and faster to use. It helps in cleaning the data and them drawing meaningful conclusions about the given data.



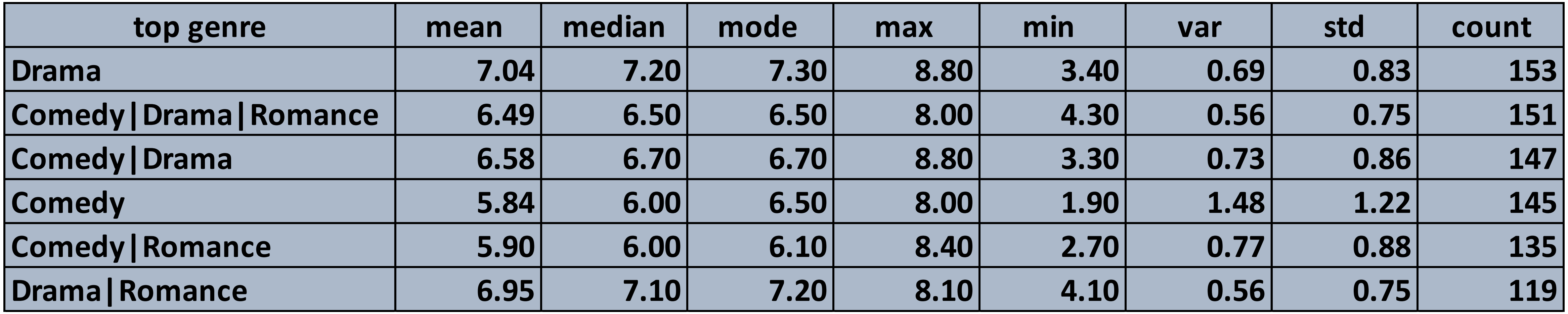
Data Cleaning

- First I examined the data and dropped the columns which were not needed.
- Columns like color, **director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio and movie_facebook_likes,** are irrelevant data. It needs to be dropped.
- Then I checked for duplicate rows and deleted the duplicates.
- I removed the null values wherever needed and filled the null values with relevant data by performing descriptive statistics in them.
- I used pivot table for various columns to find the count of unique values, I have also used COUNTIF function.
- Then I used excel’s functions like AVERAGE, MEAN, MEDIAN, MODE etc. to perform descriptive statistics.
- Lastly, I used various types of charts to visualize the insights.

FINDINGS:

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.

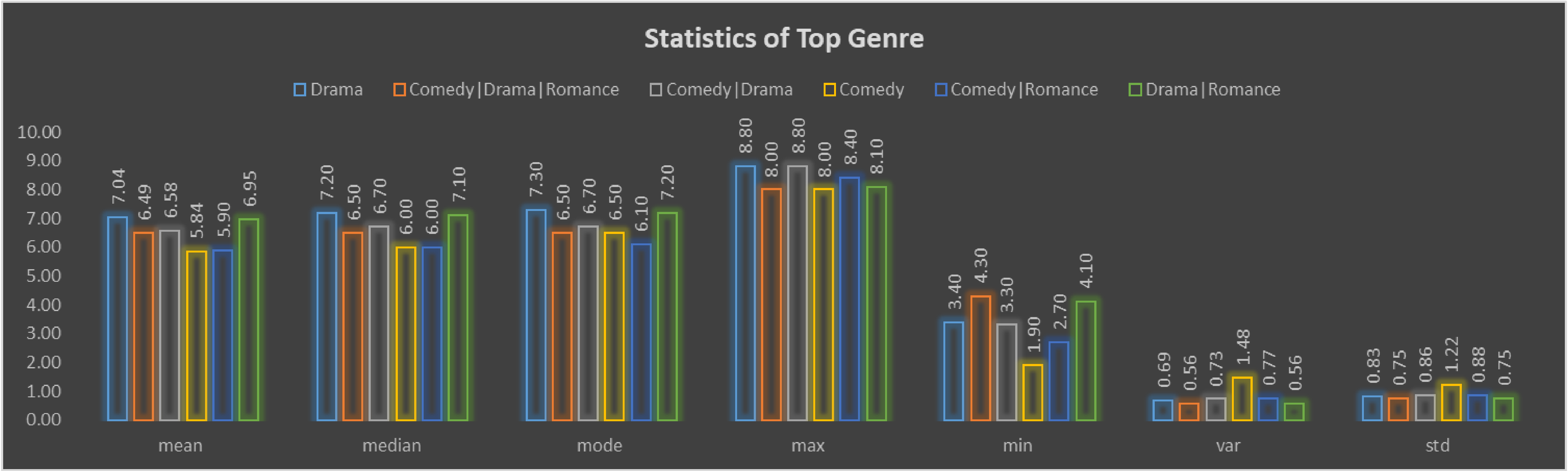
I used excel's inbuilt functions like average, median, mode, variance and standard deviation to calculate descriptive statistics of the imdb score and found that the most common genre was DRAMA followed by multiple genre COMEDY|DRAMA|ROMANCE.



top genre	mean	median	mode	max	min	var	std	count
Drama	7.04	7.20	7.30	8.80	3.40	0.69	0.83	153
Comedy|Drama|Romance	6.49	6.50	6.50	8.00	4.30	0.56	0.75	151
Comedy|Drama	6.58	6.70	6.70	8.80	3.30	0.73	0.86	147
Comedy	5.84	6.00	6.50	8.00	1.90	1.48	1.22	145
Comedy|Romance	5.90	6.00	6.10	8.40	2.70	0.77	0.88	135
Drama|Romance	6.95	7.10	7.20	8.10	4.10	0.56	0.75	119

FINDINGS:

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.

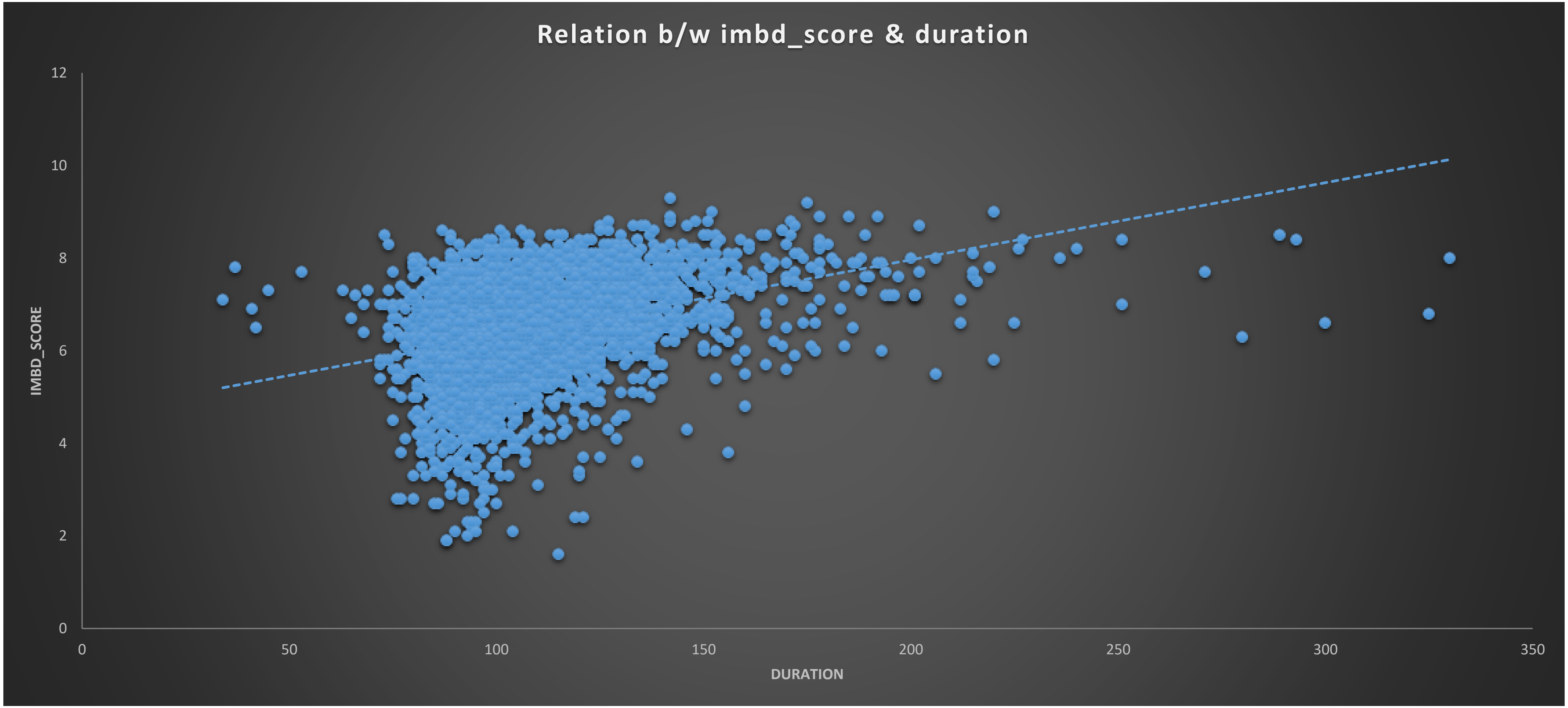


B) MOVIE DURATION ANALYSIS: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

I used excel’s inbuilt functions like average, median and standard deviation to calculate descriptive statistics of the duration of movie and found that the average movie duration is 110 approx. Most the movies with the average duration lies in 5-8 imdb score. With std deviation 22.75 and median 106.

Average	109.9052932
Median	106
Std Deviation	22.74534563

B) MOVIE DURATION ANALYSIS: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

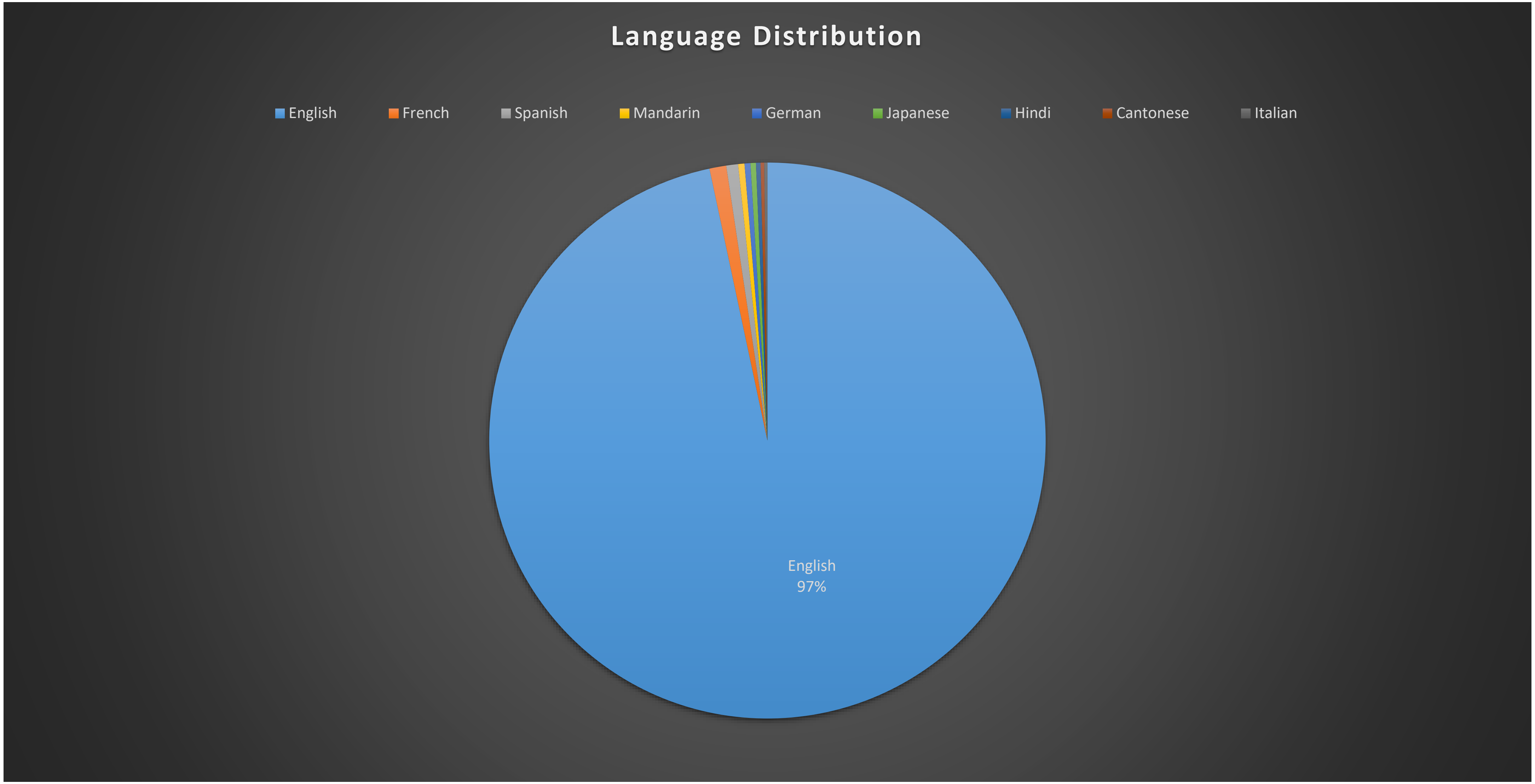


C) LANGUAGE ANALYSIS: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

I used descriptive stats, I found that 97% of movies are in the ENGLISH language , and the rest few movies are in other language, as for the imdb score, the mean for English language came up to be 6.42 and rest all language were 7 and above. As the most common language is ENGLISH, it's std deviation is 1.05.

Top Language	count	mean	median	std
English	3674	6.42	6.50	1.05
French	37	7.29	7.20	0.56
Spanish	26	7.05	7.15	0.83
Mandarin	14	7.02	7.25	0.77
German	13	7.69	7.70	0.64
Japanese	12	7.63	7.80	0.90
Hindi	10	6.76	7.05	1.11
Cantonese	8	7.24	7.30	0.44
Italian	7	7.19	7.00	1.16

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



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.

I used descriptive stats I found that the top director based on number of movies is CHIRSTOPHER NOLAN with 8 movies in total and the avg imdb score based on number of movies is 8.43. the percentile came out to be 8.60 as the highest avg score given being the percentage rank of 91%.

Large	**8.60**
Percent Rank	**0.91**
percentile	**8.60**

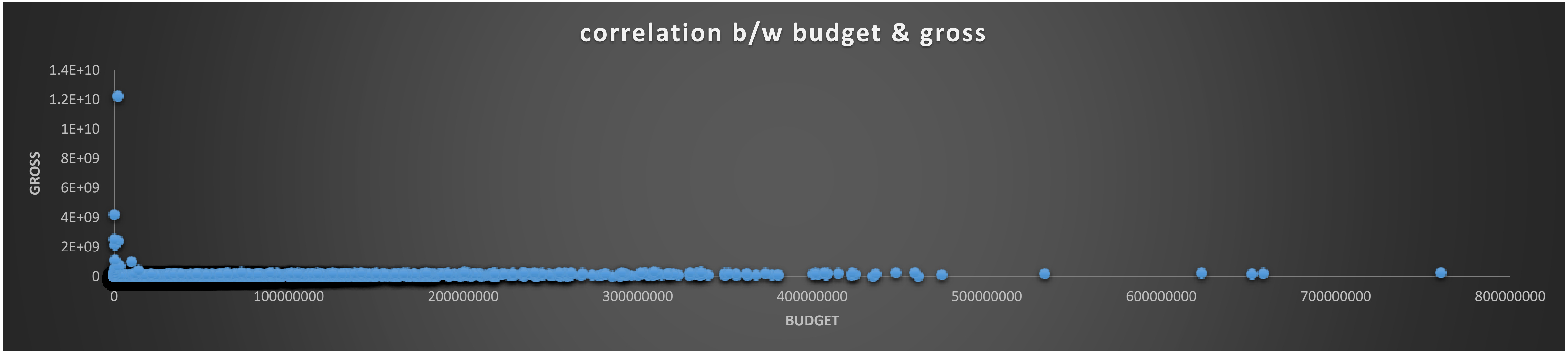
Top Director	Count of Movies	Avg_IMDB Score
Christopher Nolan	8	8.43
Sergio Leone	3	8.43
Alfred Hitchcock	1	8.50
Asghar Farhadi	1	8.40
Charles Chaplin	1	8.60
Damien Chazelle	1	8.50
Majid Majidi	1	8.50
Marius A. Markevicius	1	8.40
Richard Marquand	1	8.40
Ron Fricke	1	8.50
S.S. Rajamouli	1	8.40
Tony Kaye	1	8.60

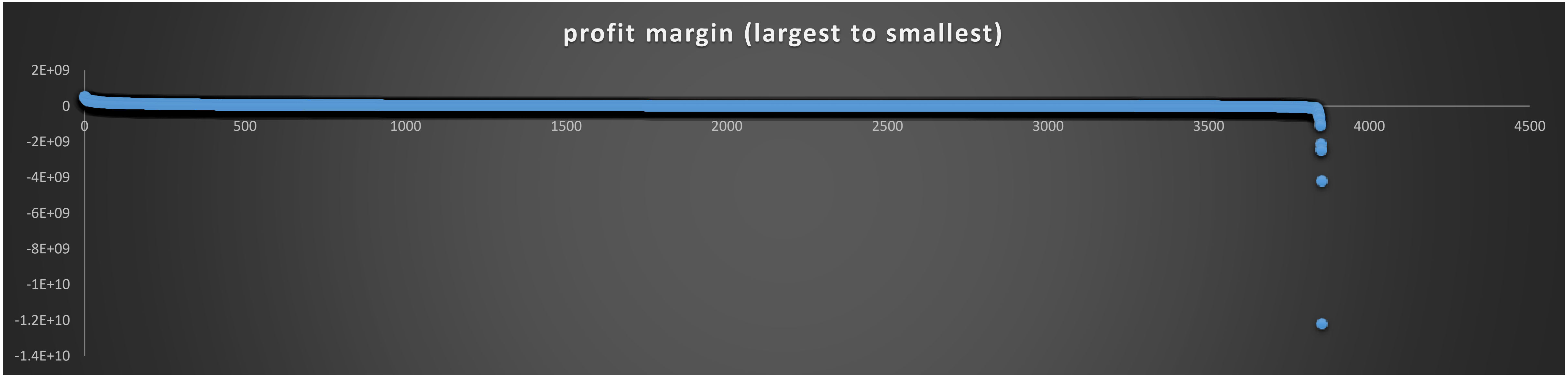
E) BUDGET ANALYSIS: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

By using excel function CORREL I found the correlation between movie budget and gross budget, it is 0.10, there is very weak relation between them as the strong relation should be between 0.80-1.00. Next, I made a column profit margin in which I filled the value of (gross-budget), to find the most profitable movie, the profit was 523505847 that I found by using MAX function and the movie title corresponding this profit is AVATAR, that I found by using FILTER function. We can also use sort function to get the value and movie title

CORRELATION	**0.10**
HIGHEST PROFIT MARGIN	**523505847**
HIGHEST PROFIT MARGIN MOVIE	**Avatar**

E) BUDGET ANALYSIS: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.





Conclusion and Insights

- The top movie genre is **DRAMA** and the avg imdb score of the it is **7.04**
- The relation between duration and imdb score is **positive linear**, most the movies lie between **100-120** duration and the average duration is **110**. There are a few movies in **300**+ duration that also did good at imdb score.
- There are **97%** movies made in **ENGLISH** language, with the average imdb score based on language is **6.42**, followed by **FRENCH**(**7.29**), **SPANISH**(**7.05**), **MANDARINE**(**7.02**), **GERMAN**(**7.69**) etc. (avg imdb score in bracket).
- Top director with avg imdb score based on number of movies is **CHIRSTOPHER NOLAN**, having made **8** movies with avg imdb score of **8.43**, followed by **SERGIO LEONE**(**8.43**) **3** movies, **ALFRED HITCHCOCK**(**8.50**) **1** movie.
- The correlation between gross and budget came out to be **0.10** which means there’s **no correlation** between them, the highest profit margin is **523505847** for the movie **AVATAR**

RESULT

This project helped me understand how the what factors influence the success of a movie on IMDB and types of questions I have to tackle to bring out the best insight form the data, to make the best data driven decisions. I used my knowledge of excel and various functions like AVERAGE, COUNTIF, MEDIAN, MODE, STD DEVIATION, VARIANCE,FILTER,PIVOT TABLE etc. as well as various types of charts to visualize the insights.